L2 voice recognition: The role of speaker-, listener-, and stimulus-related factors

Polina Drozdova, Roeland van Hout, and Odette Scharenborg

Citation: The Journal of the Acoustical Society of America **142**, 3058 (2017); doi: 10.1121/1.5010169
View online: https://doi.org/10.1121/1.5010169
View Table of Contents: http://asa.scitation.org/toc/jas/142/5
Published by the Acoustical Society of America

L2 voice recognition: The role of speaker-, listener-, and stimulus-related factors

Polina Drozdova,[a] Roeland van Hout, and Odette Scharenborg[b]
Centre for Language Studies, Radboud University Nijmegen, Erasmusplein 1, P.O. Box 9103, 6500 HD Nijmegen, the Netherlands

(Received 16 March 2017; revised 8 September 2017; accepted 20 October 2017; published online 17 November 2017)

Previous studies examined various factors influencing voice recognition and learning with mixed results. The present study investigates the separate and combined contribution of these various speaker-, stimulus-, and listener-related factors to voice recognition. Dutch listeners, with arguably incomplete phonological and lexical knowledge in the target language, English, learned to recognize the voice of four native English speakers, speaking in English, during four-day training. Training was successful and listeners' accuracy was shown to be influenced by the acoustic characteristics of speakers and the sound composition of the words used in the training, but not by lexical frequency of the words, nor the lexical knowledge of the listeners or their phonological aptitude. Although not conclusive, listeners with a lower working memory capacity seemed to be slower in learning voices than listeners with a higher working memory capacity. The results reveal that speaker-related, listener-related, and stimulus-related factors accumulate in voice recognition, while lexical information turns out not to play a role in successful voice learning and recognition. This implies that voice recognition operates at the prelexical processing level.
 https://doi.org/10.1121/1.5010169

[MS] Pages: 3058–3068

I. INTRODUCTION

Recognizing voices is a prodigious human cognitive ability. Recognition of the mother's voice in infancy has a key role in children's emotional, social, and cognitive functioning (Abrams *et al.*, 2016). The ability of adults to recognize people by voice forms a crucial social skill (Perracione *et al.*, 2011), and, in general, contributes to the perception of interlocutors' emotional states and personal identities (Nygaard, 2005; Sidtis and Kreiman, 2012). Moreover, voice familiarity has been shown to facilitate word recognition and processing (Nygaard *et al.*, 1994; Nygaard and Pisoni, 1998).

While the importance of the ability to recognize voices is beyond dispute, the factors influencing successful voice recognition are still unclear. Several types of factors have been investigated, but the obtained results were mixed. The present study groups these various factors into three categories: speaker-related, listener-related and stimulus-related, and investigates the role of these three groups of factors on speaker recognition by second language (L2) listeners, in order to shed light on their separate and combined contribution to successful voice recognition.

The first group of factors which have been shown to influence voice recognition are related to the acoustic characteristics of speakers' voices. Laver (1968) distinguished three types of information conveyed in a speaker's voice: biological (gender, age, size), psychological (emotional state), and social information (regional origin, social group, profession). This information can be expressed in diverse voice quality features (e.g., loudness, pitch, phonation types, nasalization). These features are used to a different extent in voice recognition and differentiation (see Baumann and Belin, 2010, for an overview), with fundamental frequency (F0) being the most prominent one for distinguishing voices, while the importance of other characteristics such as frequencies of the main formants (F1, F2, F3), jitter, and shimmer are dependent on the type of speaker (e.g., male or female, pathological or normal voices). In this study, we specifically investigate the contribution of two speaker-related factors: fundamental frequency (average, minimum, and maximum) and average word length.

Interestingly, speakers were shown to vary in their identifiability which depended not only on the quality of their voices, but also on which specific speech sounds were produced (Bricker and Pruzansky, 1966; Amino and Arai, 2008; Andics *et al.*, 2007). Not all sounds are equally effective in conveying speaker-specific information. Research in both automatic speech recognition (Eatock and Mason, 1994; Gallardo *et al.*, 2015) and human speech recognition (Amino and Arai, 2008; Amino *et al.*, 2006) provide a ranking of sounds contributing to talker-identification, showing that nasal consonants and vowels are more informative than other sounds for the human identification of speakers. Vowels outperform consonants due to their combination of fundamental frequency (F0) and rich harmonic structure (Owren and Cardillo, 2006), while nasals outperform other consonants due to the speaker-specific characteristics of the resonating shapes involved and the timing of the velum movements for

[a]Also at: IMPRS for Language Sciences, Nijmegen, the Netherlands, Wundtlaan 1, P.O. Box 310, 6500 AH Nijmegen, the Netherlands. Electronic mail: p.drozdova@let.ru.nl
[b]Also at: Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Kapittelweg 29, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands.

the production of nasals which is consistent within a speaker and differs among speakers (Amino *et al.*, 2007). The number of nasals and vowels in the word was investigated as one of the stimulus-related factors.

Most stimulus-related and listener-related factors are however not so easy to disentangle: stimulus-related factors relate to the actual linguistic information available in the stimuli, whereas listener-related factors relate to what extent listeners differ in their ability to use this information. Voice learning studies demonstrated that listeners can learn to recognize talkers without any access to linguistic content, e.g., in time-reversed speech (Bricker and Pruzansky, 1966; Sheffert *et al.*, 2002) or in a completely unfamiliar language (Winters *et al.*, 2008). At the same time, speaker-specific (also referred to as indexical information; Abercrombie, 1967) and linguistic information in the signal are not completely independent. On the one hand, being familiar with a speaker's voice facilitates word recognition (Bradlow *et al.*, 1999; Levi *et al.*, 2011; Nygaard and Pisoni, 1998), on the other hand, linguistic knowledge can support processing of indexical information as well (Goggin *et al.*, 1991; Bregman and Creel, 2014; Winters *et al.*, 2008).

Goggin and colleagues (Goggin *et al.*, 1991) showed that voice recognition is more accurate when listeners understand the language being spoken: monolingual English listeners identified bilingual English-German speakers better when they spoke English than when they spoke German, while the reverse pattern was true for monolingual German listeners. At the same time, Bregman and Creel (2014) demonstrated that Korean-English bilinguals are faster in learning to recognize voices speaking in their first (Korean) than in their second language (English), and that the rate of learning in recognizing voices talking in their second language is modulated by their age of acquisition. Winters and colleagues (Winters *et al.*, 2008) studied whether native English listeners trained to recognize speakers either with German or English stimuli could generalize their knowledge and correctly identify the same speakers when they spoke the language the listeners had not been trained on (English or German, respectively). While native English listeners could identify the same listeners significantly better when they spoke in English than when they spoke in German, no differences were observed for listeners trained in German. The authors concluded that listeners made use of language-dependent indexical cues to identify speakers speaking a familiar language. These studies show that although there is enough language-independent information in the signal to successfully recognize speakers when the signal lacks linguistic information, indexical information to recognize voices is not language independent. Voice recognition performance is better when listeners are (more) familiar with the language being spoken.

The advantage of a familiar language in voice recognition can possibly be explained by listeners' understanding of what is being said, or, in other words, access to lexical information. If so, lexical frequency of words and lexical knowledge of listeners could potentially influence voice recognition. School children indeed have been found to show improved sensitivity to talker-differences in highly familiar words (Levi and Schwartz, 2009). Moreover, school children showed a larger "voice familiarity" effect (i.e., better word recognition performance for familiar than unfamiliar voices) in highly familiar words than in less familiar words (Levi, 2015a). Lexical frequency highly correlated with word familiarity ratings for those children. However, no effect of lexical frequency was found for voice learning and recognition by adult listeners (Winters *et al.*, 2008) or children (Levi and Schwartz, 2013; Levi, 2015b), suggesting that phonological rather than lexical knowledge plays the most important role in the perception of speaker information. This explanation was put forward by Perrachione and Wong (2007) who explained the better performance of their listeners in their native compared to an unfamiliar language by arguing that some degree of proficiency in the language is needed to gain access to inter-talker phonetic variability. Moreover, Perrachione and colleagues (Perrachione *et al.*, 2011) observed impaired performance of dyslexic listeners in voice recognition tasks, which was correlated with their degree of phonological impairment. They concluded that dyslexic listeners fail to learn speaker-specific representations of phonetic consistency which leads to the observed voice recognition problems. Creel and Jimenez (2012) offered a similar explanation after finding differences in voice learning between adults and pre-school children. They argued that pre-school children experience problems in encoding speaker information, since they are worse than adults in recognizing speech patterns and are still learning acoustic cues mapping to speakers' identity. In discussing larger voice familiarity effects in high-familiar words for children which was attributed to word frequency, Levi (2015a) suggested that children might be less sensitive to acoustic-phonetic details in the input when they are exposed to low-familiar words, while the lack of the lexical frequency effects in adults might be connected to the fact that even low-frequent words are not unfamiliar enough. At the same time, knowledge of the phonological structure alone cannot fully explain the findings in voice recognition studies. Linguistic similarity between familiar and unfamiliar languages does not seem to modulate the language familiarity effect: Chinese listeners identified German speakers better than Spanish listeners did and they even outperformed English listeners whose native language is (far) more phonologically related to German than Chinese is (Köster and Shiller, 1997). In our experiment, lexical frequency of the word was used as a stimulus-related factor, whereas lexical proficiency in the L2 language was used as a listener-related factor.

Apart from listeners' lexical knowledge and language experience, a number of other listener-related factors have been shown to influence voice recognition performance. Since listeners have to learn to recognize previously unfamiliar speakers, working memory capacity can potentially influence the degree of voice learning and accuracy of voice recognition. Working memory is associated with the short-term storage of incoming information and its manipulation (Levi, 2015b). Previous studies (Bregman and Creel, 2014; Levi, 2015b) indeed found a positive relation between the component of working memory termed Phonological Loop

(Baddeley, 1986; Baddeley and Hitch, 1974) responsible for short-term storage of auditory information, with the speed of voice learning in bilingual listeners (Bregman and Creel, 2014) and accuracy of voice recognition in native school age children (Levi, 2015b). At the same time, another component of working memory namely Central Executive, responsible for manipulating the upcoming information and divided attention and controlling the Phonological Loop, was found to negatively impact voice recognition performance of children on the last day of voice training (Levi, 2015b). Furthermore, individuals' phonological memory and awareness have also been found to play a role in a voice recognition. This pattern was observed for both dyslexic (Perrachione *et al.*, 2011) and non-dyslexic listeners (Jimenez, 2012). Perrachione and colleagues showed that voice recognition performance of dyslexic listeners correlated with their results on the Comprehensive Test of Phonological Processing (Torgesen *et al.*, 1999), which measures phonological awareness and phonological memory. A similar correlation of phonological processing and the ability to recognize voices was demonstrated in an experiment with non-dyslexic listeners (Jimenez, 2012). The contributions of working memory capacity and phonological aptitude were investigated as listener-related factors in the present study.

Given the observed role of phonological knowledge and memory and the mixed findings about the role of lexical knowledge in voice learning, recognition and discrimination, it is surprising that there is a lack of voice-learning studies with L2 listeners. Testing this group of listeners with their incomplete L2 lexical and phonological knowledge in comparison to native listeners of the language can provide new insights into the role of the factors influencing the encoding of voice information. L2 listeners are less familiar with the sound and lexical structure of their second than their first language. Moreover, languages differ in their, partly non-linguistic, acoustic parameters, which could be used for voice recognition (Johnson *et al.*, 2011). For instance, F0 has an overall wider range in English than in Dutch (Chen *et al.*, 2004; Collins and Mees, 1999). Bregman and Creel (2014) hypothesized that in order to discriminate talkers speaking in a particular language, listeners need to be familiar with talker-varying characteristics unique to that language. It is, therefore, possible that L2 listeners have some difficulty using acoustic cues to identify voices in a non-native language. Using L2 listeners may allow us to look deeper into the role of proficiency and lexical knowledge in voice learning and recognition. As suggested in a recent study by White and colleagues (White *et al.*, 2013), weak lexical representations result in the reduction of sensitivity to phonetic detail not only in children, but also in adults. If this is the case, L2 listeners would be less able to exploit acoustic-phonetic details in low-frequent than in high frequent words, resulting in lexical knowledge and lexical frequency effects in voice learning and recognition in the L2 language. There are only two voice learning studies with L2 listeners that we are aware of. Perrachione and Wong (2007) showed that Mandarin listeners residing in the U.S., and speaking predominantly English in their daily life, recognized voices speaking in Mandarin better than voices speaking in English at the beginning of a voice learning paradigm, but this difference disappeared in the latter sessions. Bregman and Creel (2014) found that the speed of learning to recognize a voice positively correlated with the age of acquisition of the second language. However, unlike in the current study, no direct measure of second language proficiency was used in these studies.

The aim of the present study is to investigate the role of speaker-, listener-, and stimulus-related factors in voice recognition and learning in L2 listeners. To that end, a group of Dutch participants was trained to recognize four previously unfamiliar voices speaking in British English during a four-day training period (similar to Nygaard and Pisoni (1998) who used a 10-day training period). The voice recognition accuracy and learning progress per day of the listeners was measured in relation to the speakers' voice characteristics (minimum, maximum, and average fundamental frequency, and average word length for each speaker) and stimulus characteristics (lexical frequency and the number of phonemes carrying indexical information). Moreover, participants' lexical knowledge was measured with the LexTALE test (Lemhöfer and Broersma, 2012), while their phonological aptitude was measured with the Llama-D test (Meara, 2005). The computerized variant of backward Digit Span (Wechsler, 2004) with visual presentation of the stimuli and written recall was used to assess the role of working memory capacity (namely, the Central Executive component of working memory) following previous studies (e.g., Alloway *et al.*, 2004; Bull *et al.*, 2008; Levi, 2015b; Rosenthal *et al.*, 2006).

II. METHOD

A. Experimental setup

The experiment contained four sessions divided over four consecutive days. Each session combined a training and a test phase. The procedure used in this experiment follows the methodology originally developed by Nygaard *et al.* (1998) which is applied and adapted in later studies to investigate voice learning and voice familiarity effects (e.g., Levi *et al.*, 2007, 2011; Levi, 2015b). Table I gives an overview of the tasks and the number of words included in each task on each day of the experiment.

B. Materials

Seventy-six mono- and 76 bisyllabic English nouns were chosen from the SUBTLEX-UK database (van Heuven *et al.*, 2014). Both bisyllabic and monosyllabic sets contained words of different frequencies (the distribution was relatively similar for both monosyllabic and bisyllabic words): from 1.02 per million for the lowest frequency word in the set (*sob*) to 589 per million for the highest frequency word in the set (*end*). All words were content words and were judged as familiar to the L2 participant's group by the authors.

The words were recorded by 12 native British male speakers, who at the time of the experiment, were living in or visiting the Netherlands. They came from different parts

TABLE I. Experimental setup on each training day. The number of words in each task is included in brackets.

Day	1	2	3	4
Tasks	Familiarization (24) Feedback (64) Test (64) Llama LexTALE	Feedback Test Digit Span	Feedback Test	Feedback Test
Duration of session (min)	45	30	30	30

of Great Britain, and were between the ages of 21 and 33. Table II presents for each speaker the average, standard deviation and range (minimum and maximum) of the F0 as measured by Praat (Boersma and Weenink, 2009) in Hz, and the average word length. Each speaker read the word list aloud twice (second time in the opposite order). The speakers were recorded individually in a sound-proof booth with a Sennheiser ME 64 microphone at a sampling frequency of 44100 Hz. Words which were mispronounced or produced too quietly were recorded again. The words were then excised from the resulting audio files using a Matlab (The MathWorks Inc., 2013) script, and the segmentations were subsequently manually checked using Praat (Boersma and Weenink, 2009). All speakers were rewarded 5 Euro for half an hour of recording time.

In order to study the general process of voice learning, as well as to be able to include individual characteristics of voices in the analysis, the listeners were trained on different sets of speakers. This is in contrast to previous studies which provided all listeners with the same set of speakers (e.g., Nygaard and Pisoni, 1998; Winters *et al.*, 2008). Listeners were trained to recognize four speakers from the set of 12 different speakers; however, Speaker 1 was the same for all participants. This was necessary for another experiment, which is not reported here. The other three speakers were chosen from the remaining 11 speakers, in different combinations. Eleven combinations (lists) were created (e.g., list 1: Speaker 1, 5, 8, 9; list 2: Speaker 1, 11, 7, 12, etc.). Speakers 2–12 occurred three times in all the lists in different positions.

TABLE II. F0 (Hz) and word length (ms) information for the speakers used in the study calculated on the basis of the 128 words included in the training and test phases of the study.

	F0 (Hz)				Word length (ms)	
Speaker	Mean	SD	Minimum	Maximum	Mean	SD
1	107	15	89	137	585	116
2	98	16	73	129	556	122
3	153	27	115	198	490	106
4	119	13	104	143	527	118
5	90	26	73	142	516	103
6	137	19	116	162	579	137
7	114	20	94	144	437	97
8	148	15	133	174	526	110
9	156	23	103	230	569	141
10	122	21	97	155	624	124
11	115	20	93	145	431	96
12	142	11	126	165	548	119

Twenty-four words from the voice learning set were used on the first session in a familiarity phase (12 monosyllabic and 12 bisyllabic words). The remaining 128 words were semi-randomly divided over the stimuli for the feedback and test session for each day of training, so that both the feedback and test phases contained an equal number of bisyllabic and monosyllabic words with comparable frequency. Following Nygaard and Pisoni (1998), the same words were used in the voice learning part of the experiment on each day, but the speaker producing the word varied per day (e.g., if the word is pronounced by Speaker 1 on day 1, it will be pronounced by Speaker 2 on day 2). Each listener heard each word of a particular speaker only once during the course of the experiment. Different from Nygaard and Pisoni (1998), the division of the words into the set used for the feedback and the test phases differed for each day to ensure generalization of learning. Moreover, two different orders of the stimuli presentation were used (i.e., the stimuli which were presented to half of the participants on the first day of the training, were presented to the other half of the participants on the fourth day of the training).

C. Participants

Forty-five (mean age = 22.5, SD = 2.4) native speakers of Dutch with no reported history of learning or hearing disorders were recruited from the Radboud University Nijmegen subject pool. All participants had a minimum of eight years of formal training in English and possessed a "VWO" (i.e., pre-university education) diploma, meaning a B2 or higher level of English according to the European Framework of Reference. Additionally, 16 participants (2 males, mean age = 21.9, SD = 2.8) took part in the pre-test of the stimuli and experimental setup. None of the participants who participated in the pre-test took part in the main experiment. All participants received study points or 30 Euro for their participation.

D. Procedure

All participants were tested individually in a quiet sound-attenuated booth. The stimuli were presented to them binaurally through headphones. The intensity level of all the stimuli was set at 70 dB sound pressure level (SPL). The experiment was administered with Presentation software.

1. Training

In the training phase on day 1, participants were first familiarized with the speakers. They heard a sequence of five words produced by each of the four speakers (different words for different speakers) followed by one word from each of the speakers, with the name of the speaker appearing on the screen. This procedure was repeated twice. The listeners' task was to memorize the name and the voice of the speaker. The task was self-paced and listeners had to press a

button when they were ready for the next word. The familiarization phase only occurred once, on day 1.

After the familiarization phase on day 1 and at the start of days 2–4, listeners had to complete the feedback phase of the task. In this phase, participants heard a word produced by one of the four speakers, and had to choose one name from the (earlier introduced) four names which were presented on the screen. If the choice was correct participants saw "correct" appearing on the screen; if a mistake was made, the correct name appeared on the screen. Participants were instructed to press one of the four buttons on the button box corresponding to the name of the speaker. They were told to react as quickly as possible, but at the same time to minimize mistakes. The position of the names on the screen changed each day to ensure deeper learning.

2. Test

The test phase was similar to the feedback phase but without any feedback on the answers. Participants listened to the word and again had to choose one name from the four names appearing on the screen and press the button on the button box corresponding to the name of the speaker they thought they just listened to. After the response was given, participants moved to the next word. Only the responses of participants from the test phase of the experiment were analyzed.

3. Cognitive tests

After completing the three voice learning tasks on day 1, participants had to perform the Llama and LexTALE tests. At the end of the second day of the training, participants completed the backward Digit Span task.

Llama-D test is part of a battery of language aptitude tests developed by _lognostics (Meara, 2005). The test measures the ability of the listeners to learn, recognize and discriminate phonological sequences (Meara, 2005; Granena, 2013). In the Llama-D test participants listen to a set of ten (non-)words in an unfamiliar language. Their task is to listen to the words carefully since in the second part they have to decide, by pressing one of two buttons, whether the word they hear then was already presented to them in the first part. Participants both gain points for correct responses and lose points if they make an error.

LexTALE (Lemhöfer and Broersma, 2012) is a visual unspeeded lexical decision task for advanced learners of English. Participants decide by pressing one of two buttons whether the word they see is an existing word in English. The test consists of 60 trials, designed to test vocabulary knowledge of medium to highly proficient L2 speakers of English.

In the backward Digit Span task (Wechsler, 2004), which measures working memory capacity, participants see a number of digits on a screen. The digit string increases with one digit every two trials, starting from two digits and ending with a sequence length of seven digits. Each sequence of digits is presented by consecutively showing the digits on the screen for one second with a one-second-interval before the next digit of the sequence is shown. The task of the participant is to memorize the sequence and type in the digits in reverse order. Following Neger *et al.* (2014) each participant was presented with all sequence lengths irrespective of their performance on earlier trials. The visual form of digit presentation was chosen over an auditory presentation since the visual backward Digit Span task is considered optimal in multilingual settings as it allows one to tease apart non-native proficiency of the listeners and their working memory capacity (Olsthoorn *et al.*, 2014).

III. RESULTS

Due to missing data on one of the training days, data from five participants had to be excluded from the analysis. Data from the remaining 40 participants were analyzed with mixed effects logistic regression analysis (Jaeger, 2008) in R (version 3.3.2). The analysis was conducted in several steps. First, we want to establish whether the L2 listeners managed to learn to recognize previously unfamiliar voices during the training. Responses of the participants in the test phase of the experiment were coded as 1 if the response of the participant corresponded to the correct name of the speaker, and 0 if the answer was wrong. Accuracy (whether the response was correct or incorrect), thus, served as a dependent variable in the analysis. To assess the effect of the day of the training on the number of times the speaker was recognized correctly, Day (1, 2, 3, 4) was included as a fixed factor, while Subject, Word, List (the combination of speakers that the participant listened to) and Speaker were included as random factors. Additionally, a by-Subject random slope for day was introduced to account for differences over time in voice learning caused by differences across participants.

In the second step of the analysis, the effect of the various speaker-, stimulus-, and listener-related factors was investigated by adding the three types of factors group-wise to the previous best model. Each factor was added as a main factor and in interaction with Day to investigate the contribution of this factor to participants' voice recognition performance and the speed of voice learning. For the speaker-related factors, a measure of predictability of each voice based on its acoustic characteristics (average, minimum and maximum F0 and average word length) was included in the analysis. As stimulus-related characteristics, the number of phonemes in the word carrying indexical information (Eatock and Mason, 1994; Gallardo *et al.*, 2015; Amino and Arai, 2008) and lexical frequency were added to the model. Finally, the results of the listeners on the language and cognitive tests were calculated and included in the statistical model as listener-related characteristics. The new model was compared to the previous best-fitting model to evaluate whether the new model explains significantly better the variation in the data than the previous model. After evaluating the model fit, the significance of the effects is established, following Snijders and Bosker (2012). All continuous variables were centered and scaled. All steps of the analysis are explained in more detail below.

A. Voice recognition and learning in L2 listeners

Figure 1 illustrates the voice recognition accuracy of the participants in the Test phases of the four training days. As

shown by Fig. 1, listeners demonstrated improvement. This observation was corroborated in the statistical analysis: day was a significant predictor of accuracy ($\beta = 0.223$, SE $= 0.038$, $p < 0.001$). The inclusion of the factor day to the model significantly improved the model fit [$\chi^2(1) = 24.72$, $p < 0.001$]. Participants thus managed to learn the target voices and to improve their recognition scores in four-day training.

B. Speaker-related factors in voice recognition

In order to investigate to what extent acoustic voice characteristics play a role in voice learning and recognition, a measure of predictability of each voice based on its acoustics only was computed. According to earlier studies (see Baumann and Belin, 2010, for an overview), fundamental frequency (F0) is an important parameter for making judgments about the similarity between voices. To that end, the average F0, as well as the minimum and maximum F0, and the average word length per speaker (see Table II) were submitted to a multinomial logistic regression in SPSS, with Speaker as the dependent variable. Based on the classification table output of this analysis, a new variable Predicted Accuracy was computed.

Predicted Accuracy is the percentage of times the voice was correctly predicted based on the F0 measures and the average word length. Since each participant was only trained on one specific list (the combination of four speakers selected from the set of 12 speakers), the predicted accuracy was calculated for each speaker in each list of four speakers separately. Predicted Accuracy averaged across all lists is shown in Fig. 1 (black square). As shown in Table III performance of the human listeners significantly correlated with the Predicted Accuracy ($p < 0.01$), with the highest correlation obtained for the third day of the training.

Average performance of the listeners on day 1 ($M = 67.83$, $SD = 17.92$) of the training was lower than the accuracy predicted on the basis of the acoustic parameters of the voice ($M = 74.38$, $SD = 14.22$), although this difference did not reach significance [$t(81.78) = -1.90$, $p = 0.06$]. On the last day of training, this difference reversed, but was again not significant [$M = 77.91$, $SD = 12.37$; $t(84.37) = 1.24$, $p = 0.22$]. These results seem to suggest that participants succeeded using F0 and average word length to recognize the different voices from day 1 onwards.

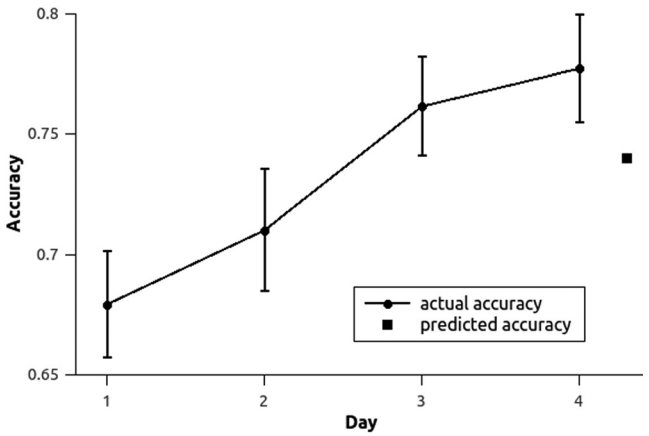

FIG. 1. Voice learning performance across four consecutive days of training, averaged over all speakers and listeners (solid line with bullets), with error bars, and predicted accuracy on the basis of a multinomial regression analysis (black square; see explanation in the section on speaker-related factors).

Predicted Accuracy for each speaker in a particular list was added as a factor to the overall analysis. The results showed that speakers that were recognized more accurately based on their acoustic characteristics by the computer were also recognized better by the human listeners ($\beta = 0.338$, $SE = 0.047$, $p < 0.001$) but at the same time Predicted Accuracy did not modify learning (the interaction Day * Predicted Accuracy was not significant). Moreover, the random factor List was excluded from the model, since it was no longer significantly improving model fit. The new model, including Predicted Accuracy and Day as fixed factors, Subject and Item as random factors (without List), and by Subject random slope for day was a significant improvement over the earlier model [$\chi^2(0) = 50.252$, $p < 0.001$].

C. Stimulus-related factors in voice recognition

To investigate the role of the amount of indexical information in the word and lexical frequency on voice recognition, an indexical information measure and the frequency value on the Zipf scale [log10(frequency per million words) + 3] from the British SUBTLEX-UK word frequency database (van Heuven *et al.*,2014) were added as fixed factors to the model of the previous subsection, as well as their interactions with the factor Day. We investigated three possible instantiations of the indexical information measure: number of syllables in the word (= number of vowels), number of nasals and vowels in the word (see the Introduction; Eatock and Mason, 1994; Gallardo *et al.*, 2015; Amino and Arai, 2008), and length of the word (= number of phonemes). To avoid multicollinearity, these factors were not included in the same model. Rather, the models including only one of the three factors were compared with one another to establish which of these factors accounted for more variation in the data. All three measures turned out to be significant predictors of general accuracy in voice recognition, and none of them interacted with the factor day. The model including the number of nasals and vowels had the lowest AIC (1637.4 against AIC = 1638.6 for the model including number of syllables as a predictor, and AIC = 1640.2 for the model including number of phonemes as a predictor). Therefore, the number of nasals and vowels in the word was included as a predictor in the subsequent analysis.

The results of the statistical analysis showed no significant effect for lexical frequency in voice recognition, neither

TABLE III. Strength of the correlation between average participants' accuracy of voice recognition on each training day and predicted accuracy.

	Participants' accuracy			
	Day 1	Day 2	Day 3	Day 4
Predicted accuracy	0.45	0.58	0.61	0.50

TABLE IV. Estimates of the best-fitting model to predict voice recognition accuracy including all significant speaker- and stimulus-related factors.

Factor	β	SE	p
Day	0.223	0.039	<0.001
Predicted accuracy	0.337	0.047	<0.001
Number of vowels and nasals	0.093	0.024	<0.001

as a main effect nor as an interaction effect with the factor Day. Inclusion of the factor Frequency in the model ($\beta = -0.021$, SE = 0.026, p = 0.425) did not significantly improve model fit [$\chi^2(1) = 0.611$, p = 0.435]. The best-fitting model, at this stage, therefore included day, Number of nasals and vowels (stimulus-related) and Predicted Accuracy (speaker-related) as fixed factors, Subject and Speaker as random factors, and a by-Subject random slope for day. The estimates of the fixed effects for this model are presented in Table IV. The stimulus-related random factor Word no longer contributed significantly to the model fit [$\chi^2(1) = 1.519$, p = 0.218] and was therefore removed.

D. Listener-related factors in voice recognition

Table V provides an overview of the scores for the two linguistic and one cognitive test, averaged over all listeners. The average score for LexTALE falls within the range of 60%–80% which corresponds to a B2 or upper-intermediate level of proficiency according to the Common European Framework of Reference (Lemhöfer and Broersma, 2012). The maximum possible score for Llama D test is 75, and the average score of the participants in the present study corresponds to an "average score" for this test as specified by Meara (2005). Following Neger *et al.* (2014), we measured percentage of correct trials in the backward Digit Span task rather than, e.g., the highest number of digit strings correctly reproduced, since some participants made errors on both trials with four digits, but reproduced trials with five or six digits correctly. Percentage of correctly reproduced trials was then a fairer measure of their performance.

To study the role of individual differences in linguistic and cognitive skills on voice learning, z-transformed scores for Llama, LexTALE, backward Digit Span and their interactions with the factor day were included in the best-fitting model from the previous subsection. Since the data for the LexTale test for one participant was missing, 39, rather than 40, participants were included in this analysis. The initial model for the analysis of the role of listener-related factors in voice recognition is presented in Table VI.

As can be seen in Table VI, neither the LexTALE nor the Llama score played a role in accurate voice recognition and learning from day 1 to day 4 of the training. Digit Span and its interaction with Day were however marginally

TABLE V. Participants' performance on the language and cognitive tests. Standard deviations are provided between brackets.

LexTALE	Llama	Digit Span
72.5 (15.7)	32.7 (16.9)	68.1 (19.5)

TABLE VI. Estimates of the initial model of voice recognition performance including the significant speaker- and stimulus-related factors and all individuals' linguistic and cognitive measures.

Factor	β	SE	p
Day	0.224	0.038	<0.001
Predicted accuracy	0.336	0.048	<0.001
Number of vowels and nasals	0.100	0.025	<0.001
Llama	−0.050	0.122	0.686
Digit Span	0.199	0.121	0.100
LexTALE	0.094	0.117	0.421
Day * Llama	0.023	0.039	0.547
Day * Digit Span	−0.072	0.038	0.059
Day * LexTALE	0.029	0.038	0.436

significant. After step-wise removal of all non-significant factors and interactions, this interaction remained marginally significant ($\beta = -0.004$, SE = 0.002, p = 0.071). Although the removal of the interaction Day * Digit Span from the model increased the AIC and log likelihood of the model, the difference between the models with and without the interaction Day * Digit Span did not reach significance [$\chi^2(1) = 3.131$, p = 0.077]. Removal of the factor Digit Span also did not significantly decrease model fit [$\chi^2(1) = 0.588$, p = 0.443]. These results seem to suggest that none of the participants' characteristics, such as lexical knowledge in the non-native language, phonetic aptitude, or Central Executive influenced voice recognition and learning.

The presence of the marginally significant interaction between Digit Span scores and Day might however indicate a different pattern of learning for the participants with high and low Digit Span scores. This hypothesis was further investigated by dividing the participants into two groups according to their Digit Span score (range 33.3%–100%; with 66.67% and higher belonging to the "high scores": 22 people). Figure 2 shows the voice recognition accuracy over the period of four training days for the participant group with the high backward Digit Span score and that of the participant group with the low backward Digit Span score.

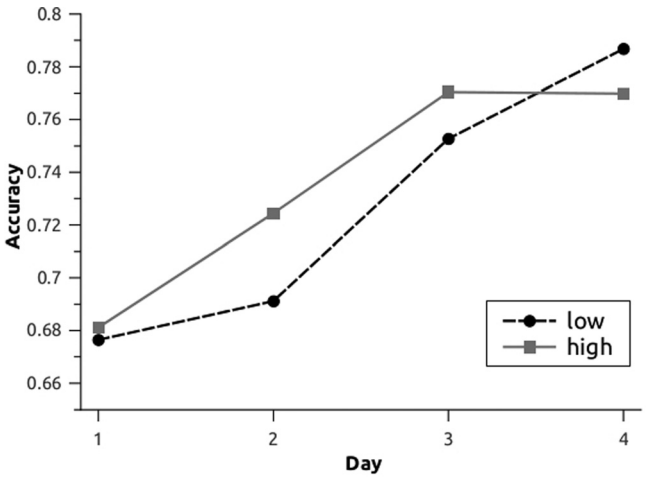

FIG. 2. The rate of improvement in voice recognition accuracy for the groups of participants with a higher and lower Digit Span score. Solid line with squares represents performance of the participants with high backward Digit Span score. Dashed line with filled circles represents performance of the participants with low backward Digit Span score.

Figure 2 seems to suggest that people with lower working memory capacity learn more slowly than people with better working memory capacity (note that the accuracy score on the test on day 1 was calculated after the familiarization and test phases thus after the first day of learning), but at the same time have more room for improvement. On the first days of learning, participants with a higher Digit Span score demonstrated a somewhat higher voice recognition accuracy than listeners with a lower Digit Span score, while on the last day this pattern reversed. Listeners with a higher Digit Span score seemed to have improved their recognition more after the second day of the training and did not show any improvement from the third to the fourth day of the training, while listeners with a lower Digit Span score improved the most on the third day of the training. It appears, therefore, that working memory capacity influences voice learning speed. Nevertheless, since removal of the interaction did not significantly decrease the model fit, we have to be careful in interpreting these results.

IV. DISCUSSION

The present study investigates the combined and separate contribution of speaker-, listener-, and stimulus-related factors to voice learning and voice recognition in L2 listening. The results suggest that voice characteristics of the speaker, expressed in average, minimum, and maximum F0, and the average word length in ms, as well as stimulus characteristics, i.e., the number of sounds in the word carrying indexical information contribute to L2 voice recognition performance. Interestingly, neither lexical frequency of an item nor lexical knowledge of the listeners (LexTALE score) played a role in voice learning and recognition performance. Moreover, no effect of phonological aptitude, expressed in the Llama test score, was observed. The results, however, seem to hint at a role of working memory capacity on the speed of learning to recognize voices.

Learning to recognize previously unfamiliar voices goes fast and seemingly effortlessly even for L2 listeners who arguably have less well developed lexical and phonological knowledge of the non-native language compared to native speakers of that language. The average voice recognition performance of the participants reached 68% correct already on the first day of the training, and significantly increased to 78% on the last day of the training. Factor Day (training) was a significant predictor of voice recognition accuracy in all the conducted analyses. The contribution of speaker-, stimulus-, and listener-related factors to the listeners' voice recognition and improvement in accuracy is discussed in detail below.

A. Speaker-related factors in voice recognition

The L2 listeners in the present study were shown to use the acoustics of the speakers' voices in voice recognition. There was a high correlation between the accuracy predicted by the multinomial logistic regression analysis for the speaker for each list and the accuracy of the listeners. Moreover, voices that were better recognized in the multinomial logistic regression were also better recognized by the listeners. When comparing the predicted scores obtained with the classification table from the multinomial logistic regression analysis and the scores obtained by the human listeners (see Fig. 1), we see that the accuracy scores obtained by the participants on the first day of training were lower than those of the regression analysis, but higher on the last day of the training. These results seem to point at an increase in listeners' sensitivity to speaker-specific acoustic characteristics due to the training. Voice learning thus seems to entail associating acoustic properties to particular voices, and doing so leads to higher recognition scores over time.

At the same time, listeners' better performance on day 4 compared to the score predicted by the multinomial logistic regression and the lower correlation between the predicted accuracy and the accuracy scores of the listeners on day 4 than on day 3 seem to indicate that listeners use additional sources, not only low-level acoustic properties of the speech signal, to learn and identify voices. This observation corroborates findings of previous studies, showing that listeners are able to identify voices based only on their acoustic properties (Winters *et al.*, 2008) and perform better if they are good at perceiving pitch differences (Xie and Myers, 2015) when the language is unfamiliar. However, when the language is familiar as in the case with the participants in the present study various additional, language-specific cues are used in voice recognition, while the ability to perceive differences in pitch no longer plays a role in voice recognition in a familiar language (Xie and Myers, 2015).

Speakers differed in how easy they were to recognize (see also Levi, 2015b, who found a significant effect of speaker in listeners' recognition accuracy). Between-speaker differences can for a substantial part be explained by the acoustic characteristics of their voices. Speaker 5 was recognized best by the listeners in the present study (the recognition accuracy reached 94%, already after the first training day). This speaker had the lowest F0 in the set of speakers (see Table II). Speaker 6, on the other hand, had the lowest accuracy score of all speakers on the last day (the recognition accuracy never got above 66%), but also had prototypical F0 measures. These findings are in line with the norm-based or prototype-based view on voice identity (Papcun *et al.*, 1989), which states that voices that are more distant from the prototype are easier to remember for listeners. This theory was further developed by Belin and colleagues (Baumann and Belin, 2010; Latinus and Belin, 2011; Yovel and Belin, 2013), who used a multi-dimensional voice space (with F0 being the primary dimension for voice-similarity judgments), in which all other voices are encoded relative to the prototypical voice. Not only voices more distant from the prototype are recognized and memorized better, they also induce greater neuronal activity in voice-sensitive cortex than more prototypical voices (Latinus *et al.*, 2013).

B. Stimulus-related factors in voice recognition

Vowels and nasals are the sounds that carry the largest amount of indexical information (Eatock and Mason, 1994). Consequently, we predicted that listeners would be more

accurate in recognizing a speaker's voice when the speaker produced a word containing a higher number of vowels and nasals. The accuracy of voice recognition was indeed found to be higher for these words. This effect of the phonetic content of the utterance (i.e., number of vowels and nasals) on listeners' voice recognition performance shows that phonological and speaker-specific information interacts in speech perception. Previous studies (see Amino and Arai, 2008 for an overview), demonstrated that speaker-specific physiological properties are reflected to a different extent in different speech sounds, which influences voice identification in native listening. Our results show that the sounds in the speech signal may enhance L2 voice recognition accuracy as well. Moreover, as suggested by Winters and colleagues (Winters *et al.*, 2008), when listening in a non-native language, vowel categories specific to the native language of the listeners might be used to a larger extent in voice recognition. Although not directly tested in the present study, this could be an interesting question for further research.

Previous results on the role of lexical frequency in voice recognition are unclear. We hypothesized that lexical frequency might play a role in voice recognition by L2 listeners since this group of listeners is assumed to have weak(er) lexical representations, of low-frequency words in particular, and therefore might be less able to exploit acoustic-phonetic details to successfully recognize voices, similar to the children in the studies by Levi and Schwartz (2009) and Levi (2015a). Lexical frequency, however, turned out not to play any role in L2 voice recognition and learning in the current experiment. This outcome corresponds however to the outcomes for native adults (Winters *et al.*, 2008; Levi and Schwartz, 2013). These results seem to suggest that lexical information is indeed not necessary for successful voice recognition (in line with Winters *et al.*, 2008; Levi and Schwartz, 2013). On the other hand, it is also possible, as suggested by Perrachione and Wong (2007), that in a more familiar language (or in more familiar words), listeners are better able to exploit phonological and acoustic cues to differentiate between voices. If we assume that L2 adults, similar to native children, are better able to exploit phonological information in high-frequent (highly familiar) than in low-frequent (low familiar) words (White *et al.*, 2013) then the absence of the effect of lexical frequency in the present study could be connected to the (relatively) high lexical knowledge of the L2 listeners and the small range but relatively high word frequencies for the materials used, so that even the "low-frequent" words were familiar to the listeners.

C. Listener-related factors in voice recognition

Different from previous voice-learning studies with L2 listeners, the present study included a measure of language proficiency (LexTALE) to investigate the role of lexical knowledge during voice learning and recognition in L2. The LexTALE score was not shown to modulate voice recognition accuracy of the listeners nor their learning over time. Hence, in the present study, lexical knowledge did not seem to play a role in voice recognition. At the same time, all listeners scored relatively high on the LexTALE test and the word frequency of the stimulus items was fairly high (see also Sec. IV B), which could have allowed them to successfully exploit acoustic-phonetic cues available in the stimuli for learning the voices.

Given the availability of phonological information in the signal and earlier findings of listeners being able to exploit it for voice recognition (Zarate *et al.*, 2015), as well as the hypothesis introduced by Levi (2015b) that learning L2 sound categories and voice learning are connected, it is perhaps surprising that no effect of phonological aptitude was observed in the present study. Previous studies that found a facilitating effect of phonological memory on voice recognition, however, did not use the Llama D task as used in this study, but either employed the Comprehensive Test of Phonological processing, including Memory for Digits and Non-Word repetition task (Perrachione *et al.*, 2011) or the auditory verbal forward Digit Span task (Bregman and Creel, 2014; Levi, 2015b) as a measure of phonological memory. The difference between the Llama D test and these measures is that while the verbal forward Digit Span and non-word repetition tasks tap into short-term phonological memory, Llama D taps into the recognition memory for phonological sequences and long-term knowledge of phonological regularities which results from that. Speciale *et al.* (2004) demonstrated that it is the combination of memory for phonological sequences (measured in our study with Llama D) and short-term phonological memory capacity (measured with forward Digit Span) that predicts both productive and perceptive L2 knowledge rather than short-term phonological memory alone, which implies that these two cognitive skills are not the same. Taken together, where individuals' short-term phonological memory seems important for learning voices, recognition memory for phonological sequences is not.

We observed a marginal, though interesting, effect of working memory capacity (more specifically, Central Executive: the ability of the listeners to simultaneously store and process information: Levi, 2015b) on voice learning, which suggests a connection between working memory capacity and voice learning speed. On the first days of learning, participants with a lower working memory span demonstrated a somewhat lower learning accuracy than listeners with a higher working memory span, while on the last day of the training, listeners with a lower backward Digit Span score outperformed those with a higher backward Digit Span score. Interestingly, similar to our finding, in the voice learning study by Levi (2015b) with school children, a significant negative effect of the score of backward Digit Span on the voice recognition performance was observed on the last day of training (in the general analysis and analysis of the performance on the first day of the training backward Digit Span did not reach significance). Levi (2015b) explained these results by suggesting that listeners with higher backward Digit Span scores used a different strategy in voice recognition and learning. The data from the current study however seem to suggest a difference in speed of learning rather than the use of a different strategy between listeners with larger and smaller working memory capacity. Since this result was only marginally significant, more research is needed to

determine the precise role (if any) of working memory in voice recognition of adult listeners.

The observation that both lexical knowledge and lexical frequency of items play no role in L2 voice recognition suggests that lexical information is not required for successful voice recognition and learning. This finding corresponds to the idea put forward by Mullennix *et al.* (1989) that information about speakers' voices is related to early perceptual processes, namely, the extraction of the acoustic-phonetic information from the speech. These processes occur at the prelexical, rather than the lexical level of speech processing (Andics, 2006). If the lexical level had been involved in voice recognition, the role of lexical frequency should have been observable, since the effect of lexical frequency occurs at the earliest moment of lexical access (Dahan *et al.*, 2001). This, therefore, implies that (L2) voice recognition operates at the prelexical level of processing. This finding is in line with other studies, showing that while access to phonological information facilitates voice recognition, lexical and semantic access are not necessary for successful recognition of voices (Perrachione *et al.*, 2011; Zarate *et al.*, 2015).

The present study is the first one to investigate the combined role of speaker-related, listener-related, and stimulus-related factors in voice recognition and learning by L2 listeners. We have shown that L2 listeners are able to learn to recognize speakers' voices while they are speaking in a language that is non-native though familiar to the listeners, and that speaker-related, listener-related, and stimulus-related factors have an accumulative effect on voice recognition. Voice recognition is better for speakers whose acoustic characteristics are more deviant from the prototype (e.g., the lowest F0 in the set), in words which contain more indexical information (words with the larger number of vowels and nasals), while working memory capacity seems to influence the speed with which listeners learn to associate acoustic properties with specific speakers.

ACKNOWLEDGMENTS

This research is supported by a Vidi-grant from the Netherlands Organization for Scientific Research (NWO; Grant Number 276-89-003) awarded to O.S.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Aldine, Chicago), p. 203.

Abrams, D. A., Chen, T., Odriozola, P., Cheng, K. M., Baker, A. E., Padmanabhan, A., and Menon, V. (**2016**). "Neural circuits underlying mother's voice perception predict social communication abilities in children," Proc. Natl. Acad. Sci. **113**, 6295–6300.

Alloway, T. P., Gathercole, S. E., Willis, C., and Adams, A. M. (**2004**). "A structural analysis of working memory and related cognitive skills in young children," J. Exp. Child Psychol. **87**, 85–106.

Amino, K., and Arai, T. (**2008**). "Differential effects of the phonemes on identification of previously unknown speakers," J. Acoust. Soc. Am. **123**, 3328–3335.

Amino, K., Arai, T., and Sugawara, T. (**2007**). "Effects of the phonological contents on perceptual speaker identification," Lect. Notes Comput. Sci. **4441**, 83–92.

Amino, K., Sugawara, T., and Arai, T. (**2006**). "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," Acoust. Sci. Technol. **27**, 233–235.

Andics, A. (**2006**). "Distinguishing between prelexical levels in speech perception: An adaptation-fMRI study," *Nijmegen CNS*, Vol. 1, pp. 47–66.

Andics, A., McQueen, J. M., and Van Turennout, M. (**2007**). "Phonetic content influences voice discriminability," in *16th International Congress of Phonetic Sciences (ICPhS 2007)*, pp. 1829–1832.

Baddeley, A. (**1986**). *Working Memory* (Clarendon, Oxford), p. 304.

Baddeley, A., and Hitch, G. (**1974**). "Working memory," Psychol. Learn. Motiv. **8**, 47–89.

Baumann, O., and Belin, P. (**2010**). "Perceptual scaling of voice identity: Common dimensions for different vowels and speakers," Psychol. Res. PRPF **74**, 110–120.

Boersma, P., and D. Weenink (**2009**). "Praat: doing phonetics by computer (version 5.1. 05) [computer program]" (Last viewed May 1, 2009).

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

Bregman, M. R., and Creel, S. C. (**2014**). "Gradient language dominance affects talker learning," Cognition **130**, 85–95.

Bricker, P. D., and Pruzansky, S. (**1966**). "Effects of stimulus content and duration on talker identification," J. Acoust. Soc. Am. **40**, 1441–1449.

Bull, R., Espy, K. A., and Wiebe, S. A. (**2008**). "Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years," Develop. Neuropsychol. **33**, 205–228.

Chen, A., Gussenhoven, C., and Rietveld, T. (**2004**). "Language-specificity in the perception of paralinguistic intonational meaning," Lang. Speech **47**, 311–349.

Collins, B., and Mees, I. M. (**1999**). *The Phonetics of English and Dutch* (Koninlijk, Leiden, The Netherlands), p. 363.

Creel, S. C., and Jimenez, S. R. (**2012**). "Differences in talker recognition by preschoolers and adults," J. Exp. Child Psychol. **113**, 487–509.

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (**2001**). "Time course of frequency effects in spoken-word recognition: Evidence from eye movements," Cognit. Psychol. **42**, 317–367.

Eatock, J. P., and Mason, J. S. (**1994**). "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Acoustics, Speech, and Signal Processing, 1994, ICASSP-94, 1994 IEEE International Conference*, Vol. 1, pp. 1–133.

Gallardo, L. F., Möller, S., and Wagner, M. (**2015**). "Importance of intelligible phonemes for human speaker recognition in different channel bandwidths," in *INTERSPEECH*, pp. 1047–1051.

Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (**1991**). "The role of language familiarity in voice identification," Mem. Cognit. **19**, 448–458.

Granena, G. (**2013**). "Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test," in *Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment*, edited by G. Granena and M. H. Long (John Benjamins, Amsterdam), pp. 105–129.

Jaeger, T. F. "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models," J. Mem. Lang. **59**(4), 434–446 (**2008**).

Jimenez, S. (**2012**). "The effect of language ability on talker identification," retrieved from http://www.cogsci.ucsd.edu/media/uploads/undergrad/honors-theses/sp12_honor_thesis_jimenez,_sofia.pdf.

Johnson, E. K., Westrek, E., Nazzi, T., and Cutler, A. (**2011**). "Infant ability to tell voices apart rests on language experience," Develop. Sci. **14**, 1002–1011.

Köster, O., and Schiller, N. O. (**1997**). "Different influences of the native language of a listener on speaker recognition," Foren. Ling. The Int. J. Speech, Lang. Law **4**, 18–28.

Latinus, M., and Belin, P. (**2011**). "Anti-voice adaptation suggests prototype-based coding of voice identity," Front. Psychol. **2**, 175–187.

Latinus, M., McAleer, P., Bestelmeyer, P. E., and Belin, P. (**2013**). "Norm-based coding of voice identity in human auditory cortex," Curr. Biol. **23**, 1075–1080.

Laver, J. D. (**1968**). "Voice quality and indexical information," Br. J. Disord. Commun. **3**, 43–54.

Lemhöfer, K., and Broersma, M. (**2012**). "Introducing LexTALE: A quick and valid lexical test for advanced learners of English," Behav. Res. Methods **44**, 325–343.

Levi, S. V. (**2015a**). "Talker familiarity and spoken word recognition in school-age children," J. Child Lang. **42**, 843–872.

Levi, S. V. (**2015b**). "Individual differences in learning talker categories: The role of working memory," Phonetica **71**, 201–226.

Levi, S. V., and Schwartz, R. G. (**2009**). "Perception of talker information by children with typical and impaired linguistic development," in *Proceedings of Meetings on Acoustics 157 ASA*, Vol. 6, No. 1, pp. 1–7.

Levi, S. V., and Schwartz, R. G. (**2013**). "The development of language-specific and language-independent talker processing," J. Speech Lang. Hear. Res. **56**, 913–920.
Levi, S. V., Winters, S. J., and Pisoni, D. B. (**2007**). "A cross-language familiar talker advantage?," Research on Speech Perception Progress Report No. 28, Speech Research Laboratory, Indiana University, Bloomington, IN, pp. 369–383.
Levi, S. V., Winters, S. J., and Pisoni, D. B. (**2011**). "Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible?," J. Acoust. Soc. Am. **130**, 4053–4062.
Meara, P. (**2005**). *LLAMA Language Aptitude Tests: The Manual* (Lognostics, Swansea, UK), p. 22.
Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**, 365–378.
Neger, T. M., Rietveld, T., and Janse, E. (**2014**). "Relationship between perceptual learning in speech, and statistical learning in younger and older adults," Front. Hum. Neurosci. **8**, 628–665.
Nygaard, L. C. (**2005**). "Perceptual integration of linguistic and nonlinguistic properties of speech," in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Oxford), pp. 390–413.
Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Atten., Percept. Psychophys. **60**, 355–376.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**, 42–46.
Olsthoorn, N. M., Andringa, S., and Hulstijn, J. H. (**2014**). "Visual and auditory digit-span performance in native and non-native speakers," Int. J. Bilingual. **18**, 663–673.
Owren, M. J., and Cardillo, G. C. (**2006**). "The relative roles of vowels and consonants in discriminating talker identity versus word meaning," J. Acoust. Soc. Am. **119**, 1727–1739.
Papcun, G., Kreiman, J., and Davis, A. (**1989**). "Long-term memory for unfamiliar voices," J. Acoust. Soc. Am. **85**, 913–925.
Perrachione, T. K., Del Tufo, S. N., and Gabrieli, J. D. (**2011**). "Human voice recognition depends on language ability," Science **333**, 595–595.
Perrachione, T. K., and Wong, P. C. (**2007**). "Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex," Neuropsychologia **45**, 1899–1910.
Rosenthal, E. N., Riccio, C. A., Gsanger, K. M., and Jarratt, K. P. (**2006**). "Digit Span components as predictors of attention problems and executive functioning in children," Arch. Clin. Neuropsychol. **21**(2), 131–139.
Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., and Remez, R. E. (**2002**). "Learning to recognize talkers from natural, sinewave, and reversed speech samples," J. Exp. Psychol.: Hum. Percept. Perform. **28**, 1447–1491.
Sidtis, D., and Kreiman, J. (**2012**). "In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication," Integr. Psychol. Behav. Sci. **46**, 146–159.
Snijders, T. A. B., and Bosker R. J. (**2012**). *Multilevel Analysis*: *An Introduction to Basic and Advanced Multilevel* Modeling (Sage, Thousand Oaks, CA), pp. 368.
Speciale, G., Ellis, N. C., and Bywater, T. (**2004**). "Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition," Appl. Psycholinguist. **25**, 293–321.
Torgesen, J., Wagner, R. K., and Rashotte, C. A. (**1999**). "Comprehensive test of phonological processing," *Pro-Ed*, Austin, TX.
Van Heuven, W. J., Mandera, P., Keuleers, E., and Brysbaert, M. (**2014**). "SUBTLEX-UK: A new and improved word frequency database for British English," Q. J. Exp. Psychol. **67**, 1176–1190.
Wechsler, D. (**2004**). *Wechsler Adult Intelligence Scale*, Dutch Version, 3rd ed. (Harcourt Test, Amsterdam).
White, K. S., Yee, E., Blumstein, S. E., and Morgan, J. L. (**2013**). "Adults show less sensitivity to phonetic detail in unfamiliar words, too," J. Mem. Lang. **68**, 362–378.
Winters, S. J., Levi, S. V., and Pisoni, D. B. (**2008**). "Identification and discrimination of bilingual talkers across languages," J. Acoust. Soc. Am. **123**, 4524–4538.
Xie, X., and Myers, E. (**2015**). "The impact of musical training and tone language experience on talker identification," J. Acoust. Soc. Am. **137**, 419–432.
Yovel, G., and Belin, P. (**2013**). "A unified coding strategy for processing faces and voices," Trends Cognit. Sci. **17**, 263–271.
Zarate, J. M., Tian, X., Woods, K. J., and Poeppel, D. (**2015**). "Multiple levels of linguistic and paralinguistic features contribute to voice recognition," Sci. Rep. **5**, 11475–11478.