Journal of Phonetics 43 (2014) 38–46

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Cross-speaker generalisation in two phoneme-level perceptual adaptation processes

Patrick van der Zande [a,*], Alexandra Jesse [b], Anne Cutler [a,c]

[a] *Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands*
[b] *Department of Psychology, University of Massachusetts Amherst, 135 Hicks Way, MA 01003, USA*
[c] *MARCS Institute, University of Western Sydney, Penrith South, New South Wales 2751, Australia*

ARTICLE INFO

Article history:
Received 12 October 2012
Received in revised form
26 September 2013
Accepted 22 January 2014

ABSTRACT

Speech perception is shaped by listeners' prior experience with speakers. Listeners retune their phonetic category boundaries after encountering ambiguous sounds in order to deal with variations between speakers. Repeated exposure to an unambiguous sound, on the other hand, leads to a decrease in sensitivity to the features of that particular sound. This study investigated whether these changes in the listeners' perceptual systems can generalise to the perception of speech from a novel speaker. Specifically, the experiments looked at whether visual information about the identity of the speaker could prevent generalisation from occurring. In Experiment 1, listeners retuned auditory category boundaries using audiovisual speech input. This shift in the category boundaries affected perception of speech from both the exposure speaker and a novel speaker. In Experiment 2, listeners were repeatedly exposed to unambiguous speech either auditorily or audiovisually, leading to a decrease in sensitivity to the features of the exposure sound. Here, too, the changes affected the perception of both the exposure speaker and the novel speaker. Together, these results indicate that changes in the perceptual system can affect the perception of speech from a novel speaker and that visual speaker identity information did not prevent this generalisation.

1. Introduction

Listeners flexibly adjust to newly encountered speech input. For example, listeners adjust the boundaries of phonetic categories to correctly include a previously ambiguous speech sound (Bertelson, Vroomen, & de Gelder, 2003; Norris, McQueen, & Cutler, 2003). In addition to this phonetic retuning in response to ambiguous speech, the perceptual system is also altered by unambiguous and clearly intelligible speech. Selective adaptation is a process whereby perceivers become less sensitive to an unambiguous sound to which they are repeatedly exposed (Diehl, 1975; Eimas & Corbit, 1973). Phonetic retuning to ambiguous sounds shows how the perceptual system deals with problematic input; selective adaptation to unambiguous speech may reflect overexposure to a particular sound. In the present experiment, we investigated the generality of these two adaptation processes. We specifically examined whether changes in the perceptual system based on the speech input from one speaker affect the subsequent processing of speech from another speaker.

Sounds can be ambiguous due to idiosyncrasies that are specific to a speaker's production. Listeners are sensitive to this variability across speakers (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Maye, Aslin, & Tanenhaus, 2008), and seem to accommodate to it. Speech is generally identified more accurately when produced by a familiar speaker than by an unfamiliar speaker (Bradlow, Nygaard, & Pisoni, 1999; Craik & Kirsner, 1974; Goldinger, 1996; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). When sounds are ambiguous, however, listeners can identify them by referring to stored lexical knowledge (Ganong, 1998). Thus if listeners hear the word *platypu*[?] (where [?] symbolises an ambiguous sound between [s] and [f]), they can correctly identify the word despite the ambiguity in the auditory signal: lexical knowledge of English tells them that *platypus* is a word while *platypuf* is not. Listeners can also disambiguate words by using information obtained from seeing the speaker talk. Visual speech input provides information that is redundant and complementary to the available auditory information (Grant, Walden, & Seitz, 1998; Jesse & Massaro, 2010; Sumby & Pollack, 1954; Summerfield, 1987; Walden, Prosek, & Worthington, 1974), and ambiguous visual speech categories may be retuned by reference to lexical information in the same way as phonetic categories (Van der Zande, Jesse, & Cutler, 2013). Disambiguating ambiguous speech input has long lasting effects on perception, as it causes shifts in listeners' phonetic category boundaries in order to assign the

[*] Corresponding author. Tel.: +31 24 352 1377.
E-mail address: Patrick.vanderZande@mpi.nl (P. van der Zande).

0095-4470/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2014.01.003

ambiguous input to the appropriate category as dictated by the disambiguating source of information. The phonetic retuning affects how listeners subsequently judge the ambiguous sounds (Baart & Vroomen, 2010; Bertelson et al., 2003; Norris et al., 2003; Samuel & Kraljic, 2009), even in the absence of any disambiguating information. Hearing an ambiguous sound between /s/ and /f/ in the context of /s/-final words, such as *platypus*, makes listeners give more [s] responses to steps along an /s/-/f/ continuum, while hearing the same sound in the context of words like *giraffe* has the opposite effect (Norris et al., 2003). Disambiguation by the visual speech signal leads to similar shifts in the auditory phonetic categories (Bertelson et al., 2003). Phonetic retuning thus facilitates the subsequent recognition of sounds and does so across the lexicon, even when word context or word-internal position is changed (Jesse & McQueen, 2011; McQueen, Cutler, & Norris, 2006; Mitterer, Chen, & Zhou, 2011; Sjerps & McQueen, 2010).

Changes in the perceptual system also occur in response to unambiguous speech input that is not difficult to process. Repeated exposure to an unambiguous sound leads to a decrease in sensitivity to the features of that particular sound (Diehl, 1975; Eimas & Corbit, 1973; Samuel, 1986; Sawusch, 1977; Sawusch & Pisoni, 1976). This reduced sensitivity to specific phonetic features is thought to be due to fatigue within the perceptual system (Samuel, 1986). Like phonetic retuning, selective adaptation affects listeners' subsequent perception of sounds, but in a manner almost opposite to the retuning effect, in that listeners give *fewer* [da] responses to the steps of a /ba/-/da/ continuum after multiple repetitions of an unambiguous /da/ than after multiple repetitions of /ba/ (Eimas & Corbit, 1973). Phonetic retuning and selective adaptation thus reflect distinct processes within the flexible perceptual system.

Selective adaptation to auditory features is driven specifically by the acoustic information in the speech signal (Blumstein, Stevens, & Nigro, 1977; Sawusch & Pisoni, 1976). Whereas phonetic retuning can be guided by visual speech input, selective adaptation is not modulated by visual speech information. Perceivers presented with McGurk stimuli (Macdonald & McGurk, 1978; McGurk & MacDonald, 1976), consisting of an auditory /ba/ and a visual /ga/, generally identify this audiovisual stimulus as /da/ even when specifically instructed to report what they had *heard* the speaker say, ignoring the visual speech signal (MacDonald & McGurk, 1978). The McGurk effect thus clearly shows the influence of the visual speech information on the perception of the auditory speech input. When presented with such incongruent McGurk stimuli, selective adaptation occurs to the auditory sound, not to the overall audiovisual percept (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). The fact that selective adaptation is in line with the acoustic signal even when it is different from the perceived identity of the audiovisual utterance suggests that selective adaptation is modality-specific and takes place before information from the auditory and visual speech signals is integrated into an overall percept (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994).

Phonetic retuning, but not selective adaptation, is affected by whether or not the auditory signal is perceived as speech, as shown in a study using sine-wave speech (Vroomen & Baart, 2009a). Sine-wave signals are stripped of much of the acoustic detail of speech, but retain overall amplitude and frequency cues. Listeners generally do not perceive sine-wave speech as speech until they are explicitly informed about the nature of the signal (Vroomen & Baart, 2009a), and typically only integrate information from sine-wave speech and accompanying visual speech if they are given information about the speech nature of the sine-wave signal (Tuomainen, Anderson, Tiippana, & Sams, 2005). In Vroomen and Baart's study, listeners heard sine-wave speech and simultaneously saw visual speech. Effects of selective adaptation were observed regardless of whether listeners were informed about the sine-wave speech signal and thus regardless of whether information from auditory and visual speech had been combined. This is in line with the idea that selective adaptation is modality-specific. Phonetic retuning, on the other hand, was only observed for informed listeners and thus appears to be dependent on the integration of the auditory and the visual speech input.

A final piece of evidence for the dissociation of the two effects is provided by the difference in the rates at which they build up and dissipate (Vroomen, van Linden, de Gelder, & Bertelson, 2007; Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004). Selective adaptation builds up slowly and persists for up to 60 consecutive categorisation trials without renewed exposure (Vroomen et al., 2004). The slow build-up suggests that it takes some time for fatigue within the perceptual system to set in. Phonetic retuning, on the other hand, is established rapidly with only a small number of exposure trials (Vroomen et al., 2007), indicating that learning occurs nearly instantly after perceiving problematic speech input. The rate of dissipation for phonetic retuning varies depending on the source of the disambiguating information during exposure. Visually guided retuning dissipates quickly and the effect is no longer observed after six categorisation trials, unless there is additional exposure (Vroomen & Baart, 2009b; Vroomen et al., 2004). The effects of lexically guided retuning are still observed after a 25-min or even a 12-h intervening period between exposure and test (Eisner & McQueen, 2006; Kraljic & Samuel, 2005), although the studies investigating visually guided and lexically guided retuning varied on more points than just the source of the disambiguating information.

Given the fact that phonetic retuning and selective adaptation are different processes of adaptation, they may also differ in the extent to which their influence affects speech from a novel speaker. Adjustments made after exposure to the speech of a single speaker could be speaker-specific, or might generalise to affect the interpretation of speech from another speaker. Whether or not selective adaptation to auditory speech generalises to a new speaker has not been investigated. It seems, however, possible to obtain such generalisation, at least if speakers produce acoustically similar realisations of a sound. The generalisation of selective adaptation is guided by acoustic similarity: selective adaptation occurs across acoustically similar phonemes (Eimas & Corbit, 1973). Exposure to /ba/ affects the subsequent perception of both a /ba/-/pa/ continuum and a /da/-/ta/ continuum (Eimas & Corbit, 1973). Generalisation is not observed across position in the syllable, however, which is attributed to the high variability of sounds across syllable positions (Ades, 1974). Generalisation of selective adaptation across speakers has been found, however, with static visual representations of speech sounds (Jones, Feinberg, Bestelmeyer, DeBruine, & Little, 2010). Exposure to still images of a speaker producing a sustained /m/ sound resulted in fewer [m] responses than exposure to an image of the speaker producing a sustained /u/ sound when images showing mouth shapes ambiguous between /m/ and /u/ were subsequently categorised. The same effect of selective adaptation to the mouth shapes was observed for the exposure speaker and for a novel speaker. It remains unclear, however, whether selective adaptation to auditory speech also generalises across speakers following exposure to natural auditory-only and audiovisual speech materials.

Unlike selective adaptation, phonetic retuning has already been shown to generalise across speakers. More specifically, lexically guided retuning generalises across speakers when the critical phonemes are plosives, but not when they are fricatives (Eisner & McQueen, 2005; Kraljic & Samuel, 2006). Exposure to an ambiguous sound between /d/ and /t/ resulted in effects of phonetic retuning regardless of whether the subsequently categorised speech was produced by the exposure speaker or by a novel speaker (Kraljic & Samuel, 2006). Generalisation across speakers was not found after exposure to an ambiguous fricative between /f/ and /s/, however (Eisner & McQueen, 2005).

The discrepancy between these findings has been attributed to differences in the phoneme contrasts that were used (Kraljic & Samuel, 2006). The voicing distinction for the plosive sounds depends, among others, on the duration of the silence before the release and the duration of vibration after the release (in both cases longer durations favour /t/). These durational cues occur on a single dimension, so while speakers may vary in their durations (Allen, Miller, & DeSteno, 2003), the nature and the direction of the effect is constant, making learning for one speaker applicable to the recognition of speech from other speakers (Kraljic & Samuel, 2005). The place distinction for fricatives is based on spectral cues, which depend on the shape of the speaker's vocal tract and vary more substantially across speakers. This variability makes learning for fricatives specific to individual speakers and learning therefore does not generalise (Eisner & McQueen, 2005; Kraljic & Samuel, 2005).

Generalisation of phonetic retuning across speakers may thus be driven by the acoustic similarity for the target phonemes. An alternative explanation might be that generalisation is influenced by the availability of speaker identity information in the input. The results of the previous studies cannot speak to which of these two factors matters, since the degree of acoustic similarity across speakers and the degree to which the speech sounds contained speaker identity information were confounded. In the current study, we investigated this problem directly by teasing apart the acoustic similarity and the availability of speaker identity information. To do so, we used audiovisual speech materials in combination with a plosive contrast. We used two plosive sounds (/b/ and /d/) in order to provide a favourable auditory context for generalisation to occur. Place of articulation was used rather than a voicing contrast, because the former but not the latter can be distinguished by listeners on the basis of visual speech (Bernstein, Demorest, & Tucker, 2000; Van Son, Huiskamp, Bosman, & Smoorenburg, 1994).

For phonetic retuning, the focus of the current study was to determine whether generalisation takes place after exposure to audiovisual speech. The auditory speech input should allow generalisation while the visual speech input contains information about the identity of the speaker, which may inhibit generalisation. In Experiment 1, participants were presented during exposure with audiovisual speech tokens containing an auditory ambiguity that was resolved by the visual speech signal. Exposure to such audiovisual materials should induce phonetic retuning. A subsequent auditory-only test phase had participants categorise continua steps produced by either the exposure speaker or by a novel speaker. If acoustic similarity drives phonetic retuning, we should see an effect of retuning for both speakers at test. No effect of generalisation is expected if the visual speaker identity information indeed affects the generalisation of phonetic retuning. In Experiment 2, the possibility of generalisation across speakers for selective adaptation was investigated in both an auditory-only and an audiovisual condition. Participants received unambiguous auditory and audiovisual speech materials during exposure, sufficient to induce selective adaptation. Since selective adaptation has been shown to be unaffected by visual speech input, generalisation across speakers is expected for both presentation conditions. If, on the other hand, information about the identity of the speaker in the visual speech input does affect generalisation, generalisation should only be observed in the auditory-only condition.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Twenty-eight native speakers of Dutch (mean age = 20; 8 males) were paid for their participation in Experiment 1. All participants reported having normal hearing and normal or corrected-to-normal vision. Three participants were excluded due to equipment failure. One further participant was excluded due to insensitivity to the auditory continuum in the calibration phase. The final data set used for analysis consisted of the data from 24 participants. Seven additional participants from the same population took part in an auditory-only pilot experiment.

2.1.2. Materials

Two male native speakers of Dutch were video recorded with a Sony DCR-HC1000E camera. Audio was recorded simultaneously with two stand-alone Sennheiser microphones. Videos showed the head and shoulders of a speaker. The recordings of the two talkers formed the basis for all materials used in Experiment 1 and in Experiment 2. Talkers produced multiple tokens of the nonsense vowel–consonant–vowel (VCV) utterances /a:ba/, /a:da/, and /a:xa/. These utterances were produced in pairs, avoiding list intonation. All possible combinations of the CVC tokens were recorded. Videos were digitised as uncompressed avi files (720 × 576 pixels in PAL format). Audio sampling rate was 44.1 kHz.

2.1.3. Auditory-only test materials

For both speakers an individual auditory-only /a:ba/-/a:da/ continuum was created using Praat (Boersma & Weenink, 2006). Auditory /a:ba/, /a:da/, and /a:xa/ tokens were selected; to avoid mismatched timing of features when combined with the visual speech input the selected tokens had feature durations as close as possible to the average durations across all recorded tokens of the same type. To create the continua, initial /a:/ sounds were first taken from the selected /a:xa/ tokens to ensure that the vowel transitions of the word-initial vowels did not contain any cues for either a following /b/ or a /d/. Parts of the steady-state portion of the initial /a:/s were removed so that the resulting sounds corresponded in duration to the average duration of /a:/ in this position across all tokens for the same speaker (approximately 265 ms and 375 ms for Speaker 1 and 2, respectively). Second, /ba/ and /da/ from the /a:ba/ and /a:da/ tokens were edited to have equal durations and pitch contours before being mixed into a 21-step continua changing from /ba/ to /da/ in equal steps. These 21 steps were then concatenated with the edited initial /a:/ token taken from /a:xa/ of the same speaker to create the final /a:ba/-/a:da/ continuum.

A pilot study with seven participants was conducted in order to test participants' sensitivity to the two resulting continua. Participants categorised 13 steps from both speakers' continua (steps 0, 3, 5, 7–13, 15, 17, 20). For each speaker, participants made 104 categorisations, being eight differently randomised sequences of the 13 continuum steps. The order of presentation of the two speakers was counterbalanced across participants. The stimuli were presented over headphones at a fixed level. The response alternatives “b” and “d” were displayed on a computer screen and participants categorised the sounds by clicking on one of these two labels. Participants were instructed to respond as quickly and as accurately as possible. Each new trial started only after participants had given a response.

Fig. 1 shows the results of the pilot study for both speakers' auditory continua. The results indicate that the percentage of [d] responses increased the more /d/-like the auditory continuum step was and that participants were thus sensitive to the continua. These pilot results were used to select an ambiguous range of steps to be used in the main experiment. This range was between step 7 and step 13 of the continuum and was the same for both speakers.

2.1.4. Audiovisual exposure materials

For both speakers, the seven steps that made up the ambiguous range (steps 7–13) were combined with the natural visual speech tokens /a:ba/ and /a:da/. In this way, multiple audiovisual $A_?V_b$ and $A_?V_d$ tokens were created for both speakers providing the possibility of selecting the most suitable audiovisual token for each individual participant. The visual-only speech tokens came from the same audiovisual tokens that provided the

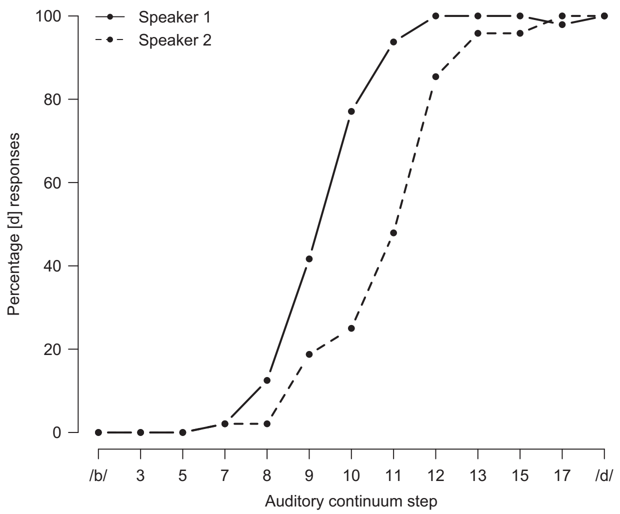

Fig. 1. Mean percentages of [d] responses as a function of /a□ba/-/a□da/ continuum steps in the auditory-only pilot. Solid lines show the results for Speaker 1 and dashed lines the results for Speaker 2.

auditory speech input used for the auditory-only continua. Each audiovisual token started and ended with 15 frames showing the face of the speaker in a neutral position and with the lips parted slightly.

2.1.5. Procedure

Participants were tested individually in a sound-attenuated booth. The auditory input was presented over Sennheiser HD280 headphones at a fixed level. The experiment consisted of three separate phases, similar to the design used by Bertelson and colleagues (2003). Participants first performed an auditory-only calibration phase and then completed 32 iterations of exposure phase/test phase sequences. Within each sequence, an audiovisual exposure phase was directly followed by an auditory-only test phase.

During the auditory-only calibration phase participants categorised 13 auditory-only continuum steps (0, 3, 5, 7–13, 15, 17, 20) of both speakers. The order of presentation was blocked by speaker and block order was counterbalanced across participants. Within each block, all continuum steps were presented eight times in a newly randomised order, giving a total of 208 calibration trials (13 steps × 8 repetitions × 2 speakers). Participants categorised the auditory continuum steps as /a:ba/ or /a:da/ by clicking with the mouse on labelled buttons on the computer screen and new trials started after a response was given. The instructions stressed the need for speed and accuracy.

The auditory-only calibration results were used to find for each individual participant the ambiguous range in the two speakers' continua. The continuum step closest to a participant's 50% cut-off point between [b] and [d] was selected ($A_?$), as were the two next steps closer to either end of the continuum ($A_{?-1}$ and $A_{?+1}$). $A_?$ was used in the audiovisual tokens for each participant in the audiovisual exposure phase. In the auditory-only test phase, participants were presented with all three auditory continuum steps.

In the audiovisual exposure phase, the most ambiguous auditory continuum step ($A_?$) was used in combination with an unambiguous visual-only token of /a:ba/ or /a:da/ to create the audiovisual tokens $A_?V_b$ and $A_?V_d$. Half of the participants received audiovisual tokens from Speaker 1, the other half from Speaker 2. In each exposure block, one of the two audiovisual tokens was presented eight times. The presentation of $A_?V_b$ and $A_?V_d$ audiovisual exposure blocks was counterbalanced across participants. No explicit task was given to participants to perform during the audiovisual exposure phase, other than that they were to pay close attention to what was said by the speaker.

Each audiovisual exposure block was followed by an auditory-only test phase in which participants categorised steps $A_{?-1}$, $A_?$, and $A_{?+1}$. The three auditory-only test tokens were presented twice in a newly randomised order, resulting in a total of six categorisation trials per test block. In each test block, participants heard either Speaker 1 or 2 and the presentation of speakers was randomised across the experiment. The auditory-only tokens were thus either produced by the speaker whom participants had perceived during exposure or by a novel speaker. Participants heard both speakers during calibration, but only saw one of the two speakers during audiovisual exposure. Participants categorised the ambiguous steps as /a:ba/ or /a:da/ as quickly and as accurately as possible by pressing the button on a button box that corresponded to the respective label shown on the computer screen ("b" or "d"). In total, participants completed 32 repetitions of these exposure-test sequences.

2.2. Analysis

Results were analysed with linear mixed-effect models in the *R* statistical package (R Development Core Team, 2007), using the lmer function of the lme4 library (Bates & Sarkar, 2007). The dependent variable was the binomial response to continuum steps (0=[b]; 1=[d]). A logistic linking function was used for the categorical dependent variable. The best-fitting model was established by systematic model comparison, using likelihood-ratio tests. We started with a full model and then gradually removed factors that did not contribute to a better model fit, from factors with the largest *p* values on. Main effects were only removed if their factors did not contribute to an interaction. The best-fitting model included participant as a random factor. Exposure condition (/b/ exposure vs. /d/ exposure) and speaker familiarity (exposure speaker vs. novel speaker) were evaluated as contrast-coded fixed factors. Auditory test token was evaluated as a numerical fixed factor, centred on $A_?$.

2.3. Results and discussion

Participants were sensitive to the fact that the three auditory test tokens formed a continuum, giving more [d] responses to the more /d/-like auditory token than to the more /b/-like token or the most ambiguous token ($\beta = 1.25$, $SE = 0.05$, $p < .001$; see Fig. 2). Overall, participants made more [d]

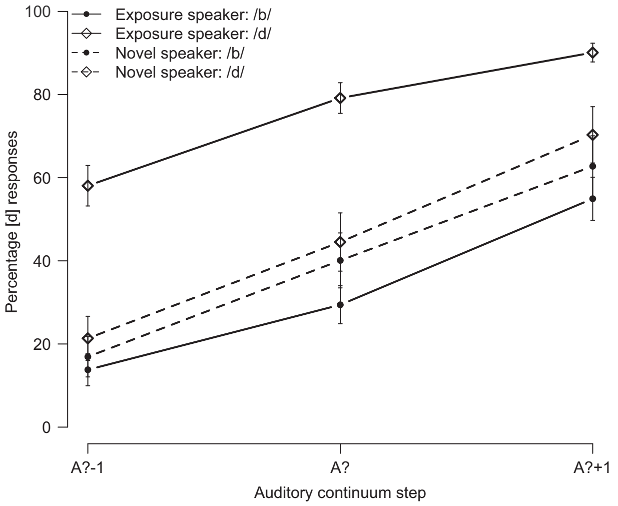

Fig. 2. Mean percentages of [d] responses as a function of auditory continuum step in Experiment 1. Solid lines show the results after exposure to $A_?V_b$ and dashed lines after exposure to $A_?V_d$. Black lines show the results for the exposure speaker at test and grey lines for the novel speaker at test. Error bars show standard error of the mean.

responses to the continuum of the exposure speaker than to that of the novel speaker ($\beta=-0.71$, $SE=0.07$, $p<.001$). Participants gave more [d] responses after /d/-exposure blocks than after /b/-exposure blocks ($\beta=1.46$, $SE=0.08$, $p<.001$), indicating that there is an effect of phonetic retuning. This effect was found for both the exposure speaker ($\beta=2.41$, $SE=0.12$, $p<.001$) and the novel speaker ($\beta=0.44$, $SE=0.12$, $p<.001$). The phonetic retuning effect size is, however, significantly smaller for the novel speaker than for the exposure speaker ($\beta=-2.29$, $SE=0.15$, $p<.001$).

Visual speech input thus results in the retuning of phonetic category boundaries relative to the learning condition (Bertelson et al., 2003). The results of Experiment 1 indicate that visually guided phonetic retuning affects the identification of speech produced both by the exposure speaker and by a different speaker even when the disambiguating visual speech signal contained information about the identity of the speaker. Generalisation was apparently not fully realised and the effect of retuning is smaller for the novel speaker than for the exposure speaker.

3. Experiment 2

The results of Experiment 1 indicate that listeners' retuned phonetic category boundaries affected their subsequent identification of speech produced by the exposure speaker as well as by the novel speaker. Generalisation across speakers thus occurred even when the disambiguating signal contained information about the identity of the speaker. Explicit knowledge about the identity of the speaker did not prevent the generalisation of visually guided phonetic retuning. As discussed above, selective adaptation reflects a different change within the perceptual system, namely one that is due to acoustic input alone and not affected by visual speech information. Recall that selective adaptation follows the acoustic input in McGurk stimuli, even when the perceived utterance differed due to the influence of visual speech information (Saldaña & Rosenblum, 1994). In Experiment 2, we tested whether the lack of modulation from visual speech means that selective adaptation also generalises to affect the perception of speech from a novel speaker.

Participants in Experiment 2 completed both auditory-only and audiovisual exposure blocks since generalisation across speakers for selective adaptation has not been investigated previously. The results for the auditory-only exposure condition will hence establish whether selective adaptation generalises across speakers. The results from the audiovisual exposure condition will inform whether this generalisation is affected by whether visual speech provides identity information about the exposure speaker. In all exposure blocks, listeners were exposed to unambiguous items before completing a categorisation task as in Experiment 1. Whereas the processing of visual speech information was necessary to disambiguate the auditory input in Experiment 1, this is not the case for the unambiguous input in Experiment 2. The effect of retuning is expected to be similar for both the exposure speaker and the novel speaker regardless of the presentation condition in the exposure phase. This finding would provide further evidence for the dissociation between phonetic retuning and selective adaptation.

3.1. Methods

3.1.1. Participants

Twenty-eight new participants (mean age = 21.5; 5 males) from the same population as in Experiment 1 were tested. One participant was excluded due to equipment failure. Another three participants were excluded due to insensitivity to the auditory-only continua. The final data set consisted of the data from 24 participants.

3.1.2. Materials

The materials for Experiment 2 were the same as in Experiment 1. The only difference was that participants were presented with unambiguous auditory-only (A_b and A_d) and audiovisual (A_bV_b and A_dV_d) versions of the exposure materials used in Experiment 1. That is, exposure materials for Experiment 2 were entirely free of conflict or ambiguity. The audiovisual stimuli consisted of the same unambiguous videos as used in Experiment 1, now combined with the endpoints of the exposure speaker's auditory continuum. The video track of the unambiguous audiovisual video tokens was replaced with a black frame in order to create the auditory-only exposure materials. Both the audiovisual and the auditory-only exposure stimuli were presented in .avi format.

3.1.3. Procedure

Experiment 2 differed from Experiment 1 in that participants were presented with unambiguous auditory-only or audiovisual versions of the stimuli during exposure. The experiment again started with an auditory-only categorisation task in which the continua for both speakers were categorised, and $A_?$ was selected for use in the auditory-only test phase. Following the pretest, participants were exposed to either the auditory-only A_b and A_d stimuli or to the audiovisual A_bV_b and A_bV_d stimuli. Presentation of the stimuli was blocked by exposure condition; blocks were presented in randomised order. Within each block, the same audiovisual or auditory-only token was presented eight times and participants had no explicit task to perform. They were, however, instructed to pay attention to what the speaker was saying at all times. Every exposure block was immediately followed by an auditory-only test block in which participants performed a categorisation task on $A_{?-1}$, $A_?$ and $A_{?+1}$. Participants completed 32 repetitions of exposure phase followed by test phase.

3.2. Analysis

Results were analysed as for Experiment 1. Exposure condition (/b/-exposure material vs. /d/-exposure material), speaker familiarity (exposure speaker vs. novel speaker), and presentation condition of the exposure material (auditory-only vs. audiovisual) were evaluated as contrast-coded fixed factors. Auditory-only continuum step was evaluated as a numerical factor centred on the middle step. Participants were included as a random factor in the best-fitting model.

The analysis of the full model revealed a four-way interaction between the fixed factors ($\beta=0.87$, $SE=0.40$, $p<.05$), indicating that the effect of selective adaptation varied as a joint function of talker familiarity, presentation condition, and auditory continuum step. We therefore report the results for the auditory-only test data separately for the auditory-only and the audiovisual exposure conditions.

3.3. Results and discussion

3.3.1. Auditory-only exposure condition

Participants gave fewer [d] responses in the auditory-only categorisation test phase after exposure to the auditory-only /d/ token than after the auditory-only /b/ token ($\beta=-1.24$, $SE=0.11$, $p<.001$; see Fig. 3) showing an effect of selective adaptation for the auditory-only materials. Overall, participants made more [d] responses to the continuum of the novel speaker than to that of the exposure speaker ($\beta=1.15$, $SE=0.11$, $p<.001$). Participants were sensitive to the auditory continua and gave more [d] responses to the more /d/-like test token than to the more /b/-like test token or the most ambiguous step ($\beta=1.14$, $SE=0.07$, $p<.001$). The difference in the effect of selective adaptation for the exposure speaker and for the novel speaker ($\chi^2(1)=3.11$, $p=.08$) was only marginally significant, showing that selective adaptation fully generalised across speakers with auditory-only exposure materials. There was also a marginally significant difference between the effect of exposure for the three ambiguous test tokens ($\chi^2(1)=3.75$, $p=.05$), indicating that the shift was larger for the middle step than for the two neighbouring steps.

3.3.2. Audiovisual exposure condition

Participants gave fewer [d] responses in the auditory-only categorisation test phase after exposure to the audiovisual /d/ token than after the audiovisual /b/ token ($\beta=-1.28$, $SE=0.11$, $p<.001$; see Fig. 4), indicating an effect of selective adaptation for the audiovisual materials. Overall, more [d] responses were again given to the novel speaker's than to the exposure speaker's continuum ($\beta=1.07$, $SE=0.10$, $p<.001$). Participants were sensitive to the auditory continua giving more [d] responses the more /d/-like the test token ($\beta=1.02$, $SE=0.07$, $p<.001$). There was no difference in the selective adaptation effect for the exposure speaker and the novel speaker (not a predictor, $\chi^2(1)=0.03$, $p=.86$), indicating that selective adaptation fully generalised across speakers even after exposure to the audiovisual materials. Cross-speaker generalisation of selective adaptation was thus not hindered by the presence of speaker identity information in the visual speech input. There was no difference in the results for the exposure speaker across the auditory-only and audiovisual exposure conditions (presentation condition not a predictor, $\chi^2(1)=0.03$, $p=.87$), which provides further evidence for the lack of influence from the visual speech input.

The results of Experiment 2 show that exposure to unambiguous auditory and audiovisual speech made participants less likely to assign ambiguous auditory tokens to the same phonetic category as the phoneme that had been encountered during exposure. In accord with the phonetic

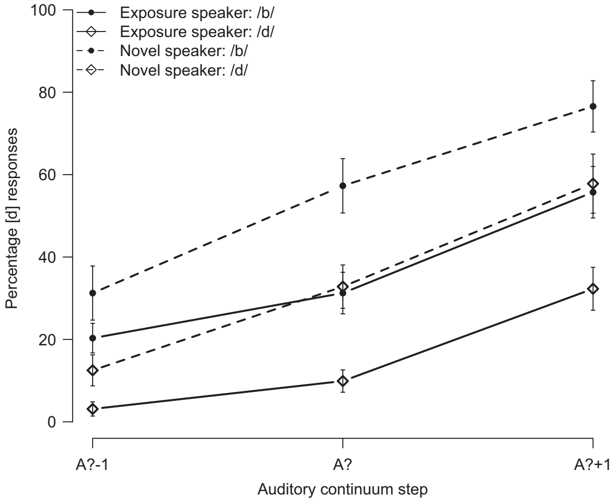

Fig. 3. Mean percentages of [d] responses as a function of auditory continuum step following auditory-only exposure in Experiment 2. Solid lines show the results after exposure to A_b and dashed lines after exposure to A_d. Black lines show the results for the exposure speaker at test and grey lines for the novel speaker at test. Error bars show standard error of the mean.

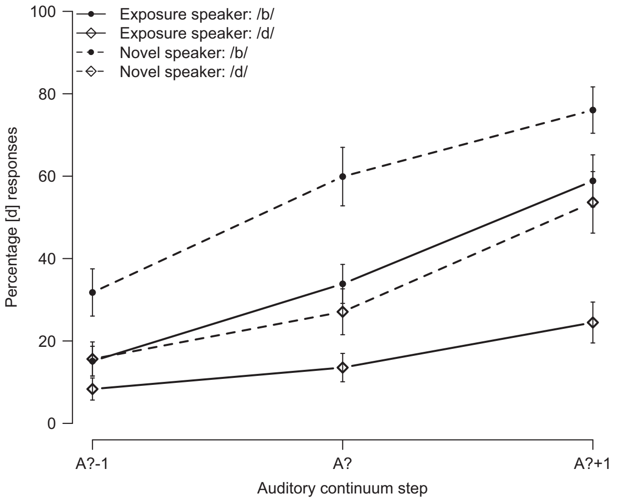

Fig. 4. Mean percentages of [d] responses as a function of auditory continuum step following audiovisual exposure in Experiment 2. Solid lines show the results after exposure to A_bV_b and dashed lines after exposure to A_dV_d. Black lines show the results for the exposure speaker at test and grey lines for the novel speaker at test. Error bars show standard error of the mean.

retuning results in Experiment 1, we find that selective adaptation affects the interpretation of speech from both the exposure speaker and a novel speaker. The effect of selective adaptation fully generalises across speakers in both the auditory-only and the audiovisual condition. This is in contrast with the results from Experiment 1, where we observed a smaller effect of phonetic retuning for the novel speaker than for the exposure speaker after audiovisual exposure. The presence of visual speaker identity information during exposure does not appear to affect the generalisation of selective adaptation, a finding that is in line with earlier studies that have shown selective adaptation to be a purely auditory phenomenon (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994).

4. Conclusions

Adjustments within the perceptual system occur after exposure to speech both when the speech in question is ambiguous and when it is unambiguous. Exposure to ambiguous, idiosyncratic speech results in shifts in listeners' category boundaries in order to incorporate ambiguous sounds into the intended categories. Unambiguous speech input results in decreased sensitivity to particular features of the input when the same sound is presented multiple times. The effects of phonetic retuning and selective adaptation both reflect changes in the perceptual system and indicate that different speech input can have very different results. The results of the current study show that changes in the perceptual system caused by phonetic retuning and selective adaptation both affect the processing of speech from the exposure speaker and from a novel speaker. Generalisation across speakers occurred despite the availability of speaker identity information in the audiovisual speech input. While the effect of selective adaptation on the processing of speech from the exposure speaker and the novel speaker was the same, the effect of phonetic retuning was smaller for the processing of speech from the novel speaker than from the exposure speaker.

That visually guided retuning of auditory phonetic categories generalises to the identification of speech from a different speaker is in line with previous results for the generalisation of lexically guided retuning (Kraljic & Samuel, 2006). Changes in phonetic categories after lexically guided retuning affect the processing of plosives produced by a different speaker (Kraljic & Samuel, 2006), but the same effect is not observed for fricatives (Eisner & McQueen, 2005). This difference in cross-speaker generalisation has been ascribed to the fact that fricatives are more variant across speakers and also provide more information about the identity of the speaker than plosives (Kraljic & Samuel, 2005, 2006). It has not been clear, however, whether it is information about the identity of the speaker or the lack of acoustic similarity that prevents generalisation.

To tease these two alternative explanations apart, our study investigated visually guided retuning rather than lexically guided retuning, and examined a place of articulation contrast for plosives as putatively offering the best chance of cross-speaker generalisation. A voicing contrast could not be used here because the visual speech input was the source of the disambiguating information and listeners are generally unable to distinguish voiced and voiceless sounds on the basis of visual speech (Bernstein et al., 2000; Van Son et al., 1994). The results of Experiment 1 show that the presence of speaker identity information, available in the visual speech input, does not prevent the generalisation of phonetic retuning across speakers. Shifts in listeners' category boundaries affected their subsequent perception of speech even when produced by a novel speaker. Also, listeners could not have disregarded the visual speech input, since it provided the only source of disambiguation for the ambiguous auditory speech input. These results thus suggest that it is acoustic similarity and not the lack of speaker identity information that allows generalisation of retuning across speakers to occur.

Although the effect of phonetic retuning was evident in the processing of speech from a novel speaker, this effect was smaller than for the exposure speaker. Phonetic retuning is thus not fully generalised to the novel speaker. One possibility for why this is could indeed be the availability of speaker identity information in the visual speech input during exposure. To fully assess this possibility, an auditory-only exposure condition would have been needed. This was not an option, however, given the setup of the present experiment wherein visual speech information was necessary for disambiguation. Prior studies on lexically guided retuning presenting auditory-only speech during exposure found full transfer of phonetic retuning to plosives (Kraljic & Samuel, 2005). There are, however, multiple other differences between this study and ours, such as differences in phoneme contrast (place of articulation vs. voicing in plosives) or disambiguating source (visual speech vs. lexical context). Future research is needed to further clarify what principles guide the transfer of phonetic retuning. Our current results nevertheless convincingly show that explicit knowledge about the identity of the speaker does not prevent generalisation as long as the auditory input from both speakers is acoustically similar.

Selective adaptation is also generalised to a new speaker, at least for plosives that are acoustically similar across speakers. The results of Experiment 2 show that selective adaptation generalises across speakers when the target sounds were plosives. The auditory-only results suggest that the decrease in sensitivity that occurs after repeated exposure to an unambiguous utterance affects the subsequent perception of speech for both the exposure speaker and a novel speaker. The change in the perceptual system is thus generally applicable and not specific to any speaker.

Selective adaptation also occurs for elements of vision (Webster, 2004; Webster & MacLin, 1999) and recent research has shown that selective adaptation can occur after exposure to static representations of speakers' mouth shapes (Jones et al., 2010). Seeing multiple repetitions of a picture showing a speaker produce a sound thus makes people less likely to perceive a more ambiguous mouth shape as representing that same sound. More interestingly, this effect of selective adaptation to visual speech was the same whether subsequent judgements were given for the exposure speaker or for a novel speaker. This second finding suggests that the selective adaptation effect in this case is not dependent on the identity of the speaker and thus reflects a more general change in the perceptual system. The results of the audiovisual exposure condition in Experiment 2 show a similar effect of generalisation across speakers for selective adaptation to audiovisual speech. Here, too, information about the identity of the speaker was available but did not affect the generalisation of selective adaptation. A decrease in sensitivity to auditory phonetic features thus affects subsequent perception of these features irrespective of the identity of the speaker.

The generalisation across speakers for selective adaptation was neither prevented nor even reduced by the presence of visual speaker identity information, which is unlike the results for phonetic retuning in Experiment 1. This difference could be due to the fact that in Experiment 2 the visual speech information was not necessary for disambiguation, since the auditory input was unambiguous. However, this finding is also in line with results from earlier studies that have found that selective adaptation is a purely auditory phenomenon and is not modulated by visual speech input (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). These studies used McGurk stimuli and found that the effect of selective adaptation was always in line with the auditory input, regardless of the fact that the perception of the combined audiovisual stimuli was different from the auditory input. The results of the audiovisual exposure condition in Experiment 2 provide further evidence that visual speech information does not affect selective adaptation. Selective adaptation was observed for both the exposure speaker and the novel speaker despite the fact that the visual speech input again contained information about the identity of the speaker.

Phonetic retuning and selective adaptation thus affect subsequent recognition of speech produced by the speaker whose speech initiated the changes in the perceptual system. Both effects also influence the recognition of speech from other speakers when the sounds they produced were acoustically similar. Where the two effects diverge is on the extent to which they are generalised to different speakers when information about the identity of the speaker is available. Phonetic retuning generalises across speakers but not fully. On the other hand, for selective adaptation, generalisation occurs to its full extent.

The perceptual system is thus flexible enough to adjust to speech input and does so regardless of whether the input is ambiguous or not. Ambiguous speech input and unambiguous speech input change the perceptual system in different directions, however, as is reflected in the effects of phonetic retuning and selective adaptation. These changes affect how listeners perceive speech on later occasions and this is true for both speech produced by the speaker for whom the original adjustments were made and for speakers who produce acoustically similar sounds. Generalisation across speakers occurs even when listeners have explicit information that the speech they are provided with is produced by a novel speaker. Generalisation for phonetic retuning may be beneficial for listeners as it can reduce processing costs. For selective adaptation, the generalisation indicates that when sensitivity to particular features of a sound is decreased it affects all sounds sharing that feature, whoever produced them. Changes in the perceptual system thus occur for various reasons and show how the system can flexibly adjust to the input it is given.

References

Ades, A. E. (1974). How phonetic is selective adaptation: Experiments on syllable position and vowel environment. *Perception & Psychophysics*, *16*(1), 61–66.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, *113*(1), 544–552.

Baart, M., & Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters*, *471*(2), 100–103.

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes [Software]. Available from ⟨http://lme4.r-forge.r-project.org/⟩.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233–252.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592–597.

Blumstein, S. E., Stevens, K. N., & Nigro, G. N. (1977). Property detectors for bursts and transitions in speech perception. *Journal of the Acoustical Society of America*, *61*(5), 1301–1313.

Boersma, P., & Weenink, D. (2006). Praat: Doing phonetics by computer (Version 5.2.40) [Software]. Available from ⟨http://www.praat.org/⟩.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*(2), 206–219.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, *116*(6), 3647–3658.

Craik, F. I. M., & Kirsner, K. (1974). Effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*(2), 274–284.

Diehl, R. L. (1975). Effect of selective adaptation on identification of speech sounds. *Perception & Psychophysics*, *17*(1), 48–52.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, *4*(1), 99–109.

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

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

Ganong, W. F. (1998). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology*, *6*(1), 110–125.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*(5), 1166–1183.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory–visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory–visual integration. *Journal of the Acoustical Society of America*, *103*(5), 2677–2690.

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention Perception & Psychophysics*, *72*(1), 209–225.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*(5), 943–950.

Jones, B. C., Feinberg, D. R., Bestelmeyer, P. E. G., DeBruine, L. M., & Little, A. C. (2010). Adaptation to different mouth shapes influences visual perception of ambiguous lip speech. *Psychonomic Bulletin & Review*, *17*(4), 522–528.

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. (2006). Generalisation in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268.

Macdonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Attention, Perception, & Psychophysics*, *24*(3), 253–257.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

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

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science*, *35*(1), 184–197.

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*(3), 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, *30*(4), 309–314.

R Development Core Team (2007). R: A language and environment for statistical computing [Software]. Available from 〈http://www.R-project.org/〉.

Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech-perception using a compelling audiovisual adapter. *Journal of the Acoustical Society of America*, *95*(6), 3658–3661.

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology*, *18*(4), 452–499.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218.

Sawusch, J. R. (1977). Peripheral and central processes in selective adaptation of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, *62*(3), 738–750.

Sawusch, J. R., & Pisoni, D. B. (1976). Response organization in selective adaptation to speech sounds. *Perception & Psychophysics*, *20*(6), 413–418.

Sjerps, M. J., & McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(1), 195–211.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212–215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In: B. Dodd, & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). London, UK: Lawrence Erlbaum.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, *96*, B13–B22.

Van der Zande, P., Jesse, A., & Cutler, A. (2013). Lexically guided retuning of visual phonetic categories. *Journal of the Acoustical Society of America*, *134*(1), 562–571.

Van Son, N. J. D. M.M., Huiskamp, T. M. I., Bosman, A. J., & Smoorenburg, G. F. (1994). Viseme classifications of Dutch consonants and vowels. *Journal of the Acoustical Society of America*, *96*(3), 1341–1355.

Vroomen, J., & Baart, M. (2009a). Phonetic recalibration only occurs in speech mode. *Cognition*, *110*(2), 254–259.

Vroomen, J., & Baart, M. (2009b). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a 24-hour delay. *Language and Speech*, *52*, 341–350.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–577.

Vroomen, J., van Linden, S., Keetels, M., de Gelder, W., & Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: Dissipation. *Speech Communication*, *44*(1–4), 55–61.

Walden, B. E., Prosek, R. A., & Worthington, D. W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech and Hearing Research*, *17*(2), 270–278.

Webster, M. A. (2004). Pattern-selective adaptation in color and form perception. In: L. Chalupa, & J. Werner (Eds.), *The visual neurosciences* (pp. 936–947). Cambridge, MA: MIT Press.

Webster, M. A., & MacLin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review*, *6*(4), 647–653.