Who is talking?

Behavioural and neural evidence

for norm-based coding in voice identity learning

ISBN: 978-90-76203-45-4

Cover illustration: Gábor Duleczky

Printed and bound by Ipskamp Drukkers b.v.

Who is talking?

Behavioural and neural evidence

for norm-based coding in voice identity learning

Een wetenschappelijke proeve

op het gebied van de Sociale Wetenschappen

Proefschrift

Ter verkrijging van de graad van doctor

aan de Radboud Universiteit Nijmegen

op gezag van de Rector Magnificus

volgens besluit van het college van decanen

in het openbaar te verdedigen

op woensdag 16 januari 2013 om 15.30 uur precies

door

Attila Andics

geboren op 6 december 1980

te Boedapest (Hongarije)

PROMOTOREN:	Prof. dr. James M. McQueen
	Prof. dr. Anne Cutler
MANUSCRIPTCOMMISSIE:	Prof. dr. David van Leeuwen
	Prof. dr. Pascal Belin
	Prof. dr. Asifa Majid

The research reported in this thesis was carried out at the Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands, the Donders Institute for Brain, Cognition and Behavior (Centre for Cognitive Neuroimaging) of the Radboud University Nijmegen, the Netherlands and the MR Research Center of the Semmelweis University, Budapest, Hungary; and was financially supported by a grant from the Max-Planck-Gesellschaft zur Förderung der Wissenschaften, München, Germany.

To my children

“My father, it was presumably my father who, with his painter's palette under his coat,
sneaked back into the museum, stole back in, to retouch the paintings he'd hung
on the wall or, at the very least, to effectuate certain emendations thereof.”

The mottoes are from Péter Esterházy, Celestial Harmonies (2000), translated by Judith Sollosy (previous page) and from Attila József, Mysteries (1937), translated by Attila Andics, inspired by translations by John Bátki and András Imrényi (next page).

“Voice of water, voice of skies,
you'll blush when you understand.
Voice of the heart, voice of eyes
flow in the wish for your hand.”

Acknowledgements

I began work on this thesis seven years ago. Just a year later, I became a father. My sons, now six and three years old, have early experiences of all the ups and downs of writing a dissertation. Or, perhaps more so, this dissertation witnessed all the ups and downs of raising two little boys. From that perspective, I find it pretty cool that I got here so fast.

I would like to express my gratitude to several people who were part of this exciting ride. First and foremost, I would like to thank my supervisor and promotor, James McQueen. James, I must have been quite a difficult student. I never stopped questioning experimental designs, analyses and interpretations we had already agreed upon, and always came up with new suggestions when you thought it's finally over. I always had one more emendation to effectuate. But you never minded, you always made it clear that you do care, and you always made me believe that I actually make sense. Thank you for always setting the bar high, and for not discouraging me when I attempted to set it even higher. Thank you for your trust in me – I think there were times when you were the only person on earth who believed that this dissertation will ever be completed. In fact, I checked your last four New Year's wishes, and you were pretty sure each time that that will be the year of my defense. I am especially grateful for your persistence, patience, positivity and perfectionism. You are the perfect supervisor, undoubtedly – and your attitude to scientific thinking, writing and mentoring makes you a wonderful role model, far beyond science as well.

I also want to thank my second promotor, Anne Cutler. Anne, thank you for teaching me that doing science is about science rather than about me, and that adding those single bricks to the big wall does not mean monotony and boredom, but creativity and joy: simply because it is me deciding how and where to fit each brick I have. Your passion for your own and others' research and your amazing speed were constant sources of motivation to me. I am also grateful for your generosity and your continuous encouragement.

Next, I would like to thank Miranda van Turennout for introducing me to the miracles of neuroimaging and brain plasticity with so much enthusiasm. Miranda, I was amazed by your ability to talk about science in a way accessible and irresistible to everyone. You, together with James, supervised my Master's thesis as well. Both that experience and your true support were critical in my decision to go for a PhD. You also co-supervised the first, crucial phase of my doctoral research. You taught me to be sure, yet always ready to

change my mind. Your readiness and courage to change remain an important example to me.

Anne and Miranda, I also thank you for giving me the opportunity to become part of your research groups. The Language Comprehension group at the MPI and the Learning and Plasticity group at the Donders Institute's Centre for Cognitive Neuroimaging became my scientific anchor points for a lifetime. I was simply astonished during every group meeting, including the Project Proposal Meetings at the Donders to see how much expertise can be squeezed into a single room. Much of what I know about how to convincingly present my results in a talk, I learned during these sessions.

For the neuroimaging studies, I also received essential guidance and support from Karl Magnus Petersson. I thank you, Karl Magnus, for making the complicated simple by always being able to satisfyingly answer long and complex technical questions on fMRI data analysis with a single monosyllabic word, and for showing how to take things easy and seriously at the same time. The solidity of your opinion was an important source of relief in periods of hesitation.

I would also like to thank the members of my reading committee, David van Leeuwen, Asifa Majid and Pascal Belin for finding the time to carefully read my manuscript. I am also grateful to Pascal Belin for the great inspiration I got from his insightful papers, from the very beginning of my doctoral studies. To be honest, Pascal, without your influence I would most probably never have come to write my dissertation on voices. Thank you.

Next, I would like to thank some of the people who contributed to my professional development and wayfinding in Hungary. I thank Valéria Csépe for introducing me to cognitive neuroscience, and for encouraging and helping me in many different ways when moving to Nijmegen and upon returning to Budapest. I am also grateful to Zoltán Vidnyánszky and Gábor Rudas at the MR Research Center of the Semmelweis University for enabling me to do part of my doctoral research in Hungary. Zoli, I also want to thank you for the intense and thoughtful discussions on the nature of neural mechanisms and on what my data might mean, and for your commitment to efficiency. I am also grateful to Ádám Miklósi for giving me the opportunity to join the colourful and dynamic MTA-ELTE Comparative Ethology Research Group, and to do truly pioneering research in a creative and joyful atmosphere.

I received specific support from a number of people. I thank Holger Mitterer for introducing me to speech manipulation methods, for many-many critical questions and comments, and for his never-ending willingness to help. I thank Benedikt Poser for creating a special sparse scanning sequence that perfectly suited my strange needs. I also thank Viktor Gál and Lajos Kozák for building up a fully functional auditory setup from scratch in the fMRI lab in Budapest. Erik, Bram, Marek and Paul, thank you for providing the best technical support one can think of, with great expertise and speed. Rian and Angela, Tildie, Sandra and Arthur, thank you for the kindness and flexibility with which you always helped me to find my way through bureaucracy, and with all sorts of typical and atypical administrative requests. I also want to thank more than hundred participants of my experiments who were enthusiastic enough to listen to voices saying words like 'mes' at least a thousand times – without their brain activities this thesis could definitely not exist.

I'd also like to thank fellow students and colleagues at the MPI, at the Donders and in Budapest who made these years so much more enjoyable and memorable. Laura, Martijn, Jasper and Thomas, thank you for introducing me to the Netherlands, doing your best to fight my emerging stereotypes of Dutch people, and your own ones of Hungarians, and for many nice and funny memories. Marieke, Barbara, Esther, Gabi, Aliette, Nina, Joost, Clemens and Jos, thank you for your always positive and supportive attitude, for the lovely group meetings and journal club dinners, and for that lunch to come. Anita, Keren, Suzanne, Eva, Hanneke, you were wonderful office mates, thank you for tolerating so kindly the annoyingly random patterns of my appearances and disappearances. Anita, when I need a moment of calm, I often recall one of those late afternoons, lights off, window open to provide access to your food in the natural fridge outside, silence, squirrels, and the nice smell of the tea you are drinking. Thanks for this memory, and for the chats about the really important stuff. Adriana, Bettina and Mirjam, thank you for your empathy and for sharing so similar perspectives on things – talking to you was always very comforting. All members of the Language Comprehension group, thank you for many nice discussions over lunch or elsewhere. Éva, Petra, Körty, Vanda, István, Balázs, Viktor, Béla and Lajos, thank you for giving me the feel of being part of a vivid fMRI research community in Budapest, and for all the joy. Márta and Anna, thank you for all those endless Sunday scannings, and for enriching my excitement for teachability with a new dimension.

Thank you, Lilla and Matthias, for being my paranymphs. Lilla, we've known each other for thirteen years or so. Thank you for your friendship, your sincerity, your deepness, for the talks we had and for the silence, for all your caring, and for hosting me during all my Nijmegen visits the last four years. Matthias, thank you for translating my quite long summary to Dutch, and for your helpfulness during this last phase of the ride. You have a great sense of support: you've helped me with dozens of things I did not even know I needed help with.

And there are the people who kept reminding me that science is not all there is. For sharing important life moments that always helped me gain new energy to move on, I am thankful to my friends in the Netherlands, in Hungary, in Estonia, Switzerland, Bosnia, the Czech Republic, Slovenia, Serbia, Germany, Turkey and other places. For helping me feel I'm as much of an educational activist as a researcher on many days of these seven years, and for not letting me forget that challenging attitudes and opening minds are among the most exciting things on earth, I am grateful to all those who dreamed and volunteered with me in EFPSA and other civil movements; and my math students in 2015A and 2016A and colleagues in the Budapesti Piarista Gimnázium.

Finally, my family in Hungary was the source of continuous, limitless, unconditional support – during these seven years as well as before, even in times that were very difficult to them. To give us some rest in those first months and years of parenthood, my parents and mother-in-law were flying back and forth as others use the local bus. My sister and brother also helped a lot with babysitting, moving, house renovation and joyful distraction whenever needed. Anyu, Apu, köszönöm a határtalan bizalmatokat, a belém vetett hiteteket, és mindazt a segítséget, ami lehetővé tette, hogy egyáltalán eszembe jusson doktorizásba fogni, majd azt, hogy az összes nehézség ellenére végül azért mégis befejezzem. Livi, Árpi, nektek is nagyon köszönöm, jófejek vagytok.

My sons, whose voices are the most frequent events in my auditory environment, you show me thousands of little wonders, joys and excitements of the world every single day, and keep reminding me that nothing is as important as these things. Thank you for teaching me to respect, experience and enjoy all colours of nature. Áron, Benedek, képzeljétek, most tényleg befejeztem azt a nagyon fontos írást, és egy igazi könyv lett belőle, ez itt. Köszönöm, hogy olyan ügyesen drukkoltatok.

Ági, we have spent together more than half of our life so far. You know me better than I know myself. You are a part of me as much as I am a part of you. Your move to the Netherlands was the true motive for me to come to Nijmegen, the town that became a very special place for us. This is the place where we first had to cook for ourselves, the place where we first felt homesick, the place where we got engaged, and the place where our first son was born. You were with me all along the way, through all those ups and downs. Without the sacrifice you make every day, this thesis would not be here. Köszönöm.

Budapest, 16 November 2012

Attila Andics

Contents

1 Introduction 1

Cues of person identity 2

Similarity-based representational spaces 3

Norm-based coding 5

Learning and re-learning 9

Voice identities 10

This thesis 15

References 18

2 Phonetic content influences voice discriminability . . . 25

Introduction 26

Method 26

Results 28

Discussion 35

References 38

3 Flexibility, cue use and abstraction in voice identity learning . . 39

Introduction 40

Experiment 1 46

Experiment 2 58

General Discussion 65

References 73

Appendices 77

4 Phonetic content shapes implicitly-learned voice categories . . 83

Introduction 84

Method 87

Results and Discussion. 90

Conclusions 95

References 98

5 Neural mechanisms for voice recognition 99

Introduction 100

Materials and Methods 103
Results 111
Discussion 122
References 130
6 Mean-based neural coding of voices 137
Introduction 138
Materials and Methods 141
Results 147
Discussion 154
Appendices 160
References 161
7 Summary and conclusions 165
Summary 166
Conclusions 171
References 177
Samenvatting en conclusies 179
Samenvatting 180
Conclusies 186
Összefoglalás és következtetések 193
Összefoglalás 194
Következtetések 200
Curriculum vitae 209
MPI Series in psycholinguistics 211

Chapter 1

Introduction

This thesis is about how human listeners represent auditory categories. The focus will be on how we extract voice identities from human speech. The behavioural and neuroimaging experiments presented here demonstrate that voice identity learning is mediated by norm-based codes. The methods applied in these experiments include a voice morphing technique, a learning and re-learning paradigm, and functional magnetic resonance imaging (fMRI).

Cues of person identity

"Hello, it's me" – says the voice in the phone, and you would like to know for sure if it is your mum, your boss, your partner or a complete stranger talking. Also, when someone smiles at you at a conference dinner, you want to be sure if it is the same person you had a chat with the previous day or somebody else. When you are wrong, you can easily find yourself in inconvenient situations. Recognizing the people we know is a very basic social ability. Whether recognition occurs from a face, a voice, or eventually a touch or a smell, the everyday significance of identifying someone from the cues available is unquestionable. Similarly, eye- and earwitnesses of a crime who can confidently identify the perpetrators are of extreme importance in forensic investigations. While much is known about the perceptual background of visual person recognition (i.e., face identification), much less is known about auditory person recognition (i.e., voice identification).

We meet a great number of people every day and succeed in recognizing many of their voices and faces. So the human perceptual system has evolved to cope with the challenge of storing and remaining ready to form new and new person identity memories. For that, the perceptual signal has to contain useful person identity cues. What constitutes a good person identity cue? First, a cue and changes in that cue have to be detectable: they have to conform to the capacities of the perceiver. For example, no visual face cues remain detectable in darkness, and no voice cues are helpful in loud noise. Second, a good person identity cue must be relatively stable across appearances of the person. Stability within person can follow from anatomical constraints (e.g., blue eyes, high-pitched voice) or from learning and choice (e.g., always with a white hat on, a mustache, a strange-sounding /s/, an accent). Third, a good person identity cue must be sufficiently variable across people. Indeed, we want to tell apart a large number of people. Cue stability within person leads to perceptual constancy; cue variability across people leads to discriminability.

All physical parameters that satisfy these three conditions can be good person identity cues. To be effective in person recognition, the perceptual system should rely on multiple, distinctive cues. And human perceivers indeed tend to use whatever cue they have at hand. So how does the human mind represent all these cues of person identity? This thesis explores perceptual mechanisms that can support person identity representations. An important question that has to be accounted for in any representation of multiple

perceptual events in some common space is how these events relate to each other, that is how similar they are. The following section introduces the idea of similarity-based representational spaces and explains the concepts related to it.

Similarity-based representational spaces

To make use of our personal database of person identities, a new appearance of a face or a new token from a voice (i.e., a new person identity event) has to be matched to old person memories to see which one it is most similar to. One way to visualize this similarity-based organization of memories is to say that the relevant person identity cues span a multidimensional representational space, one dimension for each unidimensional cue for simplicity, and we can think of individual events (a face appearance, a voice token) as points in that space, representing cue values in each relevant dimension. Crucially, in such representational spaces within-person events will be closer to each other than across-person events. This conceptualization is extremely helpful, because many important concepts of object recognition and coding directly follow from it.

Distance of two events in a representational space quantifies their (dis)similarity. The simplest decision to be made with respect to similarity and dissimilarity is if two events or stimuli are the same or different? Is there a perceived distance or not between two stimuli? For example, do those two face appearances or those two voice utterances correspond to the same person or to different persons? This is the most basic question of person and, in general, object processing.

The answer to the same versus different question depends on a number of factors, including specificity of change and perceptual sensitivity. It is possible that two person identity events differ in one cue but not in another one (cue specificity). For example, two voice tokens might clearly differ in timbre but at the same time be very similar in fundamental frequency. It is also possible that a certain information processing stage is capable of distinguishing two slightly different person identity events, while they are considered the same by another processing stage (perceptual sensitivity). For example, two tokens from a voice often differ considerably and detectably in fundamental frequency, but they still are perceived as exemplars of the same voice. Different information processing stages of the human perceptual system seem to maintain different representational spaces

with different selections of cues and with different resolutions of cue values. There can then be multiple stages of representations, with different levels of abstraction. A representational space which contains less specific cues and/or is less sensitive to fine-grained changes will then constitute a more abstract level of representation. For instance, it is possible that the processing stream for human vocalizations maintains separate representational spaces for mapping variation in acoustics, phonemic identity, talker identity, talker gender, talker emotion and so on.

Two further important properties of a similarity-based representational space are its time window and spatial window. The time window refers to the temporal length which the representational space can 'remember'. In a very short term space (e.g., with a time span of some seconds), only the last few events are stored. Positioning a new event in a short-term space is then informed by measures of its distance from the last few events' positions only. In a space with a longer term memory (e.g., minutes or days or even years), in contrast, a large number of events need to be stored. Positioning a new event in a long-term space should therefore be informed by measures of its distance from all other events' positions.

The spatial window refers to the size limitations of the representational space, that is, where its boundaries are and how large the cue variations can be to still be tolerated. For example, the human auditory system does not detect sound frequencies below approximately 15 Hz and above 20 kHz, so this imposes limitations on all auditory representational spaces. While a larger space might make it possible to accommodate events from many persons in the same space (i.e., a supra-individual space), a smaller space might contain events from one person only (i.e., an intra-individual space). Supra-individual spaces for person representations can be useful when, for instance, the perceiver has to judge the similarity of two persons. Intra-individual spaces can be helpful, for example, when we try to decide if a boy talking to us still has the cold that made his voice sound so strange yesterday, or if he is now fine again.

Taken together, the concept of similarity-based representational spaces provides a very useful heuristic when thinking about the perceptual organization of various person identity events, or, in fact, of any objects. Similarity-based spaces can be characterized by their cue specificity, sensitivity, temporal window and spatial window. An important question is: How might the perceptual system implement similarity-based representational spaces? What encodes distance and position in a space? Is it possible to trace down if a

certain representational level is encoded in a certain region of the human brain? Which of the many possible spaces are implemented neurally at all? One possibility for the implementation of similarity-based spaces, to be explored here, is norm-based coding.

Norm-based coding

In any implementation of a representational space, position in the space has to be quantified in terms of the signal values that build up the space. It has to be clear what position larger and smaller values specify in the code of a certain space. One proposal is that signal values represent perceptual distance in a polar coordinate system, with the pole as its origin. The bigger the distance of a cue value from the pole, the bigger the signal value. The question then is: what may constitute the pole, compared to which the distances are calculated?

One solution is to calculate the mean of preceding events that are within the temporal and spatial window of the representational space, and to calculate the new event's distance from that mean. This way, distance information from many previous events is packed in a single signal. This signal then gives a reasonable estimate of how far a given person identity stimulus is from some or all previously perceived person identity stimuli. As a consequence, the representational space will contain central and peripheral events. Note that temporal window of the space is a critical factor here. In an extremely short-term representational space that 'remembers' the last event only, central position corresponds to an event similar to the previous event, while peripheral position corresponds to an event different from the previous one. For example, in a short-term space that represents females' singing voices, a soprano voice will have a central position if preceded by another soprano voice that is similarly high, but it will have a peripheral position if preceded by a very different contralto voice. In a long-term space, however, central and peripheral positions correspond to events that are similar to or different from the long-term mean, respectively. For example, in a long-term space of females' singing voices, high soprano and low contralto voices will have more peripheral positions than medium mezzo-soprano voices, independently of how high or low the previously heard voice was. A long-term mean, if exists, is very informative about the specific representational space. However, averaging makes real sense only if the relevant cues are continuous, that is if intermediate values

between the extremes are equally possible. This is so for singing voices but not for eye color: there are no people with eyes halfway between brown and blue. In a space with continuous cues, the mean can be seen as the most typical exemplar of the category that this space represents, compared to stimuli far from the mean that are atypical exemplars of this category. Indeed, similarity-based spaces can be seen as category representations, with the most typical values corresponding to category centers and the least typical values to category boundaries. In other words, any representational space defines a category for which within-category variation equals to the variation that that specific representational space tolerates. The prototype of that category is then the centre (or pole) of the representational space. This way of representing events with their distance along some important dimensions from a mean value or prototype is called norm-based coding.

Norm-based representational spaces may exist on different levels of abstraction along the information processing stream, with different levels of selectivity, sensitivity, time window and spatial window. Consequently, different mean values or norms can be defined for each space. Multiple levels of norm-based maps are thus possibly maintained. For example, the processing of faces or voices may be mediated by various representational spaces, each centered around a norm: person-specific norms (e.g., a mean-Bob voice, the average of all voice events of Bob represented in that space), gender-specific norms (e.g., a mean-male voice), emotion-specific norms (e.g., a mean-happy voice), a broad but voice-specific norm (e.g., a mean-voice, the average of all voice events represented in that space) or even broader, voice-nonspecific norms (e.g., a mean-pitch representation), and so on. Which of these spaces are represented in the perceiver's brain and how, and in what ways are these different-level codes linked to each other? These are basic questions of person identity perception research. Some of these questions will be central to this thesis too.

The spatial window of a representational space can also be defined with respect to a norm. Category size limitation can then be seen as an acceptance range of variation or distance from a norm. Consider this metaphor: how far a dog can walk from its owner (the norm) depends on the length of the leash (the acceptance range). Whether a new face or voice identity event is perceived as part of Bob's identity category will depend on how far the new event is from mean-Bob, the person-specific norm face or voice, in a corresponding intra-individual representational space. Little is known about the nature of these category size limitations or acceptance ranges. For example, how much within-talker variation is

accepted in a specific dimension, or how big do the changes to a face have to be along a certain parameter for that face to be perceived as a different person's face (cf. Cabeza et al., 1999)? Are all talker categories equally big? And if not, do size differences depend on the vocal anatomy of the talker, or on what they say, or perhaps on listener biases? A good understanding of the size restrictions of person identity categories would help to characterize the processing stages of person recognition.

Norms can therefore function as natural anchor points within their representational space. But similarity-based, polar-organized representational spaces could in principle be centered around anchor points that are not norms calculated by the perceiver but special cue values inherent in the signal. Signal-inherent anchor point here simply refers to a cue value that has a special status which is independent of the cue distribution in the actual context. This special status may originate in long-term nonlinearities in the distribution of the cue, but also in long-term preferences of the perceptual system. Signal-inherent anchor points could be used as category boundaries between two neighbouring categories, replacing pole-centered acceptance ranges. Whether anchor points are calculated (and therefore relative) or signal-inherent (and therefore absolute) has great theoretical significance. If the encoding of stimulus positions in similarity-based representational spaces was supported by signal-inherent anchor points, then norms would not be needed, and, to take the consequences to their extremes, stimulus representation could possibly happen in a purely exemplar-based manner, that is, without a need for abstraction. If, however, such signal-inherent anchor points do not exist, then anchor points either have to be calculated, for example by averaging across short-term or long-term perceptual history, which is not compatible with purely exemplar-based models that assign no specific status to the average stimulus, or representational spaces have to be built up without any anchor points, which seems computationally implausible. Do there thus exist anchor points that are built-in in the signal? Or are there at least specific cue values that are preferred by the perceptual system and therefore lead to nonlinearities in category formation? For instance, are certain face appearances or voice tokens inherently better candidates for being face or voice identity category centers than other possible faces or voices? Would the members of a voice space be worse anchor points than the one around which the space is centered? Such anchor point candidates have been suggested in color perception (Anderson and Khang, 2010), but not in person identity perception.

Evidence for similarity-based representational spaces comes from both behavioural and neuroimaging findings. Short-term coding of perceptual events based on their similarity to directly preceding events is demonstrated using various techniques and different terminologies (e.g., repetition priming, neural adaptation, fMRI adaptation, mismatch negativity, mismatch field, carry-over effects, short-term repetition suppression), although different models are proposed for how the brain might code short-term stimulus similarity (see Grill-Spector et al., 2006 for a review; Aguirre, 2007; Epstein et al., 2008). A common point of these models is the observation that short-term stimulus repetition usually leads to reduced neural activity compared to the activity elicited by stimulus changes. But note that these findings can be explained without the concept of norms. In fact, many behavioural studies attempted to distinguish norm-based and exemplar-based coding, but much of the evidence presented in this old debate turned out to be compatible with both models (Valentine, 1991; Rhodes, 1996).

Behavioural evidence that seems truly compatible with norm-based but not with exemplar-based coding was shown for faces: Leopold et al. (2001) found that exposure to a face introduces a perceptual bias towards the identity that is opposite to the one presented, with respect to an average face (this phenomenon is called the face identity aftereffect or anti-face adaptation). The concept of long-term norm-based codes was also supported by recognizing its relationship to the mechanism of neural sharpening. The neural sharpening model claims that with experience, the representation of any event becomes sparser, and therefore more typical events will elicit lower overall activity than atypical ones (see Hoffman and Logothetis, 2009). Long-term norm-based neural coding for faces was then demonstrated with fMRI along these lines in both adults (Loffler et al., 2005) and four-to-six-year-old children (Jeffery et al., 2010). Norm-based coding was also found in face-responsive neurons in macaques (Leopold et al., 2006).

Long-term norm-based coding is much studied and received considerable support in the visual but not in the auditory domain, although there are some fMRI studies that indicate reduced neural activity for spoken stimuli that are more typical within some object space (Myers, 2007; Belizaire et al., 2007). There is now also behavioural support for a typicality-based representation of voices in long-term memory (Papcun et al., 1989; Bruckert et al., 2010; Mullennix et al., 2011; Latinus and Belin, 2011), but long-term norm-based neural codes for voice identities, similar to that found for faces, have not yet been

found. This thesis will present fMRI experiments that aimed to find out whether the neural coding of voices is indeed based on long-term norms.

Supported by a growing body of evidence, norm-based coding has become an important model in the research of perceptual space codes. But a vast majority of this evidence comes from the visual domain, especially from face perception. This thesis makes an attempt to identify and characterize norm-based codes in the auditory domain. Specifically, my thesis investigates norm-based coding in voice identity processing. These studies will search for evidence of norm-based codes for voice identity categories, with special attention to acceptance ranges and anchor points. Perhaps the best time to investigate a category is when it is being formed. On top of that, it is best to investigate well-defined categories, for example categories that are formed via explicit feedback. Voice categories will be observed here as they are formed, during and after voice identity training.

Learning and re-learning

We meet new people every day, so our perceptual system must cope with learning new person identities every day. But the people we already know also change (they have a new haircut, have a cold, or talk to us in a different language etc.). Therefore, we must also be able to cope with re-learning old person identities. While learning a new person identity, new voice categories are being formed. When re-learning a person identity, the corresponding representational spaces have to be adjusted. Norms for the corresponding representational spaces have to be calculated and then constantly re-calculated, to adhere to the actual sensory history of the perceiver. Norm-based coding thus has to be adaptive, to accommodate dynamically changing cues.

Evidence for perceptual learning in speech demonstrates listeners' constant readiness to update their representations for more efficient processing of incoming stimuli. Norris et al. (2003) demonstrated that listeners dynamically adjust phonemic representations, for example by expanding phonemic categories, to reflect the speech they hear. Similarly, Allen and Miller (2004) showed that listeners can learn talker-specific phonetic information (specifically: voice onset time) and this information can generalize to a novel word. Eisner and McQueen (2005) and Kraljic and Samuel (2007) presented evidence that perceptual learning might happen on different levels of abstraction, also depending on

the properties of the speech sounds. Finally, Pardo (2006) demonstrated the social validity of dynamic retuning of speech sounds by showing that social interaction increases the similarity of vowel spaces of the interacting speakers. Norm-based coding has also been shown to be adaptive for faces (Rhodes and Jeffery, 2006), and, as shown with fMRI, for other visual objects as well (Panis et al., 2011).

Mean or prototypical values in a representational space are therefore expected to change with experience. For example, teenage boys' voices deepen with vocal fold maturation, but their classmates remain able to identify them on the phone. Nevertheless, if the rules of calculating these prototypical values are the same for everyone, then little variation should be found for this prototype across perceivers with a very similar sensory history. For example, all classmates of that boy should agree along the years spent together if a certain utterance of the boy is typical or odd. Little is known about how flexible voice representations are, and how stable is a voice's perceived typicality across the population. These questions will also be investigated in this thesis. Also, if norm-based neural codes exist for voice identities, then they are expected to be adaptive and change dynamically with experience. This assumption will be used when designing fMRI experiments searching for norm-based codes of newly-learned voice identities.

Voice identities

Voices seem to be a really good choice when investigating auditory category formation. Voice signals have a special status in the auditory world. Not only are they one of the most often heard and one of the most complex of auditory stimuli, but they also carry speech, and distinctive information about the identity of the auditory source, the speaker. In this section I present results supporting the claim that voice identities are natural auditory objects: results from monkey and infant research, results about remembering voices and the specific impairments of voice-memory, and results showing that there are regions in the human brain that process voices selectively, taking them as auditory faces (Belin et al., 2004) that mediate person recognition.

Throughout this thesis, the term 'voice' is meant to refer to auditory percepts of vocalizations of human individuals. So 'voice' simply means the perceived vocal signal. Importantly, the use of this term is not restricted to cases where the corresponding vocal

signals are voiced speech sounds in contrast with voiceless speech sounds, or to spoken utterances in contrast with nonspeech vocalizations; but it is restricted to human in contrast with animal vocalizations, and to vocal in contrast to nonvocal auditory events (i.e., sounds produced without the involvement of the vocal tract). This use of the term 'voice' conforms to a growing body of literature studying the behavioural and neuroscientific aspects of human vocalization processing (e.g., Belin et al., 2004). In line with this, the term 'voice identity' is meant to refer to the voice-based percepts of person identity. It can be thought of as an analogue of the term 'face identity', as used extensively in the visual person identification literature (e.g., Calder and Young, 2005). 'Voice identity', in contrast with what the terms speaker and talker would perhaps imply, is thus thought of as a perceptual entity referring to the vocalizing person, rather than the vocalizing person him/herself. Voice identity information then simply means information in the vocal signal that is used to identify the vocalizing person (i.e., the speaker). Similarly, voice identity representations are representations that encode voice identity information; and voice identity learning means learning to recognize the vocalizing person using voice identity information; and voice identity recognition, or simply voice recognition, is a synonym for speaker (or, more precisely: vocalizer) recognition. The term talker is often used in the cognitive literature (e.g., Nygaard and Pisoni, 1998) to refer to the vocalizing person (i.e., the speaker), and will be used in this thesis interchangeably with speaker.

Undoubtedly, there is a great selective pressure motivated by social interactions to be tuned in to voice identity information. Indeed, human adult listeners use voices very efficiently for person recognition (e.g., Schweinberger et al., 1997). We can remember voices, even unfamiliar ones, with a very high accuracy, and for a long time (e.g., Papcun et al., 1989). This ability to recognize voices appeared much earlier than speech, both phylogenetically and ontogenetically: it is not unique to humans and is there from a very young age on. Rhesus monkeys are able to identify their conspecifics based on their vocalizations (Rendall et al., 1998). Newborns prefer their mothers' voices (DeCasper and Fifer, 1980), and 7-month-olds are highly skilled at voice discrimination, especially in their native language (Johnson et al., 2011). Furthermore, voice processing abilities can be impaired selectively. Van Lancker et al. (1989) reported a neuropsychological patient who had normal hearing and normal memory abilities but was unable to remember and recognize voices. The authors referred to this disability as phonagnosia (cf. Garrido et al.,

2009). In another neurophysiological study, Schacter et al. (1995) found that voice-specific auditory priming may depend on a memory system that is impaired in amnesia.

Voices also have a special status in the brain. Using functional magnetic resonance imaging (fMRI), Belin et al. (2000) demonstrated that there are cortical regions along the bilateral superior temporal sulcus (STS) that respond selectively to voices. That is, in this region of the brain voice signals elicit increased neural activity compared to non-vocal sounds. Since this milestone-study from Belin and colleagues, the findings presented in that paper were replicated and confirmed several times (von Kriegstein et al., 2003; Lattner et al., 2003; Grandjean et al., 2005; Ethofer et al., 2009; see Belin et al., 2011 for a review). Furthermore, voice-selective temporal regions were recently found in macaque monkeys using both fMRI and intracranial recordings (Ghazanfar et al., 2005; Petkov et al., 2008; Perrodin et al, 2011; Joly et al., 2012) and in infants using near-infrared spectroscopy (Grossmann et al., 2010), but cortical voice-selectivity was shown to be impaired in autism (Gervais et al., 2004). Cortical regions outside the temporal lobes, in the inferior frontal cortex (IFC) were also found to be voice-sensitive in both monkeys (Romanski and Goldman-Rakic, 2002; Romanski et al., 2005) and humans (Fecteau et al., 2005; von Kriegstein and Giraud, 2006).

Evidence for an early interaction and direct information sharing of face and voice processing regions is shown by fMRI functional connectivity of the face-selective fusiform face area (FFA) and the voice-selective STS (von Kriegstein et al., 2005), and by direct structural connections between FFA and STS regions using probabilistic tractography (Blank et al., 2011). Furthermore, Ghazanfar et al. (2005) showed in rhesus monkeys that the primate auditory cortex integrates facial and vocal signals through local field potentials in core and lateral belt regions. These findings further strengthen the claim that the primary reason why the human brain maintains voice-selective regions is to serve, together with face-selective areas, the ultimate goal of person recognition.

Human voices are most typically heard as speech. When listening to speech, we typically not only want to know who speaks but also what is said. The parallel presence of the goals of person recognition and speech recognition necessarily leads to interactions between person processing and speech processing. So when thinking about auditory person recognition, one must also consider the influence of speech perception. How distinct or common these processes are? On the one hand, these processes are clearly separate. The

influence of talker-specific detail on the performance of the listeners was repeatedly demonstrated in the last decades (e.g., Mullennix and Pisoni, 1990; Nygaard and Pisoni, 1998; Goh, 2005; for reviews on this, see Goldinger, 1998 and McQueen et al., 2006). There is neurophysiological and neuroimaging evidence for separate voice identity and speech processing mechanisms. Vongphoe and Zeng (2005) found that listeners with cochlear implants can perform well in a vowel recognition task, but had difficulties with the same stimuli in a talker recognition task. Belin and Zatorre (2003) used adaptation-fMRI to investigate which cortical regions adapt to syllable repetition and which ones to voice identity repetition. They found separate cortical regions, with a role of the right anterior STS in voice identity change detection. On the other hand, these processes are not independent. Both behavioural and neuroimaging studies found evidence for an early interaction of voice identity processing and speech processing. Lachs and Pisoni (2004) asked subjects to match visual and auditory displays of acoustically transformed speech based on the identity of the speaker and found that the acoustic signal of speech simultaneously and in parallel carries articulatory information about both the linguistic message and indexical properties of the talker. In perception experiments which used sinewave replicas of natural speech to eliminate natural voice quality and dramatically reduce non-segmental acoustic information (such as the fundamental frequency information) while preserving idiosyncratic segmental variation, Remez et al. showed that talker identification is possible on the basis of phonetic information only (Fellowes et al., 1997; Remez et al., 1997). Experiments using MEG and fMRI demonstrated the early parallel extraction of phonetic and identity information from the voice signal in the auditory cortex, and found an interaction of the processes already at preattentive perceptual stages (Knösche et al., 2002; Lattner et al., 2005). These findings indicate that separate mechanisms may underlie voice identity processing and phonetic information processing, and that these mechanisms involve multiple levels of abstraction. Although the levels of interaction are not well-established yet, it seems that these parallel processes begin to interact already in an early phase. This thesis further explores these questions from the person identification angle. Do we use the same acoustic cues for person and phoneme identification? Does phonetic content influence voice identity processing? What is the contribution of segmental and non-segmental cues to voice identity category formation?

An 'auditory face' model of cerebral voice processing was proposed by Belin et al. (2004), extending Bruce and Young's (1986) face processing model, and also building on earlier findings suggesting distinct acoustic, unimodal and multimodal steps in person identification (Ellis, 1989; Burton et al., 1990; Ellis et al., 1997; Neuner and Schweinberger, 2000). This model proposes that during vocal information processing, a general low-level auditory analysis is followed by a voice-specific structural analysis, which in turn is followed by partially dissociable functional pathways for the analysis of speech content, affective content and voice identity information, leading to the activation of unimodal voice recognition units and multimodal person identity nodes. The model proposed by Belin and colleagues (2004) offered a useful framework to study voice processing. It has been suggested that the acoustic analysis of voices is supported by mainly posterior STS regions (Belin et al., 2000; Belin et al., 2002; von Kriegstein et al., 2003), and that a more categorical level of voice identity processing might involve distinct, right anterior regions of the STS (Nakamura et al., 2001; Belin and Zatorre, 2003; von Kriegstein and Giraud, 2004; Sokhi et al., 2005). But the interpretation of these findings has often been difficult: indeed, in many of these studies, the differential response patterns of the proposed voice processing stages could be explained by between-test acoustic changes (Belin et al., 2000, 2002; Belin and Zatorre, 2003), task changes (von Kriegstein et al., 2003) or both (von Kriegstein and Giraud, 2004). Therefore, it is important to see if these differential response patterns for different processing stages persist in a setup that carefully controls for both acoustic and task changes. In this thesis I will present two fMRI experiments that do exactly that. Furthermore, the exact role of the representational stages along the cortical hierarchy of voice identity processing has remained unclear. One reason for that is that to date, very few neuroimaging studies attempted to characterize the neural coding mechanisms of voice recognition. This thesis will test the hypothesis that voice processing mechanisms make use of norm-based (neural) coding on multiple levels of abstraction: for example, on a voice-acoustic level and on a more abstract voice identity level. And if so, is the formation of norm-based voice categories affected by varying phonetic content?

This thesis

My thesis reports experiments that investigated voice identity category learning. How do we distinguish and how do we learn new voices? How are voice identity categories formed? On what levels of abstraction can we find evidence for norm-based coding in voice processing? How are these different levels of abstractions represented in the human brain? How does speech content influence voice identity processing? What factors determine acceptance ranges for voice identity category size? Is built-in category structure information present in the speech signal? What happens when within-category variation is larger than typical within-talker variation? To examine these questions, a variety of research tools were used including a voice pool, a sound morphing technique, a learning and re-learning paradigm and sparse-sampling fMRI.

A pool of thirteen voices was created using high-quality recording. Each talker said the same eight words ten times, and read a list of sentences and short stories. The stimuli in almost all experiments (Chapters 2, 3, 4 and 6) were then selected from this voice pool.

Perceptually relevant within-talker and across-talker variation have been claimed to be based on essentially the same acoustic cues (Potter and Steinberg, 1950; Nolan et al., 1997; Benzeghiba et al., 2007), so natural within-talker variability can be modeled by voice morph stimuli created across voices. It has been argued that possible auditory cues (those with well detectable across-event variation) are restricted to the frequency and time domains (Kubovy and Van Valkenburg, 2001). To systematically manipulate both frequency and time parameters of the test voices, a special sound morphing technique, STRAIGHT (Kawahara, 2006), was used in all training experiments (Chapters 3 to 6).

A learning and re-learning paradigm was used in three of the five studies (Chapters 3, 5 and 6): the basic idea here is that the same participant is trained in multiple sessions to categorize voice identities on a voice morph continuum, but the voice identity category changes across sessions and the listener is kept unaware of this change. This way the categorical properties of a voice stimulus could be varied without adding an acoustic bias to the design. This paradigm made it possible to test the flexibility of voice learning in various settings, and to investigate multiple levels of abstractions in parallel (e.g., supra-individual and intra-individual levels).

Multiple levels of abstractions need multiple levels of representations. Multiple parallel processes are very difficult to trace down with button press measures, because normally we have only a single dependent measure of information processing. This is where brain research tools can help. In two voice learning experiments presented here (Chapters 5 and 6), fMRI was used to measure all brain regions' activity at once, continuously. Using fMRI in auditory experiments is not trivial, because measurements are very loud. Special sparse imaging techniques were applied here to enable stimulus presentation in silence but to allow for enough (noisy) measurements. Norm-based coding was then searched for on different levels of neural abstraction.

This thesis is organized as follows. Chapter 2 describes a voice discrimination experiment that explored the voice pool. Same or different responses were collected for all possible pairs of thirteen voices, for eight monosyllables. This experiment investigates if voice discrimination performance is influenced by phonetic content, and if there are voices consistently perceived as prototypical or atypical. A further goal here was to define a multidimensional voice space from behavioural measures of voice distances and relate this to a space based on acoustic measurements.

Chapters 3 and 4 report behavioural experiments on voice identity learning. These training studies asked if and how explicitly trained (Chapter 3) and implicitly-learned (Chapter 4) voice identity categories are shaped by phonetic content. These experiments used button press measures and made use of across-voice sound morphing. More specifically, the experiments in Chapter 3 examined the degree of flexibility in voice identity learning, investigated the role of segmental and non-segmental cues in the formation of voice identity categories, and tested if voice learning entails abstraction. Experiment 1 of Chapter 3 applied a learning and re-learning paradigm. Participants were trained to categorize stimuli on voice A to voice B continua as one of the voices, but with different identity boundaries in different sessions, using two words and two talkers from the voice pool, based on the voice discriminability results of the experiment in the previous chapter. Then, Experiment 2 of Chapter 3 used the same continua but trained listeners to perceive different individual voice categories in an A or not-A paradigm. Chapter 4 further tested the limits of voice category formation. Here, two words from four talkers were used, and participants were trained to categorize two voice groups composited from two individual voices each. An additional question here was if any voice category size can be represented

in a norm-based space, and if the acceptance range of a voice category can vary with phonetic content.

Chapters 5 and 6 present multisession fMRI experiments investigating norm-based coding for voices. These studies combined button press measures at training and test with measures of haemodynamic activity. The category learning and re-learning paradigm tested in Chapter 3 was used again here to manipulate across-talker and within-talker typicality patterns separately in a within-participant design. The main aim here was to characterize neural coding mechanisms of voice identity processing on different levels of abstraction, by exploiting brain plasticity. Critical comparisons in these tests focused on short-term and long-term, supra-individual and intra-individual similarity spaces. Voice identity training sessions were first based on voice A or not-A categorizations (Chapter 5), and then on voice A or voice B categorizations (Chapter 6). Chapter 5 used monosyllabic (consonant-vowel) stimuli from female speakers of Hungarian. Chapter 6 used stimuli from male Dutch speakers who were selected from the voice pool and pre-tested in Chapter 3. Chapter 6 focused on category-selective regions in the STS and the IFC.

Chapter 7 provides a summary of the most important findings and discusses them in the context of norm-based coding.

References

Aguirre, G.K., 2007. Continuous carry-over designs for fMRI. NeuroImage 35, 1480-1494.

Allen, J.S., Miller, J.L., 2004. Listener sensitivity to individual talker differences in voice-onset time. Journal of the Acoustical Society of America 115, 3171-3183.

Anderson, B.L, Khang, B.G., 2010. The role of scission in the perception of color and opacity. Journal of Vision 10(5), 1-16.

Belin, P., Bestelmeyer, P.E.G., Latinus, M., Watson, R., 2011. Understanding voice perception. British Journal of Psychology 102, 711-725.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences 8, 129-135.

Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. NeuroReport 14, 2105-2109.

Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal lobe responses to vocal sounds. Cognitive Brain Research 13, 17-26.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309-312.

Belizaire, G., Fillion-Bilodeau, S., Chartrand, J.P., Bertrand-Gauvin, C., Belin, P., 2007. Cerebral response to 'voiceness': a functional magnetic resonance imaging study. NeuroReport 18, 29-33.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., 2007. Automatic speech recognition and speech variability: A review. Speech Communication 49(10-11), 763-786.

Blank, H., Anwander, A., von Kriegstein, K., 2011. Direct structural connections between voice- and face-recognition areas. The Journal of Neuroscience 31(36), 12906-15.

Bruce, V., Young, A., 1986. Understanding face recognition. British Journal of Psychology 77, 305-327.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G.A., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Current Biology 20, 116-120.

Burton, A.M., Bruce, V., Johnston, R.A., 1990. Understanding face recognition with an interactive activation model. British Journal of Psychology 81, 361-80.

Cabeza, R., Bruce, V., Kato, T., Oda, M., 1999. Prototype effect in face recognition: Extension and limits. Memory and Cognition 27, 139-151.

Calder, A.J., Young, A.W., 2005. Understanding facial identity and facial expression recognition. Nature Neuroscience Reviews 6(8), 641-653.

DeCasper, A.J., Fifer, W.P., 1980. Of human bonding: newborns prefer their mothers' voice. Science 208, 1174-6.

Eisner, F., McQueen, J.M., 2005. The specificity of perceptual learning in speech processing. Perception & Psychophysics 67(2), 224-238.

Ellis, A.W., Young, A.W., Critchley, E.M.R., 1989. Loss of memory for people following temporal lobe damage. Brain 112, 1469-1483.

Ellis, H.D., Jones, D.M., Mosdell, N., 1997. Intra- and inter-modal repetition priming of familiar faces and voices. British Journal of Psychology 88(1), 143-156.

Epstein, R.A., Parker, W.E., Feiler, A.M., 2008. Two kinds of fMRI repetition suppression? Evidence for dissociable neural mechanisms. Journal of Neurophysiology 99, 2877-2886.

Ethofer, T., Van De Ville, D., Scherer, K., Vuilleumier, P. 2009. Decoding of emotional information in voice-sensitive cortices. Current Biology 19, 1028–1033.

Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2005. Sensitivity to voice in human prefrontal cortex. Journal of Neurophysiology 94, 2251-2254.

Fellowes, J.M., Remez, R.E., Rubin, P.E., 1997. Perceiving the sex and identity of a talker without natural vocal timbre. Perception & Psychophysics 59, 839-849.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., Duchaine, B., 2009. Developmental phonagnosia: A selective deficit of vocal identity recognition. Neuropsychologia 47, 123-131.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthelemy, C., Brunelle, F., Samson, Y., Zilbovicius, M., 2004. Abnormal cortical voice processing in autism. Nature Neuroscience 7(8), 801-802.

Ghazanfar, A.A., Maier, J.X., Hoffman, K.L., Logothetis, N.K., 2005. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. The Journal of Neuroscience 25(20), 5004-5012.

Goh, W.D., 2005. Talker variability and recognition memory: instance-specific and voice-specific effects. Journal of Experimental Psychology: Learning, Memory, and Cognition 31(1), 40-53.

Goldinger, S.D., 1998. Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105, 251-279.

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., Vuilleumier, P., 2005. The voices of wrath: brain responses to angry prosody in meaningless speech. Nature Neuroscience 8(2), 145-146.

Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends in Cognitive Sciences 10, 14-23.

Grossmann, T., Oberecker, R., Koch, S.P., Friederici, A.D., 2010. Developmental origins of voice processing in the human brain. Neuron 65, 852-858.

Hoffman, K.L., Logothetis, N.K., 2009. Corical mechanisms of sensory learning and object recognition. Philosophical Transactions of the Royal Society B 364, 321-329.

Jeffery, L., McKone, E., Haynes, R., Firth, E., Pellicano, E., Rhodes, G., 2010. Four-to-six-year-old children use norm-based coding in face-space. Journal of Vision 10(5), 18.

Johnson, E.K., Westrek, E., Nazzi, T., Cutler, A., 2011. Infant ability to tell voices apart rests on language experience. Developmental Science 14(5), 1002-1011.

Joly, O., Pallier, C., Ramus, F., Pressnitzer, D., Vanduffel, W., Orban, G.A., 2012. Processing of vocalizations in humans and monkeys: A comparative fMRI study. NeuroImage 62, 1376-1389.

Kawahara, H., 2006. STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoustic Science and Technology 27(6), 349-353.

Knösche, T.R., Lattner, S., Maess, B., Schauer, M., Friederici, A.D., 2002. Early parallel processing of auditory word and voice information. Neuroimage 17(3), 1493-1503.

Kraljic, T., Samuel, A.G., 2007. Perceptual adjustments to multiple speakers. Journal of Memory and Language 56, 1-15.

Kubovy, M., Van Valkenburg, D. ,2001. Auditory and visual objects. Cognition 80, 97–126.

Lachs, L., Pisoni, D.B., 2004. Cross-modal source information and spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance 30(2), 378-396.

Latinus, M., Belin, P., 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. Frontiers in Psychology 2, 1-12.

Lattner, S., Maess, B., Wang, Y., Schauer, M., Alter, K., Friederici, A.D., 2003. Dissociation of human and computer voices in the brain: evidence for a preattentive Gestalt-like perception. Human Brain Mapping 20, 13-20.

Leopold, D.A., Bondar, I.V., Giese, M.A., 2006. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. Nature 442, 572-575.

Leopold, D.A., O'Toole, A.J., Vetter, T., Blanz, V., 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. Nature Neuroscience 4, 89-94.

Loffler, G., Yourganov, G., Wilkinson, F., Wilson, H.R., 2005. FMRI evidence for the neural representation of faces. Nature Neuroscience 8(10), 1386-1390.

McQueen, J.M., Cutler, A., Norris, D., 2006. Phonological abstraction in the mental lexicon. Cognitive Science 30(6), 1113-1126.

Mullennix, J.W., Pisoni, D.B., 1990. Stimulus variability and processing dependencies in speech perception. Perception & Psychophysics 47, 379-390.

Mullennix, J.W., Ross, A., Smith, C., Kuykendall, K., Conard, J., Barb, S., 2011. Typicality effects on memory for voice: Implications for earwitness testimony. Applied Cognitive Psychology 25, 29-34.

Myers, E.B., 2007. Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: an fMRI investigation. Neuropsychologia 45, 1463-1473.

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura A., Hatan, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., Kojima, S., 2001. Neural substrates for recognition of familiar voices: a PET study. Neuropsychologia 39, 1047-1054.

Neuner, F., Schweinberger, S.R., 2000. Neuropsychological impairments in the recognition of faces, voices, and personal names. Brain and Cognition 44(3), 342-366.

Nolan, F., 1997. Speaker recognition and forensic phonetics. In: Hardcastle, W., Laver, J. (Eds.), A Handbook of Phonetic Science. Oxford: Blackwell, 744-766.

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cognitive Psychology 47(2), 204-238.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Perception & Psychophysics 60, 355-376.

Panis, S., Wagemans, J., Op de Beeck, H.P., 2011. Dynamic norm-based encoding for unfamiliar shapes in human visual cortex. Journal of Cognitive Neuroscience 23, 1829-1843.

Papcun, G., Kreiman, J., Davis, A., 1989. Long-term memory for unfamiliar voices. Journal of the Acoustical Society of America 85, 913-925.

Pardo, J.S., 2006. On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America 119(4), 2382-2393.

Perrodin, C., Kayser, C., Logothetis, N.K., Petkov, C.I., 2011. Voice cells in the primate temporal lobe. Current Biology 21(16), 1408-15.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nature Neuroscience 11(3), 367-374.

Potter, R.K., Steinberg, J.C., 1950. Toward the specification of speech. Journal of the Acoustical Society of America 22, 807-820.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance 23, 651-666.

Rendall, D., Owren, M.J., Rodman, P.S., 1998. The role of vocal tract filtering in identity cueing in rhesus monkey (Macaca mulatta) vocalizations. Journal of the Acoustical Society of America 103, 602-614.

Rhodes, G., 1996. Superportraits: Caricatures and recognition. Hove: The Psychological Press.

Rhodes, G., Jeffery, L., 2006. Adaptive norm-based coding of facial identity. Vision Research, 46, 2977-2987.

Romanski, L.M., Averbeck, B.B., Diltz, M., 2005. Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. Journal of Neurophysiology 93, 734-747.

Romanski, L.M., Goldman-Rakic, P.S., 2002. An auditory domain in primate prefrontal cortex. Nature Neuroscience 5, 15-16.

Schacter, D.L., Church, B., Bolton, E., 1995. Implicit memory in amnesic patients: impairment of voice-specific priming. Psychological Science 6, 20-25.

Schweinberger, S.R., Herholz, A., Sommer, W., 1997. Recognizing famous voices: influence of stimulus duration and different types of retrieval cues. Journal of Speech, Language and Hearing Research 40, 453-463.

Sokhi, D.S., Hunter, M.D., Wilkinson, I.D., Woodruff, P.W.R., 2005. Male and female voices activate distinct regions in the male brain. NeuroImage 27, 572-578.

Valentine, T., 1991. A unified account of the effects of distinctiveness, inversion, and race in face recognition. Quarterly Journal of Experimental Psychology A: Human Experimental Psychology 43, 161-204.

Van Lancker, D., Kreiman, J., Cummings, J., 1989. Voice perception deficits: Neuronanatomical correlates of phonagnosia. Journal of Clinical and Experimental Neuropsychology 11, 665-674.

von Kriegstein, K., Eger, E., Kleinschmidt, A., Giraud, A.L., 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Cognitive Brain Research 17, 48-55.

von Kriegstein, K., Giraud, A.L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. NeuroImage 22, 948-955.

von Kriegstein, K., Giraud, A.L., 2006. Implicit multisensory associations influence voice recognition. PLoS Biology 4(10), e326.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., Giraud, A.L., 2005. Interaction of face and voice areas during speaker recognition. Journal of Cognitive Neuroscience 17, 367-376.

Vongphoe, M., Zeng, F. G., 2005. Speaker recognition with temporal cues in acoustic and electric hearing. Journal of the Acoustical Society of America 118(2), 1055-1061.

Chapter 2

Phonetic content influences voice discriminability

Abstract

We present results from an experiment which shows that voice perception is influenced by the phonetic content of speech. Dutch listeners were presented with thirteen speakers pronouncing CVC words with systematically varying segmental content, and they had to discriminate the speakers' voices. Results show that certain segments help listeners discriminate voices more than other segments do. Voice information can be extracted from every segmental position of a monosyllabic word and is processed rapidly. We also show that although relative discriminability within a closed set of voices appears to be a stable property of a voice, it is also influenced by segmental cues – that is, perceived uniqueness of a voice depends on what that voice says.

A version of this paper appeared as Andics, A., McQueen, J. M., Van Turennout, M. (2007). Phonetic content influences voice discriminability. In J. Trouvain, & W. J. Barry (Eds.), Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007) (pp. 1829-1832). Dudweiler: Pirrot.

Introduction

Behavioural and neuroscientific studies indicate that voice processing and speech processing are partly independent, but interact at an early stage of processing (e.g., Knösche et al., 2002). One example of this interaction is the demonstration of early voice-specific effects on fricative perception (Eisner and McQueen, 2005; Kraljic and Samuel, 2007). But the other direction of the interaction – whether voice-specific segmental information contributes to voice processing – has been studied less extensively. Remez et al. (1997), using sinewave replicas of speech, demonstrated that speaker-specific phonetic information can in certain cases be sufficient for talker identification. But does segmental information contribute to the efficiency of discrimination of natural voices?

We investigated possible segmental effects on voice discrimination from the listener's perspective and from the speaker's perspective. First, we explored whether phonetic content influences the voice discrimination performance of listeners. Second, we examined whether segmental cues influence the relative discriminability of different voices. One can find a voice that is more or less distinguishable from other voices, but does this depend on what words the voices say?

These questions were addressed in a voice discrimination experiment. Dutch listeners were presented with a list of Dutch CVC words, spoken by Dutch speakers, and were asked to decide whether each word was spoken in the same or a different voice as the preceding word. Segmental content was controlled using eight words which were made by factorially combining two onset consonants, two vowels, and two coda consonants.

Method

Participants

Twelve native Dutch listeners with no known hearing disorders participated.

Stimuli

Thirteen speakers were chosen. To reduce non-segmental (e.g., fundamental frequency) variability of the voices, the speakers were selected from a relatively homogenous group: young male non-smoking native speakers of Dutch with no

recognizable regional accents and no speech problems (age range: 18-30). Segmental overlap between the words was systematically varied using the words met [mɛt], mes [mɛs], mot [mɔt], mos [mɔs], let [lɛt], les [lɛs], lot [lɔt] and los [lɔs]. The recordings were sampled at 44100 Hz, 16 bits per sample. Average amplitude was equalized over all stimuli. Average syllable duration was 565 ms.

Procedure

Stimuli were presented via headphones binaurally, at a standard, comfortable listening level. To make the task harder, stimuli followed each other at a relatively fast pace (2400 ms between syllable onsets), and a pink noise was presented after each syllable (from 600 ms till 2400 ms after every syllable onset).

Subjects were instructed to listen to two-minute long blocks of these CVC words. A same/different forced-choice one-back task was used. Listeners had to decide whether the word they heard was pronounced by the same voice as the preceding word or by a different voice. That is, listeners had to make a decision after every syllable they heard, except for the first one within each block. Assignment of left and right index fingers to same and different buttons was balanced across subjects. The experiment lasted 51 minutes, excluding a short practice session and self-paced breaks between blocks.

Design

Stimulus presentation was blocked by word, so within one block only one of the eight words appeared. One block consisted of 53 stimuli (that is, 52 comparisons), and there were 24 such blocks. Every listener heard all possible voice pairings for each of the eight words during the experiment. To balance response biases as much as possible, half of the voice comparisons required a “same” response and half of them a “different” response. To achieve that equal distribution, every same-voice pair was presented six times per word, and every different-voice pair was presented exactly once per word. There were at most three same or different pairs in a row. To ensure that responses were based on voice processing rather than auditory change detection, six different utterances of each word from each speaker were used, each of these utterances appeared only twice during the experiment, and these two identical stimuli were always separated by at least one full block.

Stimulus ordering was otherwise random and varied across listeners. Altogether 1248 responses were collected per listener.

Results

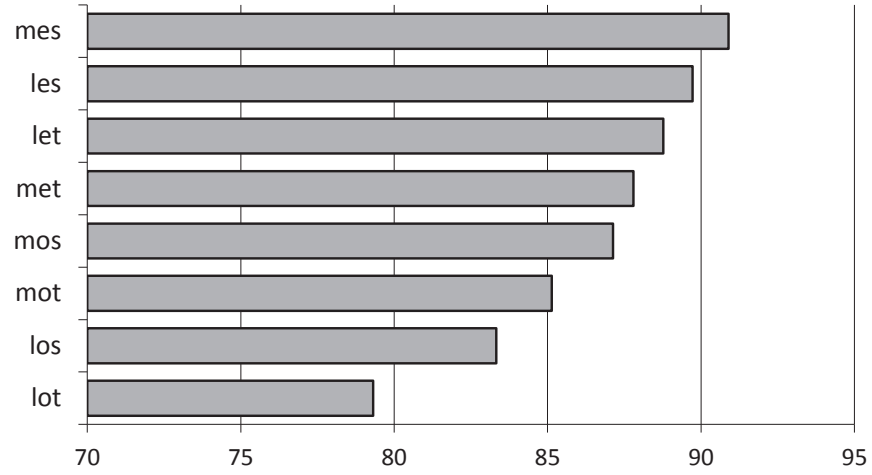
Overall performance

Overall proportion of correct responses was 87.2%, with a similarly high proportion for same-voice pairs (88%) and different-voice pairs (86.5%). Individual overall hit rates varied between 78.7% and 94.7%, ranging from a responder with a strong "same" bias (98.6% for same-voice pairs and 60.1% for different voice-pairs) to a responder with a clear "different" bias (70.1% for same-voice pairs and 98.6% for different-voice pairs). This listener bias was independent of phonetic content. Average response time was 799 ms for same-voice pairs and 855 ms for different-voice pairs.

Hit proportion per word

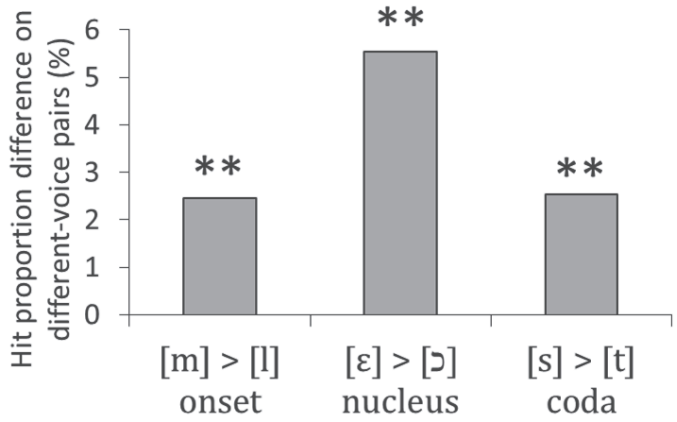
Phonetic contributions to voice discrimination performance were investigated by comparing responses for each word. There were differences in the hit proportion of responses to different-voice pairs between words (see Fig. 1), ranging from 79.3% for [lɔt] to 90.9% for [mɛs].

Fig. 1. Same or different voice? Hit proportion of responses to different-voice pairs per word (% correct).



The nature of the CVC stimuli made it possible to examine this word effect at the segmental level (i.e., segmental contributions to voice discrimination) in 2 x 2 x 2 repeated-measures ANOVAs with the factors onset position, nucleus position and coda position, on hit proportions for different-voice pairs and same-voice pairs separately. For different-voice pairs, we found a main effect for each segmental position (onset/nucleus/coda: $F(1,11) = 16.010/16.319/12.607$, $p = .002/.002/.005$), showing a benefit of [m] in onset position, [ɛ] in nucleus position and [s] in coda position over [l], [ɔ] and [t] respectively. For same-voice pairs, we found a main effect for the onset and nucleus, but not for the coda position (onset/nucleus/coda: $F(1,11) = 6.936/30.018/2.385$, $p = .023/.000/.151$), with benefits in the same directions as for different-voice pairs. Note that the effect size is largest for the nucleus position (see Fig. 2).

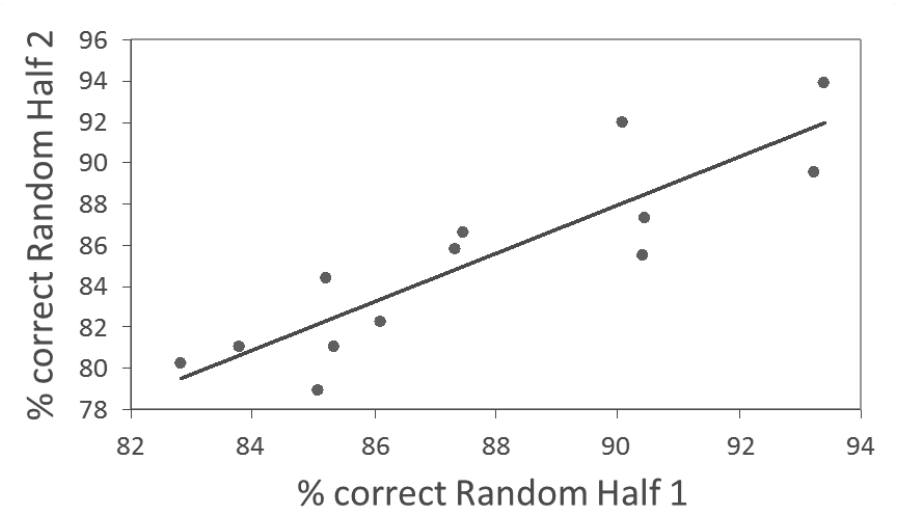
Fig. 2. Segmental contribution to voice discrimination performance.



Hit proportion per voice

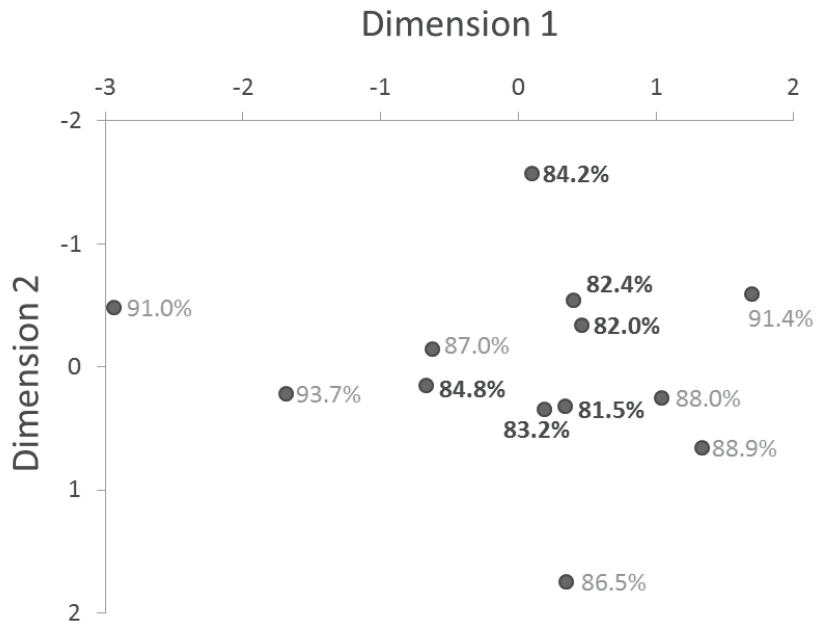
Discriminability of a voice was investigated by comparing the hit proportion of responses to different-voice pairs for each voice. This measure was calculated by collapsing different-voice trials for each voice across all pairs in which that voice was a member. This way we gained a perceptual rating of the thirteen voices, ranging from the voice which was the most difficult to distinguish from the rest (81.5% correct) to the voice which was the most easily discriminable from the other voices (93.7% correct). To check the reliability of this rating, the same perceptual measure was calculated after randomly splitting the listeners into two groups. Fig. 3 shows the high positive linear correlation of two ratings of voices based on data from these two random halves of the set of listeners ($r = + .883$, $p < .01$).

Fig. 3. Correlation of hit proportion per voice between two random halves of listeners (% correct).



Furthermore, hit proportions on same-voice and different-voice pairs were also found to be positively correlated ($r = + .66$, $p < .05$). This showed that utterances of voices that are less discriminable are also less identifiable, that is, they were perceived as the same voice less consistently than the utterances of more discriminable voices. This reduction of perceived consistency for less discriminable voices was not explained by acoustic differences: indeed, we found not smaller but greater within-speaker acoustic consistency for these less consistently perceived, less discriminable voices. For a similarity-based multidimensional scaling of all voices, perceptual distance was calculated as the proportion of hits for each voice pair (stress = .136, RSQ = .917, Fig. 4). Note that less discriminable voices are perceptually similar, and take a central position on the map.

Fig. 4. Multidimensional scaling of voices based on perceptual similarity. Each point represents a voice, data labels show discriminability (hit proportion) for each voice (black for the less discriminable, gray for the more discriminable halves of voices).

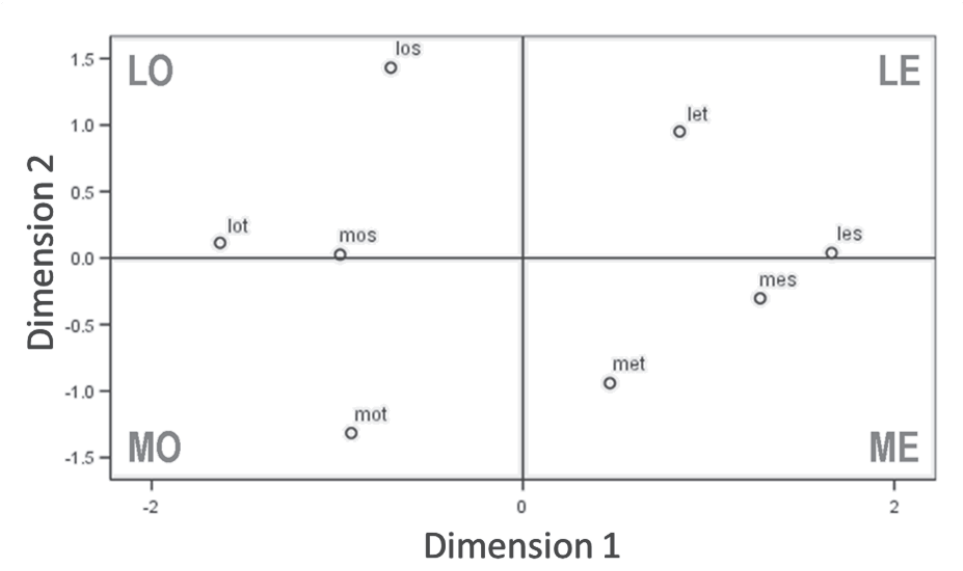


Word effects on voice discriminability

To investigate the possible effect of segmental cues on the perceived discriminability of a voice, the discriminability ratings of voices described above were also calculated separately for each word. The correlation coefficient of voice ratings for two given words was considered to be a proximity measure (the higher the correlation, the closer the ratings based on those words are). Inversion of this proximity measure results in a distance measure. Distances were calculated for every word pair (the smaller the distance, the closer the words are with respect to their contribution to voice discriminability). We then performed a multidimensional scaling of the words based on those distances (SPSS ALSCAL using a Euclidean distance model; stress = 0.098, RSQ = 0.919). Fig. 5 shows the resulting two-dimensional map. Note that dimension 1 of this map clearly distinguishes words with [ɛ] and with [ɔ] (right vs left side of the map), while dimension 2 distinguishes words with [m] and with [l] (lower vs upper part of the map). This is illustrated with the corresponding

onset-nucleus labels in the four corners of the map. That is, segmentally closer words are also closer perceptually. This suggests that voice discriminability is strongly determined by segmental properties.

Fig. 5. Multidimensional scaling of words based on the similarity of their effect on voice discriminability.



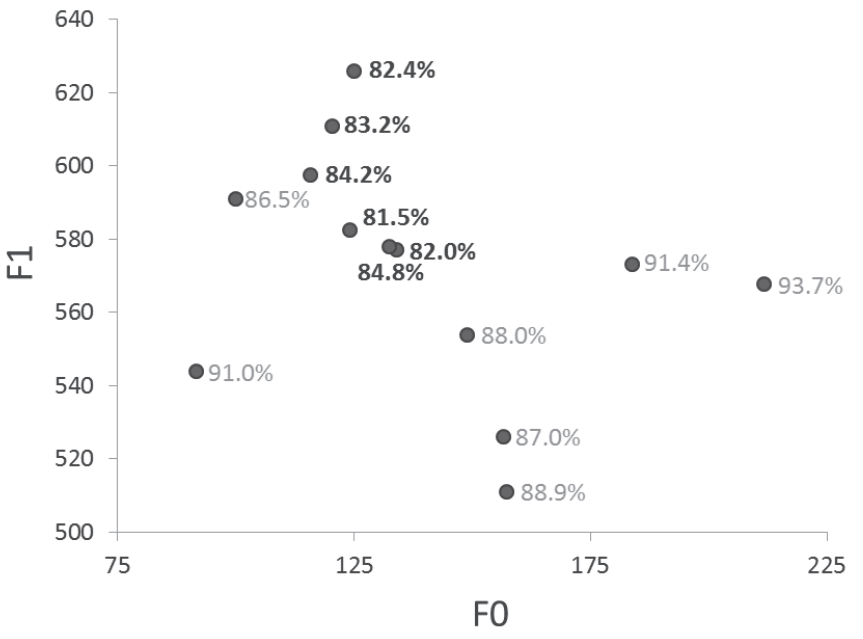
Acoustic measurements

We measured basic acoustic parameters of the segments [l] and [ɛ] in all /les/ tokens (13 speakers x 6 utterances): duration and F0, F1, F2 at segment midpoint. For the segment [ɛ], that listeners found to be more characteristic for voices than [ɔ], the across-talker mean of the across-token standard deviation (i.e., within-talker variation) was lower than the across-talker standard deviation of the across-token mean (i.e., across-talker variation) for F0, F1 and F2 as well (mean of SD, F0/1/2 = 10.050/17.264/32.841 Hz; SD of mean, F0/1/2 = 33.101/32.321/117.107 Hz), and for the segment [l], that listeners found to be less characteristic than [m], it was only so for F0, but the other way around for F1 and F2 (mean of SD, F0/1/2 = 10.289/251.357/274.400 Hz; SD of mean, F0/1/2 = 26.386/138.818/101.919 Hz).

Variation of mean and standard deviation of these acoustic parameters was investigated per speaker, between less versus more discriminable voices in independent t-

tests, on the vowel [ɛ]. These tests showed significant differences between voice groups in mean F1 (mean = 595.25 vs 552.44 Hz, t(11) = 3.138, p = .009), and in standard deviation of F0 (mean = 6.70 vs 12.92 Hz, t(11) = 3.138, p = .009) and F2 (mean = 25.14 vs 39.44 Hz, t(11) = 2.362, p = .038). The vowel [ɛ] of less discriminable voices had a higher F1, and acoustically more consistent F0 and F2 values. Fig. 5 displays a scatter plot of mean F0 and F1 values for the vowel [ɛ], one value per voice. Note the acoustic similarity of less discriminable voices, and the similarity between Fig. 4 (voice map based on perceptual distance) and Fig. 6 (voice map based on acoustic distance).

Fig. 6. A two-dimensional map of voices based on acoustic parameters (F0 and F1). Each point represents a voice, data labels show discriminability (hit proportion) for each voice (black for the less discriminable, gray for the more discriminable halves of voices).



Discussion

The naturalness of voice discrimination

Listeners were presented with blocks of voices uttering one of eight CVC words and they had to compare each words' vocal identity to that of the previously heard word. All listeners performed far above chance level. This indicates that voice discrimination is an extremely robust ability of human listeners that is readily applicable even in an attentionally demanding and unnatural task.

Interestingly, many listeners had a considerable response bias either for the "same" or for the "different" response, but this effect disappeared after collapsing data over all listeners. Therefore, this variability does not seem to be caused by an inherent biasing factor in the experimental design, but rather by individual variation in how conservative a given listener is when setting up categories for new voices.

Phonetic content influences voice discrimination performance

Phonetic contribution to listeners' performance was investigated by comparing the hit proportion of responses to different-voice pairs for each word. We found a higher proportion of correct voice discriminations for words containing an onset [m] versus [l], a vowel [ɛ] versus [ɔ] and finally a coda segment [s] versus [t]. These differences suggest that the phonetic content of speech affects the listener's voice discrimination performance, and this effect is not restricted to certain segmental positions within a CVC word.

Three important observations have to be made here. First, vowel change seems to make the greatest difference, since its effect is higher than the effect of any of the consonant changes, especially for same-voice pairs. This suggests that vowels may vary more than consonants in the amount of paralinguistic information that they can carry. Further research is required, however, to test whether the present results generalize to other vowel- and consonant-pairs.

Second, segmental variation in the coda position makes a significant difference to voice discrimination performance. This indicates that listeners do not always make their decisions based on the vowel or based on the first two segments only, but rather they use all segments of a word before making a "same voice" or "different voice" decision. If we

now put this result together with the listeners' average response times, we can see that vocal identity information extracted from the coda position is applied quite rapidly: the most acoustic energy of the coda segment is situated around 300-500 ms after syllable onset, and average response time for different-voice pairs is 855 ms, meaning that listeners are able to apply phonetic information to distinguish between voices in less than half a second.

Third, segmental cues that contributed more to voice discrimination performance (different-voice pairs), were also more helpful for voice identification (same-voice pairs). This suggests that although perceptually relevant within-talker and across-talker variation seem to be based on the same acoustic cues (Nolan et al., 1997), within-talker and across-talker acoustic variation might not be proportional. This claim is supported by our acoustic measurements: for a more characteristic segment, within-talker variation was lower than across-talker variation, while for a less characteristic segment, within-talker variation was in cases even higher than across-talker variation.

Discriminability is a stable property of a voice

By comparing proportion of responses to different-voice pairs across voices, we obtained discriminability ratings for every voice. The high correlation of these voice ratings suggest that discriminability, at least relative to other voices within a closed set, is a stable property of a voice. That is, a voice's discriminability rating is independent of individual listener's biases.

We also examined the correlation between hit proportions on same-voice and different-voice pairs. They showed that utterances of voices that are less discriminable are also less identifiable, that is, they were perceived as the same voice less consistently than the utterances of more discriminable voices. This reduction of perceived consistency for less discriminable voices was not explained by acoustic differences: indeed, we found not smaller but greater within-speaker acoustic consistence for these less consistently perceived, less discriminable voices.

We therefore suggest that the discriminability ratings reported here may reveal the prototypical organization of voices. In keeping with the nature of prototypically organized categories in for example phonetic categories (Kuhl, 1991), voices close to the hypothesized

prototype-voice are perceived as less discriminable than voices further from the prototype, independently of the individual listener.

The proposal that voices are organized around a prototype-voice is further strengthened by the similarities found between two two-dimensional voice spaces: one based on acoustic and one on perceptual similarities. Less discriminable voices (those that were perceived as more typical) took a central position on both the acoustic and the perceptual map.

Segmental cues affect the discriminability of voices

Although discriminability of a voice is relatively independent of individual listener biases, it need not be independent from the segmental information that the voice carries. Our results indicate that segmental cues do have an effect on the perceived discriminability of a voice. We presented a multidimensional scaling of the eight words that were used in the experiment, based on the similarity of their effects on the voice discriminability ratings (see Fig. 4). The distribution of the words on this map suggested that word-specific contributions to voice discriminability are at least in part structured by segmental cues. That is, certain phonetic contents make some voices more and some other voices less discriminable than what one would expect on the basis of their overall discriminability. In short, perceived typicality or uniqueness of a voice depends on what that voice says.

References

Eisner, F., McQueen, J.M., 2005. The specificity of perceptual learning in speech processing. Perception & Psychophysics 67, 224-238.

Knösche, T.R., Lattner, S., Maess, B., Schauer, M., Friederici, A.D., 2002. Early parallel processing of auditory word and voice information. NeuroImage 17, 1493-1503.

Kraljic, T., Samuel, A.G., 2007. Perceptual adjustments to multiple speakers. Journal of Memory and Language 56, 1: 1-15.

Kuhl, P.K., 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. Perception & Psychophysics 50, 93-107.

Nolan, F., 1997. Speaker recognition and forensic phonetics. In: Hardcastle, W., Laver, J. (Eds.), A Handbook of Phonetic Science. Blackwell, Oxford, pp. 744–766.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance 23, 651-666.

Chapter 3

Flexibility, cue use and abstraction in voice identity learning

Abstract

Two multi-session training experiments investigated how listeners learn to identify the voices of previously unknown talkers. We focused on a simple form of the voice-learning problem: During training, listeners heard tokens of only one word, on a voice morph continuum between endpoints spoken by two talkers, and were taught to identify the two voices. We used the same voice morph continua throughout, but systematically varied voice identity feedback in a between-session and between-experiment learning-relearning paradigm. We demonstrate that new voice identities, unlike new phonemic categories in adulthood, are easy for adults to learn, but that, like established phonemic categories, the category boundaries of new voice identities can readily be adjusted. We argue that voice identities are abstract auditory categories. Generalization of learning suggests that these abstract categories are based partly on segment-specific cues (e.g., how the talkers said /m/) and partly on non-segmental cues (e.g., the talkers' voice quality).

Andics, A., McQueen, J. M. (in preparation). Flexibility, cue use and abstraction in voice identity learning.

Parts of this work have been presented at the 11th Winter Conference of the NVP (Dutch Society for Psychonomics), in Egmond aan Zee, the Netherlands, and at the 11th Laboratory Phonology Conference, in Wellington, New Zealand.

Introduction

Recognizing a person from his or her speech under highly varying circumstances is a task that human listeners perform with astonishing ease. But it is unclear how a talker's voice comes to be represented in the listener's mind. For example, how much flexibility is there in newly-acquired voice identity representations? How quickly can a listener learn a new voice identity, and how quickly can a listener adjust voice identity knowledge in the context of new experience with that voice? And how does the speech signal inform the listener about the voice characteristics of a new talker? Does the listener form abstract voice identity categories such that learning can generalize over words? This study investigated the degree of flexibility in voice identity learning, examined the role of various speech cues in the creation of voice identity categories, and asked whether voice learning entails abstraction.

On the flexibility of speech categories

Getting to know a new talker's voice means that we begin to use information in the talker's speech signal to create representations of his or her voice identity. Voice identities (e.g., 'Bob's voice'), just like phonemes (e.g., /b/), are auditory categories informed by the temporally and spectrally continuous speech signal. It has been proposed that once an auditory category is formed, it may influence subsequent signal perception. For example, categorical representations of vowels can lead to nonlinear perception of a vowel continuum (Kuhl, 1991). An important question, therefore, is if and when such nonlinearities emerge in voice identity learning.

These kinds of nonlinearities in signal perception are well predicted by Bayesian models of distribution learning (Feldman, Griffiths and Morgan, 2009). Such models assume that listeners behave near optimally (i.e., in the sense that behavior is captured well by the predictions of Bayes' theorem). They are capable of explaining a wide range of speech perception phenomena (Norris and McQueen, 2008; Feldman et al., 2009). This suggests that having a clear idea about what optimal listener behavior would entail would be helpful for the understanding of auditory category processing, and specifically of voice identity processing.

Optimal listener behavior should involve a trade-off between the capacity to identify already-acquired categories and the capacity to learn new ones. In the case of phonemic identification in the native language, for instance, where the number of possible values (that is, phonemic categories which distinguish words) is very limited, a mechanism that weighs identification of old categories more than acquisition of new categories would be more beneficial. But in cases where the number of possible values is very large (for instance, talker-specific acoustic-phonetic categories), a mechanism that weighs recent experience more than past experience would be more useful. Over-reliance on recent experience, however, would lead to a loss in robustness.

The available evidence on flexibility in phonemic category learning is consistent with this analysis. Speech perception is sometimes rigid while at other times it is flexible. A well-known case of inflexibility to form new auditory categories from the speech signal is observed in the comparison between infant and adult speech categorization. Sensitivity to linguistically irrelevant phonetic cues and therefore the ability to form phonemic categories decreases towards the end of the first year of life (Werker and Tees, 1984). After that age, the creation of new phonemic categories becomes much harder. For example, learning phonological contrasts in a non-native language in adulthood is notoriously difficult (Logan, Lively and Pisoni, 1991). The benefit of this inflexibility is the stability of already-acquired categories in the native language.

In contrast, other aspects of phonemic category processing remain flexible. Both infants and adults are able to adjust their phonemic categories as a function of the distributional properties of the input (Maye, Werker and Gerken, 2002; Norris, McQueen and Cutler, 2003). Furthermore, perceptual learning about speech is fast (Norris et al., 2003), thorough (Sjerps and McQueen, 2010), stable over time (Kraljic and Samuel, 2005; Eisner and McQueen, 2006), generalizable to novel words (Allen and Miller, 2004; McQueen, Cutler and Norris, 2006) and can be talker-specific (Kraljic and Samuel, 2005, 2007; Eisner and McQueen, 2006) to the extent that multiple talker-specific phonemic category representations can be maintained simultaneously (Kraljic and Samuel, 2007). The benefit of this flexibility, this ability to tune in to properties of the current input, is that it allows the listener to recognize more easily the current talker's next words (Norris et al., 2003; McQueen et al., 2006).

This evidence suggests that there is an appropriate balance in plasticity in phonemic categories – speech perception is stable when you need stability; and it is flexible when you need flexibility. Much less is known about voice perception in this regard. As the number of voices in a human listener's environment and the level of variation within each voice are typically high, a reasonable hypothesis based on the above analysis is that an optimal listener would learn new voice identities easily (contrary to learning new phonemic categories) and would be able to adjust them quickly (similarly to talker-specific adjustments of phonemic categories). Thus, even though phonemic categories are different from voice categories in many ways (e.g., with respect to size of repertory, (non)linguistic function, and acoustic specification), observations about optimal listener behavior in phonemic learning can still be used to generate the above hypothesis about voice learning. The present experiments tested this hypothesis. We probed the readiness of the perceptual system to create new voice identities and to modify them. We investigated the nature of voice identity category formation, testing how listeners use distributional information about voices and how fast they create and adjust voice identities.

To achieve these ends, we used a between-session learning-relearning paradigm. We pared voice identity learning down to its bare essentials: Listeners heard only one word during training, and were taught to identify tokens of that word as being spoken by one of two previously unknown talkers. More specifically, during the training phases, listeners heard stimulus steps on a continuum made by morphing the auditory token of the word spoken by one of the talkers into a token of the same word spoken by the other talker. The listeners' task was to learn which voice identity went with which stimuli. In subsequent test phases, listeners identified the voice identity of the trained stimuli and of other morphed stimuli. Across experimental sessions (on different days), we systematically varied voice identity feedback during the training phase (i.e., which steps on the continuum were associated with which voice identity). We could thus ask how quickly the listeners learned and relearned voice identities while controlling for the acoustic characteristics of the materials (because the same voice morph continuum was used in all training phases). Control over stimulus characteristics was necessary for us to examine not only these questions concerning the flexibility of voice identity representations, but also our second question: Which sources of information in the speech signal are involved in voice learning?

On linguistic versus voice identity information in the speech signal

Perception of auditory categories (e.g., phonemes, voice identities) in the speech signal is motivated by different, distinct goals, such as understanding words or identifying talkers. But to what extent are these separate goals served by separate processes? Furthermore, does the same information get used for both linguistic and voice identity processing, or are there separate information sources?

Evidence from neuropsychological and neuroimaging studies suggests that voice identity processing and linguistic processing involve distinct neural substrates (Van Lancker, Cummings, Kreiman and Dobkin, 1988; Belin, Fecteau and Bedard, 2004). Furthermore, it appears that each process can exist without the other. On the one hand, linguistic processing can occur when voice identity processing fails: for instance, listeners with cochlear implants can perform well on a vowel recognition task but perform poorly in talker recognition given the same stimuli (Vongphoe and Zeng, 2005). On the other hand, voice identity processing may be based on processes that do not depend on the presence of linguistic information: for example, primates recognize their conspecifics (a form of processing that is at least similar to human voice recognition) from vocalizations that carry no linguistic content (Petkov et al., 2008).

One could then argue that there might be a clear-cut distinction between cue types in the speech signal. Local, segmental cues (i.e., those tied to individual segments) could dominate phonemic processing, and global, non-segmental cues (i.e., those not tied to individual segments) could dominate voice identity processing. It is plausible that listeners use cues for voice perception that vary minimally with segmental content: Such cues could be direct acoustic correlates of vocal anatomy (see Kreiman, 1997) and/or persistent characteristics of use (e.g., Nolan, 1983). Indeed, the important role of global or non-segmental cues in voice recognition has long been known: Fundamental frequency and speaking rate are found to correlate strongly with perceptual measures of talker similarity (Walden, Montgomery, Gibeily, Prosek and Schwartz, 1978). Similarly, long-time-average-spectra have been shown to give a good estimate of voice classification (Cleveland, 1977). Other dominant global voice cues include mean formant frequencies, timbre and breathiness (Klatt and Klatt, 1990).

But we know that listeners are sensitive to cue reliability (e.g., Clayards, Tanenhaus, Aslin and Jacobs, 2008), and non-segmental cues are not necessarily the most reliable cues

to voice identity. Indeed, many global cues to voices are known to be badly affected by situational context (Nolan, 1983; Vaissiere, 2005), and they are also easier to imitate. For instance, there are indications that mimicry of global properties (pitch, global speaking rate) is possible for experienced impersonators, but that of formant frequencies and of local features such as relative segment durations is very hard (Eriksson and Wretling, 1997). These results suggest that local traces of the imitator's own voice identity are much harder to remove from the signal than global ones, and/or that adding local traces of another voice identity is very hard. This in turn suggests that the use of local, segmental cues (even if those cues are not themselves entirely reliable) could contribute to the robustness of voice perception.

In addition, a considerable number of studies show that voice identity processing and linguistic processing are interdependent: Voice specific ("indexical") information is used in speech perception (Mullennix and Pisoni, 1990; Nygaard, Sommers and Pisoni, 1994; Nygaard and Pisoni, 1998; McLennan and Luce, 2005; Jesse, McQueen and Page, 2007) and linguistic (local, phonetic, segmental) information is used in voice perception (Fellowes, Remez and Rubin, 1997; Remez, Fellowes and Rubin, 1997; Johnson, Westrek, Nazzi and Cutler, 2011; Remez, Fellowes and Nagel, 2007; Andics, McQueen and van Turennout, 2007). Although some results suggest that indexical specificity might affect slow but not fast linguistic processing (McLennan and Luce, 2005), the majority of studies indicate that voice and linguistic processing interact at an early (i.e., prelexical) level. For example, there is electrophysiological evidence for the preattentive, integral parallel extraction of indexical and linguistic information types (Knösche, Lattner, Maess, Schauer and Friederici, 2002), and Jesse et al. (2007) found that a same-voice benefit in word recognition persisted for non-trained words consisting of segments repeated by the trained talker.

In a series of experiments using sine-wave replicas of speech, Remez and colleagues have shown that when non-segmental information is missing, segmental information alone is enough for listeners to identify talkers (Fellowes et al., 1997; Remez et al., 1997). These results again demonstrate that non-segmental cues may not describe voice representations exhaustively. Talker similarity judgments on these distorted stimuli correlated well with judgments on non-distorted versions of the same stimuli, suggesting that processing of these two stimulus types use similar cues (Remez et al., 2007). This result can be seen as indirect evidence for the use of segmental cues during natural voice perception.

It thus appears that both segmental and non-segmental cues contribute to voice recognition. But there is as yet no direct evidence for the use of segmental cues for voice learning in the presence of potentially stronger non-segmental cues. In Experiment 1 we therefore asked if both segmental and non-segmental cues are used in voice identity learning. We predicted that this would be the case, for the simple reason that, because both sources of information are valuable for robust voice recognition, both are likely to be used in voice learning. Such an outcome would also provide further support for the view that speech and voice identity processing are inter-dependent.

On abstraction in voice identity learning

A third key question concerns the nature of voice identity learning. Are new voice identities based solely on episodic memories or is there abstraction over those episodes? One test for abstraction is to ask if learning generalizes over words (e.g., McQueen et al., 2006): If there is transfer to materials that were not heard in the training phase, then voice identity learning must have gone beyond the mere storage of training episodes. In the test phases of Experiment 1, therefore, listeners identified not only the voice identity of stimuli from the voice morph continuum on which they were trained, but also stimuli from two other voice-morph continua. These were made from natural utterances spoken by the talkers that the listeners had been trained on. One continuum was based on new tokens of the word used in training and one was based on tokens of a different word (with different phonemes).

Would listeners be able to identify the voices of the talkers only if they heard tokens from the trained morph continuum (but different morphs on that continuum that those heard during training), or also if they heard completely new tokens of the word used in training, or even if they heard a segmentally entirely different word? It is important to note that these tests of increasing degrees of generalization were also tests of the cues that listeners use in voice identity learning. If there were abstraction to a new word, with different segments, then learning would have to entail, at least in part, the use of non-segmental cues. If segmental cues also play a role, however, performance on new tokens of the trained word (i.e., with overlapping segments) should be better than on the new word with non-overlapping segments.

Experiment 1

Experiment 1 therefore had three goals: to examine the flexibility of voice identity learning, to explore which information sources (segmental and non-segmental) are exploited in this learning process, and to test whether learning about new voices involves abstraction. As we have already outlined, we presented listeners in the training phases of the experiment with a voice-morph continuum created between two natural tokens of the same word, spoken by two previously unknown talkers. We asked the listeners to decide which of the two talkers they heard and we gave them explicit feedback during training according to an artificially defined voice identity category boundary on the voice-morph continuum. The voice-identification task remained the same in the test phases, but there was no feedback.

To test how flexible voice identity representations are, listeners were trained on two different category boundaries with a one-day delay. We expected a shift in voice identification curves as a function of these changes in the boundary settings. To explore the speed of category formation and the stability of the formed representations, voice identification performance was tested at different time points: before training, in the middle and at the end of each day's training, and also one day after training. We predicted that voice identity learning would be rapid, because listeners need to be able to learn new voice identities after only little exposure. We also predicted that voice identity representations would be flexible, because listeners need to be able to keep track of an individual's changing voice characteristics (e.g., when a talker's speaking rate or style may change in different contexts). In other words, we predicted that voice identity learning would be tuned to the computational demands of this task in everyday listening.

We examined the information sources involved in voice identity learning and the degree of abstraction in this process by manipulating the stimuli used in the test phase: the trained continuum, a new continuum based on new tokens of the word used in the trained continuum (spoken by the same talkers), and a completely new continuum based on tokens of an unrelated word (but again spoken by the same talkers). The words in the untrained continua thus had segmental content that was either overlapping or non-overlapping with the trained word (i.e., phonemically either the same or entirely different). We expected that if voice knowledge includes abstracted non-segmental information, then a training

effect on untrained tokens with no segmental overlap would be found. But we also expected that if voice identity knowledge contains abstract information specific to individual segments, then untrained tokens with complete segmental overlap would be identified better than those with no segmental overlap.

Method

Participants

Sixteen native Dutch listeners with no hearing disorders were paid to take part.

Stimuli

To minimize between-talker subphonemic phonetic differences, two talkers (Voice A and Voice B) with relatively similar voices were chosen from a set of young male non-smoking native speakers of Dutch with no recognizable regional accents and no speech problems (Andics et al., 2007). The choice was based on objective perceptual similarity measures of thirteen voices; the selected talkers were judged to be highly similar, but still discriminable: more specifically, these two voices were correctly categorized as different voice identities in a one-back voice discrimination task in 74% of all cases, while the overall hit rate for all thirteen voices was 87% (Andics et al., 2007). The voices were new to the listeners. Recordings of the Dutch CVC words *mes* (knife) and *lot* (fate) were made by both talkers; these words have no overlapping segments. The recordings were sampled at 44100 Hz, 16 bits per sample.

We then created voice morph continua in Matlab using the speech manipulation algorithms of STRAIGHT (Kawahara, 2006). STRAIGHT decomposes the speech signal into three parameters: a voice source (periodic energy), a noise source (aperiodic energy) and a dynamic spectral filter (spectral shape). Additionally, we supplied manually determined anchor points for the onsets and offsets of each of the three segments in each of the CVC words. Voice morph continua were resynthesized based on values of the three parameters for each pair of corresponding segments. More specifically, the resynthesis algorithm generated morphs between two original tokens of each word by finding analogous time points in the two tokens according to the manually determined anchor points, and then interpolating 99 equidistant intermediate values of each of the three parameters (periodic

and aperiodic energy, and spectral shape). The endpoints (levels 0 and 100) were also resynthesized.

Three morph continua were created, each by morphing one monosyllabic word into another token of the same word spoken by the other talker. In Continuum 1 (used in training and at test), *mes* spoken by Voice A was morphed into *mes* spoken by Voice B. In Continuum 2 (used only at test) , a second token of *mes* from Voice A was morphed into a second token of *mes* from Voice B. Finally, in Continuum 3 (test only), *lot* spoken by Voice A was morphed into *lot* spoken by Voice B. All training and test stimuli were morphs from one of these three continua. Average syllable duration was 565 ms. Average amplitude was equalized over all morphs. Listeners reported at the end of the experiment that they thought they had heard naturally spoken stimuli. Sound files containing all morph steps from all three morph continua are available as supplementary material (http://mpi.nl/people/andics-attila/research).

Procedure and design

Stimuli were presented via headphones binaurally, at a standard, comfortable listening level. Participants were instructed to make forced-choice decisions on talker identity after every word they heard. They were told that there were two talkers with similar voices and that they would be trained to be able to tell them apart through a variety of stimuli, some more ambiguous and some less ambiguous with respect to voice identity. To allow initial assignment of talker names (Peter and Thomas) on response buttons to voice identities (Voice A and Voice B), listeners were presented three naturally produced monosyllables from each talker before the experiment on Day 1. The assignment of talker names to voices and to dominant or non-dominant index fingers was counterbalanced across participants.

The experiment was carried out on two consecutive days with all participants. There were two 18-minute training phases on both days, each followed by a 9-minute long test (Tests 2 and 3 on Day 1, and Tests 5 and 6 on Day 2). Additionally, the experiment on each day began with a pretest that was identical to the test phases. The pretest on Day 1 (Test 1) served as a baseline; that on Day 2 (Test 4) provided a measure of consolidation over the one-day delay.

The full stimulus range was sampled both during training and at test, but there was no exact stimulus overlap between the two parts (i.e., the morph levels used at training were different from those used at test; see Table 1 and next paragraph). During training, one of the continua with the word *mes* was used (Continuum 1). The category boundary was made explicit by giving feedback according to a predefined boundary at 50% voice B morphs one day (symmetric training) and at either 30% or 70% the other day (asymmetric training). The order of symmetric and asymmetric training was counterbalanced across participants. Participants were not informed about the category boundary shift. This training manipulation was amplified by presenting more stimuli from the most ambiguous part of the continuum (see Table 1). Through selection of morph levels and how often they were repeated, it was possible to ensure that the mean of all stimuli from each voice identity category was a 10% distance from the boundary for that category (e.g., when the boundary was at 30%, the mean of all Voice A stimuli was at 20% and the mean of all Voice B stimuli was at 40%). Table 1 lists the morph levels that were used in each training condition.

Table 1. Experiment 1: Morph levels and feedback during training

Trained boundary	**Category feedback**	**Morph steps used during training**														
30%	Voice A	1	6	11	14	17	19	21	23	24	25	26	27	28	29	29
	Voice B	31	31	31	32	32	32	33	33	33	34	34	36	46	63	99
50%	Voice A	1	21	32	39	41	43	44	45	46	47	47	48	48	49	49
	Voice B	51	51	52	52	53	53	54	55	56	57	59	61	68	79	99
70%	Voice A	1	37	54	64	66	66	67	67	67	68	68	68	69	69	69
	Voice B	71	71	72	73	74	75	76	77	79	81	83	86	89	94	99

Note that some morph levels close to the boundary are listed multiple times. With respect to repetition, these levels count as if they were different stimuli.

Half of the participants had the symmetric training on Day 1, half of them on Day 2. At test three continua were used: the trained *mes* continuum (Continuum 1), the other *mes* continuum based on different tokens from the same talkers (Continuum 2), and the *lot* continuum from the same talkers (Continuum 3). Stimulus presentation at test was blocked by word continuum. Nine morph levels were used for each continuum at test: 0, 20, 30, 40, 50, 60, 70, 80 and 100% (i.e., even for Continuum 1, therefore, test stimuli were not presented in training). The tested word continuum changed after every 9-trial block. Stimuli

on consecutive trials were physically different. Stimulus ordering was otherwise random and varied across listeners.

Training trials were 3000 ms long and included visual feedback (i.e., whether responses were correct, incorrect or late), presented from 2000 to 2700 ms after trial onset. Training phases contained 360 trials (12 repetitions of 30 morph levels). At test no feedback was given; these trials had a duration of 2000 ms. Test phases contained 270 trials (10 repetitions of 9 morph levels on 3 continua). Altogether 1620 responses per listener were collected on the test trials. Reaction Times (RTs) were measured from stimulus onset. The experiment lasted 63 minutes each day, excluding self-paced breaks.

Results and Discussion

Overview

The key results are summarized in Figures 1 and 2. Fig. 1 plots the proportion of Voice B responses, collapsed over the three continua and the feedback conditions, in each of the six tests. It presents three main findings: First, the step-like categorization functions show that listeners were able to learn the voice identities; second, the steepening of the slope of the categorization functions between Tests 1 and 2 show that most of the learning took place in the first training session; and third, learning was stable over a one-day delay (i.e., there was no substantial difference between Tests 3 and 4). The weak evidence of categorization on Test 1 is likely to be due at least in part to the exposure to the talker labels at the outset of the experiment. Fig. 2 plots the proportion of Voice B responses, separately for the three continua and the feedback conditions, in the four main tests (i.e., ignoring the pretest on each day). The differences among the four feedback conditions within each panel show that voice identity learning was flexible (i.e., category boundary placement as defined by the feedback tended to be reflected in the responses). The sharpening and increasing separation of the functions between the two tests on each day show that there were improvements in learning after additional training. The global similarity between the functions on each day (top two rows vs bottom two rows), however, shows that relearning was also possible. Finally, there was generalization of learning: effects of feedback were found both for the untrained *mes* continuum (Continuum 2) and the untrained *lot* continuum (Continuum 3).

A series of ANOVAs and t-tests examined these patterns statistically (see Tables 2-6). In all ANOVAs, participants were used as random factors. Uncorrected degrees of freedom are given, but they were Greenhouse-Geisser corrected for F-score calculations. Only effects with $p < .1$ are reported in the tables. In these tables we present main effects, interactions, the linear and quadratic components of effects, up to 3-way-interactions with a polynomial contrast. We present analyses of categorization responses (i.e., the proportion of Voice B responses). The results of RT analyses supported the analyses of the categorization data and are given in Appendix A and Fig. 5. Note that the presentation of the results is organized around specific questions on flexibility, stability, abstraction and cue use in voice identity learning.

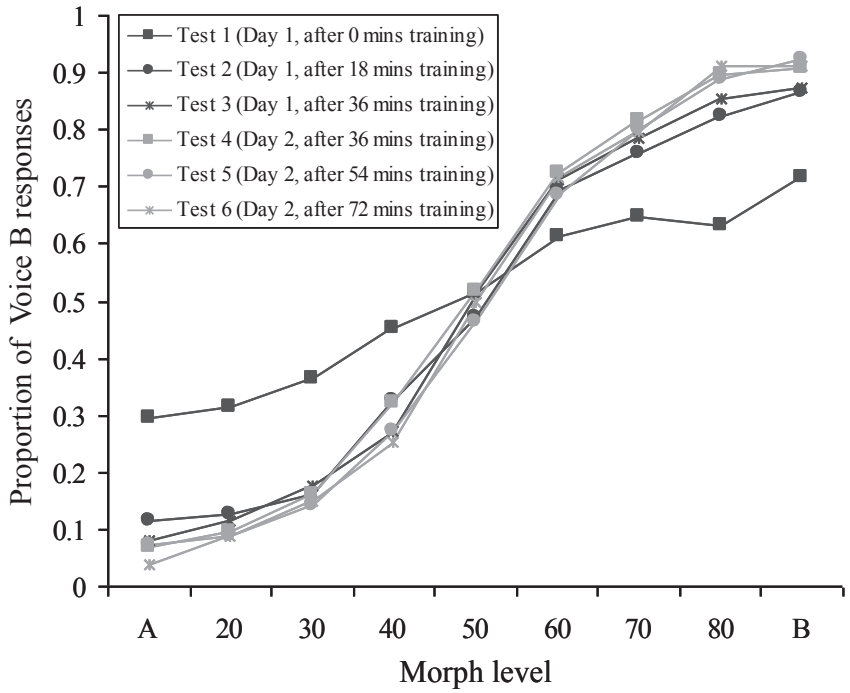
Flexibility in voice identity learning

A repeated-measure ANOVA on categorization responses over the six tests (collapsing over the three text continua) examined the effect of the amount of training. Morph level (Voice A, 20, 30, 40, 50, 60, 70, 80, Voice B) and amount of training (Tests 1-6) were within-participant factors. The effect of level ($F(8,120) = 176.32$, $p < .001$) is an initial indication that listeners were able to learn voice identities. Furthermore, categorization of stimuli on the voice morph continua became less and less ambiguous with training, as shown by an interaction of amount of training and level ($F(40,600) = 8.52$, $p < .001$). Pairwise comparisons of different amounts of voice training across morph levels indicated that most of the learning did indeed take place in the first training session (see Table 2 and Fig. 1).

Table 2. Experiment 1: Effects of amount of training on categorization responses

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Test 1 (Day 1, 0 mins)		**11.56**	**12.62**	**16.25**	**17.34**	**17.53**
		(.004)	**(.003)**	**(.001)**	**(.001)**	**(<.001)**
Test 2 (Day 1, 18 mins)			.833	**3.26**	2.98	**4.15**
			(.376)	**(.091)**	(.105)	**(.060)**
Test 3 (Day 1, 36 mins)				**5.57**	**3.69**	**9.11**
				(.032)	**(.074)**	**(.009)**
Test 4 (Day 2, 36 mins)					.23	2.70
					(.639)	(.121)
Test 5 (Day 2, 54 mins)						1.19
						(.293)
Test 6 (Day 2, 72 mins)						

Fig. 1. Experiment 1: Voice identity categorization responses collapsed across training conditions, after different amounts of training.

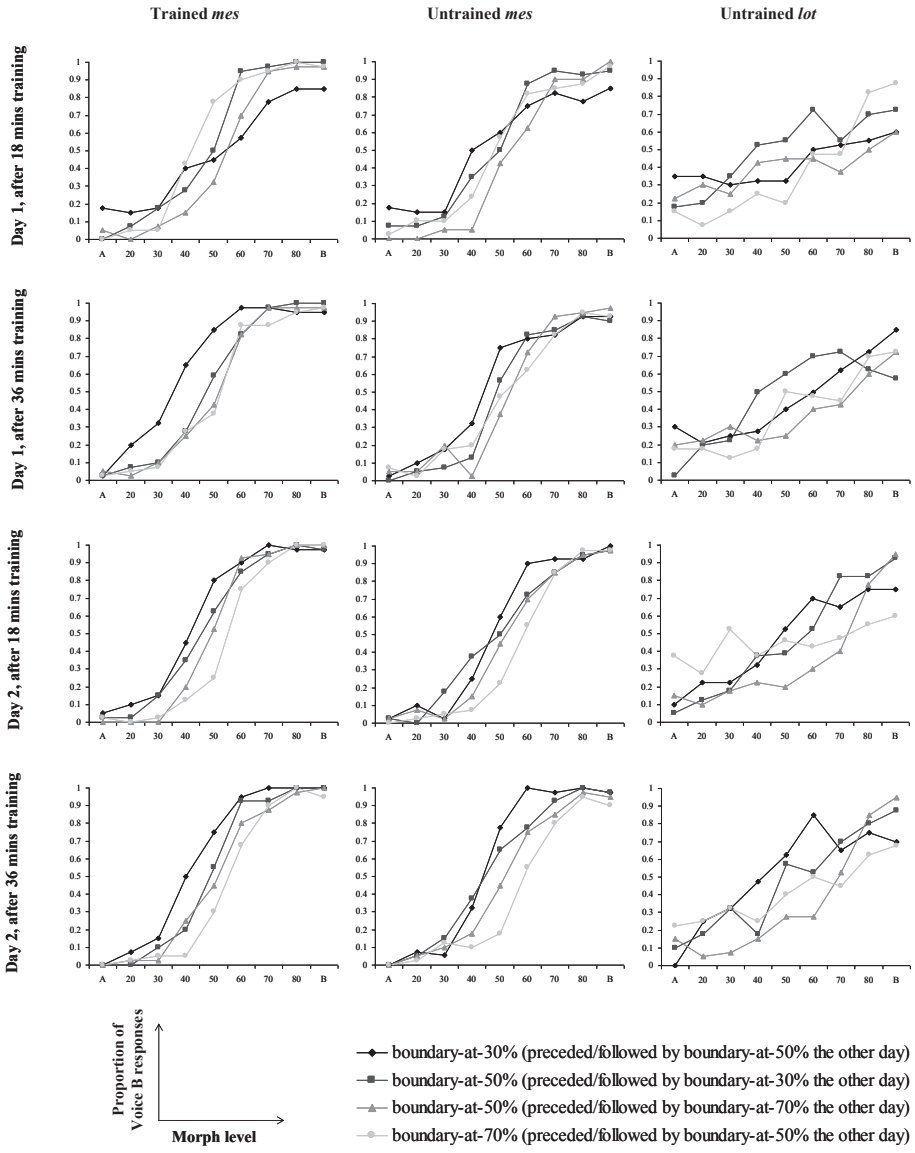


The next ANOVA (see Table 3) focused on categorization performance in the tests immediately after each of the four training sessions (i.e., the data plotted in Fig. 2). Within-participant factors included test word (trained, untrained), test day (first day, second day), test session (after 18 mins training, after 36 mins training) and morph level (Voice A, 20, 30, 40, 50, 60, 70, 80, Voice B). Training condition was a between-participant factor. Participants were coded as having an average boundary either at 40% (for those trained with boundaries at 30% and 50%) or at 60% (for those trained with boundaries at 50% and 70%). This analysis confirmed that participants were able to learn the trained voice identities: voice endpoints were categorized unambiguously after training (main effect of morph level; Table 3). This analysis also showed that learning took place over time: listeners performed better on the second than on the first day and better in the test sessions after 36 mins of training than in the sessions after 18 mins of training (main effects of day and session; Table 3).

Table 3. Experiment 1: Training effects on categorization responses in the post-training tests

	F	df	error df	p
training	17.79	1	14	0.001
level	314.28	8	112	< 0.001
[linear]	740.37	1	14	< 0.001
word	15.31	1	14	0.002
training x level	3.60	8	112	0.025
[quadratic]	12.41	1	14	0.003
training x session	3.23	1	14	0.094
day x level	2.73	8	112	0.048
[linear]	5.81	1	14	0.030
word x level	15.15	8	112	< 0.001
[linear]	30.31	1	14	< 0.001
[quadratic]	3.41	1	14	0.086
training x session x level	3.03	8	112	0.033
[quadratic]	10.69	1	14	0.006
training x day x level	2.34	8	112	0.078
[quadratic]	7.09	1	14	0.019
day x session x word	5.84	1	14	0.03
training x day x session x word	5.94	1	14	0.029
training x day x session x word x level	2.20	8	112	0.083

Fig. 2. Experiment 1: Voice identity categorization responses after training in the four boundary training conditions per test word, day and session (i.e., in Tests 2, 3, 5 and 6).



Participants were also able to re-learn the trained voice identities: the categorization curve shift followed the change in average trained boundary (main effect of training; Table 3). This re-learning effect appears to have been caused by the change in the feedback about

the boundary: the perceptual shift was largest for the morph levels around the boundary, and smallest for the endpoints (see Fig. 2 and the significant quadratic component of the training by morph level interaction in Table 3). The training-related shifts in the categorization functions were present in the asymmetric training conditions each day (compare the boundary-at-30% and boundary-at-70% functions in Fig. 2). One-tailed independent samples t-tests comparing the two asymmetric boundary settings (30% and 70%) for each of the two training days separately showed effects for Day 1 ($t(7) = 2.30$, $p = .024$) and Day 2 ($t(7) = 4.39$, $p = .002$). These effects in Day 2 thus confirm that re-learning was not blocked after participants had previously learned that the category boundary was at the 50% morph.

Stability in voice identity learning

With respect to the stability of voice identity category learning, the steepening of the categorization curve of the trained voice continuum not only persisted for at least one day after training, but also the one-day consolidation made the categorization of the trained voices more unambiguous, even without additional training (see Fig. 1 and the direct comparison of Test 3 and Test 4 across morph levels; Table 2).

Furthermore, boundary training was able to influence perception even 24 hours later: participants who received asymmetric boundary training on Day 1 were found to show a trend towards a perceptual effect consistent with that training on Day 2 (one-tailed $t(7) = 1.50$, $p = .088$). That is, listeners who received the boundary at 30% on Day 1 shifted their perception of Day 2's trained boundary at 50% towards 30%, while listeners who received the boundary at 70% on Day 1 shifted their perception of Day 2's trained boundary at 50% towards 70%. This trend suggests that effects of boundary training can persist even in the presence of feedback indicating a new boundary. Note that we found no perceptual shift in the symmetric condition on Day 1 ($t(7) = 1.02$, $p = .171$). That is, there was no initial bias for the listeners with boundary at 50% on Day 1 that might explain the difference in their performance in the 30% and 70% conditions on Day 2.

Abstraction in voice identity learning

Analyses then turned to the question whether voice learning on the trained word generalized to voice identification on untrained words. The earlier analyses (see Table 3)

already showed a main effect of word (trained vs untrained), but also effects of both day and training session (better performance in later sessions) and interactions of day, session and word, and of training, day, session and word. Subsequent analyses therefore focused on subsets of the data.

First, an ANOVA (see Table 4) was performed on categorization responses to only the untrained words (untrained *mes* and untrained *lot*) in the asymmetric training conditions on Day 2 after 36 mins of training (i.e., under the conditions where an effect of 30% and 70% feedback was most likely to show generalization of learning). The main effect of training and the interaction of training by level indicate that the training-related boundary shift generalized to both untrained word continua. Second, session-wise, word-by-word analyses (see Table 5) confirmed that there were training effects in the asymmetric training conditions for all three words on Day 2 after 36 mins of training, but only for trained and untrained *mes* on Day 2 after 18 mins of training. Table 5 also shows that there were no differences between the two groups with asymmetric training on Day 2 prior to the start of training on Day 1. Finally, independent t-tests were performed to compare categorization responses under different training conditions for each level and for each word in the asymmetric training condition after 36 mins training on Day 2 (see Table 6). They showed that the perceptual shift across categorization curves was most expressed around the trained boundaries and was not expressed at the endpoints.

Table 4. Experiment 1: Effects on categorization responses for untrained words in the final test

	F	df	error df	p
training	8.81	1	14	0.010
level	82.14	8	112	< 0.001
[linear]	298.81	1	14	< 0.001
word	3.88	1	14	0.069
training x level	3.23	8	112	0.028
[quadratic]	8.00	1	14	0.013
word x level	7.75	8	112	< 0.001
[linear]	15.45	1	14	0.002

Table 5. Experiment 1: Specific tests of categorization responses: training effect per word

Training effect	Before training (Day 1)	After 18 mins training (Day 2)	After 36 mins training (Day 2)
trained *mes*	.06 (.810)	**15.50 (.001)**	**15.57 (.001)**
untrained *mes*	.04 (.841)	**5.87 (.030)**	**6.05 (.027)**
untrained *lot*	.52 (.483)	1.41 (.255)	**4.75 (.047)**

The table displays F scores (df = 1, 14) with the corresponding p values in brackets.

Table 6. Experiment 1: Independent t-tests per word and morph level in the final test

trained *mes*	*Voice A*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*Voice B*
mean diff.	0.00	0.01	0.09	0.20	0.28	0.20	0.08	0.01	0.03
SE diff.	0.00	0.03	0.05	0.08	0.09	0.06	0.05	0.01	0.02
t(14)		0.40	1.79	2.65	2.93	3.21	1.57	1.00	1.53
p		0.693	0.095	**0.019**	**0.011**	**0.006**	0.138	0.334	0.149
untrained *mes*	*Voice A*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*Voice B*
mean diff.	0.00	0.03	-0.01	0.21	0.40	0.24	0.13	0.04	0.05
SE diff.	0.00	0.05	0.09	0.15	0.13	0.14	0.06	0.03	0.04
t(14)		0.51	-0.11	1.47	3.00	1.70	2.24	1.43	1.25
p		0.622	0.915	0.165	**0.010**	0.111	**0.042**	0.176	0.233
untrained *lot*	*Voice A*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*Voice B*
mean diff.	-0.14	0.06	0.13	0.13	0.26	0.30	0.19	0.04	-0.03
SE diff.	0.07	0.10	0.12	0.12	0.11	0.10	0.12	0.12	0.13
t(14)	-1.92	0.64	1.01	1.06	2.32	3.01	1.53	0.32	-0.19
p	0.076	0.530	0.329	0.306	**0.036**	**0.009**	0.149	0.757	0.856

Cue use in voice identity learning

The above demonstrations of generalization to a segmentally non-overlapping word suggest that voice identity learning is based, at least in part, on non-segmental cues. But not all voice identity information was transferred to the untrained words. Voice identification was more unambiguous for trained than for untrained words (see Fig. 2), and endpoints of the trained *mes* continuum were categorized with more confidence than the endpoints of the untrained word continua (main effect of word, linear component of the word by level interaction; Table 3).

The results also suggest, however, that voice identity categorization relied, at least partly, on segmental information. In the analysis of the untrained words in the final test session (Table 4), there was a main effect of word – untrained *mes* compared to untrained

lot – and an interaction of word and level, indicating better performance, at least for some morph levels, on the segmentally identical untrained word. Furthermore, as already noted, the training effect for *mes* on Day 2 was significant after 18 mins of training, while that for *lot* emerged only after 36 minutes of training (Table 5).

Summary

Experiment 1 showed that listeners are able to learn to categorize new voice identities, that they can do so rapidly, and that there is flexibility in the learning process – in particular, listeners could easily re-learn the voice identities when the feedback changed. The demonstrations of enhanced performance after a 24-hour delay, and of influences of Day 1 training on Day 2 performance, however, also indicate that there is stability in voice identity learning. Voice learning appears to be stable even when listeners are fatigued (they made responses to 1620 trials over two days of testing). Experiment 1 presented in addition evidence of abstraction in voice identity learning (generalization to untrained words), and that abstract knowledge about newly-learned voices is based on cues that are partially segment-specific and partially not segment-specific.

Experiment 2

Experiment 1 demonstrated that the exact location of the perceived category boundary on a Voice A – Voice B morph continuum could be shifted by training. That is, the same signal could sometimes be perceived as being one voice, and sometimes as another voice. But do the morphed stimuli determine any properties of perceived voice identity category structure? The results of Experiment 1 do not exclude the possibility that acoustic properties of the voice stimuli influenced categorization responses. There are at least two possible biasing phenomena. First, voice identity category centers (what counts as the most prototypical instantiation of each voice) may be coded in some way in the speech signal and preserved in the morphs. The stimuli close to voice identity category centers could acoustically be more strongly flagged as being tokens of a particular voice than stimuli far from a category centre. For example, stimuli close to category centers could contain properties which are more diagnostic of that voice than stimuli further from category centers. Second, the morphing technique might have made the middle region of the

stimulus continuum sound less natural. These phenomena in the speech signal, if present, could have contributed to the category boundary training effect by eliciting more uncertain voice identity categorization responses for the morphs that were more distant from the natural endpoint voices.

Experiment 2 was designed to test the extent to which voice identity learning is flexible by separating the effects of learnt voice identity category structure from possible stimulus-specific effects, such as built-in category structure information or voice naturalness in the morphed stimuli. We attempted to replicate Experiment 1, but with any stimulus-specific effects factored out. This was achieved by presenting the same voice morph continua as the ones used in Experiment 1, but with different feedback. Listeners were trained to perceive the middle region of the continuum (i.e., the voice morphs that were most distant from the natural voices) as a separate voice identity. They were trained to identify stimuli at both endpoints as not being exemplars of that voice. We hypothesized that if there is no built-in category structure information and no voice naturalness variation in the speech signal, then listeners would perceive the voice morph continuum in accordance with the learnt category structure, and that they would be able to do so already after a short training session.

Method

Participants

Sixteen new, native Dutch listeners with no hearing disorders were paid to take part.

Stimuli

Two stimulus continua of Experiment 1 were used: the previously trained *mes* continuum (Continuum 1) and the *lot* continuum (Continuum 3). As in Experiment 1, the voices were new to the listeners.

Procedure and design

Experiment 2 consisted of a single training phase and a single test phase. Unlike in Experiment 1, listeners here had to perform an (A, not A) categorization task (Ashby and Maddox, 2005). They were instructed to make forced-choice decisions on whether they

heard a certain talker or someone else, after every syllable they heard. Participants were not informed about the number of talkers (i.e., whether there were either two talkers, A and "not A", or three talkers, A and one for each endpoint, or some larger number). To allow initial assignment of talker identity to the trained voice, listeners were presented before the experiment with five repetitions of the training monosyllable *mes* at the 50% morph level, accompanied with a display of a face that they were told was that of the trained talker.

The critical manipulation was performed between participants. Listeners were trained according to a predefined voice identity category: for half of them this category was between the 20% and 60% morphs (the 20-60% group), while for the other half of the listeners it was between the 40% and 80% morphs (the 40-80% group). We hypothesized that listeners would categorize endpoints of the voice morph continuum unambiguously as 'other voice' stimuli, the trained voice identity category centers (40% in the 20-60% group and 60% in the 40-80% group) would be categorized unambiguously as 'trained voice' stimuli, and the trained voice identity category boundaries (20% and 60% in the 20-60% group; 40% and 80% in the 40-80% group) would be the most ambiguous.

To maximize the training effect, we slightly modified the trial settings as compared to Experiment 1. Here, trial onsets were signaled with a question mark displayed in the middle of the screen for 300 ms. The auditory stimulus (a voice morph of the word *mes*) began 200 ms after trial onset and lasted on average 565 ms. A response had to be made within 1800 ms of stimulus onset. Listeners received both visual feedback on their performance and further reinforcement of learning (visual and auditory) on every trial. First, they saw visual feedback (i.e., whether responses were correct, incorrect or late) between 2000 and 2250 ms after trial onset. Then they were presented with a picture between 2700 and 3450 ms after trial onset. If the stimulus morph fell within the trained voice identity category (in 42% of all trials), then the feedback picture was the trained talker's face (i.e., the face shown during the initial face-voice assignment). If the stimulus morph fell outside the trained voice identity category, then a scrambled picture (matched in size, color and contrast) was presented instead of the face. This visual reinforcement (cf. von Kriegstein and Giraud, 2006) was accompanied with the auditory repetition of the stimulus, temporally synchronized with the display, starting at 2700 ms after trial onset. This way, every training stimulus was immediately repeated after the listener made their choice, but for the second

time with a visually disambiguated talker identity. Note however that the reinforcement portion of the feedback was independent of the listener's choice. Trials had a duration of 5500 ms.

The training phase lasted about 27 mins. It consisted of nine 3-minute blocks (33 trials each, in total 297 trials), with self-paced breaks between the blocks. A block contained all 25 training stimuli at least once. These were morph levels at every 4% across the continuum, starting from 2%. Eight morph levels (at 18, 22, 38, 42, 58, 62, 78 and 82%) were close to the trained voice identity category boundaries across the two groups. These critical levels were presented a second time in every block. Including the auditory reinforcements, every voice morph level was therefore repeated at least 18 times, and the 8 most critical levels were presented 36 times during the training phase.

The test phase was almost identical to that in Experiment 1, but here only two continua were used: Continua 1 and 3. The 8-minute test contained 243 trials (18 repetitions of *mes* and 9 repetitions of *lot*, sampling each continuum with 9 morph levels, namely 10, 20, 30, 40, 50, 60, 70, 80 and 90%). Stimulus presentation at test was blocked by word continuum. A block of *mes* with 9 repetitions of the 9 random-ordered morph levels was followed by an analogue block of *lot*, which was then followed by another block of *mes*. This made it possible to test the possible effects of test delay by comparing the two blocks of *mes*. The task was the same as during training, but no feedback was given. Trials had a duration of 2000 ms. Stimuli in consecutive trials were physically different. Stimulus ordering was otherwise random and varied across listeners. Table 7 lists the morph levels and feedback that were used in each training condition.

Table 7. Experiment 2: Morph levels and feedback during training

Trained category	**Category feedback**	**Morph steps used during training**														
20-60%	Voice A	22	26	30	34	38	42	46	50	54	58					
	not Voice A	2	6	10	14	18	62	66	70	74	78	82	86	90	94	98
40-80%	Voice A	42	46	50	54	58	62	66	70	74	78					
	not Voice A	2	6	10	14	18	22	26	30	34	38	82	86	90	94	98

Results and Discussion

Table 8 lists effects found in overall and word-by-word ANOVAs on Voice A categorization responses. The overall ANOVA was performed with the between-participants factor training condition (20-60% training, 40-80% training) across the nine morph levels (10, 20, 30, 40, 50, 60, 70, 80, 90) and the two word continua (trained *mes*, untrained *lot*). Analyses on the trained *mes* continuum included an additional factor, namely delay (no delay: data collected immediately after training; and delay: data collected approximately 6 mins, i.e., 162 trials later).

Fig. 3 displays categorization curves (Voice A categorization responses) for each word and each training condition of Experiment 2. Listeners assigned voice identities to the same voice morph continua as in Experiment 1, but we see a completely different voice identity categorization pattern here. This follows from the differences in training between the two experiments. There were more 'trained voice' responses for the middle part of the voice morph continuum than for the endpoints, as confirmed by the significant quadratic component of the morph level effect both in the overall analysis and in each of the word-by-word analyses (see Table 8). There was no main effect of training for any of the word continua. This indicates that the proportion of 'trained voice' responses did not vary significantly between training conditions. Nevertheless, listeners with different categorization training conditions perceived the voice morph continuum differently, giving more 'trained voice' responses within the trained category than outside this category, which is visualized as a shift between the categorization curves of each group in Fig. 3, and was also shown by the linearly loaded training by level interaction in both the overall analysis and the separate analysis for the trained word *mes*. No training-related shift effect was found for the untrained *lot*, suggesting that not all category training information transferred successfully to the untrained word. The loss of categorization sharpness for the untrained word compared to the trained word can be seen in Fig. 3 as a flattening of the inverted-U-shaped curves: listeners were poorer at categorizing the voice morphs of the untrained word continuum. This flattening was indicated in the quadratic component of the interaction of word and level. The loss of training-related information was also captured in an interaction of training, word and level. Finally, there were less 'trained voice' responses for delayed categorization responses, compared to immediate responses for the trained word *mes*, suggesting that as more time after training with confirmatory feedback elapses,

listeners quickly become more conservative about categorizing voice exemplars as the trained voice. This appears to occur after only six minutes. This was shown by the linear component of the delay by level interaction. This increase of conservativism with time spent without confirmatory feedback may also be caused by the exposure to different exemplars of the same voices saying a different word (the untrained *lot* test block) in this delay period.

Fig. 3. Experiment 2: Voice identity categorization responses after training in the two training conditions per test word.

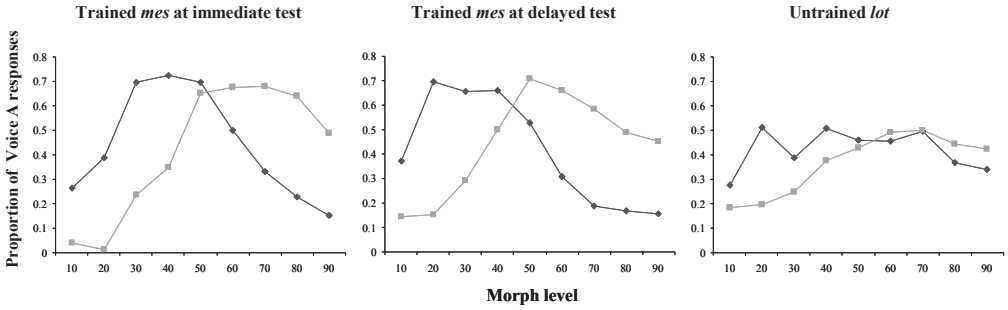


Table 8. Experiment 2: Effects on categorization responses

	F	df	error df	p
overall ANOVA				
level	8.740	8	112	< 0.001
[quadratic]	46.85	1	14	< 0.001
training x level	10.95	8	112	< 0.001
[linear]	17.46	1	14	< 0.001
word x level	2.60	8	112	0.066
[quadratic]	9.58	1	14	0.008
training x word x level	3.52	8	112	0.024
trained *mes*				
level	10.00	8	112	< 0.001
[quadratic]	89.47	1	14	< 0.001
training x level	15.150	8	112	< 0.001
[linear]	20.98	1	14	< 0.001
delay x level	3.37	8	112	0.015
[linear]	5.79	1	14	0.031
untrained *lot*				
level	2.88	8	112	0.050
[quadratic]	6.96	1	14	0.020

The ANOVAs were followed up by independent t-tests for each morph level and for each word (see Table 9). The trained category shift was reflected in significant differences between training conditions for the trained *mes* continuum for almost all but the middle level comparisons, with a change of direction of the difference at 50%. For the untrained *lot*, only the 20% level responses were significantly different across conditions, but note again the change of direction of the difference at 50%.

Table 9. Experiment 2: Independent t-tests per word and morph level

trained *mes*	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*
mean diff.	0.20	0.36	0.47	0.36	0.03	-0.20	-0.39	-0.39	-0.33
SE diff.	0.12	0.12	0.13	0.06	0.08	0.14	0.14	0.13	0.13
t	1.66	2.96	3.64	6.52	0.38	-1.46	-2.80	-3.07	-2.45
p	0.120	**0.010**	**0.003**	**0.000**	0.713	0.166	**0.014**	**0.008**	**0.028**
untrained *lot*	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*	*90*
mean diff.	0.10	0.36	0.16	0.14	0.07	-0.03	0.00	-0.09	-0.09
SE diff.	0.14	0.10	0.14	0.16	0.16	0.15	0.17	0.13	0.11
t	0.74	3.59	1.08	0.89	0.41	-0.22	0.00	-0.73	-0.83
p	0.473	**0.003**	0.298	0.391	0.690	0.828	1.000	0.475	0.419

In summary, Experiment 2 replicated several of the main findings of Experiment 1. Participants once again demonstrated that they could rapidly learn novel voice identities, and that they could to some extent generalize what they had learnt to a segmentally non-overlapping word. Importantly, although the same morph continuum was used to train voice identity learning in both experiments, completely different feedback conditions were used. The flexibility shown by participants across experiments in the placement of voice-category centers and voice-category boundaries indicates that there were no non-linearities across the training continuum, such as built-in differences in voice identity information or in voice naturalness. Such differences may well exist in fully natural spoken stimuli, but at least we can conclude that they are unlikely to have been present in the morphed stimuli used here. RT analyses (Appendix B and Fig. 6) supported the patterns observed in the categorization analyses.

General Discussion

Voice identity categories are flexible and stable

These experiments tested listeners' ability to learn and relearn voice identities. In Experiment 1 we found that listeners are able to form new voice identity categories rapidly. The shape of the voice identification curves was close to linear at the baseline test before training, but became S-shaped already after 18 minutes of training, suggesting that voice

identity representations influenced perception of the continuum quickly. We also found that further learning made the curves even steeper, suggesting that categorization of the voice morph continuum became more unambiguous with more training. In contrast, studies on teaching a nonnative phonemic contrast to adults report behavioural and neural traces of perceptual improvements in the identification of contrasting stimuli only after several days or weeks of extensive training (Lively, Pisoni, Yamada, Tohkura and Yamada, 1994; McCandliss, Fiez, Protopapas, Conway and McClelland, 2002; Zhang et al., 2009). So unlike new phonemic categories in adulthood, new voice identity categories are easy for adults to learn. For phonemes, the benefit of stability that arises from being able to recognize variable input as one of a limited inventory of native-language segments comes with the cost that it is hard to learn new segments that do not fit in that closed inventory. For voice identities, in contrast, there is no cost to expanding the inventory of known voices, so learning a new one appears to be relatively easy.

We also found evidence for stability in voice learning, however. In Experiment 1, the effects of voice training on Day 1 did not fade away in 24 hours (see Fig. 1). Day 2's tests furthermore indicated that the voice identification curves of listeners who got a 50% boundary on the second day differed as a function of the training they received on Day 1. These results thus show that voice representations, even after limited evidence, are stable after one day. Training-related shifts of voice identity categories take place quickly, and they can last as long as 24 hours. Similar patterns of stability have been found in speech perception: Trained shifts of phonemic categories are stable 12 hours after training (Eisner and McQueen, 2006). But what happens to voice representations after several days? Investigations of long-term memory for unfamiliar voices showed that listeners remember newly trained voices even 4 weeks after training, but recognition accuracy decreases as a function of delay (Papcun, Kreiman and Davis, 1989). It has been proposed that voice representations are organized in a typicality-based manner (Papcun et al., 1989; Andics et al., 2010; Bruckert et al., 2010; Mullennix et al., 2011), and that the decrease in accuracy over time can partly be caused by a general listener bias toward falsely recognizing typical-sounding voices that have not been heard previously (Mullennix et al., 2011). It remains to be determined whether these biases are only present during recognition, such that voice identity representations are only modulated by perceptual counter-evidence or whether

there is an overall tendency for voice identity representations to be shifted toward or away from a 'mean voice' over the time course of several days or weeks.

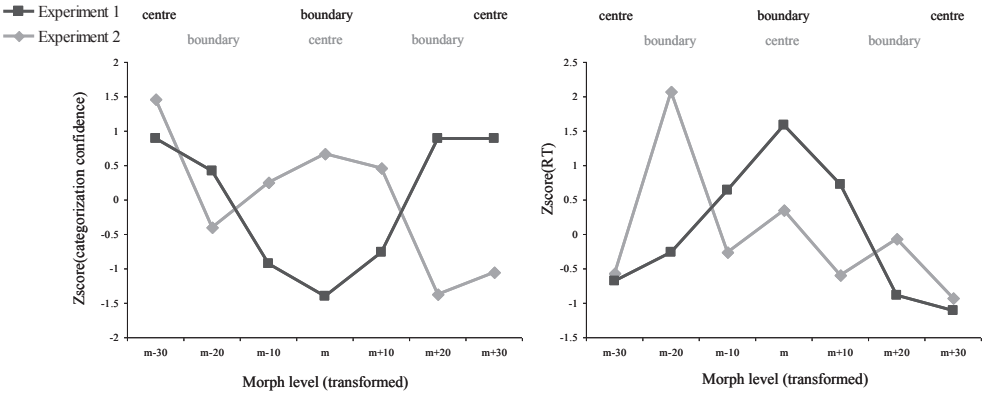
Our results have also shown that listeners readily adjust voice identity representations, including category centers and boundaries, in response to changing feedback. We predicted that listeners would have this flexibility in order to be able to deal with within-voice variability in normal listening situations. The categorization data from Experiment 1 suggest that middle points of a voice morph continuum between two talkers may be perceived as one talker's voice on one day, but as the other talker's voice the next day, even after only a short training session. Furthermore, the categorization data from Experiment 2 showed that, after only a little training, listeners were able to perceive the middle region of the same morph continuum (i.e., the category boundary region of Experiment 1) as a voice identity which is separate from both natural endpoints. Furthermore, listeners in Experiment 2 were able to perceive the same morph level in the middle of the voice morph continua as either the best (category center) or worst (category boundary) exemplar of a voice, dependent on the training condition. These findings suggest that no built-in category structure information was present in the speech signal – the same acoustic stimulus can be perceived as a voice identity category center or as a voice identity category boundary.

A similar effect of flexibility was found in the RT results (Appendices A and B). Longer RTs were assumed to correspond to more ambiguously perceived morph levels. Differences in RT pattern were found both between training conditions and between experiments. This too suggests a difference in the perceived category boundary.

Fig. 4 provides an across-experiment overview of the categorization and RT data, to illustrate the crucially different perception of the voice continua in the two experiments. These figures also demonstrate the absence of voice naturalness variation across the more and less extremely morphed steps of the voice continua: the "most morphed" middle steps of the continua were the most difficult in Experiment 1, but the least difficult in Experiment 2. Listeners in Experiment 2 apparently did not mind or notice that what they learnt as a voice identity category center was in between two natural voices. This is the first study to demonstrate that, for voices, no natural anchor points exist in the speech signal, at least not in the morphed stimuli that we used. Our results expand previous reports on the flexibility of categories in the speech signal (Maye et al., 2002; Norris et al., 2003; Sjerps and

McQueen, 2010) and suggest that just like phonemic categories, voice identity categories are flexible. For both types of acoustic category, the listener needs to be able to adjust to within-category variability.

Fig. 4. Perceived typicality of the same voice stimuli across the two experiments (left panel: categorization data; right panel: RT data). Morph level m refers to the actual middle point of voice identity categorization training (morph40, the average category boundary level during training with a boundary closer to voice A in Experiment 1, and the category center level during 20-60-training in Experiment 2; and morph60, the average category boundary level during training with a boundary closer to voice B in Experiment 1, and the category center level during 40-80-training in Experiment 2). The values are thus realigned across conditions relative to m. Only the trained *mes* trials from the test session immediately following the last training session of each experiment are used here. Categorization confidence per level was calculated as the distance of proportion of 2AFC decisions from the chance level at 0.5. Categorization confidence and RT data were normalized per experiment (using Z-scores) to control for overall task difficulty differences across experiments.



Interestingly, the perceptual boundary shifts in Experiment 1 were smaller than those encouraged by the training (e.g., the training specifying a boundary at the 30% morph resulted in a perceptual boundary closer to 40%; see Fig. 2). One simple explanation could be the following: it is known that in 2AFC categorization tasks listeners tend to respond to both possible choices (here voice identities) equally often (e.g., Repp and Liberman, 1987). As the sampling of morph levels in our tests was centrally symmetric, such a response bias could shift categorization curves toward the 50% morph level in all training conditions. But this would not explain the RT results: In the asymmetric training conditions, RTs were longer

for the 50% and 40%/60% morphs than for the trained 30%/70% boundaries (see Fig. 5; Appendix A). On Day 2, this bias towards the middle of the continuum could be explained by the long-lasting effect of the 50% boundary training on Day 1. But, interestingly, the same bias was observed in RTs on Day 1. That is, the perceived category boundaries on the voice morph continua were consistently closer to the middle of the continua than the trained boundaries. We propose that this bias reveals limits on flexibility in voice perception. The asymmetric category boundary training in Experiment 1 assumed that one of the two individual voice identity categories that the listeners built up was a very broad one, including a voice endpoint and all voice morphs which have at least 30% of this voice and therefore up to 70% of another voice. For this specific pair of voices or at least for these three pairs of tokens, this seemed to push the voice recognition system too far. We suggest that listeners' responses were biased towards the middle of the continua, because they tended not to accept oversized voice identity categories, even if explicit feedback instructed them to do so. Perceptual traces of built-in acceptance ranges for individual person categories have been described for faces (Cabeza, Bruce, Kato and Oda, 1999) but not, to our knowledge, for voices. Our findings suggest that the category structure of voices is not restricted by built-in properties of the speech signal, but that the listeners have built-in expectations on the acceptance range of individual voice identity categories, that is, on how variable or broad a voice can be.

It has to be noted that our voice morphing method focused on cues that are continuous in nature and ignored noncontinuous cues that could not be captured well by morph steps of the voice continua. Many cues in the speech signal are known to be continuous, for example, F0 (Walden et al., 1978) or voice onset time (Allen and Miller, 2004). But there may be additional variation across talkers that can only be learned by looking at effects of noncontinuous cues (e.g., British speakers are known to release intervocalic stops but Americans flap them; Scott and Cutler, 1984), but note that even most of those noncontinuous effects are graded in nature (e.g., American speakers tend to flap, and British speakers tend not to). That is, the continuous case we looked at thus seems to have been the best place to start. For this case at least, voice identity categories are stable and flexible.

Voice identity categories are abstract: Segmental and non-segmental cues

Categorization training provided knowledge about voice identity membership that was not specific to the trained stimuli. Category shift effects in both experiments were shown to generalize to untrained word continua as well, similarly to what has been found for phonemic categories (Allen and Miller, 2004; McQueen et al., 2006). This indicates that the listeners had abstracted voice knowledge rather than that they had done no more than store stimulus-specific, purely episodic memories of the morphs. In the present experiments, abstraction over training information was demonstrated on three levels. The category boundary shift generalized to non-trained utterances (1) of the same continuum (remember that the continuum steps used during test were not heard during training), (2) of a different continuum with the same word (only in Experiment 1), and (3) even to a different word with no segmental overlap. This last effect showed that the training led to knowledge about the voices that was not specific to individual segments. Our study thus adds to the existing literature by demonstrating for the first time that transfer of voice knowledge to new words is possible even after only minimal exposure, and that this generalization was based on non-segmental cues to voice identity. These cues could include the talkers' fundamental frequency and their voice quality characteristics (e.g., timbre and breathiness).

It is important to note, however, that we do not use the word 'non-segmental' in this study in its strongest possible sense. Although the phonemes of the two test words *mes* and *lot* are all different, their subphonemic properties are not. There is subphonemic overlap between, for example, the place of articulation of [s], [l] and [t], and it is thus possible that our non-segmental effects are partly based on such subphonemic cues. Thus, we do not suggest that our test words are phonetically fully independent, and the issue of subphonemic cues obviously warrants further investigation. What we mean by non-segmental effects here, instead, is simply that these effects cannot be due to knowledge that is indexed to specific phonemes, because there is no overlap between *mes* and *lot* at the phonemic level.

We predicted, however, that listeners would use not only non-segmental but also segmental cues in voice learning, since both types of cue are informative about talker identity. This prediction was confirmed. Training effects were typically stronger for the untrained *mes* continuum than for untrained *lot* continuum. Categorization confidence for the least ambiguous morph steps (i.e., at the endpoints in both experiments, and also in the

middle of the continua in Experiment 2) was lower for untrained *mes* than for *lot*. Furthermore, voice identity judgments differed for different word continua, especially around the category boundaries. For example, in Experiment 1 listeners could perceive, for example, a 50% morph from a *mes* continuum as a better exemplar of Voice B than of Voice A, while a 50% morph from the *lot* continuum was perceived as a better exemplar of Voice A than of Voice B. These findings strengthen earlier claims that voice identity categorization involves segmental information (Eriksson and Wretling, 1997; Fellowes, Remez and Rubin, 1997; Remez, Fellowes and Rubin, 1997; Andics et al., 2007). Furthermore, our results suggest that the acceptable acoustic variation within a single voice might not be constant across different segments. It is thus possible that listeners are able to maintain multiple segment-specific voice identity category representations for a single talker simultaneously, analogously to reports of multiple talker-specific phonemic category representations (Kraljic and Samuel, 2007).

Clearly, under natural circumstances, when speech is continuous, listeners become familiar with the talker's voice through all segments of the language at approximately the same time. In those cases, between-segment variation may be less relevant for voice identity processing. But it becomes important in cases when the amount of input from a specific voice is limited. For example in situations of forensic speaker comparisons, when an ear-witness needs to recognize a recently heard voice from among different voice exemplars (Nolan, 1997; French and Harrison, 2006, Mullennix et al., 2011), he or she should, according to our results, give more confident and more reliable responses if the test words are the same as those witnessed. Similarly, our findings indicate a possible segment-specific influence on decisions on whether two speech samples are consistent with having been produced by the same speaker, and also on the estimation of how distinctive two speech samples' shared features are (i.e., how probable it is that the shared features are also shared by other speakers; French and Harrison, 2006).

Finally, our findings underline earlier claims that vocal and linguistic information are highly intermixed in speech (Mullennix and Pisoni, 1990; Nygaard, Sommers and Pisoni, 1994; Fellowes, Remez and Rubin, 1997; Remez, Fellowes and Rubin, 1997; Nygaard and Pisoni, 1998; McLennan and Luce, 2005; Jesse et al., 2007; Remez, Fellowes and Nagel, 2007; Andics et al., 2007). Local or segmental cues in the speech signal are not only essential

for spoken word processing, they are also learnt as important talker characteristics, making voice recognition less dependent on global variations (cf. Eriksson and Wretling, 1997).

Conclusions

The present study showed that new voices can be quickly learned, and, importantly, that voice knowledge transfers to new utterances even after minimal exposure. Our experiments demonstrate, on the one hand, that voice identities are abstract auditory categories, and, on the other hand, that this abstract knowledge is partly based on segment-specific cues in the speech signal and partly on non-segmental cues in the same signal. This fortunate combination provides the perceptual system with the advantage that voice knowledge is flexible and stable at the same time. Furthermore, our study is the first to demonstrate that there are no natural voice anchor points in voice-morphed stimuli – the same acoustic stimulus was perceived as a voice identity category center or as a voice identity category boundary. But, while no built-in category structure information seems to have been encoded in the materials, listeners did have built-in expectations on the acceptance range of individual and also segment-specific voice identity categories. Thus, while voice identity category learning appears to be characterized by its flexibility and stability and by generalization over exposure episodes, these characteristics also appear to have their limits.

References

Allen, J.S., Miller, J.L., 2004. Listener sensitivity to individual talker differences in voice-onset time. Journal of the Acoustical Society of America, 115, 3171-3183.

Andics, A., McQueen, J.M., Petersson, K.M., Gál, V., Rudas, G., Vidnyánszky, Z., 2010. Neural mechanisms for voice recognition. Neuroimage, 52(4), 1528-1540.

Andics, A., McQueen, J.M., Van Turennout, M., 2007. Phonetic content influences voice discriminability. In J. Trouvain, W. J. Barry (Eds.), Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007) (pp. 1829-1832). Dudweiler: Pirrot.

Ashby, F.G., Maddox, W.T., 2005. Human category learning. Annual Review of Psychology, 56, 149-78.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: Neural correlates of voice perception. Trends in Cognitive Sciences, 8, 129-135.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G.A., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Current Biology 20, 116-120.

Cabeza, R., Bruce, V., Kato, T., Oda, M., 1999. Prototype effect in face recognition: Extension and limits. Memory and Cognition, 27, 139-151.

Clayards, M., Tanenhaus, M.K., Aslin, R.N., Jacobs, R.A., 2008. Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108(3), 804-809.

Cleveland, T.F., 1977. Acoustic properties of voice timbre types and their influence on voice classification. Journal of the Acoustical Society of America, 61, 1622-1629.

Eisner, F., McQueen, J.M., 2006. Perceptual learning in speech: Stability over time (L). Journal of the Acoustical Society of America, 119(4), 1950-1953.

Eriksson, A. Wretling, P., 1997. How flexible is the human voice? – A case study of mimicry. In Proceedings of EUROSPEECH 1997, Rhodes, Vol. 2, pp. 1043-1046.

Feldman, N.H., Griffiths, T.L., Morgan, J.L., 2009. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. Psychological Review, 116, 752-782.

Fellowes, J.M., Remez, R.E., Rubin, P.E., 1997. Perceiving the sex and identity of a talker without natural vocal timbre. Perception Psychophysics, 59, 839-849.

French, J.P., Harrison, P., 2006. Investigative and evidential applications of forensic speech science. In A. Heaton-Armstrong, E. Shepherd, G. Gudjonsson D. Wolchover (Eds.), Witness Testimony: Psychological, Investigative and Evidential Perspectives. Oxford: Oxford University Press.

Jesse, A., McQueen, J.M., Page, M., 2007. The locus of talker-specific effects in spoken-word recognition. In J. Trouvain, W. J. Barry (Eds.), Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007) (pp. 1921-1924). Dudweiler: Pirrot.

Johnson, E.K., Westrek, E., Nazzi, T., Cutler, A., 2011. Infant ability to tell voices apart rests on language experience. Developmental Science, 14(5), 1002-1011.

Kawahara, H., 2006. STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoustic Science and Technology, 27(6), 349-353.

Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America, 87, 820-857.

Knösche, T.R., Lattner, S., Maess, B., Schauer, M., Friederici, A.D., 2002. Early parallel processing of auditory word and voice information. Neuroimage, 17(3), 1493-1503.

Kraljic, T., Samuel, A.G., 2005. Perceptual learning for speech: Is there a return to normal? Cognitive Psychology, 51, 141-178.

Kraljic, T., Samuel, A.G., 2007. Perceptual adjustments to multiple speakers. Journal of Memory and Language, 56, 1-15.

Kreiman, J., 1997. Listening to voices: Theory and practice in voice perception research. In K. Johnson J. W. Mullennix (Eds.), Talker Variability in Speech Processing (pp. 85-108). San Diego: Academic.

Kuhl, P.K., 1991. Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. Perception Psychophysics, 50, 93-107.

Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y.I., Yamada, T., 1994. Training Japanese listeners to identify English /r/ and /l/: III. Long-term retention of new phonetic categories. Journal of the Acoustical Society of America, 96(4), 2076-2087.

Logan, J.S., Lively, S.E., Pisoni, D.B., 1991. Training Japanese listeners to identify English /r/ and /l/: A first report. Journal of the Acoustical Society of America, 89, 874-886.

Maye, J., Werker, J.F., Gerken, L., 2002. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition, 82(3), B101-B111.

McCandliss, B.D., Fiez, J.A., Protopapas, A., Conway, M., McClelland, J.L., 2002. Success and failure in teaching the [r]– [l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. Cognitive, Affective, Behavioral Neuroscience, 2, 89-108.

McLennand, C.T., Luce, P.A., 2005. Examining the time course of indexical specificity effects in spoken word recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 306-321.

McQueen, J.M., Cutler, A., Norris, D., 2006. Phonological abstraction in the mental lexicon. Cognitive Science, 30(6), 1113-1126.

Mullennix, J.W., Pisoni, D.B., 1990. Stimulus variability and processing dependencies in speech perception. Perception Psychophysics, 47, 379-390.

Mullennix, J.W., Ross, A., Smith, C., Kuykendall, K., Conard, J., Barb, S., 2011. Typicality effects on memory for voice: implications for earwitness testimony. Applied Cognitive Psychology, 25(1), 29-34.

Nolan, F., 1983. The Phonetic Bases of Speaker Recognition. Cambridge University Press, Cambridge, UK.

Nolan, F., 1997. Speaker recognition and forensic phonetics. In: Hardcastle, W., Laver, J. (Eds.), A Handbook of Phonetic Science. Blackwell, Oxford, pp. 744-766.

Norris, D., McQueen, J.M., 2008. Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review, 115(2), 357-395.

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cognitive Psychology, 47(2), 204-238.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Perception Psychophysics, 60, 355-376.

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1994. Speech perception as a talker-contingent process. Psychological Science, 5, 42-46.

Papcun, G., Kreiman, J., Davis, A., 1989. Long-term memory for unfamiliar voices. Journal of the Acoustical Society of America, 85, 913 - 925.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nature Neuroscience, 11, 367-374.

Remez, R.E., Fellowes, J.M., Nagel, D.S., 2007. On the perception of similarity among talkers. Journal of the Acoustical Society of America, 122, 3688-3696.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance, 23, 651-666.

Repp, B.H., Liberman, A.M., 1987. Phonetic categories are flexible. In S. Harnad (Ed.), Categorical Perception (pp. 89-112). Cambridge University Press.

Scott, D.R., Cutler, A., 1984. Segmental phonology and the perception of syntactic structure. Journal of Verbal Learning and Verbal Behavior, 23, 450-466.

Sjerps, M.J., McQueen, J.M., 2010. The bounds on flexibility in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 36, 195-211.

Vaissiere, J., 2005. Perception of intonation. In R. Remez D. Pisoni (Eds.), Handbook of Speech Perception (pp. 236-263). Oxford: Blackwell.

Van Lancker, D.R., Cummings, J.L., Kreiman, J., Dobkin, B.H., 1988. Phonagnosia: A dissociation between familiar and unfamiliar voices. Cortex, 24, 195-209.

Vongphoe, M., Zeng, F.G., 2005. Speaker recognition with temporal cues in acoustic and electric hearing. Journal of the Acoustical Society of America, 118(2), 1055-1061.

von Kriegstein, K., Giraud, A.-L., 2006. Implicit multisensory associations influence voice recognition. PLoS Biology 4(10): e326.

Walden, B.E., Montgomery, A.A., Gibeily, G.J., Prosek, R.A., Schwartz, D.M., 1978. Correlates of psychological dimensions in talker similarity. Journal of Speech and Hearing Research, 21, 265-275.

Werker, J.F., Tees, R.C., 1984. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Infant Behavior and Development, 7, 49-63.

Zhang, Y., Kuhl, P.K., Imada, T., Iverson, P., Pruitt, J., Stevens, E.B., Kawakatsu, M., Tohkura, Y., Nemoto, I., 2009. Neural signatures of phonetic learning in adulthood: a magnetoencephalography study. Neuroimage, 46(1), 226-40.

Appendices

Appendix A: Response Time analyses, Experiment 1

The RT data are plotted in Fig. 5. Prior to input to an ANOVA, RTs were normalized using Z-scores. Mean RTs were calculated for each cell of a matrix containing the factors participant mean, training condition, test word, test day and test session. The actual RTs were then substituted by the number of standard deviations from these specific means (Z-scores). This was done to rule out irrelevant variation caused by overall differences in participant speed and test word length. Table A1 displays effects found in an omnibus ANOVA and specific effects found in the corresponding word-by-word ANOVAs on normalized RTs with the following within-participant factors: training condition (symmetric: boundary at 50%, asymmetric: boundary at 30% or 70%), test word (trained, untrained), test session (after 18 mins training, after 36 mins training) and morph level (voiceA/voiceB, 20/80, 30/70, 40/60, 50, 60/40, 70/30, 80/20, voiceB/voiceA). To collapse data over boundary-at-30% and boundary-at-70% conditions, morph level labelling was transformed such that morph levels for the boundary-at-70% trials were flipped around the voice continuum's acoustic centre, i.e. 50% (so, for instance, the third morph level, i.e. 30/70 always referred to the actually trained boundary in these conditions).

RTs were shortest for the voice endpoints and longest for the most ambiguous stimuli (quadratic component of the morph level effect; Table A1). This effect was present in the overall analysis, for both the trained and untrained *mes*, but not for the untrained *lot*. The position of the RT peaks also shifted across boundary training conditions, and this shift followed the direction of the trained boundary (interaction of training and morph level; Table A1). This effect was present for both the trained and untrained *mes*, but not for the untrained *lot* (interaction of training and morph level; Table A1). Note that the size of the RT peak shift, like the perceptual shift in the categorization data, was smaller than the difference between boundaries as was defined in the training conditions (see Figures 2 and 5). RTs were also modulated by training condition and the amount of training for all tested word continua (trained and untrained *mes*: quadratic component of the training by morph level interaction; trained *mes*: linear and quadratic components of the session by level interaction; untrained *lot*: quadratic component of the training by session by morph level interaction; see Table A1).

Greater differences were found between endpoint and boundary RTs for the trained word than for the untrained words, suggesting clearer distinctions between easy and difficult voice decisions for the better learnt trained word continuum (word by level interaction; Table A1). This was reflected by the word by level interaction. Furthermore, longer training also made responses faster and the RT peak more expressed (linear and quadratic components of the session by level interaction). Finally, RTs for the untrained, segmentally non-overlapping word *lot* were also modulated by training condition and the amount of training (interaction of training, session and morph level).

These RT analyses strengthen the results of the main analyses. Differences in the steepness and position of the category boundaries across conditions in the categorization data are generally reflected by differences in RT across conditions.

Fig. 5 (Appendix A). Experiment 1: Response times at test in the four boundary training conditions per test word, day and session.

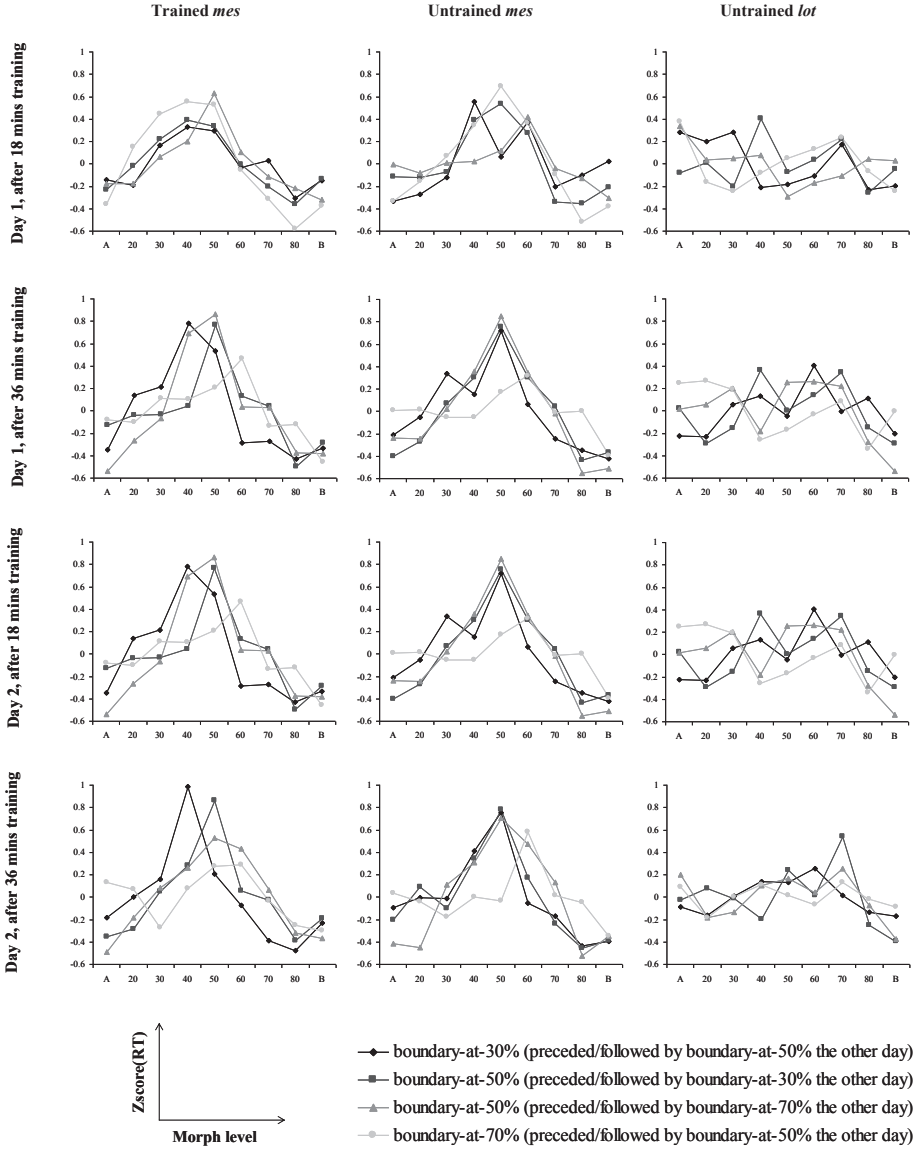


Table A1. Experiment 1: Effects on response times

	F	df	error df	p
overall ANOVA				
level	13.27	8	120	< 0.001
[quadratic]	21.97	1	15	< 0.001
training x level	2.97	8	120	0.029
[quadratic]	3.09	1	15	0.099
word x level	3.09	8	120	0.035
[quadratic]	13.60	1	15	0.002
session x level				
[linear]	6.20	1	15	0.025
[quadratic]	5.67	1	15	0.031
session x word x level				
[quadratic]	3.78	1	15	0.071
trained *mes*				
level	12.80	8	120	< 0.001
[quadratic]	27.09	1	15	< 0.001
training x level	2.46	8	120	0.052
[quadratic]	3.08	1	15	0.099
session x level				
[linear]	3.88	1	15	0.068
[quadratic]	6.23	1	15	0.025
untrained *mes*				
level	12.32	8	120	< 0.001
[quadratic]	26.51	1	15	< 0.001
training x level				
[quadratic]	3.33	1	15	0.088
untrained *lot*				
session x level				
[quadratic]	3.13	1	15	0.097
training x session x level				
[quadratic]	4.48	1	15	0.051

Appendix B: Response Time analyses, Experiment 2

The RT data are plotted in Fig. 6. Table B1 displays the ANOVAs on these data. Similarly to Experiment 1, RTs were first normalized (Z-scores) for participant mean, training condition and test word block, to rule out irrelevant variation caused by overall differences in participant speed and test word length. An overall ANOVA was then performed on Z-scores of the RTs with the between-participants factor training condition (20-60% training, 40-80% training) across the nine morph levels (10, 20, 30, 40, 50, 60, 70, 80, 90) and the two word continua (trained *mes*, untrained *lot*), that is, using the same factors that were used for the analyses of categorization responses. The overall ANOVA was followed by word-by-word ANOVAs. The analyses on the trained *mes* continuum here again included the delay factor.

By definition, no main effects were present for the normalized factors. Instead, we investigated these factors' interactions with morph level. Fig. 6 demonstrates that RTs were modulated by the training, with slower responses close to the trained category boundaries, and faster responses far from the trained category boundaries. This is confirmed by the significant, linearly loaded training by level interactions that were found not only in the overall ANOVA, but also separately for each word, including the untrained *lot*. It shows that the effects of voice identity categorization training generalized to the untrained word as well. A clearer flattening effect was caused by test delay: In trials that were presented in the final block of test, that is, 6 mins later than the block following training immediately, RT differences between the morph levels that were responded to relatively quickly versus relatively slowly were reduced, as confirmed by the significant quadratic component of the delay by level interaction for *mes*. Finally, note that no main effect of level was found. This suggests that RT differences in this experiment cannot be explained by inherent differences (e.g., in speaking rate) between the voice endpoints.

Fig. 6 (Appendix B). Experiment 2: Response times at test in the two training conditions per test word.

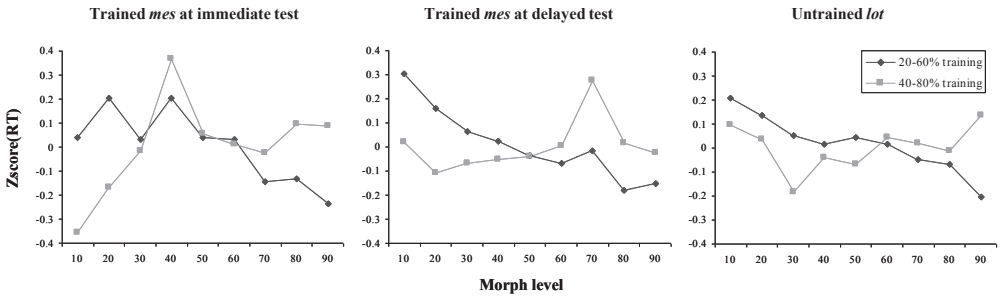


Table B1. Experiment 2: Effects on response times

	F	df	error df	p
overall ANOVA				
training x level	3.14	8	112	0.028
[linear]	9.19	1	14	0.009
word x level				
[quadratic]	3.54	1	14	0.081
trained *mes*				
training x level	3.369	8	112	0.023
[linear]	8.75	1	14	0.010
delay x level	4.71	8	112	0.001
[quadratic]	11.85	1	14	0.004
untrained *lot*				
training x level				
[linear]	5.54	1	14	0.034

Chapter 4

Phonetic content shapes implicitly-learned voice categories

Abstract

In a study on voice-category learning, Dutch listeners heard stories and isolated words spoken by two members of each of two families and were trained to identify the speakers' family membership. The listeners were then asked to identify individual voices on voice-morph continua as old or new. Voice recognition was no better for within-family than across-family morphs, but individual voice categories were learned. These findings support the view that listeners can form prototype-centered representations of voices, and define some boundary conditions of this ability. In particular, these findings suggest that formation of prototype-based representations of groups of voices does not occur even with explicit feedback, but also that representations of the voices of individuals can be formed implicitly. The study also asked if voice categories are shaped by phonetic content. The voice-morph test continua were based on two three-phoneme Dutch words (*mes*, knife, and *lot*, fate), and training included words with those six phonemes (but neither *mes* nor *lot*). *Mes* contained more talker-specific phonetic detail than *lot*, and, accordingly, voices saying *mes* were recognized better, and with more confidence, than those saying *lot*. The amount of talker-specific detail in each of the six critical phonemes thus influenced what was learned about the four speakers. Since phonetic content thus shapes prototype-based voice categories, the ease with which speaker's voices can be learned depends on the words they say.

Andics, A. & McQueen, J. M. (in preparation). Phonetic content shapes implicitly-learned voice categories.

Introduction

Human voices are among the most often heard acoustic stimuli. A fundamental task of the perceptual system is to organize these stimuli into meaningful voice categories. As new voices are learned, listeners need to distinguish between irrelevant and relevant variation in the signal, and consider it to be within-category and across-category variation, respectively. One way to do so would be to form prototype-based representations of new voices. Norm-based coding is a powerful way to represent perceptual spaces. There is now behavioural (Papcun et al., 1989; Mullennix et al., 2011; Latinus and Belin, 2011) and neuroimaging evidence (Andics et al., 2010) for the norm-based coding of voices, but the limits of this representational capacity are unknown. Here we investigated three aspects of voice category learning.

First, we asked if it is only possible for listeners to learn individual voice categories (i.e., distinguish within- and across-talker variation) or if they are also able to acquire supra-individual voice categories (i.e., distinguish within- and across-group variation). Individual voice identity categories are very useful for person recognition, a fundamental social ability. But identifying larger categories, such as a talker's group membership, can sometimes be equally important. However, little is known about listeners' ability to learn supra-individual voice categories. We are able to distinguish groups of voices when grouping is supported by obvious acoustic differences (e.g., female versus male voices; Childers and Wu, 1991). But is it possible to create voice group identities when grouping relies on acoustics and feedback? Furthermore, is it possible to learn a large voice category?

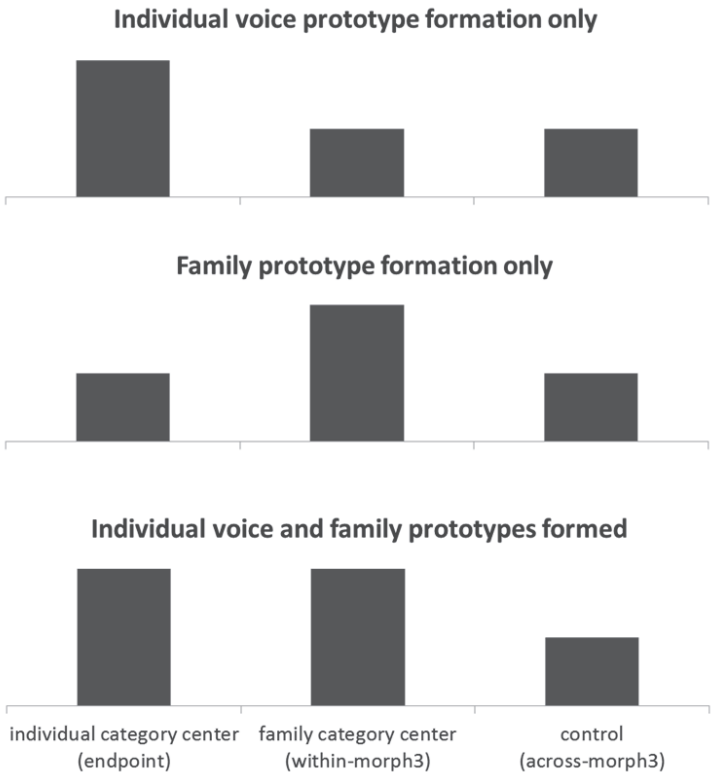
Second, is the presence of explicit feedback on voice decisions crucial, or can listeners learn new voice categories implicitly? Certain category contrasts are very hard to learn without explicit feedback. But distributional information in the sensory input alone can be sufficient for the acquisition of some categories (Goudbeek et al., 2009). So can voice identity categories be learnt without explicit feedback? To test this, we focused listeners' attention on voice identity information without giving them feedback on person identity. That is, we trained listeners via explicit feedback on group membership but not the person identity of the voices they heard. This made it possible to ask whether supra-individual voice categories are learnable (with explicit feedback) and to ask whether individual voice categories can be learnt implicitly.

Voice identity information is used in speech perception (e.g., Mullennix and Pisoni, 1990; Nygaard and Pisoni, 1998) and linguistic information is used in voice perception. For example, segmental information alone can be sufficient for listeners to identify talkers (Remez et al., 1997). Andics et al. (2007) found that phonetic content influences voice identity discriminability. They showed that changing a single segment in a CVC word could make voices less or more discriminable. For example, a word onset [m] supported voice discrimination more than a word onset [l] did, the vowel [ɛ] provided more voice identity information than the vowel [o], and [s] was more helpful than [t] in coda position. These differences seemed to be additive. Specifically, Dutch listeners were much better at discriminating talkers when the voices said *mes* (knife) than when they said *lot* (fate). These findings suggest that memory representations about voices contain suprasegmental properties such as pitch and timbre, and segmental properties. But little is known about whether and how phonetic content influences voice category learning.

Here we investigated these questions in a voice training paradigm. Dutch participants were trained through explicit feedback to identify the family membership of four Dutch talkers. At test, listeners were asked to categorize voice-morphed stimuli in-between the trained voices. They were also asked if these morphs were spoken by the trained voices or by new ones.

We hypothesized that differences in categorization performance on voice endpoints, within-family morphs and across-family morphs would reveal the structure of the category representations used. We assumed that if listeners form prototype-centered voice categories, then hit rate and categorization confidence would be higher for stimuli that were close compared to stimuli that were far from the prototype. We therefore predicted that if training on family membership leads to the formation of supra-individual family categories, then performance benefits would be found for within-family morphs (i.e., the stimuli close to the family prototypes), compared to across-family morphs or voice endpoints. Alternatively, if during family training listeners form individual voice categories through implicit learning, then performance benefits would be found for the voice endpoints (i.e., the individual voice prototypes) compared to both within- and across-family morphs. Finally, it was also possible that both individual and family prototype formation take place (see Fig. 1).

Fig. 1. Schematic bar graphs indicating expected performance (e.g., proportion of correct responses or level of confidence) for different morph levels during voice family categorization corresponding to the alternative predictions of (1) individual voice prototype formation, (2) family prototype formation, and (3) formation of both individual and family prototypes.



We also hypothesized that voice category formation would depend on phonetic content. Since voice discriminability varies across phonemes (Andics et al., 2007), it is likely that the ease with which a category is formed and the acceptance range of that category will do too. More specifically, we predicted that because Dutch listeners find voices saying *mes* more discriminable than voices saying *lot* (Andics et al., 2007), the current participants would find it easier to form voice categories based on *mes* than those based on *lot*, but also that they would accept less within-category variation for *mes* than for *lot*.

Method

Participants

Sixteen native Dutch listeners with no hearing disorders were paid to take part.

Stimuli

Eight Dutch CVC words, based on six phonemes, were selected (*mes, mos, met, mot, les, los, let, lot*). Tokens of these words and three five-minute stories were recorded by four native, young adult male speakers of Dutch with no recognizable regional accents and no speech problems (voice A, voice B, voice C and voice D). The voices were new to the listeners. The recordings were sampled at 44100 Hz, with 16-bit resolution. The stories were split into four similarly long sections. During the experiment, listeners heard complete stories, reconstructed such that consecutive sections were from different talkers.

For two of the words, *mes* and *lot*, within-word voice morph continua were created (see Fig. 2). Morphing was done in Matlab using STRAIGHT (Kawahara, 2006). STRAIGHT decomposes the speech signal into three parameters: a voice source, a noise source and a dynamic spectral filter. We supplied the algorithm manually-determined anchor points for the onset and offset each phoneme. The algorithm then generated intermediate steps between two original tokens by finding analogous time points according to the anchor points, and used the three signal parameters to generate intermediate morphs. Voice morph continua with six equidistant intermediate levels were resynthesized. Endpoints were also resynthesized. Two continua were created per word, per voice pair, using different tokens, making 16 continua in total (2 token-pairs x 2 words x 4 voice-pairs; see Fig. 2).

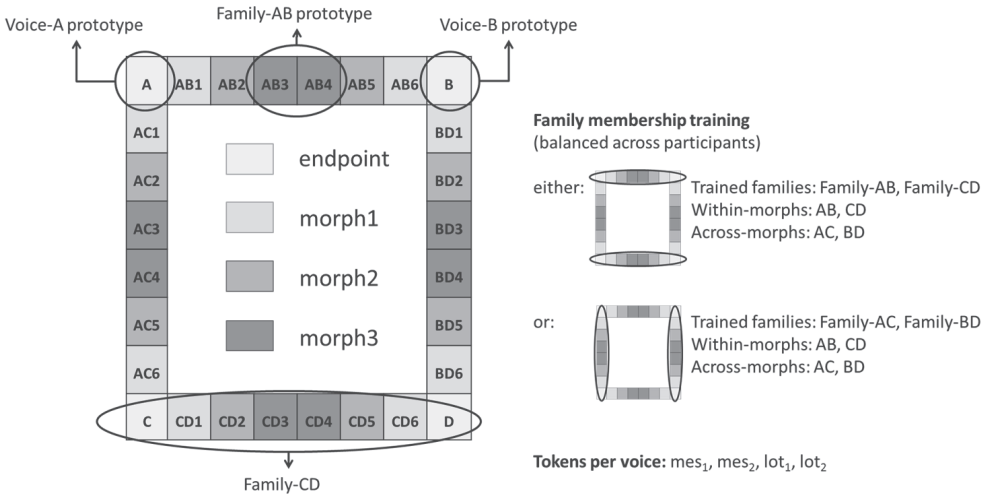
The stories and the six other words were natural. Average CVC duration was 565 ms. Average amplitude was equalized over all stimuli. Listeners reported after the experiment that they thought they had heard only natural stimuli. All word stimuli are available as supplementary material (http://www.mpi.nl/people/andics-attila/research).

Procedure and design

Stimuli were presented via headphones binaurally, at a comfortable volume. Listeners were told that they would learn the voices of several brothers from two families

and that they would later be asked to distinguish between the families. Families were determined by two voices (counterbalanced, see Fig. 2). Listeners were not informed that families included two voices only.

Fig. 2. Experimental design: voice endpoints and morphed stimuli used at tests. The squares A, B, C and D refer to the corresponding voice endpoints. Each voice endpoint is instantiated by four tokens, mes_1, mes_2, lot_1, lot_2. The sets of six numbered squares AB, CD, AC and BD refer to the six intermediate voice morph levels of the corresponding two voices. Family membership training was balanced across participants: for half of the listeners Family-AB and Family-CD were trained; for the other listeners Family-AC and Family-BD were trained. Four within-word continua were used for each family: mes_1 – mes_1, mes_2 – mes_2, lot_1 – lot_1, lot_2 – lot_2. Each voice morph is of type 1, 2 or 3, where the number indicates distance from the closest natural endpoint voice. Endpoints represent individual voice prototypes. Morph3 stimuli for trained families represent family prototypes. Within-family morphs are those sampling trained families, across-family morphs are those sampling the other two continua.



The experiment consisted of three phases, each with three parts. In Part 1 of all three phases, listeners were instructed to listen to one story (Story 1, 2 or 3) and try to memorize the voices and their family membership (training with stories). In Part 2 of all phases, listeners heard words and made two-alternative forced choice decisions on trained family membership (training with words). Six tokens of the six training words were used for each of the four voices (144 trials). Visual feedback (i.e., whether responses were correct, incorrect or late) was provided. Trial length was 2500 ms. Responses were possible until

2000 ms. Feedback was displayed from 2000 to 2250 ms. In Part 3 of all phases, listeners heard word stimuli sampling the morph continua from the non-trained words *mes* and *lot* in a test with no feedback (test with voice-morphed words). Stimuli included natural voice endpoints and six intermediate steps for the four continua (see Fig. 2). Two of the continua were within-family; two were across-family. In total, Part 3 included 112 trials (4 endpoints x 2 words x 2 tokens, plus 4 continua x 6 steps x 2 words x 2 tokens). No stimulus was repeated within any part of the experiment. Stimulus ordering was pseudorandom and varied across listeners.

There were differences across phases in Parts 1 and 3. In Phase 1, Part 1 consisted of listening to Story 1 twice, first by the two members of one family (e.g., Family-AB: voice A, B, A, B), then by the two members of the other family (e.g., Family-CD: voice C, D, C, D), with voices interleaved. In Part 3 of Phase 1 (Test 1), listeners made old-new judgments on a six-step scale after every word (1: voice surely heard before … 6: voice surely not heard before). This was a voice recollection task which did not depend on trained voice family membership. Trial length was 2500 ms. Responses were possible until trial offset.

Phases 2 and 3 were the same throughout, but they were different from Phase 1 in several aspects. In Phases 2 and 3, Part 1 consisted of listening to Story 2 or Story 3, respectively, including all four voices, with family membership interleaved (e.g., voice A_{AB}, C_{CD}, B_{AB}, D_{CD}). While listening, participants were informed on a screen about the family membership of each voice they heard. As in Phase 1, listeners heard stories and tried to memorize the voices and their family membership. In Part 3 of Phases 2 and 3 (Tests 2 and 3), listeners heard the same stimuli as in Test 1, but had a different task. They made two-alternative forced choice decisions on trained family membership, and then made a second button press reporting on a six-step scale the level of confidence the first decision was made with (1: maximally uncertain ... 6: maximally confident). Trials in Part 3 of Phases 2 and 3 were self-paced. The trial start was signaled with a 250 ms long fixation cross. Stimulus onset was at 450 ms. As soon as the first response was made, a question mark was displayed and remained on the screen until the second decision was made. The trial ended 350 ms after the second decision.

Results and Discussion

Test 1: Voice recognition

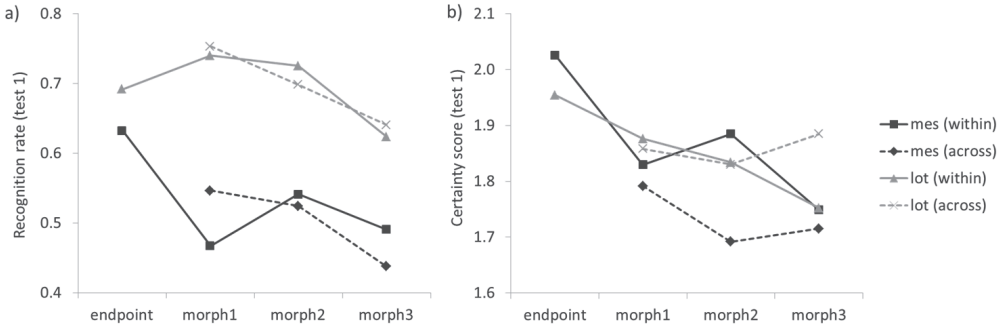
The recognition test was used to investigate if listeners perceived previously unheard stimuli as voices they had or had not heard before, and whether this was modulated by phonetic content or family training.

The recognition rate for the endpoint stimuli (see Fig. 3a) was higher than 50% for *lot* ($t(15) = 3.931$, $p = .001$) and marginally so for *mes* ($t(15) = 2.064$, $p = .057$). While *lot* morphs were rated as already-heard voices in more than half of the cases (morph1/2/3: ts = 6.634/6.220/2.650, ps = .000/.000/.018), this was not so for *mes* morphs (morph1/2/3: ts < .784, ps > .445).

A repeated-measures ANOVA on the proportion of recognized voices (i.e., rated as already heard vs not yet heard, collapsing across the 6 confidence levels into these two classes) was performed with the factors word (*lot, mes*) and level (endpoints and morphs, defined with respect to stimulus distance from the closest natural endpoint voice, measured in steps: 0 [= endpoints], 1, 2 and 3). It revealed a word effect ($F(1,15) = 13.936$, $p = .002$) and a word by level interaction ($F(3,45) = 3.331$, $p = .037$). The level effect was not significant ($F(3,45) = 2.229$, $p = .125$). Follow-up comparisons revealed that recognition ratings for *lot* were higher than those for *mes* at each morph level (morph1/2/3: ts = 3.266/2.796/2.231, ps = .005/.014/.041), but not for the endpoints ($t(15) = 1.541$, $p = .144$). Finally, no difference was found in recognition rate between within-family and across-family morphs for either word for any morph level (ts < 1.150, ps > .269).

These data demonstrate that the voices were learned and that voice knowledge transferred from trained to test words, but also that *lot* morphs were more accepted as already heard voices than *mes* morphs. This suggests that voice identity acceptance ranges vary across words, and hence that the size of individual voice identity categories is modulated by phonetic content. No difference in recognition performance between within- and across-morphs shows that the family training did not modulate the recognition of the morphs. Within-family morphs were not perceived as more familiar than across-family morphs, indicating a failure of family prototype formation.

Fig. 3. Voice recognition performance (Test 1).



Test 1: Recognition confidence

Confidence ratings were used to measure how distinctive certain tokens were. We assumed that as distinctiveness of a voice stimulus decreases, certainty of identity decisions for that voice should decrease too.

We performed a repeated-measures ANOVA with the same data as before, but now on mean confidence ratings (1: uncertain, 3: confident), with the factors word (*lot, mes*) and level (endpoint, morph1, morph2, morph3). A main effect of level ($F(3,36) = 6.664$, $p = .007$) and a word by level interaction ($F(3,36) = 3.931$, $p = .037$) were found. The word effect was not significant ($F(1,12) = .46$, $p = .511$). Direct comparisons revealed higher confidence for the endpoints than for each of the morph levels for *mes* but not for *lot* (*mes*: morph 1/2/3: ts = 2.825/3.426/4.560, ps = .015/.005/.001; *lot*: ts < 1.670, ps > .119). No difference in recognition confidence was found for either word for any morph level, except for a weak effect on morph2 of mes ($t(15) = 2.182$, $p = .045$; all other ts < 1.488, ps > .157); see Fig. 3b.

These results demonstrate that the voice endpoints were perceived as more characteristic of the voices than the morphs, but only for *mes.* Taken together with the recognition results, this suggests that a word which has a narrower voice identity acceptance range (i.e., *mes*), is more characteristic, within that range, than another word with a broader range (i.e., *lot*). Outside that range, however, there is no difference in distinctiveness between words with narrower vs broader acceptance ranges.

Tests 2&3: Family categorization

The categorization tests were used as a measure of learning during training and of prototype formation. Better categorization performance was expected for stimuli around individual voice category centers and/or family category centers.

Categorization performance was above chance for both words, for the endpoints (*mes*: t(15) = 4.719, p < .001; *lot*: t(15) = 3.217, p = .006) and for most of the morphs (*mes*: morph 1/2/3, ts = 3.166/3.839/2.050, ps = .006/.002/.058; *lot*: morph2, t(15) = 2.578, p = .021; other comparisons n.s.).

A repeated-measures ANOVA on the proportion of correct voice family category decisions in Tests 2 and 3 was performed, with the factors test (2, 3), within/across family, word (*lot, mes*) and level (morph1, morph2, morph3). Endpoints were excluded because they were all within-family stimuli. We found a main effect of test (F(1,15) = 6.875, p = .019) and a marginal main effect of level (F(2,30) = 2.860, p = .074). The proportion of correct responses was higher in Test 3, and lower for the morphs further from the endpoints (i.e., closer to the family category centers); see Table 1. No other effects were significant.

Table 1. Family categorization hit rate (Test 2&3). Means (in bold) and corresponding standard deviations are shown per condition.

			within-morph1	within-morph2	within-morph3	across-morph1	across-morph2	across-morph3
		endpoint						
mes	Test 2	**0.59**	**0.547**	**0.504**	**0.492**	**0.551**	**0.609**	**0.586**
		0.196	0.182	0.269	0.216	0.249	0.273	0.203
	Test 3	**0.719**	**0.641**	**0.625**	**0.586**	**0.633**	**0.672**	**0.563**
		0.185	0.193	0.194	0.169	0.18	0.136	0.214
lot	Test 2	**0.57**	**0.59**	**0.559**	**0.512**	**0.547**	**0.535**	**0.488**
		0.182	0.207	0.19	0.207	0.176	0.177	0.12
	Test 3	**0.609**	**0.563**	**0.617**	**0.531**	**0.539**	**0.508**	**0.523**
		0.136	0.151	0.168	0.202	0.208	0.201	0.131

ANOVA on morph3 categorization with the factors test (Test 2, Test 3), within/across family and word (*lot, mes*) found no significant effects. An ANOVA investigated the mean hit

proportion values, collapsing across the within/across factor but including the endpoints, with the factors word (*lot, mes*) and level (endpoint, morph3). There was an effect of level ($F(1,15) = 13.19$, $p = .002$), but no word effect ($F(1,15) = 2.196$, $p = .159$), and no interaction ($F<1$). Follow-up comparisons revealed benefits for endpoints compared to morph3 for both words (*mes*: $t(15) = 2.660$, $p = .018$; *lot*: $t = 1.944$, $p = .071$); see Fig. 4a.

That is, endpoints were better categorized than morphs. These data indicate that listeners learned about the endpoint voices and applied this knowledge to the surrounding morphs, that is, that they formed individual voice categories centered around the endpoints. The fact that endpoints were better categorized as family members than morphs confirms that listeners performed the family categorization task without forming prototype-centered supra-individual voice family categories. This conclusion is also supported by the lack of a within/across effect. Such categories should have been centered around the mid-continuum morphs, but no benefits were found for such morphs.

Tests 2&3: Categorization confidence

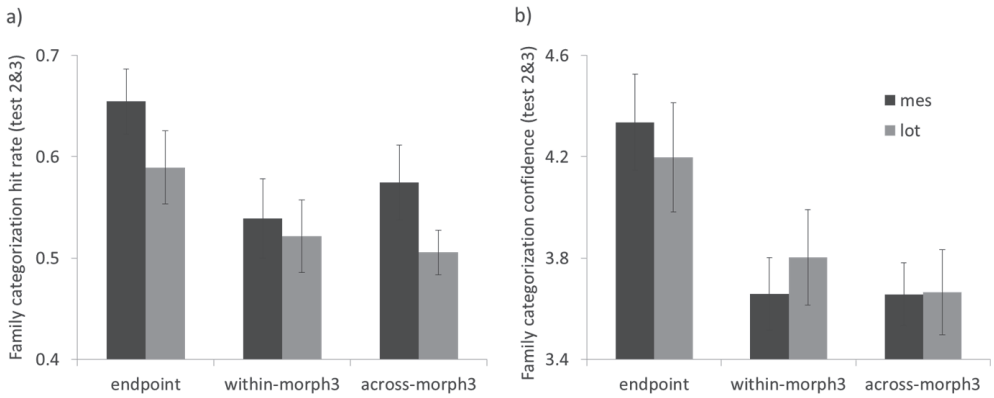
The confidence ratings were used to further characterize voice learning. A repeated-measures ANOVA was performed on the ratings (1: uncertain, 6: confident), with the factors test (2, 3), within/across family, word (*lot*, *mes*) and level (morph1, morph2, morph3 – endpoints were again excluded). A main effect of level was found ($F(2,30) = 22.878$, $p < .001$): categorization confidence was higher for endpoints than for morphs. We also found a significant interaction of test and word: while *mes* certainty increased, *lot* certainty decreased from Test 2 to Test 3 ($F(1,15)=5.893$, $p = .028$); see Table 2. The test by word interaction was carried mainly by the difference across words in Test 3, where higher certainty was found for *mes* than for *lot* ($t = 1.895$, $p = .077$, other comparisons n.s.). No other effects were significant.

Table 2. Family categorization confidence (Test 2&3), scaled from 1: maximally uncertain to 6: maximally confident. Means (in bold) and corresponding standard deviations are shown per condition.

		endpoint	within-morph1	within-morph2	within-morph3	across-morph1	across-morph2	across-morph3
mes	Test 2	**4.203**	**4.18**	**3.719**	**3.629**	**4.09**	**3.777**	**3.594**
		0.829	0.727	0.578	0.694	0.815	0.704	0.625
	Test 3	**4.469**	**4.156**	**3.867**	**3.687**	**4.203**	**3.992**	**3.719**
		0.737	0.813	0.775	0.604	0.669	0.633	0.628
lot	Test 2	**4.184**	**4.23**	**3.824**	**3.777**	**4.133**	**3.812**	**3.8**
		1.01	0.928	0.884	0.928	0.9	0.882	0.818
	Test 3	**4.211**	**3.914**	**3.797**	**3.828**	**3.883**	**3.766**	**3.531**
		0.914	0.73	0.629	0.727	0.724	0.769	0.684

Further analyses again focused on individual and family category centers (endpoints and morph3 stimuli). An ANOVA on morph3 categorization confidence, with the factors test (2, 3), within/across family and word (*lot, mes*), found no significant effects. An ANOVA on confidence scores collapsed across the within/across factor but included the endpoints, with the factors word (*lot*, *mes*) and level (endpoint, morph3). There was an effect of level ($F(1,15) = 16.575$, $p = .001$), but no word effect ($F<1$) and no interaction ($F(1,15) = 1.571$, $p = .229$). Follow-up comparisons revealed benefits for endpoints compared to morph3s for both words (*mes*: $t(15) = 4.267$, $p = .001$; *lot*: $t = 2.728$, $p = .016$); see Fig. 4b.

Fig. 4. Family categorization performance (Test 2&3). Error bars indicate the standard error of the mean.



These results corroborate the categorization hit rate findings: endpoints are categorized with greater certainty, and there is no confidence difference between within- and across-family morphs. This pattern of categorization performance is consistent with the prediction that the family categorization task was performed using individual voice categories formed around the endpoint voices, and inconsistent with the prediction that it is based on the formation of voice family prototypes. These data also indicate that more training with the voices helped listeners gain confidence for voice decisions on *mes* but not on *lot* tokens. This, together with the categorization results, suggests that learning about the voice family categories was easier through more distinctive phonemes (those in *mes)* than through less distinctive ones (those in *lot)*.

Conclusions

We investigated the formation of individual and supra-individual voice categories. Listeners were given explicit feedback on voice family membership but not on individual voice identities and then tested on within- and across-family voice-morphed stimuli based on the trained voices. We predicted that if categories are formed around individual voice prototypes, then better performance would be found for natural voice endpoints than for the morphs. But if categories are formed around trained voice family prototypes, then better performance would be found for within- than for across-family morphs. We found

that the family contrast was learned, but there was no evidence for prototype-centered supra-individual voice category formation: no post-training benefit was found for within-family compared to across-family morphs in either a categorization or a recognition task. This indicates limits on voice category formation. Although there is behavioural (Papcun et al. 1989; Mullennix et al. 2009; Latinus and Belin 2011) and neuroimaging evidence (Andics et al., 2010) for the norm-based coding of voices, none of these studies investigated category formation in cases where within-category variation was larger than typical within-talker acoustic variation, as was done here (within-family variation was larger than normal within-talker variation). Using face morphs, Cabeza et al. (1999) demonstrated that the face prototype effect (i.e., better performance on a face corresponding to the never-seen central value of a series of seen faces) tends to disappear when face exemplars are more different than what one would expect from exemplars of an individual face. We propose that the lack of a voice group prototype effect in the present study captures a similar category-size restriction for voice category formation.

Our results also confirmed that individual voice categories can easily be formed, and established that this occurs even without feedback. We found a post-training benefit at test in categorization and recognition of voice endpoints compared to intermediate morphs. We propose that these newly-formed individual voice categories are centered around the endpoints and helped listeners to categorize new voice exemplars.

Finally, we predicted that phonetic content would modulate voice category formation such that words with more distinctive phonemes would support voice learning but would make voice categories more sensitive to variation. We found that this was the case. Specifically, *lot* morphs were more often recognized as heard voices than *mes* morphs, but *mes* morphs were better categorized than *lot* morphs. Furthermore, training increased voice categorization confidence for *mes* more than for *lot*. Andics et al. (2007) showed that the phonemes of *mes* are more distinctive than those of *lot*, indicating that *mes* exemplars might contain more talker-specific detail than *lot* exemplars. We can hence conclude that less within-talker variation is accepted for a voice saying a word with phonemes with more talker-specific detail. That is, *mes* seems to determine a narrower voice identity acceptance range than *lot* does, presumably because the phonemes of the former word were more informative about talker identity during training than those in the latter word. In turn, this increased strictness in the voice identity category leads to greater distinctiveness and better

learnability for voice identity categories. The ease with which a talker's voice can be learned, and what is learnt about that voice, thus does indeed depend on the words that talker says.

References

Andics, A., McQueen, J.M., Van Turennout, M., 2007. Phonetic content influences voice discriminability. In J. Trouvain, & W. J. Barry (Eds.), Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007) (pp. 1829-1832). Dudweiler: Pirrot.

Andics, A., McQueen, J.M., Petersson, K.M., Gál, V., Rudas, G., Vidnyánszky, Z., 2010. Neural mechanisms for voice recognition. Neuroimage, 52(4), 1528-1540.

Cabeza, R., Bruce, V., Kato, T., Oda, M., 1999. The prototype effect in face recognition: Extension and limits. Memory & Cognition, 27, 139-151.

Childers, G., Wu, K., 1991. Gender recognition from speech. Part II: fine analysis. Journal of the Acoustical Society of America, 90, 1841-1856.

Goudbeek, M., Smits, R., Swingley, D., 2009. Supervised and unsupervised learning of multidimensional auditory categories. Journal of Experimental Psychology: Human Perception and Performance, 35, 1913-1933.

Kawahara, H., 2006. STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoustic Science and Technology, 27(6), 349-353.

Latinus, M., Belin, P., 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. Frontiers in Psychology. 2, 1-12.

Mullennix, J.W., Ross, A., Smith, C., Kuykendall, K., Conard, J., Barb, S., 2011. Typicality effects on memory for voice: Implications for earwitness testimony. Applied Cognitive Psychology 25, 29-34.

Mullennix, J.W., Pisoni, D.B., 1990. Stimulus variability and processing dependencies in speech perception. Perception & Psychophysics, 47, 379-390.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Perception & Psychophysics, 60, 355-376.

Papcun, G., Kreiman, J., Davis, A., 1989. Long-term memory for unfamiliar voices. Journal of the Acoustical Society of America, 85, 913-925.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance, 23, 651-666.

Chapter 5

Neural mechanisms for voice recognition

Abstract

We investigated neural mechanisms that support voice recognition in a training paradigm with fMRI. The same listeners were trained on different weeks to categorize the mid-regions of voice-morph continua as an individual's voice. Stimuli implicitly defined a voice-acoustics space, and training explicitly defined a voice identity space. The predefined centre of the voice category was shifted from the acoustic centre each week in opposite directions, so the same stimuli had different training histories on different tests. Cortical sensitivity to voice similarity appeared over different time-scales and at different representational stages. First, there were short-term adaptation effects: Increasing acoustic similarity to the directly preceding stimulus led to haemodynamic response reduction in the middle/posterior STS and in right ventrolateral prefrontal regions. Second, there were longer-term effects: Response reduction was found in the orbital/insular cortex for stimuli that were most versus least similar to the acoustic mean of all preceding stimuli, and, in the anterior temporal pole, the deep posterior STS and the amygdala, for stimuli that were most versus least similar to the trained voice identity category mean. These findings are interpreted as effects of neural sharpening of long-term stored typical acoustic and category-internal values. The analyses also reveal anatomically separable voice representations: one in a voice-acoustics space and one in a voice identity space. Voice identity representations flexibly followed the trained identity shift, and listeners with a greater identity effect were more accurate at recognizing familiar voices. Voice recognition is thus supported by neural voice spaces that are organized around flexible 'mean voice' representations.

A version of this paper appeared as Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. NeuroImage, 52, 1528-1540.

Introduction

The ecological significance of voices is reflected in the existence of regions in the primate (Petkov et al., 2008) and human cortex (Belin et al., 2000) that are specially tuned to conspecifics' vocalizations. Voices are used very efficiently for person recognition (e.g., Schweinberger et al., 1997). To do that, listeners need to link variable voice encounters to stable voice identity categories. But how the brain could represent voice identities is still largely unknown. That is the central question of this paper.

To identify mechanisms that support voice recognition, one needs to separate voice identity representations from earlier levels of voice processing. It has been suggested that a voice structural processing stage which is sensitive to voice-acoustic changes is anatomically separable from a voice identity processing stage which is sensitive to changes in voice identity (Belin et al., 2004; Campanella and Belin, 2007). Voice-acoustic analysis has been proposed to take place in voice-sensitive regions of the bilateral superior temporal sulci (Belin et al., 2000; Belin, Zatorre and Ahad, 2002; von Kriegstein et al., 2003, 2005), and voice identity analysis has been linked to regions of the right anterior temporal lobe (Nakamura et al., 2001; von Kriegstein et al., 2003, 2005; von Kriegstein and Giraud, 2004; Belin and Zatorre, 2003; Lattner et al., 2005; Sokhi et al., 2005).

Although this previous research has contributed considerably to our understanding of the separation of different voice processing stages, the precise nature of the underlying neural mechanisms at each of these stages is still unknown. One aim of this study was to address this issue. Furthermore, there is a common difficulty in the interpretation of many of the studies that have claimed to distinguish voice identity representations from earlier levels of voice processing. This is that their critical contrasts were based on acoustic manipulations (e.g., Belin and Zatorre, 2003; Belin et al., 2000; Belin et al., 2002), task changes (e.g., Stevens, 2004; von Kriegstein et al., 2003), or both (e.g., von Kriegstein and Giraud, 2004). The proposed separation of voice processing stages may possibly reflect these acoustic and/or task differences. A second aim of the present study was therefore to try to distinguish between these processing stages with acoustic and task differences controlled. Several other cortical regions have also been implicated in voice processing in both primates and humans, including the anterior insular cortex (Remedios et al., 2009; Wong et al., 2004), the ventrolateral prefrontal cortex (Romanski et al., 2005; Fecteau et al.,

2005), and paralimbic regions including the amygdala (Lloyd and Kling, 1988; Fecteau et al., 2007). A third aim was to try to clarify the role of these areas in voice recognition.

A useful voice processing mechanism positions voice stimuli in an object space. FMRI evidence on natural object processing suggests that stimuli that are more typical within an object space elicit reduced neural responses (Loffler et al., 2005; Myers, 2007; Belizaire et al., 2007). A possible neural mechanism for object space representation is based on neural sharpening: with experience, the coding of central values in relevant object dimensions becomes sparser (for a recent review, see Hoffman and Logothetis, 2009). Neural sharpening reflects long-lasting cortical plasticity and is thus suitable for positioning stimuli in an object space over the long term. Long-term neural sharpening has been demonstrated in a face space (Loffler et al., 2005). In a study on face-identity processing, reduced haemodynamic responses were found in the fusiform face area for central stimuli only when those were also central in the long-term stored face space of the viewer (referred to as 'mean face' stimuli, Loffler et al., 2005), suggesting that long-term central faces are encoded more sparsely. Based on these results and on behavioural findings that have indicated a prototype-centered representation of voices in long-term memory (Papcun et al., 1989; Bruckert et al., 2010; Mullennix et al., 2011), we can expect a typicality-based neural sharpening mechanism for voices similar to that found for faces.

But long-term neural sharpening is not the only mechanism that can explain response reduction for central stimuli. Another candidate mechanism is short-term neural adaptation: in case of fast and balanced stimulus presentation, neural response reduction for central stimuli can be a consequence of the on-average greater physical similarity of preceding events to central than to peripheral stimuli (Aguirre, 2007; Epstein et al., 2008). Short-term adaptation, just like neural sharpening, is sensitive to the object's relative position among similar objects, but in this case sensitivity is restricted to a very limited time scale. Short-term adaptation, in contrast with long-term neural sharpening, presupposes no long-term stored knowledge about the centre of the object space. But voice recognition cannot be successful without long-term stored information on person identity, that is, long-lasting voice identity representations. Voice-acoustic analysis, on the contrary, might be based on short-term mechanisms exclusively, or it might be supported by an automatically formed, long-term stored voice-acoustics space, with a 'mean voice' as its centre. No previous studies have found evidence for the existence of such 'mean voice'

representations. Here we attempted to identify long-lasting voice representations, and separate them from short-term stimulus similarity effects.

The present study evaluated two hypotheses. First, we attempted to confirm the hypothesis that person recognition from vocal information is mediated by anatomically separable stages of voice analysis (i.e., voice-acoustic analysis and voice identity analysis). Second, we tested the hypothesis that voice analysis at each of these stages is supported by neural representations of the stimulus space such that long-term stored typical values are coded more sparsely than atypical values, that is, that there are both voice-acoustic and voice identity spaces. To achieve these goals, we applied a learning-relearning paradigm. Listeners were trained to categorize the middle part of several voice-morph continua as a certain person's voice. Because perceptually relevant inter-speaker and intra-speaker variation are largely based on the same acoustic cues (Potter and Steinberg, 1950; Nolan, 1997; Benzeghiba et al., 2007), the stimuli, although they were made by morphing between voices, nevertheless modeled natural within-voice variability in the way each individual produces spoken words. The training hence simulated normal voice learning, where the same voice identity must be linked to variable tokens of words. The trained voice identity category was associated with a different interval on the voice-morph continua on each of two weeks for every listener. The voice-acoustics space was defined implicitly by the stimulus continuum used throughout the experiment, while the voice identity space was defined by explicit feedback during training. Training was followed by fMRI tests each week.

We thus investigated two equivalent contrasts with the same subjects, the same stimuli and the same task. One contrast measured voice-acoustic sensitivity and the other measured voice identity sensitivity. We predicted that if a neural region is sensitive to deviations from long-term stored typical values in either the voice-acoustic or the voice identity space, then that region will respond less strongly to acoustically central or trained identity-internal stimuli than to acoustically peripheral or trained identity-external stimuli respectively, while remaining insensitive to short-term adaptation effects. To reveal the contribution of long-term and short-term mechanisms behind these sensitivities, we separated the effect of stimulus similarity to the directly preceding voice stimulus from longer-lasting effects.

Materials and Methods

Participants

Twenty-five Hungarian listeners (14 females, 11 males, 19-31 years) with no reported hearing disorders were paid to complete the experiment. Written informed consent was obtained from all participants. One person was excluded because of a failure to perform the task during training. The analyses presented below were based on the remaining twenty-four subjects.

Stimuli

Recording. We recorded two young female non-smoking native Hungarian speakers with no speech disorders, saying the Hungarian words "bú" [sadness], "fű" [grass], "ki" [out], "lé" [liquid], "ma" [today] and "se" [neither] in standard Hungarian with no recognizable regional accent (voiceA and voiceB). These monosyllables were selected to cover various types of segmental content, with consonants varying in manner and place of articulation and in voicing, and with vowels varying in height, backness, roundedness and length. Speakers were similar in pitch (voiceA: 195 Hz, voiceB: 179 Hz), as shown by measurements averaged across the six words. Recordings were made in a soundproof booth using a Sennheizer Microphone ME62, a MultiMIX mixer panel, and Sony Sound Forge. All stimuli were digitized at a 16 bit/44.1 kHz sampling rate and were volume balanced using Praat software (Version 4.2.07; Boersma and Weenink, 2007).

Morphing. Voice morphing was then performed between the natural endpoint tokens of the two speakers, making one 100-step continuum per word (voiceA = morph0, voiceB = morph100). Intermediate steps were made using the morphing algorithms of STRAIGHT (Kawahara, 2006).

Perceptual rescaling. To ensure approximately equal perceptual distances between neighbouring steps on each of the stimulus continua, the morphs for each of the six words were subjected to perceptually-informed rescaling. A behavioural pretest was carried out in order to acquire psychophysical data which could then be used for re-labelling the morph steps. In this pretest, ten repetitions of seven steps (5, 20, 35, 50, 65, 80 and 95) of each of the six morph continua were presented, in random order, to 10 naive listeners who performed a forced-choice voiceA or voiceB categorization task (these listeners did not take

part in the main experiment). There was no training or feedback provided. The test directly followed an initial voice-to-response-button assignment, in which listeners were presented with a single repetition of all six natural endpoint tokens of each speaker. Group-averaged 'voiceB' response proportions per level for each continuum were then subjected to linear interpolation, to get estimates of how each step of each continuum would be perceived. All morph steps were then re-labelled to best match the corresponding, interpolated 'voiceB' response proportions. For example, after perceptual rescaling, morph20 for each word refers to the morph step on that word continuum whose identification proportion as 'voiceB' was closest to 20% in this pretest. Example stimuli are available as Supplementary Materials.

Training

Design. The voices were unfamiliar to all listeners. Listeners were trained to categorize the middle parts of the voice-morph continua as a certain person's voice (we call this the trained voice identity). They had to perform an A or not-A categorization task on each stimulus (Ashby and Maddox, 2005). They were asked whether the presented stimulus was an exemplar of the trained voice identity or of a different voice. A within-subject training manipulation was applied. The trained voice identity category was associated with a different interval on the voice-morph continua on each of two weeks for every listener, namely either the morph20-morph60 range or the morph40-morph80 range – these will be referred to as 'voice20-60 training' and 'voice40-80 training', respectively. The whole continuum was sampled each week, and listeners were presented with exactly the same stimuli (with a different trial order) during the two training sessions. The difference between the training conditions was restricted to the feedback that was provided. The order of the training sessions was counterbalanced: half of the listeners had voice20-60 training on the first week and voice40-80 training on the second week, while the other half of the listeners had the reverse order.

During training, 25 stimuli from each of the six 0-100 voice morph continua were presented, sampling the continua at approximately equal perceptual distances (a difference of 4 steps). The steps used were morphs 2, 6, 10, … , 90, 94, and 98. To maximize any training effect, the 8 stimulus steps that were closest to the critical 20, 40, 60 or 80 levels (i.e., those that were used at test) were presented twice as often as the rest (these steps

were 18, 22, 38, 42, 58, 62, 78, 82). There was, however, no difference in presentation frequency between central and peripheral stimuli. In each of two weeks participants received 80 mins of training over 2 days, with 4 training sessions of 16 min each on day1 and a single training block on day2. The first two blocks were blocked by word; in subsequent blocks the words were mixed. Training was followed by an fMRI test session on day2 in each week.

Procedure. Trial onsets were signaled with a question mark displayed in the middle of the screen for 300 ms. The auditory stimulus (a voice morph of one of the six words) began 200 ms after trial onset and lasted on average 456 ms. A response had to be made within 1800 ms of stimulus onset. Listeners received feedback on every trial. This feedback consisted of two parts. First, they saw an evaluation of their performance (i.e., whether the response was correct, incorrect or late) between 2000 and 2250 ms after trial onset. Second, this visual feedback was followed by auditory and visual reinforcement of learning. Listeners were presented with a repetition of the auditory stimulus, starting at 2700 ms after trial onset. This auditory reinforcement was accompanied with temporally synchronized visual reinforcement (a picture) presented between 2700 and 3450 ms after trial onset. If the stimulus morph was within the pre-defined trained voice identity category (in 42% of all trials), then this picture was a face (positive feedback). If the stimulus morph was outside the trained voice identity category, then a scrambled picture (matched in size, colour and contrast) was presented instead of the face (negative feedback). The same female face and the same scrambled picture were shown to all listeners in all training sessions on both weeks. We used the same face throughout the experiment in order to model natural voice learning, where acoustic variability in the realization of spoken words has to be mapped onto the same voice identity. The manipulation appeared to be successful in that all participants reported, after the experiment, that they thought that they had heard various exemplars of natural voices only and that they were convinced that the trained voice was an actual person's voice. The face was unfamiliar to all listeners before the experiment. They were told that it was the trained talker's face at the beginning of a half-minute long practice session on the training task which was presented before the first training session. The procedure ensured that every training stimulus was immediately repeated after the listener had made their choice, but for the second time with a visually

disambiguated talker identity. No response had to be made on the repeated stimulus. Trials had a duration of 5500 ms.

Conditions of interest. The critical stimuli in the fMRI test were morphs 20, 40, 60 and 80. The categorization training defined identity membership of these stimuli (internal, boundary and external), although these specific morph levels were not presented during training. During voice20-60 training, morph40 stimuli were category-internal, morph80 stimuli were category-external, and morph20 and morph60 stimuli were at the category boundaries. In contrast, during voice40-80 training, morph60 stimuli were internal, morph20 stimuli were external, and morph40 and morph80 stimuli were at the boundaries. Voice identity membership was trained by giving explicit feedback on every trial. Feedback was always positive for stimuli within the artificially determined trained voice identity interval, and it was always negative for stimuli outside this interval. During voice20-60 training, for example, morph steps greater than 20 but smaller than 60 were trained as internal through positive feedback, and morph steps smaller than 20 or greater than 60 were trained as external through negative feedback. An analogous procedure was used for voice40-80 training. As a consequence, out of the trained morph levels corresponding to the internal, boundary and external conditions at test, the proportion of morphs with positive feedback was 100, 50 and 0%, respectively. This defined the identity space. The stimuli therefore also differed in categorization ambiguity: it was expected that internal and external stimuli would be categorized less ambiguously and more accurately than boundary stimuli.

The critical voice morphs also differed in terms of their distributional position on the stimulus continua: morph40 and morph60 were close to the middles of the continua, while morph20 and morph80 were close to the endpoints – these morphs will be referred to as acoustic-central and acoustic-peripheral stimuli, respectively. Identity-internal stimuli were always acoustic-central, identity-external stimuli were always acoustic-peripheral, and identity-boundary stimuli were acoustic-central and acoustic-peripheral equally often. See Fig. 1a for an overview of the training and test design.

Analyses of training data. Voice category training data were collapsed across training blocks and days, and binned around the nine morph levels used at test (10, 20, …, 90) applying a +/- 4 morph step interval. This was done to enable a direct comparison of the training data to the fMRI test data (see Fig. 1b,c). The trained morph levels 2 and 98 were

not included in any bins. The non-critical morph level bins (10, 30, 50, 70, 90) comprised three stimuli that were actually used in training (the morph in the middle of the bin plus those in a 4-step distance in both directions, e.g., bin 10 comprised data corresponding to stimulus levels 6, 10 and 14). Each of these non-critical bins corresponded to 90 trials per condition, per subject (30 trials per stimulus level). The critical morph level bins 20, 40, 60 and 80 comprised two actually trained stimulus levels, in a 2-step distance in both directions (e.g., bin 20 comprised data corresponding to stimulus levels 18 and 22 – the actual morph level 20 was only presented at test). Every critical bin corresponded to 120 trials per condition, per subject (60 trials per stimulus level, as the number of repetitions on these critical stimulus levels was doubled).

fMRI test

Design and procedure. At fMRI test the task was the same as during training ("do you hear the trained voice identity or another voice?"), but no feedback was given. The 10-minute test contained 216 trials (four repetitions of six word continua, sampling each continuum with 9 morph levels, namely 10, 20, 30, 40, 50, 60, 70, 80 and 90), and a button-press response was expected after each stimulus. Trials had a duration of 2500 ms. Stimulus presentation was blocked by word continuum: all 9 levels of a word continuum were presented in each 9-trial-miniblock. Morph levels were therefore evenly distributed throughout the trial sequence. The word was different in consecutive miniblocks, and stimuli in consecutive trials were physically different. Stimulus ordering was otherwise random and varied across listeners. An example of a miniblock is: "lé"[30] -- "lé"[80] -- "lé"[10] -- "lé"[50] -- "lé"[40] -- "lé"[90] -- "lé"[20] -- "lé"[70] -- "lé"[60].

We explored the role of the task in an additional test in which subjects had to perform a word-repetition detection task by pressing a button when two consecutive words were the same. For this task the trained voice category-membership properties (i.e., whether they were exemplars of the trained voice identity or of another voice) were irrelevant. Two 9-minute runs with stimuli from the six trained word continua, sampled with the critical morph levels 20, 40, 60 and 80, were presented. At this test stimulus presentation was blocked by morph level, in 7-trial-miniblocks. Every miniblock contained each of the six words, and exactly one of them was repeated in each miniblock, in a randomly-chosen position within the block. An example of a miniblock at the irrelevant-task

test is: "ki"[40] -- "lé"[40] -- "bú"[40] -- "fű"[40] -- "fű"[40] -- "ma"[40] -- "se"[40]. A response was expected for the second "fű" stimulus but not for the other six stimuli in the block. Subjects were not informed about the frequency of word repetitions.

This irrelevant-task test preceded the relevant-task test each week. The constant order of tests was preferred to a balanced ordering because our focus was not on a direct comparison of the two tasks, but rather on a direct comparison of training effects across weeks within each test. We assumed that a constant order of tests would reduce noise caused by variation in listening history and in the amount of time already spent in the fMRI scanner.

Further tests included a single localizer run for voice-sensitive regions in the first week (including blocks of vocal and nonvocal sounds, using the stimuli from Pernet et al., 2007, with passive listening), and one for face-sensitive regions (including blocks of faces, houses, objects and matched scrambled objects, with a picture repetition detection task) in the second week.

Stimuli were presented at a standard, comfortable volume. Stimuli were controlled using Presentation software (Version 10.2; www.neurobs.com). During imaging, stimulus presentation was synchronized by a TTL trigger pulse with the data acquisition. Stimuli were delivered binaurally through MRI-compatible headphones (MR Confon, Magdeburg, Germany).

Data acquisition. MRI measurements were performed on a Philips Achieva 3T whole body MR unit (Philips Medical Systems, Best, The Netherlands) equipped with an eight-channel Philips SENSE head coil. For the main tests EPI-BOLD fMRI time series were obtained from 27 transverse slices covering temporal lobes and the inferior part of the frontal lobes with a spatial resolution of 3.5 × 3.5 × 3 mm, including a 0.5 mm slice gap, using a single-shot gradient-echo planar sequence (parallel imaging; ascending slice order; acquisition matrix 64 × 64; FOV = 224 mm; TR = 2500 ms; TA = 1763 ms (i.e., 737 ms silent gap); TE = 32.3 ms; and flip angle = 90°). That is, the acquisition of each volume was followed by a 737 ms gap when the scanner was silent. Compared to standard sparse sampling methods, this close-to-continuous sampling method not only increased statistical power by increasing the number of data points, but also made it possible to haemodynamically model each individual stimulus. At the same time it was possible to present all auditory stimuli in silence

(stimulus onset time coincided with scanner silent gap onset). The relevant and irrelevant task runs included 265 and 225 volumes respectively.

For the voice localizer there were 29 transverse slices and a longer silent gap between acquisitions (TR = 10000 ms, including 2000 ms acquisition and 8000 ms silent gap; TE = 36.5 ms). For the face localizer we used continuous scanning with 31 transverse slices (TR = 2200 ms; TE = 37 ms). The voice and face localizer runs included 63 and 200 volumes respectively. All other parameters were identical to the main test settings.

In addition to the functional time series, a standard T1-weighted three-dimensional scan using a turbo-field echo (TFE) sequence with 180 slices covering the whole brain was collected for anatomical reference at the end of the second scanning session, with 1 × 1 × 1 mm spatial resolution.

Data analysis. Image preprocessing and statistical analysis were performed using SPM5 (www.fil.ion.ucl.ac.uk/spm). The functional EPI-BOLD images were realigned, slice-time corrected (except for the voice area localizer run, where each volume acquisition was followed by a four times longer silent gap, and in this case slice-time correction is known to be more harmful than helpful, Friston et al., 2007), spatially normalized, and transformed into a common anatomical space, as defined by the SPM Montreal Neurological Institute (MNI) T1 template. Next, the functional EPI-BOLD images were spatially filtered by convolving the functional images with an isotropic 3-D Gaussian kernel (10 mm FWHM). The fMRI data were then statistically analyzed using a general linear model and statistical parametric mapping (Friston et al., 2007). For the relevant task run, every single stimulus was modeled as a separate event. For the irrelevant task run, seven consecutive stimuli, all representing the same voice morph level, were modeled as a block. Conditions in the voice and face localizer runs were also modeled as blocks.

For the main analyses, condition regressors for the relevant and irrelevant task tests were constructed per morph level. Sensitivity to voice-acoustic stimulus similarities was measured in a test contrasting continuum-central and continuum-peripheral stimuli, but controlling for category membership properties by only including stimuli that were trained as identity boundaries. After voice20-60 training, these were morphs 20 and 60; after voice40-80 training these were morphs 40 and 80 (see Fig. 1a). Voice identity sensitivity was tested in a contrast that had an identical stimulus load to that of the acoustic contrast, but those stimuli now also entailed a training-induced identity manipulation. Trained internal

stimuli were compared to external stimuli (after voice20-60 training these were morphs 40 and 80 respectively, after voice40-80 training these were morphs 60 and 20 respectively; see Fig. 1a).

To determine the role of short-term stimulus similarity-based mechanisms in the relevant task test, an additional analysis was performed. For that, critical condition regressors (corresponding to morphs 20, 40, 60, and 80) were split into more regressors, based on a oneback-distance measure, that is, the morph level distance of the actual trial from the preceding one (regressors of the new model: c10, c20, c30, c40, c50, p10, p20, p30, p40, p50, p60, p70; i10, i20, i30, i40, i50, e10, e20, e30, e40, e50, e60, e70 – where the number refers to the oneback-distance and c = acoustic-central from acoustic test, p = acoustic-peripheral from acoustic test, i = identity-internal, and e = identity-external). For example, the condition c10 involved acoustic-central stimuli as used in the acoustic test (so only identity-boundary cases are included) for which the preceding stimulus was 10 morph steps distant (e.g., after voice20-60 training, this would comprise those morph60 trials that come after morph50 or morph70). The effect of short-term similarity sensitivity was then measured by comparing trials with the minimal one-back distance to trials with the maximal one-back distance (c10 + p10 + i10 + e10 < c50 + p50 + i50 + e50; distances larger than 50 were not available for all critical conditions).

This split regressors model was also used in confirmatory follow-up tests that were aimed at distinguishing long-term from short-term effects. They did so by controlling for short-term biases in the main acoustic and identity tests. In those tests, low one-back distances were more frequent and thus overweighted among acoustic-central and identity-internal trials, while high one-back distances were more frequent and thus overweighted among acoustic-peripheral and identity-external trials. In the follow-up tests equal weights were therefore assigned to all one-back distances. The main acoustic analysis contrast c < p was substituted with c10 + c20 + c30 + c40 + c50 < p10 + p20 + p30 + p40 + p50, and the main identity analysis contrast i < e was substituted with i10 + i20 + i30 + i40 + i50 < e10 + e20 + e30 + e40 + e50.

Realignment regressors were also included for each run to model potential movement artefacts. A high-pass filter with a cycle-cutoff of 128 s was implemented in the design to remove low-frequency signals. Single-subject fixed effect analyses were followed

by whole-brain random effects analyses on the group level. Significance levels were FDR-corrected.

Results

Behavioural results

The training was successful and had long-lasting effects: Listeners learned that the voice category was located in the middle of the presented stimulus continua, and they shifted this category during re-learning on the second week (Fig. 1b). The learning effect found during training was present at fMRI test as well (Fig. 1c). Repeated-measures ANOVAs on categorization responses during the training and then at the fMRI test examined the effect of condition (voice20-60 training or voice40-80 training) across nine morph levels (10, 20, …, 90; as described above, these levels for the training phase were created by binning data around these values). We found a main effect of morph level (training: $F(8, 184) = 257.89$, $p < .001$; test: $F(8, 184) = 70.21$, $p < .001$), no main effect of condition (training: $F(1, 23) = 1.40$, $p = .250$; test: $F(1, 23) = 1.18$, $p = .289$), and a significant condition by morph level interaction (training: $F(8, 184) = 21.44$, $p < .001$; test: $F(8, 184) = 67.47$, $p < .001$). Moreover, the quadratic trend was highly significant for this interaction during training and at test (training: $F(1,23) = 643.86$, $p < .001$; test: $F(1,23) = 287.17$, $p < .001$). We also found a significant linear trend during training but not at test (training: $F(1,23) = 97.04$, $p < .001$; test: $F(1,23) < 1$). The presented degrees of freedom are uncorrected, but were Greenhouse-Geisser corrected for F score calculations.

Recognition performance accuracy during training was calculated for every listener (mean d' = .85, SD = .19). For the d primes, hit rates versus false alarm rates were calculated from responses to all stimuli with positive versus negative feedback respectively. These recognition accuracy scores were later compared to neural sensitivity scores in correlation analyses.

Decision difficulty affected both recognition accuracy and response times (see Table 1). The training stimuli corresponding to the boundary stimuli were categorized with lower recognition accuracy than those corresponding to internal and external stimuli. Response times during training were significantly longer for trials corresponding to boundary stimuli than for trials corresponding to internal/external stimuli. The same pattern was observed at

test. Note that the stimulus load contributing to the easy and difficult conditions was identical.

Table 1. Voice recognition accuracy (d') and response times (RTs) at training and test

	boundary	**internal/external**	**t(23)**
training d'	.143 (+/- .136)	1.131 (+/- .324)	16.636*
training RT (ms)	940 (+/- 155)	924 (+/- 156)	4.047*
test RT (ms)	954 (+/- 186)	931 (+/- 182)	3.783*

The values refer to group mean and to standard deviations. Significant paired t-tests ($p < .001$) are denoted with *.

fMRI results

Acoustic sensitivity. This test contrasted continuum-central and continuum-peripheral stimuli, including only identity-boundary trials in each condition (see Fig. 1a). Large regions were found in a whole-brain analysis (FDR-correction, $p < .05$). Clusters that showed response reduction for central compared to peripheral stimuli included anterior, middle and posterior parts of the bilateral superior temporal sulcus (STS; BA 21, 22), the bilateral orbitofrontal cortex extending to the anterior insula (BA 47, 11) and the bilateral posterior ventrolateral prefrontal cortex (VLPFC) along the inferior bank of the inferior frontal sulcus (BA 44, 45) (see Fig. 2 and Table 2). No clusters were found in the opposite test.

Identity sensitivity. Here we compared identity-internal to identity-external stimuli in a contrast that had an identical stimulus load to that of the acoustic contrast (see Fig. 1a). Reduced BOLD responses were found for identity-internal compared to identity-external stimuli in the bilateral middle and posterior STS (BA 21, 22) extending ventromedially to the middle temporal gyrus in the right hemisphere, and medially to the Heschl's gyrus in the left hemisphere (BA 41); the bilateral anterior temporal pole (BA 38); the left amygdala; and a left deep posterior STS region (BA 39) in the proximity of the angular gyrus and the intraparietal sulcus (see Fig. 2 and Table 2). No regions were found in the reverse contrast.

There was a partial overlap of the posterior STS clusters found in the acoustic and the identity tests, in both hemispheres. There were no voxels in any other cortical areas that were significantly active in both the acoustic and the identity tests, not even at a more liberal threshold ($p < .001$, uncorrected).

Fig. 1. Design and behavioural results. (a) Training and test design. Stimulus position with respect to identity was defined via feedback during training: internal morphs were associated with positive feedback (+) and external morphs with negative feedback (-). For critical test stimuli (morphs 20, 40, 60 and 80; in bold), which were not presented during training, stimulus position with respect to identity was in half of the cases (red boxes) internal (I) for more central and external (E) for more peripheral stimuli, while in the other, stimulus-matched half of the cases (blue boxes) stimulus position was at the voice-category boundary (?) for both central and peripheral stimuli. (b) Proportion of 'trained voice' responses across binned morph levels during training, collapsing over all training blocks in each condition. Error bars represent the standard error of the mean (n = 24). (c) Proportion of 'trained voice' responses across morph levels at fMRI test, for each training condition. Error bars represent the standard error of the mean (n = 24).

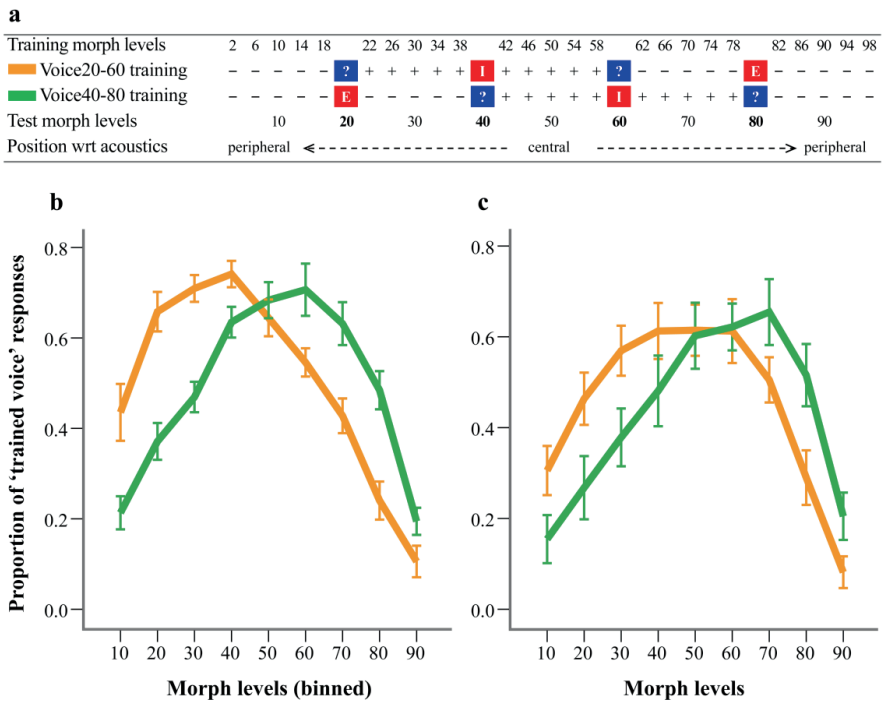


Table 2. List of regions found in the voice-acoustic and voice identity sensitivity tests

	BA	x	y	z	T	p (corr.)	mm³
acoustic sensitivity							
R anterior / middle / posterior STS	21/22	58	−38	6	8.16	0.001	11312
L anterior / middle / posterior STS	21/22	−50	−32	4	6.11	0.002	5248
R orbitofrontal cortex / anterior insula	47	42	16	−12	5.67	0.003	5184
R medial orbitofrontal cortex	11	10	20	−14	5.88	0.003	248
L orbitofrontal cortex / anterior insula	47/11	−22	14	−12	6.03	0.003	4936
R posterior VLPFC	44/45	36	6	34	6.62	0.002	4672
L posterior VLPFC	45	−44	16	28	4.21	0.009	88
identity sensitivity							
R middle / posterior STS	21/22	50	−20	−6	5.05	0.040	1416
L middle / posterior STS	21/22/41	−42	−38	4	6.01	0.037	3376
L deep posterior STS	39	−30	−58	22	5.40	0.037	304
R anterior temporal pole	21/38	48	18	−28	4.68	0.045	304
L anterior temporal pole	21/38	−54	10	−24	4.59	0.046	272
L amygdala	–	−30	−2	−20	4.86	0.041	464

A single peak per region is shown. Analyses were thresholded at t(23) > 4, cluster size > 10 voxels.

Correlation analyses. To investigate the behavioural relevance of the variation in neural activity found in the acoustic contrast and identity contrast, these tests were followed up by correlation analyses. Recognition performance accuracy during training, characterized by d-prime scores for every subject, was compared to neural sensitivity, characterized by the size of significant response reductions in regions found in either contrast. Behavioural scores were added to both the acoustic and the identity contrast's group design matrix as a regressor. In the context of the GLM, carrying out a t-test on the coefficient of this regressor is equivalent to testing the corresponding correlation.

Small volume correction analyses were performed for every activated cluster. Seven acoustic clusters and six identity clusters were investigated. Table 3 reports the local maxima and corrected p-values (corrected for the number of voxels within each cluster, but uncorrected for the number of tested clusters) for the behavioural regressor. Peaks with a significant correlation with recognition accuracy were found for identity clusters: the right middle / posterior STS (BA 21,22), the left deep posterior STS (BA 39), the right anterior temporal pole (BA 38), and the left amygdala (see Fig. 3). No significant positive correlations

were found for acoustic clusters. No significant regions showed negative correlations between acoustic or identity sensitivity and behaviour.

Table 3. Correlation of recognition accuracy and significant acoustic or identity sensitivity

	BA	x	y	z	T	p (corr.)
correlation with acoustic sensitivity	*[no clusters contained suprathreshold voxels]*					
correlation with identity sensitivity						
R middle / posterior STS	21/22	46	−14	−20	3.86	**0.020**
L middle / posterior STS	21/22	−40	−42	10	3.16	0.357
L deep posterior STS	39	−32	−60	24	3.56	**0.015**
R anterior temporal pole	38	56	8	−28	3.39	**0.030**
L anterior temporal pole	*[no suprathreshold voxels]*					
L amygdala	–	−30	2	−22	3.81	**0.028**

Correlation contrasts were thresholded at t(23) > 3. Small-volume correction was based on clusters from the corresponding main analyses, thresholded at t(23) > 4.

Fig. 2. Coronal and axial slices and sagittal views display significant acoustic sensitivity (blue), identity sensitivity (red) and short-term effects (green), thresholded at t(23) > 4.

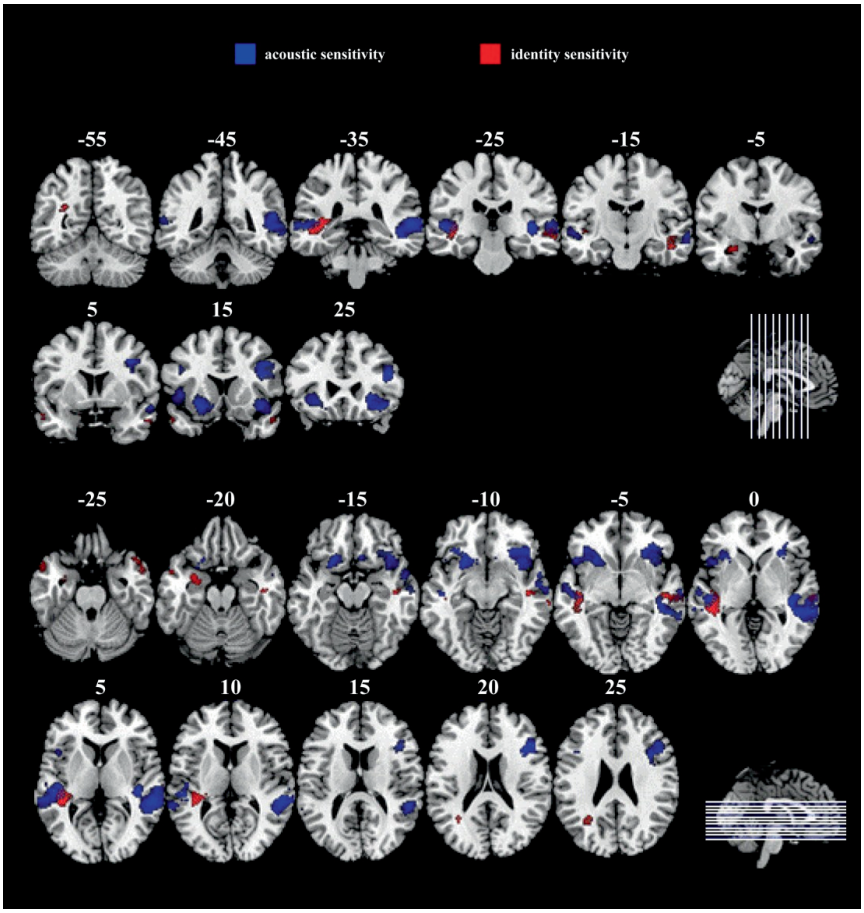
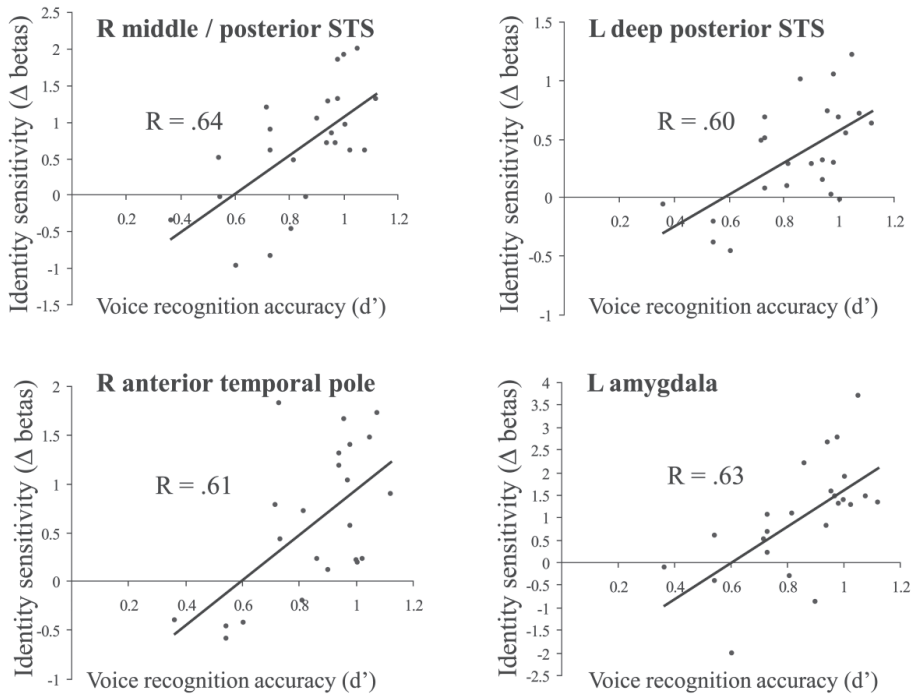


Fig. 3. Significant correlations between voice recognition accuracy scores and neural identity sensitivity for the peak coordinates defined in the correlation analyses (dots denote individuals).



Short-term effects. To determine whether the acoustic and identity effects could be caused by short-term perceptual similarity-based mechanisms, an additional analysis was performed. The short-term effect was measured in a contrast orthogonal to the acoustic and identity tests, by taking all critical conditions and comparing trials with the minimal distance between the stimulus and the immediately preceding stimulus (10 morph steps) to trials with the maximal distance between stimuli (50 morph steps). We expected that in regions sensitive to short-term stimulus similarities we would see an effect of one-back distance. Reported results were thresholded at the whole-brain level ($t > 4$, see Table 4, Fig. 2 and Fig. 4). Reduced BOLD responses were found for minimal-distance compared to maximal-distance stimuli in the bilateral middle / posterior STS (BA 21, 22), extending medially to the Heschl's gyrus (BA 42), and in the right hemisphere also ventromedially to the inferior temporal gyrus (BA 20). A further cluster was found in the right posterior

ventrolateral prefrontal cortex (BA 44). The bilateral temporal clusters overlapped with the bilateral STS clusters of both the acoustic and the identity test. The right VLPFC cluster also overlapped with that found in the main acoustic test (see Table 6). This suggests that the STS and right VLPFC clusters detected in the main acoustic analyses and the STS clusters found in the main identity analyses are findings that can at least partially be explained by short-term adaptation effects. No regions were found in the reverse contrast.

Table 4. List of regions found in the short-term acoustic similarity-sensitivity test

	BA	x	y	z	T	p (corr.)	mm³
short-term similarity-sensitivity							
R middle / posterior STS, ITG	20/21/22/42	48	-32	-6	6.04	0.026	6256
R posterior VLPFC	44	46	14	22	5.23	0.026	640
L posterior STS	22/42	-66	-38	12	5.22	0.026	592
L middle / posterior STS	21/22	-62	-26	-4	5.02	0.026	1136

A single peak is shown per region. The analysis was thresholded at t(23) > 4, cluster size > 10 voxels.

Long-term effects. We have seen that some but not all of the acoustic and identity effects could be explained by short-term similarity-based mechanisms. To confirm that brain regions with acoustic or identity sensitivity but without a sensitivity to short-term similarities were indeed based on long-term mechanisms, we followed up on the acoustic and identity tests in a confirmatory analysis. ('Long-term' here and throughout the paper refers to a time interval that is longer than the distance between two consecutive trials.) We used contrasts that were parallel to the main acoustic and identity analysis contrasts, but we defined the contrasts with separate regressors for each distance (10, 20, 30, 40, 50 morph steps) from the preceding stimulus, to control for short-term stimulus similarity effects.

Results were thresholded at t(23) > 3 and small-volume corrected for each of the corresponding main analysis clusters (seven acoustic or six identity clusters, thresholded at t(23) > 4, see Fig. 4). Table 5 reports the local maxima and corrected p-values (corrected for the number of voxels within each cluster, but uncorrected for the number of tested clusters) for the long-term acoustic and identity sensitivity tests. Long-term acoustic sensitivity (response reduction to short-term controlled central compared to short-term controlled peripheral stimuli) was found in the right orbital/insular cortex (BA 47, 11); and in the

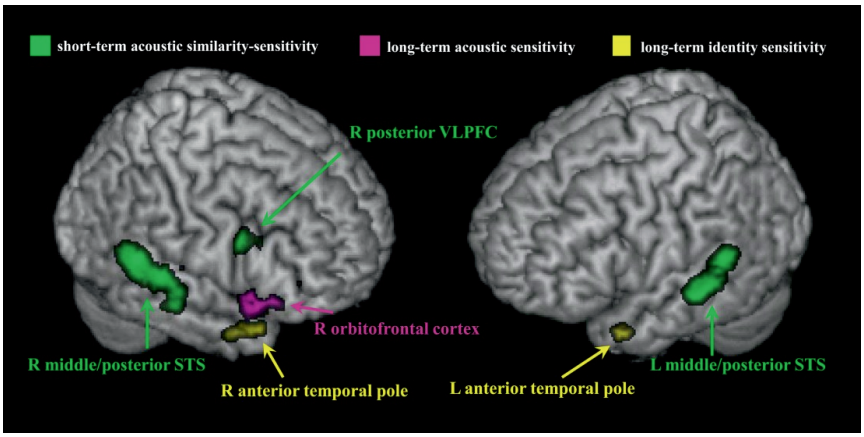
posterior medial portion of the right STS cluster, close to the junction of BA 20, 37 and 41. No significant long-term acoustic sensitivity was found in the left STS cluster, the VLPFC clusters and the left orbital/insular cluster. Long-term identity sensitivity (response reduction to short-term controlled identity-internal compared to short-term controlled identity-external stimuli) was found in the bilateral anterior temporal pole (BA 38); in the left deep posterior STS region (BA 39) and in the left amygdala. No significant long-term identity sensitivity was found in the middle/posterior STS clusters in either hemisphere. No clusters were found in the opposite tests. Although these confirmatory analyses are based on functionally non-independent small-volume corrections that can possibly result in false positives, they are nevertheless strict tests, since the largest STS clusters found in the main analyses did not survive them. These analyses thus suggest that activity in most of the brain regions that was found in the main acoustic and identity analyses, and that remained insensitive to short-term stimulus similarities, can indeed be explained by long-term mechanisms.

Table 5. List of regions found in the long-term acoustic and identity sensitivity tests

	BA	x	y	z	T	p (corr.)
long-term acoustic sensitivity						
R posterior medial temporal cortex	21/22	40	−26	0	4.79	0.012
R orbitofrontal cortex / anterior insula	47	44	18	−16	4.79	0.002
R medial orbitofrontal cortex	11	8	18	−16	3.53	0.009
long-term identity sensitivity						
L deep posterior STS	39	−30	−62	24	5.09	< 0.001
R anterior temporal pole	21/38	48	18	−28	4.39	0.002
L anterior temporal pole	21/38	−52	14	−28	4.73	0.001
L amygdala	–	−20	−8	−18	3.02	0.034

A single peak is shown per region. Long-term sensitivity contrasts were thresholded at t(23) > 3. Small-volume correction was based on clusters from the corresponding main analyses, thresholded at t(23) > 4.

Fig. 4. Sagittal views display short-term effect (green), thresholded at t > 4; long-term acoustic sensitivity effect (purple) and long-term identity sensitivity effect (yellow), thresholded at t(23) > 3 and masked by the corresponding main analyses thresholded at t(23) > 4.



Voice- and face-sensitivity. Voice-sensitivity was measured with a functional localizer (Pernet et al., 2007) using a contrast of voice stimuli versus matched non-voice stimuli. Face-sensitivity was measured with another functional localizer using a contrast of faces versus matched scrambled objects. The localizer activities were thresholded at t > 4 and narrowed down for the activated clusters of the acoustic and the identity test (Table 6). Among acoustic test clusters, a high proportion of voxels within the STS clusters showed voice-sensitivity, and the posterior part of the right STS also showed considerable face-sensitivity. Part of the right posterior VLPFC region from the acoustic test was also shown to be sensitive to voices but not to faces. In identity test clusters, the overwhelming majority of activated voxels in the bilateral middle / posterior STS and anterior temporal pole showed voice-sensitivity, but none showed face-sensitivity. On the contrary, the left amygdala as found in the identity test showed clear face-sensitivity but almost no voice-sensitivity. Interestingly, the left deep posterior STS region of the identity test which was also well correlated with recognition accuracy did not contain any voice- or face-sensitive voxels.

Table 6. Overlapping regions in main analyses and additional independent tests

	short%	voice%	face%
acoustic sensitivity			
R anterior / middle / posterior STS	29	89	28
L anterior / middle / posterior STS	4	95	< 1
R orbitofrontal cortex / anterior insula			
R medial orbitofrontal cortex			
L orbitofrontal cortex / anterior insula		3	
R posterior VLPFC	4	12	
L posterior VLPFC			
identity sensitivity			
R middle / posterior STS	14	92	
L middle / posterior STS	< 1	71	
L deep posterior STS			
R anterior temporal pole		100	
L anterior temporal pole		97	
L amygdala		2	90

The columns short%, voice% and face% show the proportion of voxels in each acoustically sensitive or identity-sensitive cluster that were also differentially active in the (1) short-term effect test (minimal-distance < maximal distance), (2) voice area localizer (non-vocal stimuli < voices) and (3) face area localizer (scrambled objects < faces) respectively (thresholded at t(23) < 4).

Lateralization. To directly compare hemispheric contributions to the two contrasts, lateralization indices were calculated from voxel values for the temporal lobes, where large clusters were found in both tests. Individual maps were thresholded at $p < .05$ uncorrected. Activity in the identity test was left-lateralized (mean(LI) = -.141, SD(LI) = .392), but in the acoustic test it was right-lateralized (mean(LI) = .182, SD(LI) = .406). There was a significant difference of individual lateralization indices in the temporal lobes between tests ($p = .025$, paired t-test).

The role of decision difficulty. To explore direct effects of decision difficulty on critical stimuli, a test comparing difficult and easy trials was performed. Difficult trials included the ambiguous identity boundary stimuli, that is, all stimuli of the acoustic test. Stimulus-matched easy trials included the unambiguously trained identity-internal and identity-

external stimuli, that is, all stimuli of the identity test. No significant voxels were found in either direction of the comparison (whole-brain analysis, FDR-correction, $p < .05$).

The role of the task. As noted in the Methods, there was a test where listeners performed a word repetition detection task instead of voice recognition, on the same stimuli. In an analysis of the fMRI data for this word repetition task, no significantly active regions were found for the same acoustic and identity contrasts as were used in the main analysis.

Discussion

Voice identity processing is separable from voice-acoustic processing

It has been proposed that the neural substrates for the recognition of voice identities are separable from general acoustic processing regions (see Belin et al., 2004 for a review). This view has been strengthened by reports on cortical regions that are differentially active in voice recognition tasks (Nakamura et al., 2001; von Kriegstein et al., 2003, 2005; Belin and Zatorre, 2003; Lattner et al., 2005; Stevens, 2004), and on selective deficits of voice identity recognition abilities (Van Lancker et al., 1988; Garrido et al., 2009). Nevertheless, until now there were few attempts to describe the neural mechanisms underlying voice identity representations. We identified identity-sensitive regions that are both functionally and anatomically distinct from acoustic-sensitive regions. While temporal lobe activity in the acoustic contrast was right-lateralized, it was left-lateralized in the identity contrast. This lateralization difference suggested that these stimulus-matched and task-matched contrasts indeed measure different functions. Identity-sensitive but not acoustically sensitive regions involved the voice-sensitive bilateral anterior temporal pole; the face-sensitive left amygdala; and a left deep posterior STS region which was not found in either of the functional localizer tests.

Voice identity but not voice-acoustic sensitivity was found to covary with person identification performance. This covariation suggests that the identity sensitivity we described is indeed useful for voice recognition: listeners with a greater neural sensitivity for voice identities are more accurate at recognizing familiar voices. Covariation between significant identity sensitivity and behaviour was found for voice-sensitive regions (the middle/posterior STS and the anterior temporal pole) in the right but not in the left

hemisphere. Right hemisphere biases in voice recognition have been reported both in imaging (Nakamura et al., 2001; von Kriegstein et al., 2003; von Kriegstein and Giraud, 2004) and in clinical studies (Van Lancker and Kreiman, 1987; Ellis et al., 1989; Van Lancker et al., 1989; Gainotti et al., 2003). Covariation was also found between neural and behavioural identity sensitivity in regions that were not differentially sensitive to voices in the voice-localizer test, namely the amygdala and the deep posterior STS in the left hemisphere. These covariations not only validate our identity test but are also among the first demonstrations of the direct behavioural relevance of voice identity representations. In addition, the fact that we did not find any significant covariation between neural sensitivity in acoustic regions and performance further strengthens our claim that identity processing is separable from acoustic processing.

Short-term similarity effects

Auditory stimuli that are similar to other, just presented stimuli are expected to elicit more reduced neural responses than dissimilar stimuli, in cortical regions that are sensitive to those auditory changes. This neural mechanism is known as the short-term carry-over effect (Aguirre, 2007), or, in its purest form in same versus different tests, as rapid fMR-adaptation (Grill-Spector and Malach, 2001). To reveal the possible contribution of short-term stimulus similarity-based mechanisms behind the sensitivities measured by our acoustic and identity tests, we separated the effect of stimulus similarity to the directly preceding voice stimulus from longer-lasting effects. Extensive regions were found in and around the bilateral middle/posterior STS (BA 21, 22) in both the acoustic and the identity tests. These were the only brain regions that were found to be differentially active in both main tests. Neural sensitivity in the right STS, as measured in the voice identity test but not in the voice-acoustic test, was even found to covary with person identification performance. Furthermore, we demonstrated that these temporal regions were involved in short-term similarity processing. These regions are very similar to the temporal voice areas (Belin et al., 2000) that have been found to respond differentially to voice stimuli in healthy subjects but not in autism (Gervais et al., 2004). The present findings confirm short-term stimulus similarity-sensitivity in the voice-tuned middle/posterior STS, and that better short-term sensitivity may lead to better voice recognition performance.

Only one further region, the right VLPFC, showed sensitivity to short-term stimulus similarity processing. This posterior ventrolateral prefrontal region on the inferior bank of the inferior frontal sulcus (BA 44, 45) was found bilaterally, but with strong right-hemisphere dominance in the main acoustic but not in the identity sensitivity test. The right ventrolateral prefrontal region, just as the bilateral STS, was also differentially sensitive to voice stimuli in general. This prefrontal region involves Broca's area in the left hemisphere and is known to be crucial for linguistic processing. Its right-hemisphere counterpart has been shown to be more active in nonverbal memory tasks with environmental sounds (Opitz et al., 2000). Additionally, right ventrolateral prefrontal regions have been proposed to be involved in voice analysis in both primates (Romanski et al., 2005) and humans (Fecteau et al., 2005). Our findings suggest that this right VLPFC region, similarly to the voice-tuned STS regions, participates in short-term voice-acoustic change detection.

Short-term sensitivity to acoustic similarities between voice stimuli in the middle/posterior STS and in the VLPFC confirms these areas' responsiveness to acoustic changes within the stimulus set. However, an area's involvement in a short-term cortical mechanism does not exclude its involvement in mechanisms based on long-term representations. The STS is a region that is highly heterogeneous functionally (e.g., Beauchamp et al., 2004), and the middle/posterior STS was proposed to be crucial for different stages of voice identity processing (von Kriegstein and Giraud, 2004; Warren et al., 2006). Recent findings also suggested VLPFC involvement in the representation of long-term stored objects (Latinus et al., 2009). It was therefore somewhat surprising that in our confirmatory analyses we found no evidence suggesting that STS or VLPFC regions would mediate long-term voice memory (except for a small right posterior medial temporal region close to the junction of BA 20, 37 and 41). One explanation is that, contrary to these earlier claims, the neural substrates of long-lasting object space representations, including acoustic-mean or category-mean voice representations, are located elsewhere. Alternatively, it is possible that long-term effects were indeed present in the STS and VLPFC, but were masked by co-existing short-term effects in the present design. Further investigations are needed to resolve this issue.

Voice-acoustics space representation

The acoustic sensitivity test contrasted acoustically central and peripheral stimuli. This contrast tested the hypothesis that during listening to stimuli from a voice morph continuum, an implicit prototype-formation process takes place in the voice-acoustics space, resulting in the creation of a long-term stored 'acoustic mean voice' representation and hence in long-lasting neural sharpening for acoustically central stimuli. This hypothesis was confirmed. Although some regions found in this test, including the STS and the VLPFC, were shown to be biased by covarying short-term similarity, other regions, including the bilateral orbitofrontal cortex extending to the anterior insula (BA 47, 11) did not exhibit short-term stimulus similarity-sensitivity. Furthermore, there was no difference in presentation frequency between central and peripheral stimuli at either training or test to motivate a long-term bias without an 'acoustic mean voice' representation. So the orbital/insular cortex activity in the acoustic sensitivity test can best be described as long-term stimulus similarity sensitivity. This claim was further supported by a confirmatory test looking for long-term acoustic space sensitivity: the bilateral orbital/insular cortex was found in this test but the STS and VLPFC regions were not (except for a small right posterior medial temporal region close to the junction of BA 20, 37 and 41). The anterior insula has been implicated in the processing of sound and more specifically speech information (Wong et al., 2004), and it has also been proposed to possibly play a role in processing vocal paralinguistic information such as vocal emotion or vocal identity (Remedios et al., 2009; Watson, 2009). Our findings do not confirm that the insula handles vocal identity information; instead, the response reduction for voice stimuli that were most versus least similar to the acoustic mean of all preceding stimuli suggests that 'acoustic mean voice' representations exist and that they may be created in the orbital/insular cortex. This acoustic mean voice seems to be created independently from any representation of trained voice-identities. Our results thus show that a perceptual typicality-based organisation arises automatically for voice representations, similarly to what has been reported for faces (Loffler et al., 2005).

Voice identity space representation

We hypothesized that voice analysis at the stage of identity processing is also supported by neural representations of the stimulus space in which long-term stored typical

values are coded more sparsely than atypical values. Our findings support this hypothesis. We found response reduction for identity-internal versus identity-external stimuli in regions (including the voice-tuned ATP, the amygdala and the deep posterior STS) that showed no response reduction for the same stimulus contrast when it was free from the identity manipulation. The response pattern of regions with an identity effect but no acoustic effect can be explained as a long-term neural sharpening effect induced by the explicit categorization feedback during training. These results and the finding of significant covariation between neural identity-sensitivity and behavioural sensitivity in almost all identity-sensitive clusters (except for the left ATP) therefore argue for the existence of a neural voice identity space and of 'trained category-mean voice' representations. This explanation is further supported by our additional analyses that confirmed the presence of long-term identity representations but found no effects of short-term stimulus similarity-sensitivity in the bilateral ATP, the left deep posterior STS and the left amygdala.

The finding of voice identity representations in the anterior temporal pole confirms existing reports about the anterior temporal lobe's role in voice identity processing (Nakamura et al., 2001; von Kriegstein et al., 2003, 2005; Belin and Zatorre, 2003; Lattner et al., 2005, Sokhi et al., 2005) and seems to support the idea that this region corresponds to the unimodal voice recognition module in the model proposed by Belin and colleagues (Belin et al., 2004; Campanella and Belin, 2007). The novelty of our ATP finding is that we demonstrated this voice-tuned region's involvement in the representation of a category mean-centered voice identity space, and showed the effect of individual identity space sensitivity on voice recognition performance. Anterior temporal lobe regions, however, have also been shown to be involved in person identity recognition for different modalities (von Kriegstein and Giraud, 2006), in the multimodal integration of person information (for a review, see Olson et al., 2007; but see also Turk et al., 2005) and in the 'what' processing pathway (Scott and Johnsrude, 2003; Belin and Zatorre, 2003). Furthermore, clinical reports suggest that voice identity recognition and supramodal person identity recognition can be selectively impaired after degeneration of the anterior temporal lobe (e.g., Hailstone et al., 2009). The location of anterior temporal lobe findings in the present study [48, 18, -28; -52, 14, -28] is in-between previously reported coordinates of supra-modal person recognition in the temporal pole (slightly superior to e.g., [46, 16, -40; -44, 16, -40] in Sugiura et al., 2006) and those of unimodal voice recognition in the anterior STG/STS (slightly inferior and

anterior to e.g., [57, 9, -21; 54,12,-15; 48, 6, -18] in von Kriegstein et al., 2003, or to [58, 2, -8] in Belin and Zatorre, 2003). We therefore cannot exclude the possibility that our anterior temporal pole findings correspond instead to a different stage in Belin and colleagues' model (Belin et al., 2004; Campanella and Belin, 2007), namely to the supramodal person identification stage. Note that other, non-neuroimaging research has also suggested that there may be distinct acoustic, unimodal and supramodal steps in person identification (Ellis et al., 1997; Neuner and Schweinberger, 2000). Further clarification of the distinction between unimodal and supramodal processing regions within the anterior temporal lobe will probably require a direct experimental comparison of these person identification steps. Furthermore, earlier studies have created some uncertainty with respect to whether voice identity processing in ATP regions is restricted only to the right hemisphere or is present bilaterally. Our results, although remaining inconclusive, offer a better view on this issue: we found identity-sensitivity in the ATP bilaterally, but voice recognition was shown to reflect only the right ATP sensitivity.

Voice identity representations were also found in a left deep posterior STS region (BA 39) in our study. Our knowledge about the possible role in object recognition of the deep posterior STS region is very limited. Brodmann area 39 is often considered to be part of the Wernicke's area (Wise et al., 2001), an important centre for speech processing. Sensitivity to biological motion (Grossman et al., 2000) and audiovisual integration of voice and face information (Kreifelts et al., 2007) has been found for close but more lateral parts of the posterior superior temporal gyrus. Additionally, the left but not the right angular gyrus and medial parietal regions were found to be sensitive to voice familiarity in a prosopagnosic patient with bilateral damage (Arnott et al., 2008). Neighbouring, but more medial brain regions of the precuneus/retrosplenial cortex have shown sensitivity to person familiarity (Shah et al., 2001), and have been proposed as possible loci of cross-modal person identity nodes (Campanella and Belin, 2007). We suggest that this deep posterior STS region close to the angular gyrus and the intraparietal sulcus may contribute to a modality-nonspecific person identity representation.

We also found the identity effect in the amygdala, with significant covariation between neural and behavioural sensitivity. The amygdala activity persisted in our confirmatory long-term identity effect test. The amygdala has been suggested to be involved in the processing of socially relevant stimuli such as faces (Breiter et al., 1996;

Morris et al., 1996; Whalen et al., 1998) and voices (Fecteau et al., 2007; Campanella and Belin, 2007), but the specific role of this region is debated. Belin et al. (2004, Campanella and Belin, 2007) proposed that during voice analysis distinct neural processing streams are responsible for the recognition of speech categories, emotions and identities, and that the amygdala is responsible for vocal emotion processing. But recent findings suggest an important role for the amygdala also in the processing of emotionally neutral face stimuli both in monkeys (Gothard et al., 2007) and in humans (Kleinhans et al., 2009). Recently, Kleinhans et al. (2009) found reduced neural habituation in the amygdala for neutral facial stimuli in autism, a complex developmental disorder characterized by deficits in social interaction. It has also been proposed that there is a paralimbic network including both the amygdala and the anterior temporal pole which is specialized for person identification (Olson, 2007). The amygdala seems to be tuned to emotional stimuli more than to neutral stimuli, and to faces more than to voices, but our results indicate that it nevertheless participates in the representation of person identity given neutral voice stimuli. This finding is in line with psychophysical and electrophysiological evidence suggesting that voice analysis modules are not fully independent (Campanella and Belin, 2007), for example, speech perception has been shown to influence voice perception (Remez et al., 1997; Perrachione and Wong, 2007; Perrachione et al., 2010), and vocal emotions have been shown to modulate early sensory processing (Spreckelmeyer et al., 2009). A better understanding of the amygdala's role will clearly help to clarify the interplay of different voice analysis modules and the separability of neural substrates for different object types conveyed by voice and face stimuli.

Interestingly, no regions with identity sensitivity were found when, in an additional test, listeners had to perform a voice-irrelevant word repetition detection task. This indicates that identity sensitivity requires the presence of a relevant task, confirming earlier reports that specified similar brain regions responsible for voice identity processing by manipulating task relevance but not stimuli (von Kriegstein et al., 2003, von Kriegstein and Giraud, 2004).

Flexibility in voice representation

Finally, this study demonstrates the dynamics of voice processing. Voices, although carrying information about an anatomically defined vocal tract, are modulated by less

permanent factors such as language, dialect, speech style, emotions, volume, speed, health situation etc. that are known to influence talker identification (Nolan, 1997; Perrachione and Wong, 2007; Perrachione et al., 2010). Indeed, speakers dynamically tune their voices to the situation they find themselves in (e.g., in phonetic convergence, speakers tend to talk more like their interlocutors as a conversation progresses; Pardo, 2006). Therefore, the human perceptual ability to adapt flexibly to dynamic object changes (Kourtzi and DiCarlo, 2006; Jiang et al., 2007) is especially important for voice stimuli (cf. Schweinberger et al., 2008). Consequently, neural representations of voice identities need to be highly plastic to support voice recognition. Our findings demonstrate listeners' flexibility in learning and representing voice identities. On the first week of the experiment, listeners rapidly learned a new voice identity and then, when a week later a different voice morph interval was associated with the same identity, they dynamically adapted their representations. Neural sharpening for a long-term stored 'category mean voice' followed the trained shift and therefore retuned the neural representation of the voice identity space.

Conclusion

Our results are in line with the proposal that voice recognition is supported by a categorical level of processing that is anatomically separable from voice structural processing (Belin et al., 2004). Our findings also confirm that there exist dissociable neural mechanisms for short-interval versus long-interval fMRI repetition suppression (Epstein et al., 2008). More specifically, we have argued for the existence of dynamic, long-lasting 'mean voice' representations at both voice-acoustic and voice identity stages of processing. In accordance with recent findings in behavioural studies of voice processing (Papcun et al., 1989; Bruckert et al., 2010; Mullennix et al., 2011) and with those in the face processing domain (Loffler et al., 2005), our demonstrations of neural 'mean voice' representations constitute the first neuroimaging evidence that voice representations are centered around prototypes in long-term memory.

References

Aguirre, G.K., 2007. Continuous carry-over designs for fMRI. NeuroImage 35, 1480-1494.

Arnott, S.R., Heywood, C.A., Kentridge, R.W., Goodale, M.A., 2008. Voice recognition and the posterior cingulate: an fMRI study of prosopagnosia. Journal of Neuropsychology 2(1), 269-286.

Ashby, F.G., Maddox, W.T., 2005. Human category learning. Annual Review of Psychology 56, 149-178.

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., Martin, A., 2004. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nature Neuroscience 7(11), 1190-1192.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences 8, 129-135.

Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. NeuroReport 14, 2105-2109.

Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal lobe responses to vocal sounds. Cognitive Brain Research 13, 17-26.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309 –312.

Belizaire, G., Fillion-Bilodeau, S., Chartrand, J.P., Bertrand-Gauvin, C., Belin, P., 2007. Cerebral response to 'voiceness': a functional magnetic resonance imaging study. NeuroReport 18, 29-33.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., 2007. Automatic speech recognition and speech variability: A review. Speech Communication 49(10-11), 763-786.

Bergerbest, D., Ghahremani, D.G., Gabrieli, J.D.E., 2004. Neural correlates of auditory repetition priming: reduced fMRI activation in the auditory cortex. Journal of Cognitive Neuroscience 16(6), 966-977.

Boersma, P., Weenink, D., 2005. Praat: Doing phonetics by computer (Version 4.2.07.) [Computer programme]. <http://www.praat.org/>

Breiter, H.C., Etcoff, N.L., Whalen, P.J., Kennedy, W.A., Rauch, S.L., Buckner, R.L., Strauss, M.M., Hyman, S.E., Rosen, B.R., 1996. Response and habituation of the human amygdala during visual processing of facial expression. Neuron 17, 875-887.

Bruce, V., Young, A., 1986. Understanding face recognition. British Journal of Psychology 77, 305-327.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G.A., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Current Biology 20, 116-120.

Campanella, S., Belin, P., 2007. Integrating face and voice in person perception. Trends in Cognitive Sciences 11, 535-543.

Ellis, A.W., Young, A.W., Critchley, E.M.R., 1989. Loss of memory for people following temporal lobe damage. Brain 112, 1469-1483.

Ellis, H.D., Jones, D.M., Mosdell, N., 1997. Intra- and inter-modal repetition priming of familiar faces and voices. British Journal of Psychology 88(1), 143-156.

Epstein, R.A., Parker, W.E., Feiler, A.M., 2008. Two kinds of fMRI repetition suppression? Evidence for dissociable neural mechanisms. Journal of Neurophysiology 99, 2877-2886.

Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2005. Sensitivity to voice in human prefrontal cortex. Journal of Neurophysiology 94, 2251-2254.

Fecteau, S., Belin, P., Joanette, Y., Armony, J. L., 2007. Amygdala responses to nonlinguistic emotional vocalizations. NeuroImage 36(2), 480-487.

Friston, K.J., Ashburner, J., Kiebel, S.J., Nichols, T.E., Penny, W.D. (eds), 2007. Statistical Parametric Mapping: The analysis of functional brain images. Academic Press, London.

Gainotti, G., Barbier, A., Marra, C., 2003. Slowly progressive defect in recognition of familiar people in a patient with right anterior tempooral atrophy. Brain 126, 792-803.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., Duchaine, B., 2009. Developmental phonagnosia: A selective deficit of vocal identity recognition. Neuropsychologia 47, 123-131.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthelemy, C., Brunelle, F., Samson, Y., Zilbovicius, M., 2004. Abnormal cortical voice processing in autism. Nature Neuroscience 7(8), 801-802.

Gothard, K.M., Battaglia, F.P., Eickson, C.A., Spitler, K.M., Amaral, D.G., 2007. Neural responses to facial expression and face identity in the monkey amygdala. Journal of Neurophysiology 97, 1671-1683.

Gougoux, F., Belin, P., Voss, P., Lepore, F., Lassonde, M., Zatorre, R.J., 2009. Voice perception in blind persons: A functional magnetic resonance imaging study. Neuropsychologia 47(13), 2967-2974.

Grill-Spector, K., Malach, R., 2001. fMR-adaptation: a tool for studying the functional properties of human cortical neurons. Acta Psychologica (Amsterdam) 107, 293-321.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., Blake, R., 2000. Brain areas involved in perception of biological motion. Journal of Cognitive Neuroscience 12(5), 711-720.

Hailstone, J.C., Crutch, S.J., Vestergaard, M.D., Patterson, R.D., Warren, J.D., 2010. Progressive associative phonagnosia: A neuropsychological analysis. Neuropsychologia 48(4), 1104-1114.

Hoffman, K.L., Logothetis, N.K., 2009. Corical mechanisms of sensory learning and object recognition. Philosophical Transactions of the Royal Society B 364, 321-329.

Jiang, X., Bradley, E., Rini, R.A., Zeffiro, T., VanMeter, J., Riesenhuber, M., 2007. Categorization training results in shape- and category-selective human neural plasticity. Neuron 53, 891-903.

Kawahara, H., 2006. STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoustic Science and Technology 27(6), 349-353.

Kleinhans, N.M., Johnson, L.C., Richards, T., Mahurin, R., Greenson, J., Dawson, G., Aylward, E., 2009. Reduced neural habituation in the amygdala and social impairments in autism spectrum disorders. The American Journal of Psychiatry 166, 467-475.

Kourtzi, Z., DiCarlo J.J., 2006. Learning and neural plasticity in visual object recognition. Current Opinion in Neurobiology 16, 1-7.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., Wildgruber, D., 2007. Audiovisual integration of emotional signals in voice and face: An event-related fMRI study, NeuroImage 37(4), 1445-1456.

Latinus, M., Crabbe, F., Belin, P., 2009. fMRI investigations of voice identity perception. NeuroImage 47(Supplement 1), S156. Organization for Human Brain Mapping 2009 Annual Meeting, July 2009.

Lattner, S., Meyer, M.E., Friederici, A.D., 2005. Voice perception: sex, pitch, and the right hemisphere. Human Brain Mapping 24, 11-20.

Lloyd, R.L., Kling, A.S., 1988. Amygdaloid electrical activity in response to conspecific calls in squirrel monkey: Influence of environmental setting cortical inputs and recording site. In: Newman, J.D. (Ed.), The Physiological Control of Mammalian Vocalization, Plenum Press, New York, 137-151.

Loffler, G., Yourganov, G., Wilkinson, F., Wilson, H.R., 2005. FMRI evidence for the neural representation of faces. Nature Neuroscience 8(10), 1386-1390.

Morris, J.S., Frith, C.D., Perrett, D.I., Rowland, D., Young, A.W., Calder, A.J., Dolan, R.J., 1996. A differential neural response in the human amygdala to fearful and happy facial expressions. Nature 383, 812–815.

Mullennix, J.W., Ross, A., Smith, C., Kuykendall, K., Conard, J., Barb, S., 2011. Typicality effects on memory for voice: Implications for earwitness testimony. Applied Cognitive Psychology 25, 29-34.

Myers, E.B., 2007. Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: an fMRI investigation. Neuropsychologia 45, 1463-1473.

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura A., Hatan, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., Kojima, S., 2001. Neural substrates for recognition of familiar voices: a PET study. Neuropsychologia 39, 1047-1054.

Neuner, F., Schweinberger, S.R., 2000. Neuropsychological impairments in the recognition of faces, voices, and personal names. Brain and Cognition 44(3), 342-366.

Nolan, F., 1997. Speaker recognition and forensic phonetics. In: Hardcastle, W., Laver, J. (Eds.), A Handbook of Phonetic Science. Oxford: Blackwell, 744-766.

Olson, I.R., Plotzker A., Ezzyat, Y., 2007. The Enigmatic temporal pole: a review of findings on social and emotional processing. Brain 130, 1718-1731.

Opitz, B., Mecklinger, A., Friederici, A.D., 2000. Functional asymmetry of human prefrontal cortex: Encoding and retrieval of verbally and nonverbally coded information. Learning and Memory 7, 85-96.

Papcun, G., Kreiman, J., Davis, A., 1989. Long-term memory for unfamiliar voices. Journal of the Acoustical Society of America 85, 913-925.

Pardo, J. S., 2006. On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America 119(4), 2382-2393.

Pernet, C., Charest, I., Bélizaire, G., Zatorre, R.J., Belin, P., 2007. The Temporal Voice Areas: spatial characterization and variability. 13th International Conference on Functional Mapping of the Human Brain, Chicago, USA, NeuroImage 36, Supp1.

Perrachione, T.K., Chiao, J.Y., Wong, P.C.M., 2010. Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. Cognition 114(1), 42-55.

Perrachione, T.K., Wong, P.C.M., 2007. Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. Neuropsychologia 45(8), 1899-1910.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nature Neuroscience 11(3), 367-374.

Potter, R.K., Steinberg, J.C., 1950. Toward the specification of speech. Journal of the Acoustical Society of America 22, 807-820.

Remedios, R., Logothetis, N.K., Kayser, C., 2009. An auditory region in the primate insular cortex responding preferentially to vocal communication sounds. The Journal of Neuroscience 29, 1034-1045.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance 23, 651-666.

Romanski, L.M., Averbeck, B.B., Diltz, M., 2005. Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. Journal of Neurophysiology 93, 734-747.

Schweinberger, S.R., Casper, C., Hauthal, N., Kaufmann, J.M., Kawahara, H., Kloth, N., Robertson, D.M.C., Simpson, A.P., Zaske, R., 2008. Auditory adaptation in voice perception. Current Biology 18(9), 684-688.

Schweinberger, S.R., Herholz, A., Sommer, W., 1997. Recognizing famous voices: influence of stimulus duration and different types of retrieval cues. Journal of Speech, Language and Hearing Research 40, 453-463.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends in Neurosciences 26, 100-107.

Shah, N.J., Marshall, J.C., Zafiris, O., Schwab, A., Zilles, K., Markowitsch, H.J., Fink, G.R., 2001. The neural correlates of person familiarity. A functional magnetic resonance imaging study with clinical implications. Brain 124(4), 804-815.

Sokhi, D.S., Hunter, M.D., Wilkinson, I.D., Woodruff, P.W.R., 2005. Male and female voices activate distinct regions in the male brain. NeuroImage 27, 572-578.

Spreckelmeyer, K.N., Kutas, M., Urbach, T., Altenmuller, E., Munte, T.F., 2009. Neural processing of vocal emotion and identity. Brain and Cognition 69(1), 121-126. Stevens, A.A., 2004. Dissociating the cortical basis of memory for voices, words and tones. Cognitive Brain Research 18, 162-171.

Sugiura, M., Sassa, Y., Watanabe, J., Akitsuki, Y., Maeda, Y., Matsue, Y., Fukuda, H., Kawashima, R., 2006. Cortical mechanisms of person representation: recognition of famous and personally familiar names. Neuroimage 31(2), 853-860.

Turk, D.J., Rosenblum, A.C., Gazzaniga, M.S., Macrae, C.N., 2005. Seeing John Malkovich: the neural substrates of person categorization. NeuroImage 24, 1147-1153.

Van Lancker, D., Kreiman, J., 1987. Voice discrimination and recognition are separate abilities. Neuropsychologia 25, 829-834.

Van Lancker, D., Kreiman, J., Cummings, J., 1989. Voice perception deficits: Neuronanatomical correlates of phonagnosia. Journal of Clinical and Experimental Neuropsychology 11, 665-674.

Van Lancker, D.R., Cummings, J.L., Kreiman, J., Dobkin, B.H., 1988. Phonagnosia: a dissociation between familiar and unfamiliar voices. Cortex 24, 195-209.

von Kriegstein, K., Eger, E., Kleinschmidt, A., Giraud, A.L., 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Cognitive Brain Research 17, 48-55.

von Kriegstein, K., Giraud, A.L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. NeuroImage 22, 948-955.

von Kriegstein, K., Giraud, A.L., 2006. Implicit multisensory associations influence voice recognition. PLoS Biology 4(10), e326.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., Giraud, A.L., 2005. Interaction of face and voice areas during speaker recognition. Journal of Cognitive Neuroscience 17, 367-376.

Warren, J., Scott, S., Price, C., Griffiths, T., 2006. Human brain mechanisms for the early analysis of voices. NeuroImage 31, 1389-1397.

Watson, R., 2009. Selectivity for conspecific vocalizations within the primate insular cortex. The Journal of Neuroscience 29(21), 6769-6770.

Whalen, P.J., Rauch, S.L., Etcoff, N.L., McInerney, S.C., Lee, M.B., Jenike, M.A., 1998. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. The Journal of Neuroscience 18, 411-418.

Wong, P.C., Parsons, L.M., Martinez, M., Diehl, R.L., 2004. The role of the insular cortex in pitch pattern perception: The effect of linguistic contexts. The Journal of Neuroscience 24, 9153-9160.

Chapter 6

Mean-based neural coding of voices

Abstract

The social significance of recognizing the person who talks to us is obvious, but the neural mechanisms that mediate talker identification are unclear. Regions along the bilateral superior temporal sulcus (STS) and the inferior frontal cortex (IFC) of the human brain are selective for voices, and they are sensitive to rapid voice changes. Although it has been proposed that voice recognition is supported by prototype-centered voice representations, the involvement of these category-selective cortical regions in the neural coding of such "mean voices" has not previously been demonstrated. Using fMRI in combination with a voice-learning paradigm, we show that voice-selective regions are involved in the mean-based coding of voice identities. Voice typicality is encoded on a supra-individual level in the right STS along a stimulus-dependent, identity-independent (i.e., voice-acoustic) dimension, and on an intra-individual level in the right IFC along a stimulus-independent, identity-dependent (i.e., voice identity) dimension. Voice recognition therefore entails at least two anatomically separable stages, each characterized by neural mechanisms that reference the central tendencies of voice categories.

Andics, A., McQueen, J. M., Petersson, K. M. (submitted). Mean-based neural coding of voices.

Parts of this work have been presented at the 14th Annual Meeting of the Organization for Human Brain Mapping (HBM), in Melbourne, Australia.

Introduction

Human listeners can recognize individuals from their voices alone and can rapidly learn new voices. Cortical regions involved in voice recognition have been mapped out, but it is not yet known how those regions represent voice knowledge. Here we test the hypothesis that in category-selective regions voices are represented in a prototype-centered voice processing hierarchy. In particular, we ask whether and how cortical activity reflects typicality in newly-learned voice categories. We will refer to this as mean-based neural coding of voices.

Two cortical regions have been reported to be sensitive to conspecifics' vocalizations. These regions are intriguingly similar in the primate and human brain and include regions along the superior temporal sulcus (STS) (in macaques: Petkov et al., 2008; in humans: Belin et al., 2000, 2011; Grandjean et al., 2005; Ethofer et al., 2009b) and the inferior frontal cortex (IFC) (in macaques: Romanski and Goldman-Rakic, 2002; Romanski et al., 2005; in humans: Fecteau et al., 2005; von Kriegstein and Giraud, 2006). Strong anatomical and functional connections have been found between the STS and the ipsilateral IFC in both primates (Hackett et al., 1998; Romanski et al., 1999) and humans (Ethofer et al., 2012). Furthermore, STS and IFC are not only voice-selective but also sensitive to short-term voice stimulus similarity, as demonstrated in rapid fMRI adaptation and carryover effects (STS: Belin and Zatorre, 2003; Andics et al., 2010; Latinus et al., 2011; IFC: Andics et al., 2010; Latinus et al., 2011). Short-term sensitivity here refers to mechanisms typically active within the range of a few seconds (cf., short-term repetition suppression, Epstein et al., 2008). This short-term sensitivity for voice similarity is an important requirement for the ability to tune in to voice stimuli, but it is not sufficient for the representation of long-term voice knowledge. Long-term here refers to processes relying on representations that need to be stored for longer than a few seconds (cf., long-term repetition suppression, Epstein et al., 2008). We adopt this definition in the present study. Neural storage of voice knowledge in the much longer term (e.g. weeks, months) is a topic for future research. Although it seems plausible that category-selective cortical regions are there to represent category knowledge for more than a few seconds, there is little evidence so far that the voice-selective STS and IFC contribute to representing this kind of long-term voice knowledge.

This study asks whether the STS and IFC perform this function and elaborates on the recent proposal that long-term voice knowledge is represented in the human brain in a prototype-centered way. Mean-based neural coding appears to be a powerful way to represent individual stimuli in a category space (e.g., Panis et al., 2011). A possible mechanism for mean-based coding is neural sharpening (Hoffman and Logothetis, 2009): the coding of central values in relevant object dimensions becomes sparser with more experience. Neural sharpening reflects long-lasting cortical plasticity and so could be used for positioning stimuli in long-term object spaces. For faces, mean-based coding was found behaviourally (Leopold et al., 2001, Rhodes and Jeffery, 2006), in primates (Leopold et al., 2006), and also with human fMRI localizing the mechanism in face-selective fusiform regions (Loffler et al., 2005). Recent behavioural (Papcun et al., 1989; Latinus et al., 2009; Mullennix et al., 2009; Bruckert et al., 2010; Latinus and Belin, 2011) and neuroimaging studies (Andics et al., 2010) also suggest mean-based coding for voices. In other words, voice representations appear to be centered around prototypes in long-term memory.

Long-term mean-based coding for voices has nevertheless not yet been demonstrated in voice-selective cortical regions. Andics et al. (2010) found mean-based coding for voices in several regions, but some of these regions (the deep posterior STS and the orbital/insular cortex) are not voice-selective. Other regions (the amygdala and the anterior temporal pole) appear to be involved in the multimodal integration of person identity rather than in pure voice identity processing (Andics et al., 2010; Latinus et al., 2011; Belin et al., 2011). Although recent findings suggested IFC involvement in the representation of long-term stored objects (Latinus et al., 2009), to date there is thus no evidence for long-term mean-based voice encoding in the core category-selective cortical regions, namely the STS and the IFC.

It has been proposed that voice recognition involves not only mean-based voice encoding but also separate processing stages for voice-acoustic and voice identity analysis (Scott and Johnsrude, 2003; Belin et al., 2004, 2011; Charest et al., 2012; Bestelmeyer et al., 2012). This proposal, however, has received little direct support so far in the form of functional-anatomical correspondences between voice-processing stages and voice-selective regions. In the framework of mean-based coding, voice-acoustic analysis corresponds to an identity-independent, supra-individual representation of voice typicality,

while voice identity analysis corresponds to an identity-dependent, intra-individual representation of voice typicality. These definitions will be adopted in the present study. Note that typicality is thus defined here with respect to the materials in the experiment, and not judgments of typicality collected, for example, in a rating study.

Recently, Latinus et al. (2011) attempted to dissociate acoustic from identity effects in voice processing, but their design focused on short-term effects of acoustic and identity changes. Short-term acoustic processing was found in both the STS and the IFC and short-term identity processing was found in the IFC only. These short-term effects may be indicators of long-term voice processing mechanisms, but those mechanisms have not yet been tested directly. The present study therefore tested the hypothesis that long-term mean-based voice encoding is present both at voice-acoustic (supra-individual) and at voice identity (intra-individual) levels of processing, and aimed to specify the role of the two core voice-selective cortical regions in these two levels.

We performed an fMRI experiment using a within-subject voice-training paradigm. Listeners were trained on two consecutive weeks to categorize voice stimuli on a voice morph continuum as belonging to either of two talkers characterized by the two continuum endpoints (morph0, morph100). During training the entire continuum was sampled and the acoustic centre of the trained stimulus space was identical across weeks (morph50). The feedback during training on week1 and week2 specified different voice identity category boundary locations on each week (morph36 or morph64). After each training session, we could separately manipulate two perceptual properties of the voice stimuli: their perceived acoustic centrality (i.e., degree of prototypicality defined by the acoustic space, independent of identity feedback) and their perceived identity centrality (i.e., degree of prototypicality of a new voice identity, as defined by a voice-training procedure, independent of acoustic properties). Our design also allowed us to separately test for short-term effects (e.g., rapid adaptation indicating stimulus similarity sensitivity in the 0-5 seconds range) and long-term effects (e.g., neural sharpening indicating norm-based coding in the > 5 seconds range) within a single experiment.

We hypothesized that cortical representations of the voice-acoustic space are organized along an acoustically central to acoustically peripheral dimension, and thus should not be modulated by voice identity feedback. Acoustically central stimuli should have

sharper neural coding than acoustically peripheral stimuli and hence we predicted there should be less activity for central than for peripheral stimuli in voice-acoustic regions. We also hypothesized that voice identity representations are organized along a feedback-defined typical to atypical dimension, and that this typicality is fully independent of voice-acoustic properties. According to the predictions of neural sharpening, the activity of voice identity representations generated by identity-typical stimuli should therefore be less than the activity generated by atypical stimuli.

Materials and Methods

Participants

Eighteen Dutch female listeners (19-24 years) with no reported hearing disorders were paid to complete the experiment. Written informed consent was obtained from all participants. One person was excluded because of a failure to perform the task during training. Two further participants were excluded because of poor learning performance during training (i.e., voice categorization performance per morph level did not significantly differ from the 50% chance level in the final training block before scanning, one-sampled, two-tailed $t(14) < 1$, $p > .4$). The analyses presented here were based on the remaining 15 subjects.

Stimulus material

We recorded two young male nonsmoking adult native speakers of Dutch with no recognizable regional accents and no speech problems saying the Dutch word *mes* (knife). The voices were unfamiliar to the listeners. Recordings were made in a soundproof booth using a Sennheizer Microphone ME62, a MultiMIX mixer panel, and Sony Sound Forge. All stimuli were digitized at a 16 bit/44.1 kHz sampling rate and were volume balanced using Praat software (Boersma and Weenink 2007).

We then created a voice morph continuum using the speech manipulating algorithms of STRAIGHT (Kawahara, 2006). The speech signals were decomposed into three parameters: an interference-free spectrogram, an aperiodicity map and a fundamental

frequency (F0) trajectory. These parameters were then interpolated segment by segment. Finally, a 100-step stimulus continuum with equidistant intermediate levels was resynthesized. The endpoints (levels morph0 and morph100) were also resynthesized. Average syllable duration was 487 ms (audio samples can be found at http://mpi.nl/people/andics-attila/research).

Training design

Listeners received multiple-phase voice identity training on two consecutive weeks. During the entire course of training, listeners were presented with words from the voice morph continuum and were instructed to make forced-choice decisions on talker identity after every word they heard. To allow initial assignment of talker names (Peter and Thomas) on response buttons to voice identities (voice A and voice B), listeners were presented three naturally produced monosyllables from each talker before the experiment. The whole continuum was sampled each week. The assignment of talker names to voices and to dominant or non-dominant index fingers was counterbalanced across participants. The full stimulus range was sampled both during training and at test, but there was no exact stimulus overlap between the two parts (i.e., the morph levels used at training were different from those used at test; see below). Two training conditions were used: listeners were trained on different voice identity boundaries (morph36 or morph64) on the first and second week. The category boundary was made explicit by giving feedback according to a predefined boundary at 36% voice B morphs one week and at 64% the other week. Therefore, morphs between the two boundaries were trained to be categorized as voice A one week (when the boundary was at 64%), but as voice B the other week (when the boundary was at 36%). This training manipulation was amplified by presenting more stimuli from the most ambiguous parts of the continuum (Appendix A): The mean of all stimuli from each voice identity category was a 10% distance from the category boundary. The order of training conditions was counterbalanced across participants. Participants were not informed about the category boundary shift.

Training procedure

Stimuli were presented via headphones binaurally, at a comfortable listening level. In each of two weeks participants received 72 min of training over 2 days, with 3 training sessions of 18 min each on day1 and a single training block of 18 min on day2. Training was followed by an fMRI test session on day2 in each week. Stimuli on consecutive trials were physically different. Stimulus ordering was otherwise random and varied across listeners. Training trials were 3000 ms long and included visual feedback (i.e., whether responses were correct, incorrect or late), presented from 2100 to 2400 ms after trial onset. Training phases contained 360 trials (12 repetitions of 30 morph levels). The manipulation appeared to be successful in that all participants reported, after the experiment, that they thought that they had heard various exemplars of natural voices only and that they were convinced that the trained voices were two actual persons' voices.

Conditions of interest

The critical stimuli in the fMRI test were morphs05, 33, 67 and 95. The categorization training defined identity membership of these stimuli (belonging to voice identity A or B), although these specific morph levels were not presented during training. Morph05 and morph33 always belonged to voice A, while morph67 and morph95 always belonged to voice B. The critical voice morphs also differed in terms of their distributional position on the stimulus continuum: Morph05 and morph95 were close to the endpoints, while morph33 and morph67 were close to the middle of the continuum – these morphs are referred to as peripheral and central stimuli, respectively. The trained voice identity and the centrality of these critical stimuli did not change across training sessions. But, crucially, the perceived typicality of the central voice morph stimuli changed as a function of the training condition. During voice identity boundary 36% training, morph67 was a typical exemplar of voice B (i.e., far from the identity boundary), and morph33 was an atypical exemplar of voice A (i.e., close to the identity boundary); but during voice identity boundary 64% training, morph33 was a typical exemplar of voice A, and morph67 was an atypical exemplar of voice B. These morphs, dependent on whether they were far from (> 30 morph steps) or close to (= 3 morph steps) the actual voice identity boundary, are referred to as typical and atypical stimuli, respectively. Note that acoustically peripheral stimuli were always far from

the trained voice identity boundary, so they were always typical for one of the voices. Therefore, all critical stimuli fall into one of three types: peripheral-typical, central-typical or central-atypical. To control for the distance from the trained voice identity boundary across all typical stimuli when comparing these conditions, only those peripheral-typical stimuli were considered whose distance from the boundary matched central-typical stimuli's distance from the boundary (= 31 morph steps). The conditions of main interest are summarized in Table 1.

Table 1. Characterization of conditions

Condition	**Critical morphs**		**Distance from acoustic centre**	**Distance from identity boundary**	**Decision difficulty**
	boundary= morph36	boundary= morph64			
peripheral-typical	05	95	45 morph steps	31 morph steps	easiest (96%)
central-typical	67	33	17 morph steps	31 morph steps	medium (88%)
central-atypical	33	67	17 morphs steps	3 morph steps	hardest (81%)

fMRI test: design and procedure

Every listener was tested twice with fMRI. Stimuli consisted of pairs of tokens, each voice morphs of *mes*. The tokens used in the fMRI tests were morphs05, 33, 50, 67 and 95. There was an onset delay of 800 ms between tokens. Listeners were instructed to ignore the first voice and identify the second one (no feedback was given). FMRI tests were identical across the two weeks, but the pairs could fall into different condition categories on week1 and week2 depending on the identity boundary training. Each test session included 13 token pair types (Appendix B), with 20 repetitions of each type. A silent condition with 40 repetitions was also added. Token pair types were evenly distributed: each chunk of 15 consecutive trials included one of each token pair type and two silent trials. Consecutive trials were always physically different, and also different with respect to the corresponding experimental condition (Appendix B), but stimulus ordering was otherwise random.

Identical morph pairs were used to test for long-term adaptation (or neural sharpening) effects. We tested acoustically central and peripheral stimuli, and identity-

typical and -atypical stimuli, all defined with respect to their positions in the constant acoustic space and the training-varied identity space (Table 1). Short-term adaptation effects were controlled in the tests of long-term effects because the pairs of morphs in each condition were always identical, and consecutive morph pairs were sufficiently distant (> 5 seconds). Short-term effects of voice similarity were tested by comparing responses to identical versus non-identical morph pairs. We assumed that, in voice-selective cortical regions, identical pairs elicit reduced activity compared to non-identical pairs, due to rapid adaptation in response to stimulus repetition. Within non-identical pairs, we further differentiated between coarse and fine within-pair changes, determined by distance in morph steps.

Voice selective regions were defined in a separate localizer run with blocks corresponding to (1) vocal sounds (verbal and nonverbal), (2) non-vocal sounds (animals, sounds from the environment, music) matched for number of sources, in duration, and overall energy and (3) silence. Participants were instructed to passively listen to the stimuli. Stimuli were controlled using Presentation software (www.neurobs.com). During imaging, stimulus presentation was synchronized by a trigger pulse with the data acquisition. Stimuli were delivered binaurally through MRI-compatible headphones (Commander XG, Resonance Technology Inc., Northridge, CA).

fMRI data acquisition

Measuring auditorily induced haemodynamic changes with fMRI remains a technical challenge: While continuous sampling methods suffer from scanner noise interference, sparse sampling methods have to cope with a decrease in signal-to-noise ratio caused by the disturbance of steady-state magnetization and subsequent loss of statistical power. We used a 3T Siemens scanner and an in-house modified scanning protocol with scan-on periods for functional data acquisition and scan-off periods for stimulus presentation. For scan-off periods, gradient switching was removed to reduce scanner noise, but slice selective excitation pulses were played out to keep the magnetization in the steady state (see Schwarzbauer et al., 2006 for a similar protocol). Stimuli were always presented during scan-off periods. To further reduce scanner noise in all periods and to minimize period length at the same time, parallel imaging was used and no fat suppression was applied. A TR

of 1200 ms was used. Trial onset-to-onset delay (i.e., the time between trials) was 8400 ms. Five functional volumes were acquired for each trial. For the main tests EPI-BOLD fMRI time series were obtained from 24 transverse slices covering temporal lobes and the inferior part of the frontal lobes with a spatial resolution of 3.5 × 3.5 × 3.5 mm, including a 0.5 mm slice gap (TE = 30ms, ascending slice order; 300 trials; GRAPPA 2; sequence = SCAN-SCAN-SCAN-SCAN-SCAN-SILENT-SILENT; slice nr = 24; jittering: stimulus1 starts 200-800ms after silent pulse onset). In total, each test session included 300 trials. The test was conducted as a single run lasting 45 min, including 4 half-minute breaks after each 8.4 min.

For the voice localizer there were 39 transverse slices and a longer silent gap between acquisitions (TR = 2000ms; sequence = SCAN-SILENT-SILENT-SILENT-SILENT). Stimulus blocks of 8s, corresponding to vocal sounds, non-vocal sounds and silence were presented after each volume. In total there were 20 blocks of each type (62 volumes including one dummy scan at the beginning and one extra scan at the end). All other parameters were identical to the main test settings. In addition to the functional time series, a standard T1-weighted three-dimensional scan using a turbo-field echo (TFE) sequence with 180 slices covering the whole brain was collected for anatomical reference at the end of the second scanning session, with 1 × 1 × 1 mm spatial resolution.

fMRI data analysis

Image preprocessing and statistical analysis were performed using SPM5 (www.fil.ion.ucl.ac.uk/spm). Phantom image files were added before normal preprocessing to fill missing volume gaps (created by scan-offs). These phantom images were removed again after design specification but before model estimation by editing the design matrices. The functional EPI-BOLD images were realigned, slice-time corrected, spatially normalized, and transformed into a common anatomical space, as defined by the SPM Montreal Neurological Institute (MNI) T1 template. Next, the functional EPI-BOLD images were spatially filtered by convolving the functional images with an isotropic 3D Gaussian kernel (10 mm FWHM). The fMRI data were then statistically analyzed using a general linear model and statistical parametric mapping (Friston et al., 2007). Every token pair was modeled as a separate event, using constant epochs corresponding to the average token length, starting from the onset of the second token. To account for differences in response times (RT), we

also performed an a-posteriori confirmatory analysis modeling each event (i.e. token pair) with an epoch length equal to the RT specific to that trial, using the variable epoch approach as described by Grinband et al. (2008). As in the main analysis, the onset of each epoch was positioned at the onset of the second token (also corresponding to response time onset). For the main and confirmatory analyses, condition regressors were constructed per token pair type (Appendix B).

Regressors for silent trials and, to model potential movement artifacts, realignment regressors for each run were also included. A high-pass filter with a cycle-cutoff of 128 s was implemented in the design to remove low-frequency signals. Single-subject fixed effect analyses were followed by random effects analyses on the group level. An initial uncorrected threshold of $p < .001$ was applied for all tests. The whole-volume functional localizer run's statistical test was family-wise-error (FWE) corrected at the cluster level ($p < .05$). The main run's statistical tests were small-volume corrected using the three significant clusters of the functional localizer as regional masks, and FWE-corrected at the voxel level ($p < .05$).

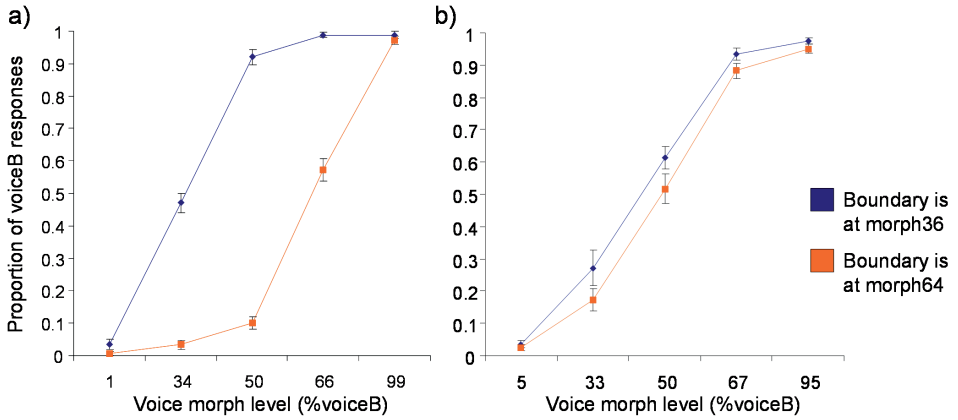
Results

Flexibility in voice learning

Participants improved in identification accuracy from 61% (block 1, week1) to 70% correct (block 4, week2). Behavioural responses at training and during fMRI confirmed that the boundary manipulation (i.e., whether the trained voice identity category boundary was at morph 36 or at morph 64) led to a training-related shift in voice identity judgments for ambiguous levels of the voice morph continuum (Fig. 1; training: boundary $F(1,14) = 855$, $p < .001$, level $F(4,56) = 730$, $p < .001$, boundary x level $F(4,56) = 146$, $p < .001$, linear component of the interaction $F < 1$, quadratic component of the interaction $F(1,14) = 738$, $p < .001$; test: boundary $F(1,14) = 19.3$, $p = .001$, level $F(4,56) = 330$, $p < .001$, boundary x level $F(4,56) = 2.61$, $p = .089$, linear component of the interaction $F < 1$, quadratic component of the interaction $F(1,14) = 10.4$, $p = .006$). The proportion of correct decisions was used to judge decision difficulty per condition. We found that at test peripheral-typical trials were easier than central-typical trials (mean difference (%correct) = 7.66, $t(14) = 4.78$, $p < .001$); and

central-typical trials, in turn, were easier than central-atypical trials (mean difference (%correct) = 7.54, t(14) = 3.07, p = .008; Table 1). These differences in decision difficulty were also reflected in RTs during fMRI. Responses for peripheral-typical trials were faster than those for central-typical trials (mean difference (RT) = 107 ms, t(14) = 4.10, p < .001); and responses for central-typical trials, in turn, were faster than those for central-atypical trials (mean difference (RT) = 38 ms, t(14) = 2.55, p = .023).

Fig. 1. Voice categorization per voice identity boundary training condition during training and at test. (a) Training: categorization performance in the final training block of each training session, data for morph levels matched to those used at test (e.g., morph50 refers to the average of two trained morph levels neighbouring morph50). (b) Test: categorization during scanning sessions, data for morph pairs with no change. Error bars represent standard error of the mean.



Voice selective regions

Voice selective regions were defined in a separate localizer run (Belin et al., 2000), contrasting vocal and non-vocal sounds (see Methods). Four regions survived an uncorrected $p < .001$ threshold ($t(14) > 3.79$): the bilateral STS and the bilateral IFC, but the left IFC region did not reach a cluster-level family-wise error (FWE) corrected level of significance (Table 2). These findings confirmed that the voice-selective regions include both superior temporal and inferior frontal regions.

Table 2. Voice sensitive regions as determined by the functional localizer.

Voice > non-voice	**size (voxels)**	**p (cluster-corr)**	**t(14)**	**x**	**y**	**z**
Right STS	2647	< 0.001	11.87	48	-32	4
			10.81	60	0	-8
			9.16	56	-20	-2
Left STS	2350	< 0.001	8.96	-60	-16	4
			8.57	-44	10	-24
			8.32	-58	-44	16
Right IFC	467	0.002	6.24	56	18	24
			5.14	42	14	32
			4.94	48	6	34
Left IFC	30	0.785	4.98	-52	32	6

Height threshold was $p < 0.001$ ($t(14)=3.79$). For each cluster, the table displays at most 3 local maxima more than 8.0 mm apart.

Mean-based coding of acoustic properties

The effect of "distance from acoustic centre" (i.e., distance from morph50) was investigated by contrasting acoustically peripheral and acoustically central stimuli. We predicted that, in regions that code acoustic centrality, peripheral stimuli would elicit greater activity than central stimuli, independently of how typical those stimuli are in the feedback-driven identity space (i.e., peripheral-typical > central-typical = central-atypical; Table 1). We found that only a single voice-sensitive cluster in the right STS was sensitive to stimulus position in the acoustic space set by the experiment (Table 3). In this region response reduction was found for acoustically central compared to peripheral voice stimuli. As this contrast controlled for short-term adaptation effects (by presenting no-change

morph pairs in each of the contrasted conditions), we propose that the response reduction found in the STS was caused by a neural sharpening mechanism acting on a long-term stored representation of the voice-acoustics space organized around the acoustic centre. This finding of mean voice representations in the right STS is analogous to proposed mean face representations in the fusiform face region (Loffler et al., 2005).

The long-term stored representation of the voice-acoustic space was further investigated to see whether activity in the space was modulated by voice identity training. We found no evidence suggesting that this was the case, that is, there was no stronger response in the right STS or anywhere else to morph33 for the test sessions where listeners were trained on morph64 as the identity category boundary (i.e., to central-typical stimuli) compared to the test sessions where listeners were trained on morph36 (i.e., to central-atypical stimuli). This suggests that the acoustic space representation was independent of voice identity feedback.

A confirmatory analysis that modeled trial-specific RTs using a variable epoch approach (Grinband et al., 2008; see Methods) yielded very similar results for the same contrasts (Table 4), but note that in one of tests acoustic centrality in the voice-sensitive STS was found bilaterally. This suggests that the STS findings cannot be explained by across-condition differences in voice identity decision difficulty, as reflected in the RTs.

Table 3. Significant BOLD effects in the main analysis.

Contrast	ROI	p	t(14)	x	y	z
Long-term acoustic centrality						
peripheral-typical > central-atypical	Right STS	0.003	6.63	64	-26	0
peripheral-typical > central-typical	Right STS	0.008	5.79	66	-34	4
Long-term identity centrality						
central-atypical > central-typical	Right IFC	0.021	4.16	44	16	30
central-atypical > peripheral-typical	Right IFC	0.022	4.07	48	8	36
Short-term similarity						
coarse change > no change	---					
coarse change to central	Right STS	0.050	4.47	66	-36	2
> no change, central						
	Left STS	0.022	4.98	-64	-20	0
coarse change to peripheral	---					
> no change, peripheral						
fine change between identities	---					
> no change (matched)						
fine-change within identity	---					
> no change (matched)						

ROIs were defined using the voice localizer run's voice vs nonvoice contrast, thresholded at $p < .001$ (uncorrected). Contrasts were thresholded at $p < .001$ ($t(14)=3.79$). The table displays FWE-corrected p values where significant. No significant effects were found with these contrasts for other ROIs, nor with any further contrasts (e.g., with the reversed tests) for any of these ROIs.

Table 4. Significant BOLD effects in the confirmatory analysis accounting for RTs.

Contrast	ROI	p	t(14)	x	y	z
Long-term acoustic centrality						
peripheral-typical > central-atypical	Right STS	0.035	4.65	50	-28	6
peripheral-typical > central-typical	Right STS	0.029	4.73	54	-26	4
	Left STS	0.015	5.11	-58	-10	8
Long-term identity centrality						
central-atypical > central-typical	Right IFC	0.004	5.15	50	8	38
central-atypical > peripheral-typical	Right IFC	0.032	3.67[+]	46	4	34
Short-term similarity						
coarse change > no change	Right STS	0.059	4.38	66	-22	8
	Left STS	0.016	5.24	-62	-24	16
coarse change to central > no change, central	---					
coarse change to peripheral > no change, peripheral	---					
fine change between identities > no change (matched)	---					
fine-change within identity > no change (matched)	---					

ROIs were defined using the voice localizer run's voice vs nonvoice contrast, thresholded at $p < .001$ (uncorrected). Contrasts were thresholded at $p < .001$ ($t(14)=3.79$). The table displays FWE-corrected p values where significant. No significant effects were found with these contrasts for other ROIs, nor with any further contrasts (e.g., with the reversed tests) for any of these ROIs.

[+]: Thresholded at $p < .002$ ($t(14)=3.44$).

Mean-based coding of voice identity

The effect of "distance from identity boundary" (i.e., distance from morph36 or morph64) was tested by contrasting identity-atypical and typical stimuli. We predicted that in regions that code identity centrality, identity-atypical would elicit greater activity than identity-typical stimuli, independently of how central or peripheral those stimuli are in the acoustic space (i.e., central-atypical > central-typical = peripheral-typical; Table 1). We found that only a single voice-sensitive cluster in the right IFC was modulated by voice identity training (Table 3). In this IFC region response reduction was found for the same voice stimuli when trained as more prototypical versus less prototypical encounters of a talker. As this contrast only included conditions with no-change morph pairs and was thus controlled for short-term adaptation effects, we propose that the response reduction found in the IFC was caused by a neural sharpening mechanism acting on long-term stored, prototype-centered representations in a voice identity space. Importantly, this response reduction was found for acoustically distant identity-typical voice stimuli that were associated with different person identities. A repeated-measures ANOVA on percent signal change values in the peak coordinate of the central-atypical vs central-typical test in the right IFC [44, 16, 30] was also performed with the factors voice identity (A, B) and identity centrality (identity-typical, identity-atypical). Beyond an obvious main effect of identity centrality ($F(1,14) = 16.95$, $p = .001$), we found no main effect of voice identity ($F<1$) and no interaction of the two factors ($F<1$). These data confirm that the identity centrality effect in IFC is equally present for each of the two voice identities we tested. This suggests that IFC maintains separate prototype-centered voice identity spaces for each voice identity.

Further analyses confirmed that the IFC findings are not caused by across-condition differences in decision difficulty. First, no IFC modulation was found for an analogue contrast with a similar difference in decision difficulty (Table 1) but without a difference in the distance from the trained category boundary (namely, for the central-typical > peripheral-typical contrast). Second, a confirmatory analysis that accounted for RT differences on a trial-by-trial basis yielded the same pattern of results (Table 4).

Rapid adaptation for voice changes in the STS

Further tests included non-identical morph pairs with coarse or fine voice changes that, through comparison to identical morph pairs, were used for investigating short-term adaptation effects. We demonstrated short-term adaptation for voice stimuli in voice-sensitive regions of the STS. Response reduction was found bilaterally in the STS for identical voice stimulus pairs compared to voice pairs with a coarse voice change, but no adaptation effect was found with a finer voice change. The loss of adaptation effect with finer voice changes was not modulated by voice identity properties (i.e., we found no adaptation in voice-selective regions for either fine between-identity changes or for fine within-identity changes). This pattern of activity indicates short-term coarse acoustic processing in the voice-selective STS. Interestingly, however, the adaptation effect with coarse voice changes was only present when no-change stimuli were acoustically central, and disappeared when no-change stimuli were acoustically peripheral. That is, short-term adaptation was modulated by long-term acoustic centrality in the voice-sensitive STS (Table 3). Note, that the RT-modulated follow-up analysis confirmed the presence of the adaptation effect with coarse voice changes, but not that it was modulated by acoustic centrality (Table 4).

Discussion

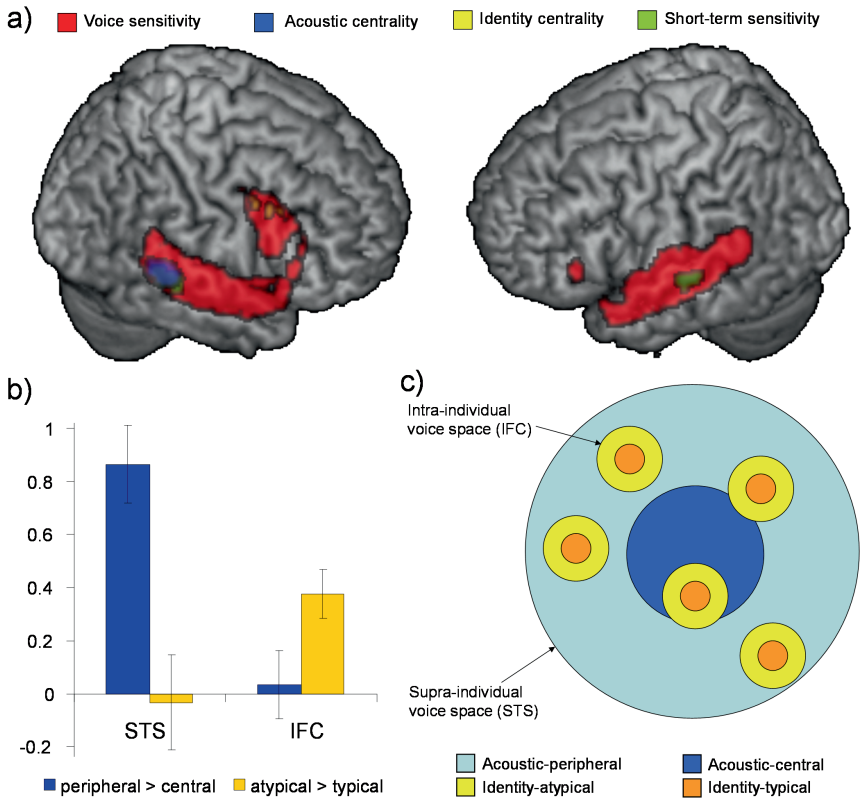
We aimed at specifying the role of voice-selective cortical regions in maintaining long-term voice knowledge. Earlier studies have indicated that voices may be represented in prototype-centered voice spaces (Papcun et al., 1989; Latinus et al., 2009; Mullennix et al., 2009; Bruckert et al., 2010; Andics et al., 2010; Latinus and Belin, 2011) and that the STS (Belin and Zatorre, 2003; Andics et al., 2010; Latinus et al., 2011) and IFC (Andics et al., 2010; Latinus et al., 2011) are core voice processing regions, showing voice selectivity and short-term sensitivity to voice similarity. But these voice-selective regions of the STS and the IFC have not previously been shown to be involved in long-term mean-based voice coding, and indeed there has to date been no other evidence of long-term neural coding of voice prototypes. Here we performed an auditory fMRI study combined with a training manipulation. Listeners were trained on the same voice morph continuum but with different voice identity category feedback on two consecutive weeks, each time followed by scanning. After each training session, we could separately manipulate two perceptual

properties of the voice stimuli: their perceived acoustic centrality (independent of identity feedback) and their perceived identity centrality (independent of acoustic properties). The main results are: (1) there is long-term encoding of acoustic centrality of voices in the right STS, and (2) there is long-term encoding of identity centrality in the right IFC (Fig. 2a,b). We also confirmed that the bilateral STS is sensitive to short-term acoustic similarity of voices.

The present study therefore not only supports a hierarchical model of voice recognition, that is, that there exist distinct voice processing functions with distinct anatomical locations (Belin et al., 2004), but, critically, it also characterizes the neural mechanisms of these processing stages: our results provide evidence that both long-term acoustic and identity processing mechanisms are based on mean-based neural coding, and that these long-term codes are maintained in voice-selective regions of the STS and the IFC.

With respect to the role of the STS, previous work has established that regions of the bilateral (but right-lateralized) STS are voice-selective and play a key role in voice recognition (Belin et al., 2000; von Kriegstein and Giraud, 2004; Gervais et al., 2004; Warren et al., 2006; Formisano et al., 2008; Andics et al., 2010; Latinus et al., 2011). Even though there is agreement that the STS is a functionally highly heterogenous region (Beauchamp et al., 2004), with distinct subregions having different properties and functions, even within the domain of voice processing (von Kriegstein and Giraud, 2004), its exact role in the hierarchical model of voice recognition is still debated. Crucially, there are differing views on whether the voice-selective right STS is also involved in identity processing of voices (Warren et al., 2006), or whether it is involved in acoustic processing exclusively (Andics et al., 2010, Latinus et al., 2011). In other words, does STS keep track of who is speaking or does it only encode how the voice sounds in relation to other voices? Andics et al. (2010) found that listeners' individual sensitivity to voice similarities in a right mid STS region correlated with pre-scan voice recognition performance, but they suggested that this measure reflected sensitivity to short-term acoustic similarity rather than long-term identity similarity. The present results show that the STS is involved in both short-term acoustic processing and in long-term acoustic processing (with a clear right-hemisphere dominance), but not in long-term identity processing.

Fig. 2. Acoustic centrality and identity centrality representations of voices. (a) Contrast maps overlaid on a rendered brain, displaying voice sensitivity: voice vs nonvoice localizer (red), acoustic centrality: peripheral-typical vs central-typical (blue), identity centrality: central-atypical vs central-typical (yellow) and short-term sensitivity: coarse change (to central) vs no change (central) (green) contrasts. (All tests are thresholded at $p < .001$, $t(14)=3.79$; and masked by the voice localizer, thresholded at $p < .001$, $t(14)=3.79$). b) Bar graph displaying percent signal change in the peak coordinate of the acoustic centrality test (peripheral-typical vs central-typical) in the right mid STS [66, -34, 4] and in the peak coordinate of the identity centrality test (central-atypical vs central-typical) in the right IFC [44, 16, 30]. Error bars represent standard error of the mean. (c) A schematic illustration of mean-based representations of acoustic and identity properties in intra-individual and supra-individual voice spaces.



We also tested for short-term identity sensitivity, but found no significant regions. Previous studies claiming to have found short-term identity processing in the STS have possible acoustic confounds. Warren et al. (2006) found that regions along the bilateral STS responded more strongly to change than to no change of speaker. They argued that the STS

is therefore crucial for voice identity processing. However, this contrast had possible acoustic biases, since the changing speaker condition necessarily contained greater acoustic variation than the fixed speaker condition. So these findings may be evidence of short-term acoustic processing. The mid STS certainly appears to be a crucial stage of the voice recognition pathway, but we suggest that it does not encode person identity (i.e., intra-individual voice typicality) information. Based on the present findings we can make the case that the voice-selective right mid STS encodes acoustic centrality by maintaining a supra-individual, feedback-independent, norm-based acoustical voice space.

With respect to the role of the rIFC, the importance of prefrontal regions in the processing of voices has been demonstrated only recently, in extracellular recording experiments with primates (Romanski and Goldman-Rakic, 2002; Romanski et al., 2005). These studies showed that neurons in the macaque ventrolateral prefrontal cortex respond stronger to conspecifics' vocalizations than to nonvocal auditory stimuli. An analogue region with a similar response pattern was identified in the human brain (Fecteau et al., 2005), responding more strongly to speech and to nonlinguistic vocalizations than to non-voice stimuli, and to emotional than to neutral vocalizations. Other studies have also suggested that the IFC is involved in voice processing (Stevens, 2004; von Kriegstein and Giraud 2004, 2006; Ethofer et al., 2009a; Andics et al., 2010; Latinus et al., 2011; Bestelmeyer et al., 2012; Charest et al., 2012), that IFC responses to voices are enhanced after learning more about the voices (von Kriegstein and Giraud, 2006), and that the IFC is sensitive to short-term voice-acoustic (Andics et al., 2010; Latinus et al., 2011) and voice identity changes (Latinus et al., 2011). The present study provides the first demonstration that individual voice identities are represented in a prototype-referenced manner in the human prefrontal cortex. A single region in the right IFC responded more strongly to identity-atypical than to identity-typical stimuli when all acoustic properties of the stimuli were controlled. Our results thus suggest that the right IFC contributes to long-term voice knowledge. More specifically it appears to encode voice identity centrality (i.e., how far a given voice stimulus is from an average of the listener's memory of that specific person's voice). Recent findings in voice gender and voice attractiveness processing come to similar conclusions. Charest et al. (2012) proposed that the IFC reflects stimulus ambiguity and long-term voice gender representations. Bestelmeyer et al. (2012) demonstrated that less attractive voices elicit

greater IFC activity, independently of acoustic properties. These studies and the present findings converge on the claim that the voice-sensitive IFC is involved in linking voice representations to basic, long-term social concepts such as person identity, person gender and person attractiveness.

Recently, Latinus et al. (2011) made an attempt to dissociate acoustic from identity effects in voice processing, using a training paradigm with voice morph continua, but despite these similarities there are major design differences between it and the present study. First, the study by Latinus and colleagues focused on short-term sensitivity effects but was not designed to capture long-term effects. Stimulus relations were systematically manipulated within morph pairs, but there were no long-interval comparisons across the different types of pairs. Their contrasts, however, were not free of long-term acoustic effects. In the present study, however, the multi-level manipulation of conditions (i.e., both within and across morph pairs) allowed us to identify effects of short-term and long-term similarity sensitivity simultaneously. Second, the acoustic and identity contrasts in the Latinus et al. study were not fully independent. In the present study, in contrast, the within-subject, multi-session training paradigm allowed us to test for identity effects with acoustic variation fully controlled. In spite of these design differences, our results can easily be reconciled with those of Latinus et al. (2011). In our view, the results of both studies converge in suggesting that the STS is involved in short-term acoustic similarity processing. Latinus et al.'s findings also indicate that the IFC is involved in short-term processing of either acoustic or identity similarities of voices and in Andics et al. (2010) it was found to be involved in short-term acoustic processing. In the present study, however, the IFC was not found to be involved in short-term identity processing. We therefore suggest that to date there is no convincing evidence for the involvement of the IFC, and, in fact, of any other cortical regions, in short-term identity processing. Instead, IFC appears to support short-term acoustic processing and, critically, long-term voice identity processing.

Andics et al. (2010) found that several other cortical regions contribute to long-term identity-based voice knowledge, including a deep posterior STS region, the anterior temporal poles and the amygdala – but, unlike in the present study, not the voice-selective IFC. Andics et al. (2010) also found short-term acoustically driven adaptation effects in IFC, but here we could not demonstrate short-term sensitivity in this region. One explanation for

this discrepancy is that co-existent short-term and long-term effects may exist in the same brain region, and they might mask each other. Short-term adaptation effects are known to be extremely sensitive to design details such as time gap between adaptor and target stimulus (Grill-Spector et al., 2006), possible carry-over from earlier trials (Aguirre et al., 2007), task (Wagner et al., 2000, Cohen Kadosh et al., 2010), cross-modal associations during a pre-test training (Latinus et al., 2011), and attention or expectation effects (Summerfield et al., 2008; Larsson and Smith, 2012).

The short-term results of the present study show that short-term adaptation is modulated by long-term acoustic centrality in the voice-sensitive STS. This can be interpreted as evidence for an interaction of short-term and long-term acoustic effects, indicating that the same STS region is involved in both short-term and long-term processing. But it is also possible that short-term adaptation is stronger for acoustically central (i.e., more expected) than for acoustically peripheral (i.e., less expected) stimuli: this latter interpretation is in accordance with recent findings demonstrating greater short-term repetition suppression for expected than for non-expected stimuli in category-selective regions (Summerfield et al., 2008).

Finally, it is worth noting that we found mean-based voice coding almost exclusively in the right hemisphere. This converges with clinical (Van Lancker and Canter, 1982) and neuroimaging studies (Belin and Zatorre, 2003; von Kriegstein and Giraud, 2004) reporting greater sensitivity for talker-related features of voice stimuli on the right side of the brain.

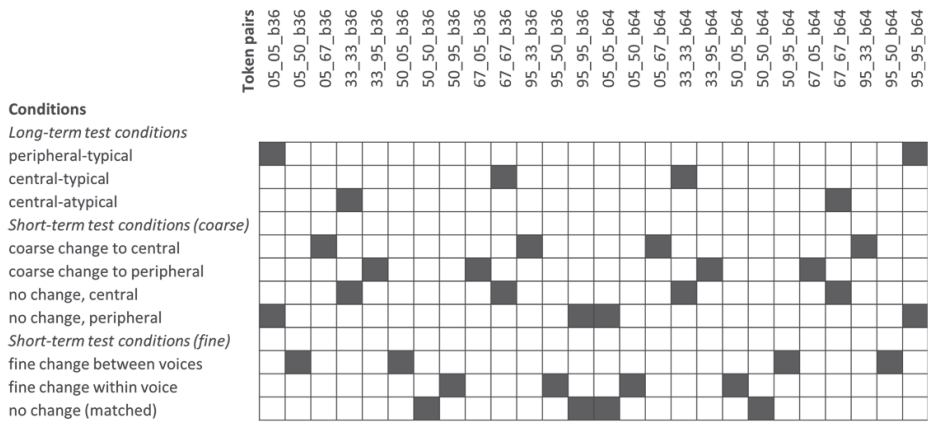
In conclusion, we propose that the right middle STS processes incoming voice stimuli with respect to their distance from the representation of a supra-individual "mean voice" category (i.e., the average across talkers of the listener's recent voice-acoustic history). This representation does not seem to be biased by voice identity information, rather it collapses across individual voices. The right IFC, in contrast, processes voice stimuli with respect to their distance from representations of "individual mean voices" that are the average of the listener's memories of the voices of specific individuals. According to this view, the IFC maintains multiple "individual mean voice" representations, one for each voice remembered (see Fig. 2c for a schematic illustration of the proposed representations). In this study, we presented the first evidence for this multilevel long-term mean-based coding in voice-selective cortical regions.

Appendices

Appendix A. Training stimuli

Trained voice identity	Trained identity boundary	Mean of all trained morphs	Stimulus morph levels used during training														
A	36	26	1	10	17	22	25	27	28	29	30	31	32	34	34	35	35
B	36	46	37	37	37	38	38	38	39	39	39	40	42	46	55	66	99
A	64	54	1	34	45	54	58	60	61	61	61	62	62	62	63	63	63
B	64	74	65	65	66	66	68	69	70	71	72	73	75	78	83	90	99

Appendix B. Experimental conditions as defined by token pair types of the fMRI tests. For example, '05_50_b36' refers to the token pair type in which the first stimulus was morph 05, the second stimulus was morph 50, and the trained identity boundary was at morph 36.



References

Aguirre, G.K., 2007. Continuous carry-over designs for fMRI. Neuroimage 35, 1480-1494.

Andics, A., McQueen, J.M., Petersson, K.M., Gál, V., Rudas, G., Vidnyánszky, Z., 2010. Neural mechanisms for voice recognition. Neuroimage 52, 1528-1540.

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., Martin, A., 2004. Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nature Neuroscience 7, 1190-1192.

Belin, P., Bestelmeyer, P.E.G., Latinus, M., Watson, R., 2011. Understanding voice perception. British Journal of Psychology 102, 711-725.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences 8, 129-135.

Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14, 2105-2109.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309-312.

Bestelmeyer, P.E.G., Latinus, M., Bruckert, L., Rouger, J., Crabbe, F., Belin, P., 2012. Implicitly perceived vocal attractiveness modulates prefrontal cortex activity. Cerebral Cortex.

Boersma, P., Weenink, D., 2007. Praat: doing phonetics by computer (Version 4.2.07). [Computer program]

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G.A., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Current Biology 20, 116-120.

Charest, I., Pernet, C., Latinus, M., Crabbe, F., Belin, P., 2012. Cerebral processing of voice gender studied using a continuous carryover fMRI design. Cerebral Cortex.

Cohen Kadosh, K., Henson, R., Cohen Kadosh, R., Johnson, M., Dick, F., 2010. Task-dependent activation of face-sensitive cortex: An fMRI adaptation study. Journal of Cognitive Neuroscience 22, 903-917.

Epstein, R.A., Parker, W.E., Feiler, A.M., 2008. Two kinds of fMRI repetition suppression? Evidence for dissociable neural mechanisms. Journal of Neurophysiology 99, 2877-2886.

Ethofer, T., Bretscher, J., Gschwind, M., Kreifelts, B., Wildgruber, D., Vuilleumier, P., 2012. Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. Cerebral Cortex 22, 191-200.

Ethofer, T., Kreifelts, B., Wiethoff, S., Wolf, J., Grodd, W., Vuilleumier, P., Wildgruber, D., 2009a. Differential influences of emotion, task, and novelty on brain regions underlying the processing of speech melody. Journal of Cognitive Neuroscience 21, 1255-1268.

Ethofer, T., Van De Ville, D., Scherer, K., Vuilleumier, P., 2009b. Decoding of emotional information in voice-sensitive cortices. Current Biology 19, 1028-1033.

Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2005. Sensitivity to voice in human prefrontal cortex. Journal of Neurophysiology 94, 2251-2254.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322, 970-973.

Friston, K.J., Ashburner, J., Kiebel, S.J., Nichols, T.E., Penny, W.D., editors. 2007. StatisticalParametricMapping: The Analysis of Functional Brain Images. London (UK): Academic Press.

Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Sfaello, I., Barthelemy, C., Brunelle, F., Samson, Y., Zilbovicius, M., 2004. Abnormal cortical voice processing in autism. Nature Neuroscience 7, 801-802.

Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., Vuilleumier, P., 2005. The voices of wrath: brain responses to angry prosody in meaningless speech. Nature Neuroscience 8, 145-146.

Grill-Spector, K., Henson, R., Martin, A., 2006. Repetition and the brain: neural models of stimulus-specific effects. Trends in Cognitive Sciences 10, 14-23.

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., Hirsch, J., 2008. Detection of time-varying signals in event-related fMRI designs. Neuroimage 43:509–520.

Hackett, T.A., Stepniewska, I., Kaas, J.H., 1998. Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. The Journal of Comparative Neurology 394, 475-495.

Hoffman, K.L., Logothetis, N.K., 2009. Cortical mechanisms of sensory learning and object recognition. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 321-329.

Kawahara, H., 2006. STRAIGHT, exploration of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds. Acoustical Science and Technology 27, 349-353.

Larsson, J., Smith, A.T., 2012. FMRI repetition suppression: neuronal adaptation or stimulus expectation? Cerebral Cortex 22, 567-576.

Latinus, M., Belin, P., 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. Frontiers in Psychology 2, 1-12.

Latinus, M., Crabbe, F., Belin, P., 2009. FMRI investigations of voice identity perception. Organization for Human Brain Mapping 2009 Annual Meeting, July 2009: Neuroimage 47(Supplement 1), S156.

Latinus, M., Crabbe, F., Belin, P., 2011. Learning-induced changes in the cerebral processing of voice identity. Cerebral Cortex 21, 2820-2828.

Leopold, D.A., Bondar, I., Giese, M.A., 2006. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. Nature 442, 572-575.

Leopold, D.A., O'Toole, A.J., Vetter, T., Blanz, V., 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. Nature Neuroscience 4, 89-94.

Loffler, G., Yourganov, G., Wilkinson, F., Wilson, H.R., 2005. FMRI evidence for the neural representation of faces. Nature Neuroscience 8, 1386-1390.

Mullennix, J.W., Ross, A., Smith, C., Kuykendall, K., Conard, J., Barb, S., 2011. Typicality effects on memory for voice: Implications for earwitness testimony. Applied Cognitive Psychology 25, 29-34.

Panis, S., Wagemans, J., Op de Beeck, H.P., 2011. Dynamic norm-based encoding for unfamiliar shapes in human visual cortex. Journal of Cognitive Neuroscience 23, 1829-1843.

Papcun, G., Kreiman, J., Davis, A., 1989. Long-term memory for unfamiliar voices. Journal of the Acoustical Society of America 85, 913-925.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nature Neuroscience 11, 367-374.

Rhodes, G., Jeffery, L., 2006. Adaptive norm-based coding of facial identity. Vision Research 46, 2977-2987.

Romanski, L.M., Averbeck, B.B., Diltz, M., 2005. Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. Journal of Neurophysiology 93, 734-747.

Romanski, L.M., Bates, J.F., Goldman-Rakic, P.S., 1999. Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. The Journal of Comparative Neurology 403, 141-157.

Romanski, L.M., Goldman-Rakic, P.S., 2002. An auditory domain in primate prefrontal cortex. Nature Neuroscience 5, 15-16.

Schwarzbauer, C., Davis, M.H., Rodd, J.M., Johnsrude, I.S., 2006. Interleaved silent steady state (ISSS) imaging: A new sparse imaging method applied to auditory fMRI. Neuroimage 29, 774-782.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends in Neurosciences 26, 100-107.

Stevens, A.A., 2004. Dissociating the cortical basis of memory for voices, words and tones. Cognitive Brain Research 18, 162-171.

Summerfield, C., Trittschuh, E.H., Monti, J.M., Mesulam, M.M., Egner, T., 2008. Neural repetition suppression reflects fulfilled perceptual expectations. Nature Neuroscience 11, 1004-1006.

Van Lancker, D., Canter, G.J., 1982. Impairment of voice and face recognition in patients with hemispheric damage. Brain and Cognition 1, 185-195.

von Kriegstein, K., Giraud, A.L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22, 948-955.

von Kriegstein, K., Giraud, A.L., 2006. Implicit multisensory associations influence voice recognition. PLoS Biology 4, e326.

Wagner, A.D., Koutstaal, W., Maril, A., Schacter, D.L., Buckner, R.L., 2000. Task-specific repetition priming in left inferior prefrontal cortex. Cerebral Cortex 10, 1176-1184.

Warren, J., Scott, S., Price, C., Griffiths, T., 2006. Human brain mechanisms for the early analysis of voices. Neuroimage 31, 1389-1397.

Chapter 7

Summary and conclusions

Summary

Recognizing a person from his or her speech is a basic social ability. This dissertation aimed at providing a better understanding of how voice identities are learned and what the principles of perceptual and neural organization of voice representations are. These questions were investigated in a series of behavioural and neuroimaging experiments.

Behavioural experiments

The experiment presented in Chapter 2 investigated segmental contributions to voice discriminability, and the correspondence of perceptual and acoustic similarity of voices. Participants heard a continuous stream of voices (several tokens of different Dutch words from multiple male speakers) and had to decide if the person they heard saying a word was the same or different from the person saying the previous word. It was found that listeners are very good at discriminating voices, but they vary considerably in what they perceive as within-voice versus across-voice variation. Voice discrimination performance was not independent from segmental content: words with the phonemes /m/, /e/ and /s/ helped voice discrimination more than words with the phonemes with /l/, /o/ and /t/ in onset, nucleus and coda positions respectively. These segmental benefits were reflected in relatively lower within-voice and higher across-voice acoustic variations for more distinctive segments – this is exactly what made the cues in these segments good person identity cues. Listeners were quick to use information in all three segment positions of the words. Furthermore, listeners agreed in which voices are more and less discriminable. Less discriminable voices were also less identifiable, despite lower within-voice acoustic variability, thus supporting the view that voices are organized in a prototype-based way. The distribution of voices on a perceptual discriminability-based distance map showed a great similarity to their distribution along formant-based acoustic dimensions, suggesting that voice typicality can be relatively well explained by simple spectral cues. A map of words was based on how similar their contributions are to voice typicality. The segment-based organization of this map indicated the presence of segment-specific prototype voices.

Chapter 3 described two multisession training experiments investigating the flexibility and the specificity of voice identity learning. The same voice morph continua between two selected talkers saying *mes* (knife) and *lot* (fate) were used with systematically

varied voice identity category feedback in a between-session and between-experiment learning and re-learning paradigm. In Experiment 1, listeners were trained to categorize voices in a 'person A or person B' task on two consecutive days, but they were unaware of a feedback-determined shift in voice identity boundary across days. The results showed that new voice identity categories are learned quickly and learning is stable even after a day. Listeners were flexible to learn and re-learn artificially defined voice identity boundaries, but this flexibility also had its limits: asymmetric category feedback leading to an oversized identity for a voice was not fully tolerated. This suggested that listeners have built-in expectations on the acceptance range of individual voice categories. Much of the voice knowledge generalized to untrained words with and without segmental overlap, but the transfer was not full for voice identity centers. Furthermore, performance was better for an untrained word that was segmentally overlapping with the trained word than for one that was segmentally unrelated to the trained word. The effect of word on voice categorization responses also indicated the presence of segment-specific representations, and that the acceptance range of within-voice variation is segment-specific. These findings demonstrated that voice knowledge entails abstraction, and suggested a role for both non-segmental and segmental cues in voice identity processing.

In Experiment 2 of Chapter 2, listeners were trained to categorize stimuli from the same voice morph continua as in Experiment 1, but now in a 'person A or not person A' task. The person A category was trained to be in between the two natural voices. So, what was a category boundary in Experiment 1 became a category center in Experiment 2. Here, category position also varied across listeners. The results showed that listeners readily learned these voice identity categories. This demonstrated that no built-in voice identity category structure information is encoded in the speech signal, and that morphing did not make the stimuli sound less natural. Again, as in Experiment 1, some of the trained voice knowledge transferred to an untrained word, but with a great loss of categorization sharpness, confirming the role of both non-segmental and segmental information. Finally, less 'person A' decisions were made after a short delay compared to no delay after training, suggesting that voice identity acceptance ranges may become narrower over time spent without reassuring evidence.

Chapter 4 presented a voice learning experiment testing the perceptual limitations of voice category formation. Listeners were trained to form categories for groups ('families') of

individual voices saying /mes/ and /lot/. Trained within-category variation was thus larger than typical within-talker acoustic variation. Listeners were then presented with both within-family and across-family voice morph stimuli, and they were asked if they had heard the voice before or not, and which family the voice was a member of. The prediction was that prototype formation for the voice families would benefit within-family morphs over across-family morphs, while prototype formation for individual voices would benefit voice endpoints over morphs. Endpoint benefits were found over morphs in both categorization responses and in recognition confidence, but no difference was found between within- and across-family morphs, suggesting that while individual voice prototypes are easily formed even implicitly, voice family prototypes are not formed, despite explicit feedback, and despite the fact that the family categories were learned. This demonstrates a built-in category size restriction for voice prototype formation, similarly to what was found for faces (Cabeza et al., 1999). It was also shown that voices saying /lot/ are more readily recognized as known voices than those saying /mes/. This is yet another demonstration of segment-specific acceptance ranges. Also, family categorization confidence increased with the amount of training more for /mes/ than for /lot/. Taken together with the finding from Chapter 2 that the phonemes of /mes/ are more distinctive than those of /lot/, these results suggest that phonetic content modulates voice category formation such that words with more distinctive phonemes support voice learning but make voice category representations more sensitive to variation.

Neuroimaging experiments

Chapter 5 described a multisession training study investigating the neural mechanisms of voice recognition with functional magnetic resonance imaging (fMRI). Hungarian listeners were trained to categorize stimuli from a voice morph continuum between two talkers saying the Hungarian words "bú" [sadness], "fű" [grass], "ki" [out], "lé" [liquid], "ma" [today] and "se" [neither]. As in Experiment 2 of Chapter 3, the trained category was in the middle of the continuum, and participants had a ‘person A or not person A’ task during training. As in Experiment 1 of Chapter 3, to manipulate perceived voice category structure properties of the stimuli (i.e., category-internal, category boundary, category-external) within-participant and across tests, feedback determined different category boundary positions on different days. Here, a one week delay was used between

the two fMRI tests performed, each preceded by extensive training over two days. At fMRI tests, listeners heard a series of word stimuli and had to perform either a voice recognition or a word repetition detection task. A fast sparse scanning sequence was applied that combined the advantages of close-to-continuous data sampling and presenting stimuli in silence. Crucially, the trained category was learned, and the trained difference in the category boundary across weeks was still there at each test. By taking into account the relationship between the actual and the preceding stimulus, the effects of short-term acoustic similarity sensitivity (found in bilateral middle/ posterior STS, and right IFC) could be separated from the effects of neural sharpening of long-term stored typical values. Furthermore, the analyses revealed two anatomically separable types of typicality-based long-term voice representation: one in a voice-acoustic space (central vs peripheral; right orbital / insular cortex, right posterior medial STS) and one in a voice identity space (identity-internal vs identity-external; bilateral anterior temporal pole, left deep posterior STS, left amygdala). This study is the first to provide neuroimaging evidence for the existence of flexible 'mean voice' representations, demonstrating the norm-based organization of neural voice spaces. Voice identity categorization performance was found to correlate with neural sensitivity to voice identity similarity (right middle / posterior STS, left deep posterior STS, right anterior temporal pole, left amygdala): listeners with a greater neural sensitivity were better at recognizing voices. This finding demonstrated the direct behavioural relevance of norm-based neural representations of voice identities. It was also found that these neural patterns were not modulated by decision difficulty. Nevertheless, no neural similarity sensitivity was found when listeners had a different task (word repetition detection) that diverted their attention away from voice identities. This indicated the role of attentional enhancement of fMRI repetition suppression effects for voices.

Finally, Chapter 6 presented a second multisession fMRI study that focused on the neural coding of voice identities in voice-selective cortical regions. Here, an in-house modified sparse scanning protocol was applied. As in the experiment in Chapter 5, the learning and re-learning paradigm was used to separately manipulate across-talker and within-talker typicality patterns in a within-participant design (two fMRI tests with a one week delay, each preceded by extensive training over two days). But now, as in Experiment 1 of Chapter 3, Dutch talkers saying /mes/ were used, and listeners performed a ‘person A or person B’ task. At fMRI tests, listeners heard pairs of words. They had to perform a voice

categorization task on the second word of the pair. The results showed that the trained categories were learned and that trained category boundary changes were still present at fMRI tests. Voice-selective regions were specified with a functional localizer (Belin et al., 2000), and included the bilateral STS and the IFC (lateralized to the right hemisphere). The analyses revealed two anatomically separable levels in the voice-processing hierarchy, both coding long-term mean voices: a supra-individual level coding an acoustic average voice (central vs peripheral; right STS) and an intra-individual level coding the identity-mean of specific voices (typical vs atypical; right IFC). Interestingly, these two voice-selective regions could also be identified by using the very same test with different directions: central-atypical < peripheral-typical revealed right STS, central-atypical > peripheral-typical revealed right IFC. Follow-up tests confirmed that these findings were not caused by changes in decision difficulty. Furthermore, short-term similarity sensitivity to coarse but not to fine acoustic changes was found in the bilateral STS. This short-term sensitivity effect was present for central but not for peripheral stimuli, indicating an interesting influence of long-term acoustic centrality on short-term processing. Advancing on recent findings from behavioural studies of voice processing (Papcun et al., 1989; Bruckert et al., 2010; Mullennix et al., 2011; Latinus and Belin, 2011) and convergent with those in the face processing domain (Loffler et al., 2005), this study provides the first evidence of the typicality-based organization of neural voice representations in voice-selective cortical regions.

Conclusions

The experiments presented in this thesis shed new light on various aspects of voice identity learning and, more generally, auditory object processing. Important conclusions can be drawn about the adaptivity of voice representations, and about the types and levels of abstraction in talker identity processing. These points will be discussed in turn in the following sections.

Adaptivity in voice identity learning

This dissertation investigated the nature of category formation for voice identities. A series of behavioural and neural experiments demonstrated that voice identity coding is adaptive, similarly to what has been found for faces (Rhodes and Jeffery, 2006). Adaptivity means readiness to change and robustness in a changing environment. This section draws some general conclusions based on the evidence presented here on flexibility and stability in voice identity learning.

Voice identity categories, unlike phonetic categories in adulthood (Logan et al., 1991), are quickly learned (Chapter 3), even implicitly (Chapter 4). Neural response patterns also showed evidence of implicit prototype formation for supra-individual representational spaces (Chapters 5 and 6). Furthermore, voice identities are quickly re-learned after a category shift, just like phonetic categories (Norris et al., 2003; Chapter 3). This re-learning is supported by plasticity in neural coding, as exemplified by dynamic adjustments of cortical response patterns to changes in voice identity typicality (Chapters 5 and 6). Anchor points, such as category centres and category boundaries for voice identities thus do not seem to be determined by nonlinearities in the speech signal. If voice identities were determined by nonlinearities, then there ought not to be such plasticity.

There are also dynamic changes for the amount of variation that is tolerated for a given talker. Voice identity acceptance ranges are narrower for cues based on more distinctive segments or words, that is, those with lower within- and higher across-voice variability (Chapters 2 and 4). It might have been this increased sensitivity to variation that made voice identity learning based on these more distinctive segments more efficient (Chapter 4). Acceptance ranges also vary across listeners: there are more conservative and more liberal voice perceivers (Chapter 2). Finally, this conservativism seems to change over

time: trained voice identity acceptance ranges become narrower even after a short delay (Chapter 3).

But flexibility in voice processing also has its limits, and these are most apparent in category size restrictions. People have built-in limitations for what size can be accepted for a person identity category, both for faces (Cabeza et al., 1999) and for voices. Oversized individual categories, where within-voice changes exceed typical intra-individual variation, are not learned, despite explicit training (Chapter 3). This does not mean that only individual voice categories are represented in the human brain: supra-individual voice spaces are also maintained (Chapters 2, 5 and 6). But, supporting the size restriction claim, no prototype appears to be formed for trained voice family categories (Chapter 4). The voice processing system may have a preference for representational spaces with the funcionally most relevant sizes, such as the size of an individual voice space (around a prototype of e.g., Bob's voice; cf. person identity nodes) or the size of a species-specific voice space (around a prototype of all human vocalizations; cf. voice-selective brain regions), in contrast with functionally less relevant sizes, such as the size of a two-person voice space (e.g., around a voice family prototype).

Despite all this flexibility, voice identity representations are relatively stable over time (Chapter 3). Multiple person identity cues are used, including segment-specific cues, making voice processing less fragile in case of unexpected variations. Indeed, different person identity cues are affected by different situations. For instance, having a cold mainly influences nasal sounds, while trying to imitate another person's voice typically distorts non-segmental cues (Eriksson and Wretling, 1997). Interestingly, voice identity representations are also relatively stable across listeners: the perceived typicality of a voice does not depend on the perceiver (Chapter 2). This does not mean that listeners have a built-in prototype-voice, it rather means that listeners with similar perceptual histories build up similar representational spaces. So there seems to be little difference in what cues various listeners use and how they use them.

Multiple levels of abstraction in voice recognition

Abstraction is a fundamental concept of human perception, but a concept that researchers use with various meanings, pointing to different key phenomena in information processing. Abstraction may refer to zooming in and out to extract relevant information

from the signal, to the calculation of the average across a distribution of values, and to advancement of the represented information in the processing hierarchy. I argue here that the findings in this thesis revealed multiple levels of abstraction in voice recognition, in all of these three senses.

The first meaning of abstraction refers to scaling. This elaborates on the idea that the similarity-based representational spaces we use in object processing (cf. Valentini, 1991) may vary in their cue specificity, sensitivity, time window and size. In this sense, a more abstract representation refers to a space with less specific cues, with lower sensitivity to variation, with a larger time window or with a bigger size. The experiments presented here provided evidence for multiple levels of scaling in voice identity processing for each of these characteristics. These will be discussed in turn.

Similarity-based representational spaces are assumed to vary in what cues they use. The presented studies suggested that perceived voice similarity can be well described by basic spectral cues (F0, F1, F2; Chapter 2), but that both segment-specific and more abstract, non-segmental cues are involved in voice identity learning. Note that segment-specific cues were not token-specific, so already they entailed abstraction (Chapter 3). Further indications of cue specificity differences were found with fMRI. The right anterior temporal pole seemed to be involved in a modality-specific representation of vocal identity, while the deep posterior STS was suggested to maintain modality-nonspecific person identity representations (Chapter 5; Campanella and Belin, 2007).

Differences in sensitivity to certain changes were also demonstrated in the fMRI studies. The voice-selective bilateral STS was sensitive to coarser but not to very fine acoustic changes (Chapter 6). Fine change detection is thought to take place in the primary auditory cortex, an area which is not specialized for voices (Belin et al., 2000).

Another variable property of representational spaces are their time windows. It was shown that voice-selective brain regions maintain both short-term and long-term representational spaces. Short-term spaces were found to be sensitive to how similar a voice token is to another token heard immediately before. Long-term spaces, in contrast, were found to be sensitive to how similar a voice token is to the central value of a series of previously heard tokens (Chapters 5 and 6). These different time windows appeared to correspond to two different kinds of fMRI repetition suppression (cf. Epstein et al., 2008).

Finally, representational spaces were also found to vary in size. Evidence was shown for large spaces representing voice tokens corresponding to different voice identities (Chapters 2, 5 and 6), and for spaces with narrower acceptance ranges that did not exceed intra-individual variation (Chapters 3, 4, 5 and 6). So there seem to exist voice spaces that can encode many or all human vocalizations in a single category, and additional voice spaces for each talker separately. Furthermore, intra-individual representational space varied with phonetic content: for example, the voice identity spaces based on the word /lot/ had broader acceptance ranges than those based on /mes/ (Chapter 4). These spaces seemed to fit to differences in natural variation: indeed, phonemes in the words corresponding to narrower voice identity acceptance ranges were shown to have relatively lower within-talker and higher across-talker variability (Chapter 2).

The second meaning of abstraction relates to averaging. It has been argued that similarity-based representational spaces are organized around norms. This is called norm-based coding (cf. Valentine, 1991). Abstraction in this sense refers to the creation of this norm by calculating the average of the values in the specific space. This abstractionist model of object processing is countered by exemplar-based models. In this theoretical contrast, exemplars are the representations of the observed events, while the norms are average, calculated values. As discussed below, this thesis presented evidence for norm-based coding of voice identities, and for the differential coding of more central and more peripheral values in neural voice spaces.

The first piece of evidence for the typicality-based organization of talker identities was that the voices that are difficult to distinguish from other voices for all listeners consistently are exactly the voices for which different tokens are less readily accepted as tokens of the same voice, although within-voice acoustic variability was not higher but lower for them (Chapter 2). It has been argued that narrower acceptance ranges around less distinctive, close-to-the-average exemplars are an indication of prototype-based organization (e.g., Kuhl, 1991, Loffler et al., 2005). Furthermore, voice group categorization benefits were found for stimuli around individual voice category centers compared to stimuli that were far from these centers, despite no explicit training for those voice identities (Chapter 4). I have also argued that the typicality-based spaces revealed in the present experiments were not organized around acoustically defined, absolute anchor points, but around means defined relative to the actual voice space: indeed, voice identity

means were shown to dynamically follow the trained category shifts (Chapter 3). This was also supported by the fMRI experiments, which provided the most convincing evidence for norm-based coding. Neural sharpening was found for typical compared to atypical exemplars of individual voices. These changes in neural activity could not be explained by acoustic changes of the voice signal, but only by changes in perceived typicality, therefore demonstrating that the anchor points of the neural spaces representing voice identity are not absolute values but quickly adapt their position to new perceptual evidence. For that, the voice identity norm had to be calculated and a special status had to be assigned to it, exactly as proposed by norm-based but not by exemplar-based coding models (Valentine, 1991; Jeffery et al., 2011; Chapter 5 and 6).

Voice processing thus seems to entail abstraction in terms of both scaling and averaging. Taken together, this suggests that multiple norm-based representational spaces exist for voices, each with its own norm. Consequently, we should for example have segment-specific norms for voice identities, or at least specific norms for each relevant cue that may be present in only a subset of segments. This was illustrated by the apparently segment-based organization of a distance map of several words calculated from how similar each word's contribution was to voice typicality (Chapter 2), and by word effects in the voice identity learning studies (Chapter 3).

The third meaning of abstraction concerns the advancement of information through a processing hierarchy. It is used in relation to hierarchical models of object perception (e.g., Bruce and Young, 1986; Belin et al., 2004, 2011) that postulate serially organized processing stages. In this sense, a more abstract level means a higher, more advanced stage in the processing hierarchy. The experiments in this thesis demonstrated multiple levels of advancement in cortical hierarchy for voices: as overviewed below, functionally and anatomically distinct stages were found in voice identity processing.

As we have already seen, multiple norm-based representational spaces seem to play a role in voice perception. The important contribution of neuroimaging to this is the finding that these multiple spaces are implemented at anatomically distinct locations in the human brain. Regions sensitive to long-term acoustic centrality were found in middle and posterior parts of the right STS, while regions sensitive to identity centrality were found in anterior temporal regions (ATP, Chapter 5) and in voice-selective inferior-frontal regions (IFC, Chapter 6). These voice-selective regions were proposed to be stages of the auditory 'what'

pathway (Belin et al., 2004; Ahveninen et al., 2006), with the STS having direct and strong structural connections downwards to the primary auditory cortex (Kumar et al., 2007) and upwards to both anterior temporal and inferior frontal regions (Ethofer et al., 2012). These different types of neural sensitivity at anatomically distinct locations thus seem to be cortical instantiations of specific stages in a voice processing hierarchy.

Taken together, abstraction is present on many levels and in many ways in voice recognition. It has been argued that different stages of the voice processing hierarchy are responsible for acoustic and identity processing (Belin et al., 2004, 2011). But this thesis also suggested that acoustic and identity sensitivity, while indeed being distinct both anatomically and functionally, can also be implemented by a single neural coding mechanism for similarity-based representational spaces that only differ in space size (i.e., a large supra-individual space, and narrow intra-individual spaces). In a broader perspective, a structure that contains multiple levels does not necessarily use complicated mechanisms. Fractals in mathematics are well-known examples of complex structures that are created with very simple rules. But the key there is that those simple rules are used again and again for various parts of the whole. After all, abstraction at different levels may be the means to build up a complex architecture from a small set of simple rules. I have argued that this appears to be the case for human voice processing.

In this dissertation I have shown that person recognition from a talker's voice is based on multiple, segmental and non-segmental cues, and that these cues all contribute to perceived voice typicality in specific ways. Voice identities have proved to be natural auditory objects in the speech signal, with built-in presuppositions on what may constitute an individual voice category. Talker identities were found to be represented by multiple, adaptive norm-based neural codes, on functionally and anatomically distinct, hierarchically organized levels in the human brain. These levels include a supra-individual voice space in voice-selective regions along the superior temporal sulcus, and intra-individual voice spaces in anterior temporal and inferior frontal regions of the right hemisphere.

References

Ahveninen, J., Jääskeläinen, I.P., Raij, T., Bonmassar, G., Devore, S., Hämäläinen, M., Levänen, S., Lin, F.-H., Sams, M., Shinn-Cunningham, B.G., Witzel, T., Belliveau, J.W., 2006. Task-modulated "what" and "where" pathways in human auditory cortex. Proceedings of the National Academy of Sciences of the United States of America 103, 14608–14613.

Belin, P., Bestelmeyer, P.E.G., Latinus, M., Watson, R., 2011. Understanding voice perception. British Journal of Psychology 102, 711-725.

Belin, P., Fecteau, S., Bedard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences 8, 129-135.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309-312.

Bruce, V., Young, A., 1986. Understanding face recognition. British Journal of Psychology 77, 305-327.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G.A., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Current Biology 20, 116-120.

Cabeza, R., Bruce, V., Kato, T., Oda, M., 1999. Prototype effect in face recognition: Extension and limits. Memory and Cognition 27, 139-151.

Campanella, S., Belin, P., 2007. Integrating face and voice in person perception. Trends in Cognitive Sciences 11, 535-543.

Epstein, R.A., Parker, W.E., Feiler, A.M., 2008. Two kinds of fMRI repetition suppression? Evidence for dissociable neural mechanisms. Journal of Neurophysiology 99, 2877-2886.

Eriksson, A., Wretling, P., 1997. How flexible is the human voice? – A case study of mimicry. Proceedings of EUROSPEECH 1997, Rhodes 2, 1043-1046.

Ethofer, T., Bretscher, J., Gschwind, M., Kreifelts, B., Wildgruber, D., Vuilleumier, P., 2012. Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. Cerebral Cortex 22, 191-200.

Jeffery, L., Rhodes, G., McKone, E., Pellicano, E., Crookes, K., Taylor, E., 2011. Distinguishing norm-based from exemplar-based coding of identity in children: evidence from face

identity aftereffects. Journal of Experimental Psychology: Human Perception and Performance 37(6), 1824-1840.

Kuhl, P.K., 1991. Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. Perception & Psychophysics 50, 93-107.

Kumar, S., Stephan, K.E., Warren, J.D., Friston, K.J., Griffiths, T.D., 2007. Hierarchical Processing of Auditory Objects in Humans. PLoS Computational Biology 3(6), e100.

Latinus, M., Belin, P., 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. Frontiers in Psychology 2, 1-12.

Loffler, G., Yourganov, G., Wilkinson, F., Wilson, H.R., 2005. FMRI evidence for the neural representation of faces. Nature Neuroscience 8(10), 1386-1390.

Logan, J.S., Lively, S.E., Pisoni, D.B., 1991. Training Japanese listeners to identify English /r/ and /l/: A first report. Journal of the Acoustical Society of America 89, 874-886.

Mullennix, J.W., Ross, A., Smith, C., Kuykendall, K., Conard, J., Barb, S., 2011. Typicality effects on memory for voice: Implications for earwitness testimony. Applied Cognitive Psychology 25, 29-34.

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cognitive Psychology 47(2), 204-238.

Papcun, G., Kreiman, J., Davis, A., 1989. Long-term memory for unfamiliar voices. Journal of the Acoustical Society of America 85, 913-925.

Rhodes, G., Jeffery, L., 2006. Adaptive norm-based coding of facial identity. Vision Research, 46, 2977-2987.

Valentine, T., 1991. A unified account of the effects of distinctiveness, inversion, and race in face recognition. Quarterly Journal of Experimental Psychology A: Human Experimental Psychology 43, 161-204.

Samenvatting en conclusies[1]

[1] De lijst met referenties voor dit hoofdstuk staan op pagina's 177-178.

Samenvatting

Mensen kunnen herkennen aan de hand van hun spraak is een basale sociale vaardigheid. Dit proefschrift is er op gericht beter te begrijpen hoe stemidentiteit wordt aangeleerd, en wat de onderliggende principes van perceptuele en neurale organisatie zijn. Deze vragen zijn onderzocht in een aantal gedragsmatige en neuroimaging experimenten.

Gedragsmatige experimenten

De experimenten die in Hoofdstuk 2 beschreven staan onderzochten de bijdrage van spraaksegmenten aan stemonderscheiding, en de overeenkomst tussen perceptuele en akoestische gelijkenis tussen stemmen. Proefpersonen hoorden een continue stroom aan stemmen (een aantal opnames van verschillende Nederlandse woorden, gesproken door verschillende sprekers) en moesten steeds aangeven of de persoon die ze hoorden dezelfde individu was als de persoon die het woord daarvoor had uitgesproken. De resultaten lieten zien dat luisteraars erg goed zijn in het onderscheiden van stemmen, maar ook dat luisteraars onderling aanzienlijke verschillen vertonen in wat ze waarnemen als variatie binnen een persoon en variatie tussen personen. Het vermogen om stemmen te onderscheiden was ook afhankelijk van het specifieke segment: woorden met de fonemen /m/, /e/ en /s/ leidden tot een beter onderscheidend vermogen dan woorden met de fonemen /l/, /o/ en /t/ in respectievelijk onset, nucleus en coda positie. Deze segmentele voordelen werden weerspiegeld in relatief lagere variatie binnen een stem en meer variatie tussen stemmen - dit is precies wat de informatiebronnen in deze segmenten goede indicatoren maakte voor stemidentiteit. Luisteraars konden snel gebruik maken van informatie in alle drie de segmentposities van de woorden. Luisteraars waren het bovendien vaak eens over welke stemmen beter of juist minder goed te onderscheiden zijn. Stemmen die minder goed te onderscheiden waren, waren ook minder goed te identificeren, ondanks lagere akoestische variatie binnen stemmen, hetgeen ondersteuning biedt aan het idee dat stemmen mentaal georganiseerd zijn volgens prototypen. De distributie van stemmen op een perceptueel georganiseerde afstandskaart (gebaseerd op de mate waarin stemmen te onderscheiden waren) vertoonden sterke gelijkenis met de distributie van stemmen in dimensies die akoestische formantwaarden volgen. Dit suggereert dat de mate waarin een stem karakteristiek is, goed verklaard kan worden door simpele spectrale

informatiebronnen. Verder werd ook een kaart van woorden gemaakt, gebaseerd op de mate waarin hun bijdrage aan stemkararakteristiekheid gelijkenis vertoonde. Het feit dat de organisatie van deze kaart voornamelijk gebaseerd was op segmenten wees op het bestaan van segmentspecifieke stem prototypen.

Hoofdstuk 3 beschreef twee trainexperimenten bestaande uit een aantal delen. Deze onderzochten de flexibiliteit en de specificiteit van het leren van stemidentiteit. Eenzelfde stemcontinuüm tussen twee sprekers werd gebruikt, zowel voor het woord "mes" als "lot". Deze continua werden aangeboden met systematische variatie in feedback over stemidentiteit, zowel met een inter-sessie en een inter-experiment leer en herleer paradigma. In Experiment 1 werden luisteraars in twee dagen getraind om twee stemmen als "persoon A" of "persoon B" te categoriseren. Ze waren zich echter niet bewust van een feedback-gebaseerde manipulatie waardoor de grens van stemidentiteit verschilde tussen de dagen. De resultaten lieten zien dat nieuwe stemidentiteitcategorieën snel aangeleerd kunnen worden en stabiel blijven, zelfs na een dag. Luisteraars bleken vrij flexibel in het leren en herleren van artificieel gedefinieerde stemidentiteit grenzen. Deze flexibiliteit was echter niet onbeperkt: wanneer feedback tot een te grote identiteitscategorie leidde, werd dit niet compleet getolereerd. Dit suggereert dat luisteraars ingebouwde verwachtingen hebben met betrekking tot de acceptabele grootte van stemcategorieën. Een groot deel van de aangeleerde kennis over stemmen generaliseerde naar ongetrainde woorden, zowel met als zonder segmentele overeenkomsten, maar de overdracht was niet compleet voor identiteitscentra. Verder waren de prestaties beter voor een ongetraind woord dat grote segmentele overlap had met het getrainde woord dan voor een woord dat qua segmenten ongerelateerd was aan het getrainde woord. Het effect van woord op stemcategorisatie antwoorden liet ook zien dat er segmentspecifieke representaties aanwezig zijn, en dat de range van acceptabele items binnen een stem segmentspecifiek is. Deze bevindingen lieten zien dat kennis over stemmen abstracte informatie bevat, en suggereert dat er een rol is voor niet-segmentele en segmentele informatiebronnen bij de verwerking van stemidentiteit.

In Experiment 2 van Hoofdstuk 3 werden luisteraars getraind om stimuli te categoriseren van hetzelfde stemcontinuüm als in Experiment 1, maar nu in een "persoon A" of "niet persoon A" taak. Er werd proefpersonen aangeleerd dat de persoon-A categorie tussen twee natuurlijke stem categorieën lag. Wat een categoriegrens was in Experiment 1

was dus een categoriecentrum in Experiment 2. De positie van de categorieën varieerde per luisteraar. Luisteraars leerden de stemcategorieën direct. Het spraaksignaal bevat dus geen ingebouwde informatie over categoriestructuur, en het vervormen leidde niet tot onnatuurlijke spraak. Wederom (net als in Experiment 1) generaliseerde een deel van de kennis over stemcategorieën naar nieuwe woorden. Er was echter aanzienlijke afname van specificiteit van categorie grenzen, wat duidt op de invloed van zowel niet-segmentele als segmentspecifieke invloeden. Bovendien werden minder 'persoon A' antwoorden gegeven na een korte pauze dan zonder pauze na de training. Dit suggereert dat de categorieën waarbinnen stemmen acceptabel zijn kleiner worden naarmate de tijd verstrijkt.

Hoofdstuk 4 onderzocht de grenzen van het vermogen om stemcategorieën te leren. Proefpersonen leerden stemgroepen te identificeren ('families') van mensen die "mes" en "lot" uitspraken. De aangeleerde variatie binnen een groep was dus groter dan de variatie die normaal gesproken optreedt binnen een stem. Daarna beluisterden ze gemixte stem stimuli van zowel binnen en buiten de categoriegrenzen. Ze moesten aangeven of ze de stem eerder hadden gehoord, en tot welke familie de stem behoorde. De voorspelling was dat de formatie van prototypen binnen de families beter zou zijn voor gemixte stimuli binnen een familie, en dat prototype formatie voor individuele stemmen beter zou zijn voor de uiteinden van continua dan de gemixte stemmen halverwege de continua. Betere scores voor uiteinden dan gemixte stimuli werden zowel in categorisatie als in antwoordzekerheid gevonden. Echter, vergelijkbare scores werden gevonden voor binnen- en tussen-categorie gemixte stimuli. Dit suggereert dat individuele stemprototypen makkelijk gevormd worden, ook impliciet, maar dat familieprototypen niet makkelijk gevormd worden, ondanks expliciete feedback, en ondanks het feit dat familiecategorieën wel werden aangeleerd. Er is dus een ingebouwde beperking aan de grootte die stemprototype categorieën kunnen krijgen. Dit is vergelijkbaar met bevindingen bij gezichtscategorieën (Cabeza et al., 1999). Verder werden stemmen die /lot/ hadden uitgesproken makkelijker herkend dan stemmen die /mes/ hadden uitgesproken. Dit laat wederom zien dat de grenzen van categorieën afhankelijk zijn van specifieke spraaksegmenten. Verder waren proefpersonen zekerder over hun antwoord bij familiecategorisatie naarmate ze meer met /mes/ trainden dan met /lot/. Samen met de bevindingen uit Hoofdstuk 2 suggereren de resultaten dat fonetische context de formatie van stemcategorieën beïnvloedt, zodanig dat specifiekere fonemen het leren

van stemmen verbetert, maar dat het de stemcategorie representaties ook gevoeliger maakt voor variatie.

Neuroimaging experimenten

Hoofdstuk 5 beschrijft een multisessie trainingstudie waarin de neurale mechanismen van stemherkenning onderzocht werden met zogenaamde "functional magnetic resonance imaging" (fMRI). Hongaarse luisteraars leerden stimuli van een stemcontinuüm categoriseren van twee Hongaarse sprekers die de woorden "bú" [verdriet], "fű" [gras], "ki" [uit], "lé" [vloeistof], "ma" [vandaag] and "se" [geen] uitspraken. Net als in Experiment 2 van Hoofdstuk 3 lag de getrainde categorie in het midden van het continuüm, en konden proefpersonen antwoorden met een 'persoon A' en een 'niet persoon A' categorie. Net als in Experiment 1 van Hoofdstuk 3 bepaalde de feedback dat de grens tussen categorieën op verschillende plekken lag tijdens de verschillende testdagen. Tussen de testsessies zat een week, en de testsessies werden voorafgegaan door uitgebreide training op de twee voorafgaande dagen. Tijdens de fMRI test beluisterden proefpersonen een aantal woorden en moesten ze een stemherkenning taak of een woord-herhalings-herkenning taak uitvoeren. Om bijna continu te kunnen scannen en de stimuli toch in stilte te presenteren werd een zogenaamde "sparse scanning sequence" gebruikt. De resultaten lieten zien dat de categorie kon worden aangeleerd, en dat het aangeleerde verschil in positie van de categoriegrenzen ook tijdens de test nog aanwezig was. Door effecten van zowel de stimulus als de voorgaande stimulus in ogenschouw te nemen, konden de effecten van korte termijn akoestische perceptie (bilateraal in de middelste/posteriore STS en de rechter IFC) onderscheiden worden van de effecten van toename in de neurale specificiteit van typische waarden die over een langere termijn zijn opgeslagen. Twee soorten lange-termijnrepresentaties gebaseerd op specificiteit werden onderscheiden: een in een stemakoestische ruimte (centraal vs. perifeer; rechter orbital / insular cortex, rechter posteriore mediale STS), en een in een stemidentiteit ruimte (identiteitsintern vs. identiteitsextern; bilaterale anteriore temporale pole, linker diepe posteriore STS, linker amygdala). Deze studie verschaft als eerste neuroimaging bewijs voor het bestaan van 'gemiddelde stem'-representaties, wat duidt op een normgebaseerde organisatie van neurale stemruimtes. Individuele scores voor stemcategorisatie correleerde met de neurale gevoeligheid voor gelijkenis tussen stemidentiteiten (rechter middelste / posteriore STS,

linker diepe posteriore STS, rechter anteriore temporale pole, linker amygdala): luisteraars met een grotere neurale gevoeligheid konden ook beter stemmen herkennen. Dit laat de directe gedragsmatige relevantie zien van normgebaseerde neurale representaties van stemidentiteit. De neurale patronen waren niet afhankelijk van de complexiteit van de beslissing. Echter, er was geen neurale gevoeligheid voor gelijkenis wanneer luisteraars een taak uitvoerden welke de aandacht van stemidentiteit afhield (woordherhaling detectie). Dit duidt op aandachtsafhankelijkheid van fMRI "repetition supression" effecten voor stemmen.

Tenslotte beschrijft Hoofdstuk 6 een tweede multisessie fMRI studie waarin de neurale codering van stemidentiteit in stemspecifieke regionen van de cortex onderzocht werd. Hiervoor werd een speciaal gemaakt "sparse scanning" protocol gebruikt. Net als in Hoofdstuk 5 werd een leer-herleer paradigma gebruikt om afzonderlijk tussen-spreker en binnen-spreker specificiteit patronen te manipuleren in een "within subject design" (twee fMRI tests met een week ertussen, beide voorafgegaan door twee training dagen). Echter dit keer werden, net als in experiment 1 van Hoofdstuk 3, Nederlandse sprekers gebruikt die /mes/ zeiden, en luisteraars voerden een 'persoon A' of 'persoon B' taak uit. Tijdens de fMRI test hoorden luisteraars woordparen. Ze voerden een categorisatie taak uit op het tweede woord van het paar. De resultaten lieten zien dat de getrainde categorieën waren aangeleerd en dat de getrainde verandering in de locatie van de categorie grenzen ook tijdens de fMRI test nog aanwezig waren. Stemselectieve regionen werden bepaald met een functioneel lokalisatie paradigma (Belin et al., 2000), en deze behelsde de bilaterale STS en de IFC (rechts gelateraliseerd). De analyses toonden twee anatomisch afzonderlijke niveaus in de stemverwerking hiërarchie, en beide vertoonde codering van lange termijn gemiddelde stemmen: een supra-individueel niveau dat een akoestisch gemiddelde stem codeerde (in de centrale vs. perifere rechts gelateraliseerde STS) en een intra-individueel niveau dat het gemiddelde van een identiteit codeerde voor specifieke stemmen (typisch vs. atypisch; in de rechter STS). Het was opmerkelijk dat deze twee stemselectieve regionen ook geïdentificeerd konden worden door dezelfde test te gebruiken met verschillende richtingen: centraal-ataypisch < perifeer-typisch verscheen in de rechter STS, centraal-atypisch > perifeer-typisch verscheen in de rechter IFC. Verdere tests bevestigden dat deze bevindingen niet werden veroorzaakt door veranderingen in de moeilijkheid van de beslissing. Bovendien, gevoeligheid voor korte termijn gelijkenis op grove, maar niet op fijne, akoestische verschillen werd gevonden in de bilaterale STS. Dit effect van korte-

termijngevoeligheid werd gevonden voor centrale maar niet voor perifere stimuli, wat suggereert dat er een invloed is van lange termijn akoestische centraliteit op de korte-termijnverwerking. In overeenkomst met recente bevindingen in het gebied van gezichtsverwerking (Loffler et al., 2005), en voortbouwend op bevindingen in gedragsmatig onderzoek naar stemverwerking (Papcun et al., 1989; Bruckert et al., 2010; Mullennix et al., 2011; Latinus and Belin, 2011), laat dit onderzoek als eerste zien dat er organisatie bestaat op basis van specifiekheid bij neurale representaties in stemselectieve corticale regionen.

Conclusies

De resultaten die gepresenteerd zijn in dit proefschrift verschaffen nieuwe inzichten over een aantal aspecten van het aanleren van stemidentiteit en over de verwerking van auditieve objecten in het algemeen. Er kunnen een aantal belangrijke conclusies worden getrokken over het aanpassingsvermogen van stemrepresentaties en over de verschillende typen en niveaus van abstractie in de verwerking van stemidentiteit. Deze punten zullen één voor één besproken worden in de volgende secties.

Aanpassingsvermogen in het leren over stemidentiteit

Dit proefschrift onderzocht de vorming van categorieën voor stemidentiteit. Een serie gedragsmatige en neuroimaging experimenten liet zien dat de codering van stemidentiteit adaptief is, net als de codering voor gezichten (Rhodes and Jeffery, 2006). Aanpassingsvermogen betekent hier een gereedheid tot verandering en robuustheid in een veranderende omgeving. In deze sectie zullen een aantal algemene conclusies worden getrokken die zijn gebaseerd op het bewijs dat hier is gepresenteerd over flexibiliteit en stabiliteit in het aanleren van stemidentiteit.

Stemidentiteitcategorieën worden snel aangeleerd (Hoofdstuk 3), zelfs impliciet (Hoofdstuk 4), en dit in tegenstelling tot fonetische categorieën. Neurale activatiepatronen lieten ook bewijs zien voor impliciete formatie van prototypen voor supra-individuele representatieruimtes (Hoofdstuk 5 en 6). Stemidentiteiten worden bovendien snel herleerd na een verschuiving van categorie grenzen, net als bij fonetische categorieën (Norris et al., 2003; Hoofdstuk 3). Dit type herleren wordt ondersteund door plasticiteit in de neurale codering, zoals werd aangetoond door de dynamische aanpassingen van corticale activiteitspatronen bij veranderingen in specifiekheid van stemidentiteit (Hoofdstuk 5 en 6). Ankerpunten, zoals de centra van categorieën en de grenzen van categorieën van stemidentiteiten lijken dus niet te worden bepaald door nonlineariteiten in het spraaksignaal. Als stemidentiteiten werden bepaald door nonlineariteiten in het spraaksignaal, dan zou er geen dergelijke plasticiteit moeten zijn.

Er zijn ook dynamische veranderingen voor de hoeveelheid aan variatie die geaccepteerd wordt voor een bepaalde spreker. Het gebied waarbinnen stemmen

acceptabel waren bleek kleiner voor informatiebronnen die gebaseerd zijn op distinctieve segmenten. Dat wil zeggen, voor segmenten met lagere binnen- en hogere tussen-stem variatie (Hoofdstuk 2 en 4). Het is mogelijk dat deze verhoogde gevoeligheid voor variatie er toe heeft geleid dat het leren van stemidentiteit efficiënter was wanneer het gebaseerd was op deze distinctievere segmenten (Hoofdstuk 4). Gebieden waarbinnen een stem een acceptabel exemplaar was verschillen ook tussen sprekers: sommige luisteraars zijn conservatievere en andere juist liberalere stemwaarnemers (Hoofdstuk 2). Dergelijk conservatisme lijkt ook veranderen over de tijd: het gebied waarbinnen getrainde stemidentiteiten acceptabel zijn wordt kleiner, zelfs na een korte pauze (Hoofdstuk 3).

Echter, er zijn ook grenzen aan de flexibiliteit in stemverwerking, en deze zijn vooral zichtbaar in de grenzen aan de grootte van categorieën. Luisteraars hebben ingebouwde grenzen wat betreft de grootte van categorieën voor een individu, zowel voor gezichten (Cabeza et al., 1999) als voor stemmen. Te grote individuele categorieën, waarbij de binnen-stem veranderingen groter zijn dan normale intraindividuele variatie, worden niet aangeleerd, ondanks expliciete training (Hoofdstuk 3). Dit betekent niet dat slechts individuele stemcategorieën worden opgeslagen in het brein: supra-individuele stemruimtes worden ook gerepresenteerd (Hoofdstukken 2, 5 en 6). Echter, in overeenstemming met de claims wat betreft de beperkingen aan categoriegrootte, worden er geen categorieën gevormd voor stemfamiliecategorieën (Hoofdstuk 4). De mechanismen voor stemverwerking hebben mogelijk een voorkeur voor stemrepresentatieruimtes die overeenkomen met grootte die over het algemeen functioneel is, zoals de grootte van die van individuen (rond het prototype van bijvoorbeeld de stem van Bob), of de grootte die past bij een diersoort (rond een prototype van alle mensen; ofwel, stemselectieve hersenregionen). Dit in tegenstelling tot functioneel minder relevante categorieën zoals een twee-personenruimte (bijvoorbeeld behorend tot één familie).

Ondanks al deze flexibiliteit zijn stemidentiteitrepresentaties relatief stabiel over tijd (Hoofdstuk 3). Meerdere informatiebronnen voor persoonidentiteit worden gebruikt, waaronder segmentspecifieke informatiebronnen, waardoor stemverwerking minder fragiel is bij onverwachte variatie. Verschillende persoonidentiteitinformatiebronnen worden anders beïnvloed in verschillende situaties. Een verkoudheid beïnvloedt bijvoorbeeld met name nasale spraakgeluiden, terwijl het nadoen van een andere persoon vaak verstoring oplevert in nonsegmentele informatiebronnen (Eriksson and Wretling, 1997). Verder zijn

stemidentiteit representaties relatief stabiel over verschillende luisteraars: the waargenomen specificiteit van een stem is niet afhankelijk van de luisteraar (Hoofdstuk 2). Dit betekent echter niet dat luisteraars een ingebouwde prototypestem hebben, maar dat luisteraars met vergelijkbare percentuele ervaringen ook vergelijkbare representatieruimtes creëren. Er lijkt dus weinig verschil te zijn tussen welke informatiebronnen luisteraars gebruiken en hoe ze deze gebruiken.

Meerdere niveaus van abstractie in stemherkenning

Abstractie is een fundamenteel concept in menselijke perceptie, maar ook een concept dat onderzoekers met verschillende betekenissen gebruiken, waarbij gedoeld wordt op verschillende belangrijke fenomenen in informatieverwerking. Abstractie kan refereren aan het in- of uitzoomen om relevante informatie op te nemen uit het signaal, aan het berekenen van een gemiddelde over een distributie van waarden, en aan de voortschrijding van informatie in de verwerkingshiërarchie. Hier beredeneer ik dat de bevindingen in dit proefschrift meerdere niveaus van abstractie laten zien, voor elk van de drie betekenissen die hierboven beschreven staan.

De eerste betekenis van abstractie refereert aan schalen, dit gaat door op het idee dat de representatieruimtes die gebaseerd zijn op gelijkenis die we gebruiken in objectverwerking (zie Valentini, 1991) kunnen variëren in de specificiteit van informatiebronnen, gevoeligheid, tijdsduur en grootte. In die zin refereert een meer abstracte representatie aan een ruimte met minder specifieke informatiebronnen, verminderde gevoeligheid voor variatie, grotere tijdsduur, of een ruimte die groter is. De experimenten die hier gepresenteerd zijn verschaffen bewijs voor stemidentiteitverwerking op verschillende schaal, voor elk van deze karakteristieken. Ik zal deze hieronder één voor één bespreken.

Van representatieruimtes die gebaseerd zijn op gelijkenis wordt aangenomen dat ze verschillen in welke informatiebronnen ze gebruiken. De studies die hier werden gepresenteerd lieten zien dat de perceptuele gelijkenis van stemmen goed beschreven kan worden door spectrale informatiebronnen (F0, F1, F2; Hoofdstuk 2), maar dat zowel segmentgebaseerde als meer abstracte, niet-segmentele informatiebronnen, betrokken zijn bij het leren van stemidentiteit. Segmentspecifieke informatiebronnen waren niet token-specifiek, dus ook hierbij speelde een zekere abstractie een rol (Hoofdstuk 3). Verdere

indicaties van de verschillen in specificiteit van informatiebronnen werden met fMRI gevonden. De rechter anteriore temporale pole leek betrokken te zijn in een modaliteitspecifieke representatie van stemidentiteit, terwijl van de diepe posteriore STS gesuggereerd is dat ze modaliteit non-specifieke persoonlijke identiteitrepresentaties bevat (Hoofdstuk 5; Campanella and Belin, 2007).

In de fMRI studies werden ook verschillen in de gevoeligheid voor bepaalde veranderingen gevonden. De stemselectieve bilaterale STS was gevoelig voor grovere maar niet erg verfijnde akoestische veranderingen (Hoofdstuk 6). Van het detecteren van kleine akoestische verschillen wordt aangenomen dat ze plaatsvinden in de primaire auditieve cortex, een gebied dat niet gespecialiseerd is in de verwerking van stemmen (Belin et al., 2000).

Een ander variabel aspect van representatieruimtes behelst hun tijdsspanne. Er bleek dat stemselectieve hersenregionen zowel korte- als lange-termijnrepresentatieruimtes behelsden. Korte-termijnruimtes bleken sensitief voor de gelijkenis tussen een stemtoken, en het token dat er vlak voor werd gehoord. Lange-termijnruimtes echter, bleken sensitief voor de gelijkenis tussen een stemtoken en de gemiddelde waarde van een serie daarvoor gehoorde tokens (Hoofdstuk 5 en 6). Deze verschillende tijdsspannen leken in relatie te staan tot twee verschillende soorten van fMRI repetitie-onderdrukking (zie Epstein et al., 2008).

Representatieruimtes bleken ook te variëren in grootte. Er was ook bewijs voor grote ruimtes waarin stemtokens werden gerepresenteerd die correspondeerden met verschillende stemidentiteiten (Hoofdstukken 2, 5 en 6), en voor ruimtes met krappere acceptatiegebieden die niet groter waren dan intra-individuele variatie (Hoofdstukken 3, 4, 5 en 6). Er lijken dus stemruimtes te bestaan die vele of zelfs alle menselijke vocalisaties kunnen encoderen in een enkele categorie, en verdere stemruimtes voor elke aparte spreker. Bovendien, intra-individuele representatieruimtes varieerden met fonetische inhoud: bijvoorbeeld, de stemidentiteitsruimtes gebaseerd op het woord /lot/ hadden bredere acceptatiegebieden dan de ruimtes gebaseerd op /mes/ (Hoofdstuk 4). Deze ruimtes lijken aan te sluiten bij natuurlijke variatie: het was inderdaad zo dat de fonemen die hoorden bij woorden met krappere acceptatie gebieden relatief ook minder binnen-spreker variatie, en meer tussen-spreker variatie hadden (Hoofdstuk 2).

De tweede betekenis van abstractie refereert aan middeling. Er wordt beweerd dat representatieruimtes gebaseerd op gelijkenis georganiseerd zijn rond normen. Dit wordt normgebaseerde codering genoemd (zie Valentine, 1991). In die zin refereert abstractie aan het formeren van de norm door het berekenen van het gemiddelde van de waarden in die specifieke ruimte. Dit op abstractie gebaseerde model van object verwerking wordt weersproken door exemplaargebaseerde modellen. In dit theoretische contrast zijn exemplaren de representaties van individuele geobserveerde stimuli, terwijl de norm gemiddelde, berekende waarden zijn. Zoals hieronder zal worden besproken presenteert dit proefschrift bewijs voor normgebaseerde codering van stem identiteiten, en voor de verschillende codering tussen meer centrale en meer perifere waarden in neurale stemruimtes.

Het eerste stuk bewijs voor organisatie op basis van specifiekheid van spreker identiteit was dat de stemmen die moeilijk te onderscheiden zijn van andere stemmen voor alle luisteraars, ook diegene zijn voor welke verschillende tokens minder snel worden geaccepteerd als tokens van dezelfde stem, terwijl binnen-stem akoestische variabiliteit niet groter was (Hoofdstuk 2). Er is beargumenteerd dat krappere acceptatieruimtes rond minder distinctieve, dicht bij het gemiddelde, exemplaren een indicatie zijn van organisatie op basis van prototypen (e.g., Kuhl, 1991, Loffler et al., 2005). Voordelen in categorisatie van stemgroepen werden gevonden voor stimuli rond individuele stemcategoriecentra vergeleken met stimuli die ver van deze centra aflagen, ondanks het feit dat er geen expliciete training was voor die identiteiten (Hoofdstuk 4). Ik heb ook beargumenteerd dat ruimtes op basis van specifiekheid die hier werden aangetoond niet georganiseerd waren rond akoestische ankerpunten, maar rond gemiddelden relatief tot de eigenlijke stemruimte: stemidentiteitgemiddelden bleken de getrainde categorieverschuivingen dynamisch te volgen (Hoofdstuk 3). Dit werd verder onderbouwd door de fMRI experimenten, waarin het meest overtuigende bewijs voor normgebaseerde codering gevonden werd. Vergeleken met atypische stemmen werd een toename in neurale specificiteit gevonden voor typische exemplaren van stemmen. Deze veranderingen in neurale activiteit konden niet verklaard worden door akoestische veranderingen van het stemsignaal, maar alleen door veranderingen in de waargenomen typischheid. Dit demonstreert dat de ankerpunten van de neurale ruimtes waarin stemidentiteiten worden gerepresenteerd, geen absolute waarden bevatten maar in plaats daarvan hun positie snel

aanpassen aan het nieuwe perceptuele bewijs. Om dit te bewerkstelligen moest de stemidentiteitnorm berekend worden en er moest een bijzondere status aan worden toegekend, precies zoals werd voorgesteld door normgebaseerde maar niet-exemplaargebaseerde modellen van codering (Valentine, 1991; Jeffery et al., 2011; Hoofdstuk 5 en 6).

Stemverwerking lijkt dus abstractie te behelzen in termen van schaal en middeling. Samengenomen suggereert dit dat meerdere normgebaseerde representatieruimtes bestaan voor stemmen, elk met hun eigen norm. Als gevolg hiervan moeten we bijvoorbeeld segmentspecifieke normen hebben voor stemidentiteiten, of ten minste specifieke normen voor elke relevante informatiebron die aanwezig is in een subset van de segmenten. Dit werd geïllustreerd door de blijkbaar segmentgebaseerde organisatie van een afstandskaart van verschillende woorden, berekend op basis van hoe vergelijkbaar de bijdrage van elk woord was aan de mate waarin een stem typisch was (Hoofdstuk 2), en door woord effecten in de stemidentiteit training studies (Hoofdstuk 3).

De derde betekenis van abstractie behelst de voortgang van informatie door de verwerkingshiërarchie. Het wordt gebruikt in relatie tot hiërarchische modellen van object perceptie (e.g., Bruce and Young, 1986; Belin et al., 2004, 2011) welke stellen dat verwerkingsniveaus serieel georganiseerd zijn. Een meer abstract niveau van verwerking betekent in dit geval een hoger stadium in de verwerkingshiërarchie. De experimenten in dit proefschrift beschreven meerdere stadia van verwerking in de corticale hiërarchie voor stemmen: zoals hieronder beschreven, werden functioneel en anatomisch verschillende niveaus van verwerking gevonden.

Zoals we al eerder zagen zijn er meerdere normgebaseerde representatieruimtes die een rol spelen in stemverwerking. De belangrijke bijdrage van neuroimaging is hierin de bevinding dat deze verschillende ruimtes geïmplementeerd zijn in anatomisch verschillende locaties in het menselijke brein. Regio's die gevoelig bleken voor lange-termijn centraalheid werden gevonden in de middelste en posteriore delen van de STS, terwijl structuren die gevoelig waren voor identiteit centraalheid juist gevonden werden in de anteriore temporale regio's (ATP, Hoofdstuk 5), en in stemselectieve inferiore frontale regio's (IFC, Hoofdstuk 6). Over deze stemselectieve regio's is gesuggereerd dat ze onderdeel uitmaken van de auditieve "wat" route (Belin et al., 2004; Ahveninen et al., 2006), waarbij de STS directe en sterke structurele verbindingen omlaag richting de primaire auditieve cortex zou

hebben (Kumar et al., 2007) en omhoog naar zowel de anteriore temporele en inferiore frontale regio's (Ethofer et al., 2012). Deze verschillende typen van neurale gevoeligheid op anatomisch verschillende locaties lijken dus corticale verschijningsvormen te zijn van specifieke niveaus in een stemverwerking hiërarchie.

Abstractie lijkt dus aanwezig op meerdere niveaus en op meerdere manieren in stemherkenning. Er is beargumenteerd dat verschillende niveaus van de stemverwerking hiërarchie verantwoordelijk zijn voor akoestische en identiteitverwerking (Belin et al., 2004, 2011). Maar in dit proefschrift suggereer ik ook dat gevoeligheid voor identiteit, ook al zijn ze functioneel en anatomisch verschillend, geïmplementeerd kunnen zijn door een enkel coderingsmechanisme voor gelijkenisgebaseerde representatieruimtes die alleen verschillen in de grootte van de ruimte (d.w.z., een grote supra-individuele ruimte, en krappere intra-individuele ruimten). Vanuit een breder perspectief, een structuur die meerdere niveaus behelst, hoeft niet per definitie ook ingewikkelde mechanismen te gebruiken. Fractals in de wiskunde zijn bekende voorbeelden van complexe structuren die gecreëerd worden met behulp van simpele regels. Het is hier belangrijk dat deze simpele regels steeds opnieuw worden gebruikt voor verschillende onderdelen van het geheel. Abstractie op verschillende niveaus zou de manier kunnen zijn om een complexe architectuur op te bouwen op basis van een kleine set simpele regels. Ik heb beargumenteerd dat dit het geval lijkt te zijn voor de verwerking van stemmen in het menselijke brein.

In dit proefschrift heb ik laten zien dat de herkenning van personen op basis van hun stem gebruik maakt van meerdere (segmentele en niet-segmentele) informatiebronnen, en dat deze informatiebronnen samen bijdragen aan de waargenomen typischheid van een stem op specifieke manieren. Stemidentiteiten bleken natuurlijke auditieve objecten in het spraaksignaal, met ingebouwde aannames over wat een mogelijke individuele stemcategorie kan zijn. Sprekeridentiteiten bleken gerepresenteerd te zijn door middel van meerdere, flexibele normgebaseerde neurale codes, op functioneel en anatomisch verschillende hiërarchisch georganiseerde niveaus in het menselijke brein. Deze niveaus behelsden een supra-individuele stemruimte in de stemselectieve regio's van de temporele sulcus, en intra-individuele stemruimtes in anteriore temporele en inferiore frontale regio's van de rechter hersenhelft.

Összefoglalás és következtetések[2]

[2] A fejezethez tartozó irodalomjegyzéket lásd a 177-178. oldalon.

Összefoglalás

Alapvető szociális készség, hogy felismerjünk valakit a beszéde alapján. E disszertáció célja az, hogy hozzájáruljon a beszélőhangi identitások tanulásának és a beszélőhang-reprezentációk perceptuális és neurális szerveződési elveinek jobb megértéséhez.

Viselkedéses kísérletek

A 2. fejezetben bemutatott kísérlet a beszélőhangi diszkriminálhatóság szegmentális összetevőit és a beszélőhangok perceptuális és akusztikus hasonlósága közti kapcsolatot vizsgálta. A résztvevők egy szófolyamot hallottak (különböző holland szavak néhány változatát több férfi beszélőtől), és szavanként el kellett dönteniük, hogy a szót kiejtő személy azonos vagy különböző attól, aki a megelőző szót mondta. Az eredmények alapján az alanyok nagyon jók a beszélőhang-diszkriminációban, de jelentősen eltérnek abban, hogy mit észlelnek beszélőn belüli, illetve beszélők közti változásnak. A beszélőhang-diszkriminációs teljesítmény nem volt független a szegmentális tartalomtól: a szótagkezdet, mag és zárlat pozíciókban /m/, /e/ és /s/ fonémákat tartalmazó szavak rendre jobban támogatták a beszélőhang-diszkriminációt, mint az azonos pozíciókban /l/, /o/ és /t/ fonémákat tartalmazó szavak. Ezek a szegmentális előnyök relatíve alacsonyabb beszélőn belüli és magasabb beszélők közti akusztikai variabilitással jártak a disztinktívebb szegmensekre nézve – pontosan ez tette az ezekben a szegmensekben jelenlévő ismertetőjegyeket személyazonosításra jól használható ismertetőjegyekké. Az alanyok a szavak mindhárom szegmentális poziciójában lévő információt gyorsan tudták használni. Továbbá, egyetértés volt az alanyok közt abban, hogy mely beszélőhangok diszkriminálhatóak jobban és melyek kevésbé. A kevésbé diszkriminálható beszélőhangok egyúttal kevésbé is voltak azonosíthatóak, az alacsonyabb beszélőn belüli variabilitás ellenére, ez pedig azt a nézetet támogatta, hogy a beszélőhangok tipikussági alapon szerveződnek. A beszélőhangok eloszlása egy perceptuális diszkriminálhatóságon alapuló, valamint egy formáns-alapú akusztikai dimenziók mentén tekintett távolságtérképen nagy hasonlóságot mutatott, azt sugallva, hogy a beszélőhang-tipikusság viszonylag jól leírható egyszerű spektrális ismertetőjegyekkel. Egy másik távolságtérkép az egyes szavakat a beszélőhang-tipikussághoz való hozzájárulásuk hasonlósága alapján pozícionálta. E térkép

szegmens-alapú szerveződése szegmens-specifikus beszélőhang-prototípusok jelenlétére utalt.

A 3. fejezet két többüléses tréningkísérletet írt le, melyek a beszélőhangi identitás tanulásának flexibilitását és specificitását vizsgálták. Ingerként a "mes" [kés] és "lot" [sors] szavakat kiejtő két kiválasztott beszélő közti beszélőhangmorf-kontinuumok szerepeltek, a beszélőhangi identitáskategóriáról adott, szisztematikusan változtatott visszajelzés mellett, egy ülésközi és kísérletközi tanulási-újratanulási paradigmában. Az első kísérletben a résztvevők a kétnapos tréning során beszélőhangokat kategorizáltak egy "A személy vagy B személy" típusú feladatban, de nem tudtak a beszélőhangi identitások közti határ két nap közötti, visszajelzésekbe épített eltolásáról. Az eredmények alapján az új beszélőhangi identitáskategóriák tanulása gyors és a tanultak egy nappal később is stabilak. Az alanyok rugalmasan megtanulták és újratanulták a mesterségesen definiált beszélőhangi identitáshatárokat, de a rugalmasságuknak is volt határa: nem tolerálták teljes mértékben azokat az asszimmetrikus kategória-visszajelzéseket, amelyek túlméretezett beszélőhangi identitáskategóriához vezettek. Ez azt mutatta, hogy az alanyoknak beépített elvárásai voltak az egyéni beszélőhang-kategóriák elfogadási tartományát illetően. A beszélőhangokról szerzett tudás nagy része generalizálódott a nem-tréningezett szavakra is, akár volt szegmentális átfedés a tréningezett szóval, akár nem, de a beszélőhangi identitások centrumaira vonatkozóan a transzfer nem volt teljes. Továbbá, jobb volt a teljesítmény egy, a tréningezett szóval szegmentálisan átfedő, de nem tréningezett szóra, mint egy, a tréningezett szótól szegmentálisan független szóra. A beszélőhang-kategorizációs válaszokban talált szóhatás szegmens-specifikus reprezentációk jelenlétére is utalt, és arra, hogy a beszélőn belüli variabilitás elfogadási tartománya szegmens-specifikus. Ezek az eredmények azt demonstrálták, hogy a beszélőhangokról szerzett tudás absztrakcióval jár, valamint, hogy a beszélőhangi identitás feldolgozása során a nemszegmentális és a szegmentális ismertetőjegyek egyaránt szerephez jutnak.

A 3. fejezet második kísérletében a résztvevők a tréning során ugyanannak a beszélőhangmorf-kontinuumnak az ingereit kategorizálták, mint az első kísérletben, de most egy "A személy vagy nem A személy" feladatban. Az A személy kategória a tréning alapján a két természetes beszélőhang között volt. Vagyis ami kategóriahatár volt az első kísérletben, kategóriacentrummá vált a második kísérletben. Itt a kategória pozíciója szintén változott az alanyok közt. Az eredmények azt mutatták, hogy az alanyok könnyedén megtanulták ezeket

a beszélőhangi identitáskategóriákat. Ez azt demonstrálta, hogy a beszédjel nem tartalmaz beépített információkat a beszélőhangi identitáskategóriák szerkezetéről, a morfolás pedig nem rontott az ingerek természetes hangzásán. Az első kísérlethez hasonlóan, a beszélőhangokról a tréning során szerzett tudás egy része generalizálódott egy nem tréningezett szóra is, de a kategorizáció pontossága jelentősen romlott, megerősítve, hogy a nemszegmentális és a szegmentális információknak is van szerepük. Végül, kevesebb "A személy" döntés született egy rövid késleltetést követően, mint a tréninget késleltetés nélkül követő tesztben, azt sugallva, hogy a beszélőhangi identitások elfogadási tartománya szűkülhet a megerősítés nélkül telő idő alatt.

A 4. fejezet egy beszélőhang-tanulási kísérletet mutatott be, ami a beszélőhangi kategóriaalkotás korlátait tesztelte. Az alanyoknak a tréning során egyéni beszélőhangok csoportjaira ('családok') kellett kategóriákat alkotniuk. A tréningezett kategórián belüli variabilitás így nagyobb volt, mint a tipikus, beszélőn belüli akusztikus variabilitás. Ezután a résztvevők mind családon belüli, mind családok közti beszélőhang-morf ingereket hallottak, és arra kellett válaszolniuk, hogy hallották-e már korábban az adott beszélőhangot, illetve, hogy melyik családhoz tartozhat a beszélő. A predikció az volt, hogy a beszélőhang-családokra vonatkozó prototípusképzés a családon belüli morfokat előnyhöz juttatja a családok közti morfokkal szemben, míg az egyéni beszélőhangokra vonatkozó prototípusképzés a beszélőhang-kontinuumok végpontjait juttatja előnyhöz a morfokkal szemben. Valóban volt különbség a végpontok javára a morfokkal szemben mind a kategorizációs válaszokban, mind a felismerési bizonyosságban, de nem volt különbség a családon belüli és családok közti morfok között, és ez arra utalt, hogy miközben az egyéni beszélőhangokra vonatkozó prototípusok megalkotása még impliciten is könnyedén végbemegy, a beszélőhang-családokra vonatkozó prototípusok nem alakulnak ki, még explicit visszajelzések ellenére sem, még úgy sem, hogy a családra vonatkozó kategóriák tanulása eközben sikeresen megtörténik. Ez az eredmény egy beépített kategóriaméret-megkötés létezését demonstrálja a beszélőhangokra vonatkozó prototípusképzésben, hasonlóan az arcokra találtakhoz (Cabeza és mtsai, 1999). Az eredmények azt is megmutatták, hogy a /lot/ szót kiejtő beszélőhangok könnyedébben felismerhetőek már hallott beszélőhangokként, mint a /mes/ szót kiejtők. Ez egy további demonstrációja a szegmens-specifikus elfogadási tartományoknak. Továbbá, a családi kategorizáció bizonyossága jobban növekedett a tréningezés mértékével a /mes/, mint a /lot/ esetében. A

2. fejezet eredményeivel együtt, melyek szerint a /mes/ szó fonémái disztinktívebbek, mint a /lot/-éi, mindezek arra utalnak, hogy a fonetikai tartalom modulálja a beszélőhangokra vonatkozó kategóriaképzést, mégpedig úgy, hogy a disztinktívebb fonémákat tartalmazó szavak jobban támogatják a beszélőhang-tanulást, de szenzitívebbé teszik a változásokra a beszélőhang-kategóriákra vonatkozó reprezentációkat.

Agyi képalkotásos kísérletek

Az 5. fejezet egy többüléses tréningvizsgálatot mutatott be, amely a beszélőhang-felismerés neurális korrelátumait vizsgálta funkcionális mágneses rezonanciás képalkotás (fMRI) segítségével. A tréning során magyar résztvevők kategorizálták a "bú", "fű", "ki", "lé", "ma" és "se" szavakat kiejtő két beszélő közti beszélőhangmorf-kontinuum ingereit. A 3. fejezet második kísérletéhez hasonlóan a tréningezett kategória a kontinuum közepén helyezkedett el, és az alanyok egy “A személy vagy nem A személy” tréningfeladatot kaptak. A 3. fejezet első kísérletéhez hasonlóan – az ingerek észlelt beszélőhangi kategóriaszerkezetére vonatkozó tulajdonságok (úgy, mint kategórián belüli, kategóriahatár, kategórián kívüli) alanyon belüli és tesztek közötti manipulálására – a visszajelzések eltérő kategóriahatár-pozíciókat definiáltak az egyes napokon. Ebben a kísérletben egyhetes szünet volt a két fMRI teszt között, és mindkét tesztet kétnapos, intenzív tréning előzte meg. Az fMRI tesztek során a résztvevők szóingerek egy sorozatát hallották, és vagy egy beszélőhang-felismerési, vagy egy szóismétlés-detekciós feladatot kellett végezniük. Egy gyors ritmusú, ritkás szkennelési szekvencia került alkalmazásra, ami egyesítette a közel folyamatos adatgyűjtés és a csendben történő ingerbemutatás előnyeit. Lényeges, hogy az alanyok megtanulták a tréningezett kategóriát, és a kategóriahatár tréningezett eltérése a két hét közt megfigyelhető volt a tesztek során is. Az aktuális és az azt megelőző inger közti kapcsolat figyelembevételével el lehetett különíteni a rövid távú, akusztikai hasonlóságra való érzékenység hatásait (a kétoldali középső és poszterior szuperior temporális szulkusz (STS) és jobboldali inferior frontális kérgi (IFC) területeken) a hosszú távon tárolt tipikus értékekre vonatkozó neurális élesedés hatásaitól. Továbbá, az elemzések két anatómiailag elkülönülő típusát fedték fel a tipikussági alapú, hosszú távú beszélőhang-reprezentációknak: az egyiket egy beszélőhang-akusztikai térben (centrális vs perifériás; jobboldali orbitális / inzuláris kéreg, jobboldali poszterior mediális STS), a másikat pedig egy beszélőhang-identitási térben (identitáson belüli vs identitáson kívüli; kétoldali anterior

temporális pólus, baloldali mélyen poszterior STS, baloldali amygdala). Ez a tanulmány elsőként mutatott agyi képalkotásos bizonyítékokat flexibilis 'átlag-beszélőhang' reprezentációk létezésére, demonstrálva ezzel a neurális beszélőhang-terek norma-alapú szerveződését. A beszélőhangi identitás kategorizációjára vonatkozó teljesítmény korrelált a beszélőhangi identitások hasonlóságára való neurális érzékenységgel (jobboldali középső és poszterior STS, baloldali mélyen poszterior STS, jobboldali anterior temporális pólus, baloldali amygdala): a nagyobb neurális érzékenységű alanyok jobbak voltak a beszélőhang-felismerésben. Ez az eredmény demonstrálta a beszélőhang-identitások norma-alapú neurális reprezentációinak direkt viselkedéses relevanciáját. Az eredmények alapján ezeket a neurális mintázatokat nem modulálta a döntés nehézsége. Mindazonáltal az eredmények nem mutattak hasonlóságra való neurális érzékenységet akkor, amikor az alanyok egy másféle (szóismétlés-detekciós) feladatot kaptak, amely elvonta a figyelmüket a beszélőhang-identitásoktól. Ez azt jelezte, hogy a figyelem is modulálhatta az fMRI-vel mért ismétléses szuppressziós hatásokat a beszélőhangok esetében.

Végül, a 6. fejezet bemutatott egy második többüléses fMRI tanulmányt, amely a beszélőhangokra szelektíven érzékeny kérgi területeken vizsgálta a beszélőhang-identitások neurális kódolását. Ebben a vizsgálatban egy házilag módosított ritkás szkennelési protokoll került alkalmazásra. Az 5. fejezetben bemutatott kísérlethez hasonlóan ez a vizsgálat is a tanulási-újratanulási paradigmára építve alanyon belül manipulálta külön-külön a beszélők közti és beszélőn belüli tipikussági mintázatokat (két fMRI teszt egyhetes szünettel, mindkettő előtt kétnapos intenzív tréning). De itt, a 3. fejezet első kísérletéhez hasonlóan az ingereket holland beszélők által kiejtett /mes/ szavak adták, és a résztvevőknek egy "A személy vagy B személy" feladatot kellett végezniük. Az fMRI tesztek során az alanyok szópárokat hallottak. Feladatuk a szópár második tagjának beszélőhang-kategorizációja volt. Az eredmények azt mutatták, hogy az alanyok megtanulták a tréningezett kategóriákat, a tréningezett kategóriahatár-változás pedig az fMRI tesztek alatt is mérhető maradt. A beszélőhangokra szelektíven érzékeny területek meghatározása egy funkcionális lokalizációs teszt (Belin és mtsai, 2000) során történt, a kapott területek a kétoldali STS-ben és a (jobboldali lateralizációjú) IFC-ben voltak. Az elemzések a beszélőhangok feldolgozási hierarchiájának két anatómiailag elkülönülő szintjét fedték fel, melyek mindegyike hosszú távon kódol átlag-beszélőhangokat: egy szupraindividuális szintet, ami egy akusztikus átlag-beszélőhangot kódol (centrális vs perifériás; jobboldali STS), és egy intraindividuális szintet,

ami az egyes beszélőhangi identitások átlagos értékét kódolja (tipikus vs atipikus; jobboldali IFC). Érdekes módon e két, beszélőhangokra szelektíven érzékeny kérgi terület egyazon teszt kétféle irányításával is azonosítható volt: a centrális-atipikus < perifériás-tipikus kontrasztból a jobboldali STS, a centrális-atipikus > perifériás-tipikus kontrasztból pedig a jobboldali IFC adódott. A további elemzések megerősítették, hogy ezeket az eredményeket nem a döntési nehézség változásai okozták. Továbbá a kétoldali STS rövid távú, hasonlóságra való érzékenységet mutatott a nagyobb akusztikai változások esetében, de a kisebbekében nem. Ez a rövid távú érzékenység jelen volt a centrális ingerekre, de a perifériásakra nem, azt jelezve, hogy egy inger hosszú távú akusztikai centralitása hatással lehet a rövid távú feldolgozásra is. A beszélőhang-feldolgozás viselkedéses vizsgálataiból származó friss eredményekre építve (Papcun és mtsai, 1989; Bruckert és mtsai, 2010; Mullennix és mtsai, 2011; Latinus és Belin, 2011), és az arcfeldolgozásos eredményekkel összhangban (Loffler és mtsai, 2005), a jelen tanulmány talált először evidenciát a neurális beszélőhang-reprezentációk tipikussági alapú szerveződésére a beszélőhangokra szelektíven érzékeny kérgi területeken.

Következtetések

A jelen disszertációban bemutatott kísérletek új megvilágításba helyezték a beszélőhangi identitás tanulásának, és általánosságban a hallási tárgyak feldolgozásának számos aspektusát. Fontos következtetések vonhatók le a beszélőhang-reprezentációk adaptivitásáról, és a beszélői identitás feldolgozása során szerepet játszó absztrakció formáiról és szintjeiről. Az alábbiakban ezek áttekintése következik.

Adaptivitás a beszélőhangi identitás tanulásában

E disszertáció a beszélőhangi identitásokra vonatkozó kategóriaalkotás természetét vizsgálta. Egy viselkedéses és agyi képalkotásos kísérletsorozat demonstrálta, hogy a beszélőhangi identitások kódolása adaptív, hasonlóan az arcoknál találtakhoz (Rhodes és Jeffery, 2006). Az adaptivitás egyfelől változásra való készséget jelent, másfelől robosztusságot a változó környezetben. Ez az alpont néhány általános következtetést von le a beszélőhangi identitás tanulásának flexibilitására és stabilitására vonatkozó, itt bemutatott evidenciák alapján.

A beszélőhangi identitáskategóriákat, szemben a fonetikai kategóriákkal felnőttek esetében (Logan és mtsai, 1991), gyorsan megtanuljuk (3. fejezet), még implicit módon is (4. fejezet). Neurális válaszmintázatok is igazolták az implicit prototípusalkotás tényét szupraindividuális reprezentációs terek esetében (5. és 6. fejezet). Továbbá, a beszélőhangi identitások gyorsan újratanulhatók egy kategória-eltolódás után, hasonlóan a fonetikai kategóriákhoz (Norris és mtsai, 2003; 3. fejezet). Ezt az újratanulást a neurális kódolás plaszticitása is támogatja, ahogy azt a beszélőhangi identitások tipikusságának változásaira adott, dinamikus módosuló agykérgi válaszmintázatok is demonstrálták (5. és 6. fejezet). Úgy tűnik tehát, hogy az olyan horgonypontokat, mint a beszélőhangi identitások esetében a kategóriacentrumok és kategóriahatárok, nem a beszédjel nonlinearitásai határozzák meg. Ha a beszélőhangi identitásokat nonlinearitások határoznák meg, akkor nem lenne helye az efféle plaszticitásnak.

Dinamikusan változik az is, hogy egy adott beszélő esetében milyen mértékű variabilitás tolerálható. A beszélőhangi identitásokra vonatkozó elfogadási tartományok szűkebbek azoknak az ismertetőjegyeknek az esetében, melyek disztinktívebb szegmensekhez vagy szavakhoz kötődnek, azaz, amelyeknél kisebb a beszélőn belüli és

nagyobb a beszélők közti variabilitás (2. és 4. fejezet). Lehetséges, hogy éppen ez a változásra való megnövekedett érzékenység teszi hatékonyabbá a disztinktívebb szegmensekre épülő beszélőhangi identitás-tanulást (4. fejezet). Az elfogadási tartományok szintén egyénenként változóak: vannak konzervatívabb és liberálisabb beszélőhang-észlelők (2. fejezet). Ez a konzervativizmus azonban változik az idő múlásával: a tréningezett beszélőhangi identitások elfogadási tartományai már egy rövid késleltetéstől is szűkebbé válnak (3 fejezet).

De a beszélőhang-feldolgozás flexibilitásának korlátai is vannak, amelyek a kategóriaméretre vonatkozó megkötések esetén a legszembeötlőbbek. Az embert beépített korlátok segítik abban, hogy milyen az elfogadható méretű személyidentitási kategória, arcok (Cabeza és mtsai, 1999) és beszélőhangok esetében egyaránt. Túlméretezett egyéni kategóriákat, ahol a beszélőn belüli változások túllépik a tipikus intraindividuális variabilitást, még explicit tréningezés ellenére sem tanulunk meg (3. fejezet). Ez nem jelenti azt, hogy az emberi agy csak egyéni beszélőhang-kategóriákat tárol: szupraindividuális beszélőhang-terek is reprezentálódnak (2., 5. és 6. fejezet). De, a méretkorlátozásra vonatkozó állítást erősítve, úgy tűnik, hogy nem történik prototípusalkotás a tréningezett beszélőhang-családok kategóriáira vonatkozóan (4. fejezet). Lehetséges, hogy a beszélőhangok feldolgozásáért felelős rendszer az olyan reprezentációs tereket preferálja, amelyek funkcionális szempontból a legrelevánsabb méretekkel rendelkeznek, például egy egyéni beszélőhang-tér méretével (pl. Bob hangjának a prototípusa körül; lásd még a személyidentitási csomópontok fogalmát) vagy egy fajspecifikus beszélőhang-tér méretével (pl. az összes emberi vokalizáció prototípusa körül; lásd még a beszélőhangokra szelektíven érzékeny kérgi területeket), szemben olyan funkcionális szempontból kevésbé releváns méretekkel, mint például egy kétszemélyes beszélőhang-tér mérete (pl. egy beszélőhang-család prototípusa körül).

Mindezen flexibilitás ellenére a beszélőhangi identitásokra vonatkozó reprezentációk viszonylag stabilak maradnak az idő múlásával is (3. fejezet). Számos személyidentitási ismertetőjegyet használunk, köztük szegmens-specifikus ismertetőjegyeket, ez teszi a beszélőhang-feldolgozást kevésbé sérülékennyé váratlan változások esetén is. Valóban, különböző helyzetek különböző személyidentitási ismertetőjegyekre vannak hatással. Például, ha megfázunk, az főként a nazális beszédhangjainkat érinti, míg ha egy másik személy hangját próbáljuk imitálni, az jellemzően nemszegmentális ismertetőjegyek

torzításával jár (Eriksson és Wretling, 1997). Érdekes módon a beszélőhangi identitásra vonatkozó reprezentációk egyének közt is viszonylag stabilak: a beszélőhangok észlelt tipikussága nem függ az észlelőtől (2. fejezet). Ez nem azt jelenti, hogy van az emberek fejében egy beépített beszélőhang-prototípus, inkább azt, hogy a hasonló perceptuális előtörténettel rendelkező észlelők hasonló reprezentációs tereket építenek fel. Vagyis úgy tűnik, kevés eltérés van abban, hogy az egyes észlelők milyen ismertetőjegyeket használnak és mi módon.

Absztrakciós szintek a beszélőhangok felismerésében

Az absztrakció az emberi észlelés egyik alapvető fogalma. Azonban ezt a fogalmat a kutatók több különböző értelemben használják, az információ-feldolgozás különféle kulcsjelenségeinek leírására. Az absztrakció jelentheti a ráfókuszálás-kifókuszálás folyamatát, melynek célja a releváns információ kinyerése a jelből; jelentheti egy adott eloszlás átlagának kiszámítását; és jelentheti a reprezentált információ következő szintre továbbítását a feldolgozási hierarchián belül. Amellett fogok érvelni, hogy a jelen disszertáció eredményei a beszélőhang-felismerés több absztrakciós szintjét fedték fel a szó mindhárom értelmében.

Az absztrakció első jelentése a fókuszálás. Ez azt a gondolatot bontja ki, hogy a tárgyfeldolgozás során használt, hasonlósági alapú reprezentációs terek (lásd Valentini, 1991) eltérhetnek a használt ismertetőjegyekre vonatkozó specificitásukban, szenzitivitásukban, időablakukban és méretükben is. Ebben az értelemben egy absztraktabb reprezentáció egy olyan térre vonatkozik, melyet kevésbé specifikus ismertetőjegyek, alacsonyabb szintű változásérzékenység, nagyobb időablak és nagyobb méret jellemeznek. Az itt bemutatott kísérletek ezen jellemzők mindegyikére nézve több fókuszálási szint jelenlétére találtak evidenciát a beszélőhangi identitás feldolgozásában. Az alábbiak sorra áttekintik ezeket.

Feltehető, hogy a hasonlósági alapú reprezentációs terek eltérnek egymástól abban, hogy milyen ismertetőjegyeket használnak. A bemutatott tanulmányok eredményei szerint az észlelt beszélőhangi hasonlóság jól leírható egyszerű spektrális ismertetőjegyekkel (F0, F1, F2; 2. fejezet), de szegmens-specifikus és absztraktabb, nemszegmentális ismertetőjegyek is szerephez jutnak a beszélőhangi identitások tanulása során. Azt is érdemes megjegyezni, hogy a szegmens-specifikus ismertetőjegyek sem voltak specifikusak

egy adott kiejtett változatra, vagyis már ezek használata is absztrakcióval járt (3. fejezet). Az ismertetőjegyek specificitásának különbözőségére utaló további eredményeket az fMRI vizsgálatok szolgáltatták. Ezek arra utaltak, hogy a jobboldali temporális pólus a beszélőhangi identitás modalitás-specifikus reprezentációjában, míg a mélyen poszterior STS a személyidentitás modalitás-nemspecifikus reprezentációjában vesz részt (5. fejezet; Campanella és Belin, 2007).

Az fMRI vizsgálatok demonstrálták továbbá, hogy egyes változástípusok esetén különbség van a szenzitivitásban is. A beszélőhangokra szelektíven érzékeny kétoldali STS szenzitív volt a nagyobb akusztikai változásokra, de a kisebbekre nem (6. fejezet). Más vizsgálatok szerint a kis változások detekciója az elsődleges hallókéregben történik, egy olyan területen, amely nem specializálódott a beszélőhangokra (Belin és mtsai, 2000).

A reprezentációs terek egy másik változó tulajdonsága az időablakuk. A jelen eredmények szerint a beszélőhangokra szelektíven érzékeny agyterületek egyaránt fenntartanak rövid távú és hosszú távú reprezentációs tereket. A rövid távú terek arra voltak érzékenyek, hogy mennyire hasonlít egy adott beszélőhang-inger egy azt közvetlenül megelőzően hallott ingerhez. A hosszú távú terek ezzel szemben arra voltak érzékenyek, hogy mennyire hasonlít egy adott beszélőhang-inger a megelőzőleg hallott ingersorozat centrális értékéhez (5. és 6. fejezet). Ezek a különböző időablakok úgy tűnt, hogy az fMRI-vel mért ismétléses elnyomás két különböző típusához kapcsolódnak (lásd még Epstein és mtsai, 2008).

Végül, a jelen eredmények szerint a reprezentációs terek méretükre nézve is változóak. Vannak nagyobb méretű terek, melyek több különböző beszélőhang-identitáshoz kapcsolódó beszélőhang-ingert reprezentálnak (2., 5. és 6. fejezet), és vannak kisebb elfogadási tartományú terek, amelyek nem nőnek túl az intraindividuális variabilitáson (3., 4., 5. és 6. fejezet). Úgy tűnik tehát, hogy léteznek olyan beszélőhang-terek, amelyek képesek több, vagy akár minden emberi vokalizációt egyetlen kategórián belül kódolni, és léteznek további beszélőhang-terek külön-külön az egyes beszélőkre is. Továbbá, az intraindividuális reprezentációs terek a fonetikai tartalommal változtak: például a /lot/ szón alapuló beszélőhangi identitásterek elfogadási taromány tágabbnak bizonyult, mint a /mes/ szón alapulóké (4. fejezet). Úgy tűnt, hogy e terek különbségei jól illeszkednek a természetes variabilitásban lévő különbségekhez: a kisebb beszélőhang-identitási elfogadási

tartományhoz kapcsolódó szavak fonémái viszonylag kisebb beszélőn belüli és nagyobb beszélők közti variabilitást mutattak (2. fejezet).

Az absztrakció második jelentése az átlagolás. Egy javaslat szerint a hasonlósági alapú reprezentációs terek normák köré szerveződnek. Ezt nevezzük norma alapú kódolásnak (lásd Valentini, 1991). Ebben az értelemben az absztrakció e norma megalkotását jelenti oly módon, hogy kiszámolásra kerül az adott téren belüli értékek átlaga. A tárgyfeldolgozás ezen absztrakcionista modelljét a példány alapú modellek ellenpontozzák. Ebben az elméleti kontrasztban a példányok a megfigyelt események reprezentációi, míg a normák átlagolt, számított értékek. Ahogy az alábbiakból kitűnik, a jelen disszertáció bizonyítékokat mutatott be a beszélőhangi identitások norma alapú kódolása, és a neurális beszélőhang-terek centrálisabb és perifériásabb értékeinek elkülönülő kódolása mellett.

Az első bizonyíték a beszélői identitások tipikussági alapú szerveződésére az, hogy a más beszélőhangoktól mindenki számára konzisztensen nehezen megkülönböztethető beszélőhangok éppen azok, amelyeknél kevésbé könnyedén fogadható el, hogy a különböző kiejtett változatok egyazon beszélőhanghoz tartoznak, noha ezek esetében a beszélőn belüli variabilitás nem nagyobb volt, hanem kisebb (2. fejezet). Egy javaslat szerint a kisebb elfogadási tartományok a kevésbé disztinktív, az átlagoshoz közelebbi példányok körül a prototípus alapú szerveződés indikátorai (pl. Kuhl, 1991; Loffler és mtsai, 2005). További bizonyíték, hogy jobb volt a teljesítmény beszélőhangi csoportok kategorizációjakor azoknál az ingereknél, amelyek egyéni beszélőhangi kategóriacentrumok közelében voltak, szemben az ilyen centrumoktól távoli ingerekkel, még úgy is, hogy ezek a beszélőhangi identitások nem voltak explicit módon tréningezve (4. fejezet). Amellett is érveltem, hogy azok a tipikussági alapú terek, melyek létére a jelen kísérletek rámutattak, nem akusztikailag definiált, abszolút horgonypontok köré szerveződtek, hanem az aktuális beszélőhang-térhez képest definiált átlagértékek köré: valóban, a beszélőhangi identitások átlagai dinamikusan követték a tréningezett kategória-eltolódásokat (3. fejezet). Ezt az fMRI kísérletek is alátámasztották, ezek szolgáltatták a legmeggyőzőbb bizonyítékokat a norma alapú kódolásra. Az eredmények neurális élesedést mutattak egyéni beszélőhangok tipikus példányaira az atipikus példányokal szemben. Ezeket a neurális aktivitásbeli eltéréseket nem okozhatták a beszélőhang-jelek akusztikai változásai, hanem csak az észlelt tipikusság változásai, és ez azt demonstrálta, hogy a beszélőhangi identitásokat reprezentáló neurális terek horgonypontjai nem abszolút értékek, hanem pozíciójukat gyorsan és adaptívan

módosítják a friss perceptuális evidenciák alapján. Ehhez pedig ki kellett számítani a beszélőhangi identitásokhoz kapcsolódó normákat és különleges státuszúvá kellett tenni őket, éppen ahogy a norma alapú kódolás modelljei javasolják, szemben a példány alapú kódolás modelljeivel (Valentine, 1991; Jeffery és mtsai, 2011; 5. és 6. fejezet).

A beszélőhang-feldolgozás tehát absztrakcióval jár a szó mind fókuszálási, mind átlagolási értelmében. Együttvéve ez azt jelzi, hogy számos norma alapú reprezentációs tér létezik a beszélőhangokra, s mindegyiknek külön normája van. Következésképpen, például szegmens-specifikus normáink kéne legyenek a beszélőhangi identitásokra, vagy legalább specifikus normák minden egyes releváns ismertetőjegyre, ami esetleg a szegmenseknek csak egy részhalmazában fordul elő. Ezt illusztrálta az a látványosan szegmens alapú szerveződést mutató távolságtérkép, amelyhez az egyes szavak helye a beszélőhang-tipikussághoz való hozzájárulásuk hasonlósága alapján került kiszámításra (2. fejezet), valamint a beszélőhangi identitások tanulását vizsgáló tanulmányok szóhatásai is (3. fejezet).

Az absztrakció harmadik jelentése a reprezentált információ feldolgozási hierarchián belüli következő szintre továbbítására vonatkozik. Ennek a jelentésnek a használata a tárgyfeldolgozás hierarchikus modelljeihez kapcsolódik (pl. Bruce és Young, 1986; Belin és mtsai, 2004, 2011), amelyek egymást követő feldolgozási szintek létét tételezik fel. Ebben az értelemben egy absztraktabb szint magasabb, később következő állomást jelöl a feldolgozási hierarchián belül. Az itt bemutatott kísérletek az agykérgi hierarchia több egymásra épülő szintjének létét igazolták a beszélőhangokra: ahogy az alábbiak összegzik, funkcionálisan és anatómiailag is elkülönülő állomások vesznek részt a beszélőhangi identitás feldolgozásában.

Ahogy már láttuk, számos norma alapú reprezentációs tér szerepet játszik a beszélőhang-észlelésben. Az agyi képalkotó eljárások lényegi hozzájárulása mindehhez az az eredmény, hogy e különböző terek anatómiailag is elkülönülten reprezentálódnak az emberi agyban. Hosszú távú akusztikai centralitásra érzékeny kérgi területek találhatóak a jobboldali STS középső és poszterior részein, míg identitásra vonatkozó centralitásra érzékeny területek találhatóak az anterior temporális pólusban (ATP, 5. fejezet) és a beszélőhangokra szelektíven érzékeny inferior frontális kéregben (IFC, 6. fejezet). Ezek a beszélőhangokra szelektíven érzékeny területek az auditoros ‘mi’ pálya állomásai lehetnek (Belin és mtsai, 2004; Ahveninen és mtsai, 2006), ahol az STS közvetlen, erős strukturális kapcsolatokkal rendelkezik lefelé, az elsődleges hallókéreg irányába (Kumar és mtsai, 2007),

és felfelé, mind az anterior temporális és az inferior frontális területek irányába (Ethofer és mtsai, 2012). Úgy tűnik tehát, hogy a neurális szenzitivitás e különböző, anatómiailag is elkülönülő típusai a beszélőhang-feldolgozási hierarchia specifikus állomásainak agykérgi megvalósulásai.

Mindent egybevetve, az absztrakció számos szinten és módon van jelen a beszélőhangok felismerésében. Egy javaslat szerint a beszélőhang-feldolgozási hierarchia különböző szintjei felelősek az akusztikai és az identitásra vonatkozó feldolgozásért (Belin és mtsai, 2004, 2011). Azonban a jelen disszertáció arra is rámutatott, hogy az akusztikai és identitási szenzitivitások, miközben valóban elkülönülnek anatómiailag és funkcionálisan egyaránt, mégis implementálhatóak egyetlen, hasonlósági alapú reprezentációs terekre vonatkozó neurális kódolási mechanizmus segítségével, ahol csak a terek méretében van különbség (azaz, van egy nagyobb szupraindividuális tér, és vannak kisebb intraindividuális terek). Szélesebb perspektívából nézve ez azt jelzi, hogy egy többszintes struktúra nem feltétlenül használ bonyolult mechanizmusokat. A fraktálok a matematikában jól ismert példái az olyan komplex struktúráknak, amelyek nagyon egyszerű szabályok alkalmazásával jönnek létre. Ott az a titok nyitja, hogy ezeket az egyszerű szabályokat újra és újra alkalmazni kell az egész különböző részeire. A több szinten működő absztrakció lehet az eszköz ahhoz, hogy egy összetett architektúra épülhessen fel néhány egyszerű szabályból. Amellett érveltem, hogy éppen ez történhet az emberi beszélőhang-feldolgozás esetében is.

A jelen disszertációban azt mutattam be, hogy a beszélő hangjából történő személyfelismerés számos szegmentális és nemszegmentális ismertetőjegyen alapul, és hogy ezek az ismertetőjegyek mind egyedi módokon járulnak hozzá az észlelt beszélőhang-tipikussághoz. Az itt leírt vizsgálatok igazolták, hogy a beszélőhangi identitások a beszédjel természetes hallási tárgyai, beépített előfeltevésekkel arra vonatkozóan, hogy mi alkothat egy egyéni beszélőhang-kategóriát. A beszélői identitásokat több, adaptív norma alapú neurális kód segítségével reprezentáljuk, melyek funkcionálisan és anatómiailag elkülönülő, hierarchikusan szerveződő szinteken vannak jelen az emberi agyban. E szintek között van egy szupraindividuális beszélőhangi tér, a beszélőhangokra szelektíven érzékeny szuperior temporális szulkusz területeken; és vannak köztük intraindividuális beszélőhangi terek is, a jobb agyfélteke anterior temporális és inferior frontális területein.

Curriculum vitae

Attila Andics was born in 1980 in Budapest, Hungary. He attended secondary education in the Németh László Gimnázium in Budapest, after which he studied psychology and mathematics at the Eötvös Loránd University in Budapest, and became a psychologist (MA), a teacher in psychology (MA), and a teacher in mathematics (MSc) in 2006. In the meantime he also studied cognitive neuroscience at the Radboud University Nijmegen and obtained a MSc in 2005. His master thesis research project was carried out at the Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging and examined the neural stages of prelexical speech processing by means of functional magnetic resonance imaging. He was awarded a scholarship from the German-Max-Planck-Gesellschaft to prepare his Ph.D. thesis at the Max Planck Institute for Psycholinguistics in Nijmegen. In his doctoral research, presented in this book, he studied the neural mechanisms of voice recognition, and the interplay of speech and talker identity processing. From 2009 to 2011 he had a position at the MR Research Centre of the Semmelweis University in Budapest and studied the neural effects of expectation for voices. Between 2010 and 2012 he also worked as a mathematics teacher in the Piarist High School, Budapest. Since January 2012 he works in the Comparative Ethology Research Group of the Hungarian Academy of Sciences and the Eötvös Loránd University, where he investigates social and affective aspects and the species-specificity of learning and person perception, using neuroimaging, psychophysical and comparative methods. In the last fourteen years he has been involved as facilitator, coordinator or advisor in a number of civil organisations and projects that aimed at raising the level of psychological awareness in various interpersonal and learning contexts. He is married to Ágnes Andics, they have two sons, Áron (2006) and Benedek (2009).

MPI series in psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*
17. Moving eyes and naming objects. *Femke van der Meulen*
18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja H. de Jong*

21. Fixed expressions and the production of idioms. *Simone A. Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*
29. At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Broersma*
35. Retrieving semantic and syntactic word properties. *Oliver Müller*
36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*
37. Sensitivity to detailed acoustic information in word recognition. *Keren B. Shatzman*
38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*
39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*
40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich*
41. Cognitive cladistics and the relativity of spatial cognition. *Daniel B.M. Haun*
42. The acquisition of auditory categories. *Martijn Goudbeek*
43. Affix reduction in spoken Dutch. *Mark Pluymaekers*

44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Kooijman*
45. Space and iconicity in German Sign Language (DGS). *Pamela Perniss*
46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*
47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*
48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*
49. Phoneme inventories and patterns of speech sound perception. *Anita Wagner*
50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*
51. A grammar of Savosavo, a Papuan language of the Solomon Islands. *Claudia Wegener*
52. Prosodic structure in speech production and perception. *Claudia Kuzla*
53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*
54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*
55. Processing the fine temporal structure of spoken words. *Eva Reinisch*
56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*
57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*
58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*
59. Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*
60. Processing casual speech in native and non-native language. *Annelie Tuinman*
61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Robinson*
62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*
63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*
64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*

65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*
66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*
67. The relevance of early word recognition: Insights from the infant brain. *Caroline Junge*
68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias J. Sjerps*
69. Structuring language: contributions to the neurocognition of syntax. *Katrien Rachel Segaert*
70. Infants' appreciation of others' mental states in prelinguistic communication: a second person approach to mindreading. *Birgit Knudsen*
71. Gaze behavior in face-to-face interaction. *Federico Rossano*
72. Sign-spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space. *Connie de Vos*
73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*