LUNDS UNIVERSITET

DEPARTMENT OF PSYCHOLOGY

Recognition of Speakers from Oral Non-Linguistic and Linguistic Sounds

Louise Owetz

Master Thesis
Spring Term 2014

Supervisor: Prof. Geoffrey Patching

Abstract

Voice recognition plays an important role in human communication and there is increasing interest in so called ‘earwitness’ testimony in the courtroom. However, research on voice identification or earwitness identification by way of short linguistic as compared to non-linguistic sounds remains in its infancy. To address this issue, the present study set out to examine speaker recognition from short linguistic and non-linguistic vocal sounds. A total of 45 participants were tested individually on a binary choice experiment to assess their ability to identify a previously unfamiliar voice when two other voices acted as lures. Participants first completed a familiarization session in which they heard either a male or female target voice describe their apartment and read a poem. This was followed by a session with 18 sounds from the target voice and two lures; each sound was repeated three times. The process was then repeated, with the target voice from the opposite sex. There was a difference in recognition abilities between one, two, and three-syllable words, where two and three-syllable words improved voice recognition. Performance with non-linguistic sounds was generally worse than that with linguistic sounds, but participants were able to identify target voices on the bases of sounds made by sighing, pondering, and hocking.

Keywords: Linguistic sounds, non-linguistic sounds, voice recognition, familiar voice, unfamiliar voice, sex differences.

Recognition of Speakers from Oral Non-Linguistic and Linguistic Sounds

The human voice is a unique characteristic of people that conveys information through speech for social communication and ultimately survival. Non-verbal information, for example, a speaker's gender, emotional state, and identity, is conveyed through a person's voice and may be perceived in parallel to verbal information (Amino & Arai, 2007; Cook & Wilding, 1997; Kriegsteina, Egera, Kleinschmidt, & Giraud, 2003; Mullennix, Johnson, Topcu-Durgun, & Farnsworth, 1995; Nygaard & Pisoni, 1998). Neuropsychological findings indicate that a partial neuroanatomical difference exists between non-verbal and verbal processing, (Kriegsteina et al., 2003) and between linguistic (content) and non-linguistic speaker identity information (Wong, Nusbaum, & Small, 2004). However, the precise extent to which a speaker can be identified on the basis of short linguistic and non-linguistic sounds has yet to be fully determined.

During a lifetime individuals obtain much detailed information and knowledge about numerous speakers (Sheffert, Pisoni, Fellowes, & Remez, 2002). Nygaard, Sommers and Pisoni (1994) and Nygaard and Pisoni (1998) studied how a listener becomes familiar with a speaker's voice. Results showed that repeated or long exposure to a voice increases the listener's sensitivity to a speaker's vocal characteristics that consequently facilitate the ability to distinguish familiar and unfamiliar voices from one another.

Auditory recognition expertise signifies a person's ability to correctly and swiftly identify individual sound sources in a set of similar stimuli (Chartrand, Peretz, & Belin, 2008). Voice and speaker recognition or voice and speaker identification can therefore be defined as an individual's ability to identify one voice over other voices, and these definitions will be used as synonyms in this thesis.

In this study an objective is to identify the types of vocal information that can be used for human voice identification, in order to facilitate a better understanding of human communication and, ultimately, human perception and memory. One aim is to identify basic aspects of vocal cues used for voice recognition, by empirical investigation of precisely what kinds of vocal sounds can be used as identification cues for previously encoded voice patterns. A further aim is to determine if individuals are able to identify previously unfamiliar voices based on non-linguistic auditory cues such as breathing, laughter, screaming, sighing ("Aaa" sound), pondering ("Hmm" sound), or the "Ahem" sound of the speaker clearing their throat; here termed hocking.

Familiar and Unfamiliar Voices

What is the difference and implication of familiar and unfamiliar voices? The recognition of a familiar voice includes focusing on the attributes that a particular voice has. For instance, voices of family, friends, colleagues, acquaintances and some celebrities are considered familiar because these voices have been heard speaking many times. A voice is considered unfamiliar to a listener if it has never been heard before. Research findings suggest that recognition of familiar voices is a primary process. To illustrate, individuals with focal brain damage can recognise familiar voices but not discriminate between unfamiliar voices (Lancker, Cummings, Kreiman, & Dobkin, 1988; Lancker & Kreiman, 1987; Neuner & Schweinberger, 2000). The indication is, therefore, that recognition of voices need not be dependent on discrimination.

In addition, humans are more accurate at voice recognition when the heard voice is speaking a language that they comprehend (Goggin, Thompson, Strube, & Simental, 1991; Perrachione & Wong, 2007; Philippon, Cherryman, Bull, & Vrij, 2007). Goggin et al. (1991) propose that listeners employ schemata when identifying voices and these are language-based and, therefore, include all aspects of a language (syntax, phonology, lexicon). So differentiating a voice from another based on dialect or recognising a voice based on unique vocal characteristics may be aligned to two different mechanisms. Relations have been found between low identification difficulty and atypical phonetics, such as differences in dialect (Blatchford & Foulkes, 2006; Foulkes & Barron, 2000; Goggin et al., 1991). Blatchford and Foulkes's (2006) experiment investigated lay listeners' ability to identify familiar voices from shouted voice sounds, and if two words provide enough voice information for identification. They conducted two experiments with 14 undergraduate students who all knew each other. Correct identification of shouts and short word samples from a familiar voice were found to depend on the listener and on variations of the speaker's voice, with 81% correct identification for long utterances and 52% correct for shorter utterance. Blatchford and Foulkes argue that there is considerable variation in identification across listeners and individual voices, for instance level of familiarity and dialect. Therefore, the inconsistency in listeners' performance identifies a need for so called 'earwitnesses' in court cases to undergo formal assessment of their voice identification abilities (Blatchford & Foulkes, 2006). Voice familiarity may not increase voice identification abilities, and this should be brought to any relevant court case's attention.

How well do individuals distinguish between linguistic and non-linguistic sounds if the sound originates from a familiar speaker? Participants in Yarmey's (2004) study listened

to tape-recorded words and non-linguistic sounds and were asked to indicate whether the sounds where made by a familiar or unfamiliar speaker. If the sound was identified as being from a familiar voice, participants were requested to name the speaker. Familiar and unfamiliar speakers produced utterances such as, 'help me', 'hello', and other non-linguistic sounds such as, a short scream, hocking, cough, moan, grunts, and laughter. The short verbal utterances of – 'hello' and 'help me' made by familiar speakers were correctly identified 50% of the time, and the results of Yarmey's (2004) study indicate that listeners find it more difficult to detect familiar voices compared to unfamiliar voices when the vocalisations were moans, sighs, grunts, and coughs. The findings also suggest that laughter can be better identified as coming from a familiar and unfamiliar speaker as compared to screaming and sighing. On the basis that listeners found it more difficult to categorise non-linguistic vocalisations as familiar compared to unfamiliar, Yarmey (2004) suggested that a stranger's voice is more unique compared to a familiar voice and therefore facilitates detection. However, Yarmey (2004) did not compare directly non-linguistic information with linguistic information. Consequently, on the basis of this study alone, it is impossible to know whether there really are any differences in the listener's ability to identify non-linguistic and linguistic sounds as familiar or unfamiliar.

More recently, Amino and Arai (2007) performed a listening speaker identification test to study effects of speaker-listener familiarity and what sounds are effective for speaker identification. In particular they wanted to determine if characteristics of nasal one-syllable sounds could be found in identification of previously unfamiliar speakers. A total of four speakers out of ten available from a previous voice identification experiment (Amino, Sugawara, & Arai, 2006) were used. These four speakers were selected based on typical fundamental frequencies, and 16 novice listeners participated in the listening speaker identification test. Participants listened to the speakers uttering one sentence each in the familiarization session and they could listen until they felt confident that they could recognise all the unfamiliar speakers, this training took on average 15 minutes. The speakers' names were provided to participants beforehand and speakers were identified on an answer sheet after each trial (Amino & Arai, 2007). Amino and Arai (2007) compared these results with their previous study (Amino et al., 2006) of four familiar speakers and five participants. Their findings showed better performance for voice identification of familiar speakers, compared to unfamiliar speakers. It was found nasality is one crucial speaker attribute participants use for voice recognition of both familiar and unfamiliar speakers. Effects of stimulus content were comparable between familiar and unfamiliar listeners. In contrast to Yarmey (2004), Amino

and Arai (2007) argued that individual voice characteristics facilitate familiarity and consequently correct identification of familiar voices. Likewise, Blatchford and Foulkes's (2006) findings highlight individual differences such as dialect differences on participants' ability to recognise and identify a speaker's voice. Nonetheless, overall effects of stimulus content were found to be similar between familiar and unfamiliar listeners. Cognitive processes may be responsible for listeners differing abilities in identifying familiar and unfamiliar voices. One suggestion (cf., Amino & Arai, 2007; Yarmey, Yarmey, Yarmey, & Parliament, 2001) is that pattern recognition is involved in the identification of familiar speakers, while identification of unfamiliar speakers depends more on feature analysis. Consequently, because feature analysis is more complex to execute, familiarity provides more precise identification (Yarmey et al., 2001). In summary, it appears that recognition of familiar voices on the basis of short linguistic and non-linguistic sounds is better than correct identification of sounds as coming from unfamiliar voices, although there maybe some exceptions to this rule (e.g., Yarmey, 2004).

Phillippon, Randall and Cherryman (2013) studied the effect on voice identification performance with the presence of verbal and non-verbal information, i.e., laughter with two target voices and six lures. Phillippon et al. found that participants' performance is superior if they are exposed to speakers who were talking and laughing than either talking or laughing alone. This supports the idea that laughter is an important facet that facilitates discrimination between voices because variability exist in voices and laughter (Philippon et al., 2013; Yarmey, 2004). According to Phillippon et al. (2007) laughter often constitutes a part of daily conversations because it appears to be a natural part of speech. Consequently, other non-linguistic sounds such as sighing "Aaa" sounds, pondering "Hmm" sounds, and hocking "Ahem" sounds, which are similarly part of everyday conversations, may also facilitate voice recognition of familiar from unfamiliar voices.

Speaker Attributes

Precisely, what auditory information about a speaker's voice is required for subsequent recognition of that voice? People appear to perform well at recognising a repeated word (Luce & Lyons, 1998; Palmeri, Goldinger, & Pisoni, 1993; Sheffert & Fowler, 1995), especially when the speaker is the same on the first and second occasion (Sheffert et al., 2002). But does this extend to correct recognition of the speaker? Participants in Nygaard, Sommers and Pisoni's (1994) experiment were trained over nine days to recognise a group of 10 voices from one-syllable words. The 19 people in the experimental group were asked to

identify words by a fixed number of speakers at four signal-to-noise ratios, while the 19 people in the control group identified the same words but from different speakers. The majority of participants improved their recognition ability over the nine days from one-syllable words, suggesting that information from one-syllable words is sufficient for voice recognition. Nygaard et al. (1994) claim that this study was the first to demonstrate "that experience identifying a talker's voice facilitates perceptual processing of the phonetic content of that speaker's novel utterances" (p. 44). It appears that knowledge of particular features of a speaker's voice can influence phonetic perception and recognition of spoken words. Familiarity with acoustic characteristics of a speaker's voice seems to assist and be related to the analysis and recognition of spoken words. Likewise the study expresses how long-term memory is involved in speech perception and recognition of spoken words (Nygaard et al., 1994). In sum, the study conducted by Nygaard et al. (1994) highlights the impact speaker variability has on an individual's ability to correctly identify different voices when based on linguistic sounds. However, Nygaard et al. (1994) focused exclusively on linguistic sounds and did not test non-linguistic sounds, so the precise extent to which their findings can be generalized to non-linguistic sounds remains to be determined.

It is of further relevance to question what speaker attributes make a voice memorable? Yarmey (1991) suggests that differences between speakers, such as differences in age, rate of speech, and fundamental frequency measures (pitch, period, length), along with the similarity /dissimilarity of these voice attributes to other voices, can facilitates or hinder listeners voice recognition. Cook and Wilding (2001) highlight that short speech recordings will be less representative or may even lack important information about the speaker's speech range and attributes. Nevertheless, Hollien (2002) argue that short linguistic utterances or non-linguistic sounds can be enough for a listener to identify a speaker. On these grounds, the present study includes short speech recordings, such as one-syllable words and screams to determine precisely how much linguistic or non-linguistic information is required for recognition of an earlier heard speaker.

Memory Structures

Longer processing time and more ways to encode speaker-specific properties can result in more unique episodic representations (Armony, Chochol, Fecteau, & Belin, 2007; Goldinger, Pisoni, & Logan, 1991). Studies like Armony et al. (2007) used a two-stimulus discrimination task and demonstrated that individuals are superior in their memory performance concerning emotionally laden sounds like laughing and crying, when compared

with more emotionally neutral sound of yawning. These findings strengthen the notion that episodic memory is reinforced by auditory emotional expression, and subsequently raises questions about the role other types of non-linguistic sounds on voice recognition.

Roebuck and Wilding (1993) highlight the difference observed for memory of a voice heard once before and a more familiar voice, and the role of memory for the actual voice and the spoken words. More vowel words increased identification abilities but longer sentences showed no effect. Nevertheless, the number of male and female distracter voices were seven, giving a total of 14 distracters per trial, so it is likely that interference effects were present (Roebuck & Wilding, 1993). Contrary to Roebuck and Wilding (1993), Cook and Wilding's (1997) experiment demonstrated an effect of length of utterances rather than improved memory because of variety in sentences used for a once-heard voice. Their experiments were similar to that of forensic situations where witnesses attend a line-up of voices days after hearing a perpetrators voice. It is therefore interesting to investigate what effect length has on voice recognition for a previously unfamiliar voice. Consequently, one, two, and three-syllable words are included in the current study as well as non-linguistic sounds. In addition, free speech and the reading of a poem are the two ways participants are able to encode speaker specific properties with the aim of achieving more unique episodic representations (after Goldinger et al., 1991).

Sex Differences

It is of further relevance to question whether any effects of sex exist in voice recognition abilities. Individuals perceive the sex of a voices based on acoustic factors such as fundamental frequency and breathiness (Klatt & Klatt, 1990). According to Mullennix (1995) the sex of a voice is not stored in abstract separate representations for males and females but rather the sex of a voice is stored in auditory-based perceptual representations, which contain information concerning acoustic voice parameters that are related to the sex of the individual voice. However, synthetic speech was used in this study and so the findings may therefore not generalise to natural voices (Mullennix et al., 1995). Therefore, the current study includes a familiarization session where, in one part, male or female target speakers speak freely so that natural speech can be captured. In Roebuck and Wilding's (1993) study, male voices were more easily recognised than female voices and a same sex interaction was found, whereas Cook and Wilding's (1997) findings did not show any differences in men and women's ability to discriminate male and female speakers. Nevertheless, women were somewhat better at identifying female speakers (Cook & Wilding, 1997). To investigate sex differences further

both male and female voices are included in the present study and the interaction between the sex of the listener and the sex of the speaker in listeners' recognition of the speaker's voice is examined.

Musicians

Further questions remain as to whether musicians are superior to non-musicians at recognising and identifying voices. Not much is known about auditory experts beyond musicians who have been studied extensively (Chartrand et al., 2008; Cohen, Evans, Horowitz, & Wolfe, 2011). Musicians are viewed as auditory experts because they use exclusive sound information in order to recognise sound sources at a minor level (Chartrand et al., 2008). Cohen et al. (2011) demonstrated by comparing musicians and non-musicians auditory and visual memory that musicians have better auditory memory compared to non-musicians. Research suggests that musicians rely on different and more complex encoding techniques as compared to non-musicians who most often rely on one encoding strategy (Williamson, Baddeley, & Hitch, 2010). On these grounds it is argued that musical training may improve cognitive function such as speech perception (Parbery-Clark, Skoe, Lam, & Kraus, 2009) and analytical listening abilities (Oxenham, Fligor, Mason, & Kidd, 2003). Therefore, information about participant's musical training was also collected, and analysed in relation to their voice abilities, in the present study.

Pilot Study

In the first instance a pilot study was conducted in which participants were presented with a binary choice experiment to identify a target voice. First, a familiarization session was conducted whereby one female speaker read the poem "Nordanvinden". Next, participants heard either the target female voice or a female distracter voice speak one, two, or three-syllable words taken from the poem and were required to answer whether the words were uttered by the target voice or not. The two female target and distracter voices available were randomised and vocal sounds consisted of 324 sound trials, (108 stimuli x 3 trials). One voice distracted the identification of the other voice recording. It was found that voice familiarity and number of syllables in a word significantly increases voice recognition abilities. In short, the findings of this pilot study provide an initial indication that familiar voices can be discriminated from unfamiliar voices on the basis of one, two, and three-syllable words.

Experiment

The present study builds on the aforementioned pilot study by the inclusion of non-linguistic vocal sounds, such as breathing, laughter, screams, sighing "Aaa" sounds, pondering "Hmm" sounds, and "Ahem" clearing of the throat sounds (hocking). In addition, the present study incorporates both male and female speakers and participants, and includes measures of participants' musical experience, and auditory imagery. As compared to earlier studies, a major advance is assessment of the effect of six different oral non-linguistic sounds not presented in the familiarization session but only in the test phase, and six oral linguistic sounds presented in both the familiarization session and the test phase.

On the basis of earlier studies (Amino & Arai, 2007; Cook & Wilding, 1997; Nygaard et al., 1994; Roebuck & Wilding, 1993; Yarmey, 2004) it is hypothesised that voice information obtained from words increases recognition abilities, and that the number of syllables will increase voice recognition. For example, two and three-syllable words increases recognition abilities more than one-syllable words. Second, it is hypothesised that voice recognition will change as a function of vocal utterances in the target voice. Specifically, the study will investigate whether vocal sounds like one, two and three syllable words, breathing, laughter, screams, sighing "Aaa" sounds, pondering "Hmm" sounds, and hocking "Ahem" sounds, are positively or negatively related to voice recognition (target voice) and voice recognition abilities. This will shed new light on whether more vocal information provides the listener with a wider speech range that consequently facilitates recognition, and what vocal sounds are most useful for voice recognition. Third, it is hypothesised that there is a same sex interaction where individuals are superior at identifying speakers of their own sex, and finally, on the basis of work conducted by (Chartrand et al., 2008; Cohen et al., 2011; Parbery-Clark et al., 2009; Williamson et al., 2010) suggesting musicians are superior in auditory recognition compared to non-musicians, it is hypothesised that there is a relation between years sung in a choir or years played an instrument or clarity of auditory imagery, and voice recognition abilities.

Method

Participants. Forty-five participants—25 men and 20 women — between the ages 19 and 45 years (mean 26.6 and 26.85 years respectively)—took part in the Experiment. All participants were recruited from Lund University's student population by way of poster and email advertisement and received two lottery tickets for their participation. All claimed to be fluent Swedish speakers, and none reported any hearing problems. 16 participants reported

that they had sung in a choir (mean 1.17 years) and 26 participants reported that they regularly played an instrument (mean 3.8 years).

Apparatus. A microcomputer (Fujitsu Esprimo Mobile M9410, Fujitsu Limited, Tokyo, Japan) running MATLAB (The MathWorks, Inc.) controlled the experiment. Auditory stimulus presentation and timing were controlled using the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997). Sound Devices 788T digital audio, a NEXUS Brüel & Kjær type 2690 A 0S4 conditioning amplifier, a binaural head and torso simulator Brüel & Kjær type 4100 with two with two microphones type 4190 and two pre-amplifiers type 2669, were used to record the stimuli. The stimuli were digitally edited using Audacity 2.0.5 (Audacity Development Team, http://audacity.sourceforge.net/) and Sound Forge 8.0 software (Sony Corporation, Tokyo, Japan).

Auditory stimuli were presented through Beyerdynamic DT 231 PRO headphones (Beyerdynamic GmbH & Co. KG, Heilbronn, Germany), adjusted to fit each participant comfortably. All participants were tested individually in a quiet testing room and responded by way of two numeric response keys, '0' and '1', positioned on the upper row of the computer keyboard, placed at a convenient distance from participants.

Stimuli. The stimuli were six linguistic and six non-linguistic sounds as made by nine women between the ages 18 to 33 years (mean 24.5 years) and nine men between the ages 20 to 43 years of age (mean 29.2 years). All speakers were recruited from Stockholm University's student population, and all were resident in Stockholm at the time of the recording. Speakers received two movie vouchers for their contribution to the current study. The recording were made in a soundproof room at the Department of Psychology, Stockholm University. Speakers were requested to sit on a chair facing the binaural head and torso simulator with their mouth 110cm distance from the microphone. The vocal recordings consisted of the following recordings, one take of the description of the speaker's home, three takes of reading the poem "Nordanvinden", and recordings of the non-linguistic sounds, deep breathing - both inhaling and exhaling from the mouth and nose, laughter, screams, sighing "Aaa" sounds, pondering "Hmm" sounds, and hocking "Ahem" sounds. All sounds were repeated five times during one recording and recorded once. A mixture of one, two, and three-syllable words, were selected from the description of participants' home, and likewise from one take of the poem "Nordanvinden". Two out of five takes in each vocal sound file were selected for experimental use. A recording of one out of five deep breathing, laughter,

screams, sighing “Aaa” sounds, pondering “Hmm” sounds, and hocking “Ahem” sounds - were also selected from the 18 speakers’ voice recordings. This resulted in 12 stimuli per voice and a total of 216 stimuli for the 18 voices.

Design. The experiment employed a 12 x 2 x 2 within participant and between participants design. One independent variable was the voice stimuli, six linguistic (one, two and three-syllable words) and six non-linguistic sounds (deep breathing, laughter, scream, sighing “Aaa” sound, pondering “Hmm” sound, and hocking “Ahem” sound), and was used within-participants. The second independent within-participant variable was the sex of the speaker (male, female). The between participants factor was sex of the listener. The dependent variable was correct answers, measured in percentage by correctly answered questions to stimuli (in signal detection terms, hit rate). Voices were randomised so that each participant received a new mix of three male and three female voices, one target voice and two distracters per sex, resulting in six vocal sound trials with two different target voices (Three trials per target voice). Voices were randomly selected from 18 voices available. The task on each trial was to respond with a ‘Yes’ or ‘No’ answer on the keyboard whether the sound belonged to the target voice they previously heard or not.

Clarity of Auditory Imaginary Scale. The auditory imaginary scale developed by Willander and Baraldi (2010) was used to determine whether individuals ability to clearly imagine auditory information is linked to their voice recognition skills. This scale comprises 16 items that participants were requested to rank on a 5-point scale, on the basis of the following guidelines. “Imagine the sounds listed below one at the time. Subjectively, how clearly do you hear the sounds of . . .(1 = not at all; 5 = very clear)”. Willander and Baraldi report Cronbach’s alpha = .88, which they claim is satisfactory. In the current study the Cronbach’s alpha = .90.

Data Collection Procedures. In the first instance, participants were informed that the study is about voice recognition and that their task is to learn two voices, one male and one female voice. All participants were informed that their data would be used anonymously and confidentially, and that they were under no obligation to take part in the study and could withdraw at any time, without prejudice, if they should wish. After which, all participants voluntarily consented in writing to take part in the experiment.

Participants first listened to one out of two familiarization sessions, i.e., one of the speakers describing their apartment and reading the poem “Nordanvinden”. The session continued with three trials of intermixed verbal and non-verbal sounds before the second familiarization session began with the opposite sex describing their apartment, and reading the poem “Nordanvinden” following which the three new trials of each sound started.

During the experimental trial, participants were asked after the first session to take off their headphones and put them back on. This was asked because incoming sound may differ depending on how the headphones are positioned, which was not of interest in the present study and considered a random factor. The aforementioned process was repeated with all participants.

Data analyses. Signal detection theory (SDT) was adopted to assess participants’ sensitivity to the auditory signals in terms of *d*–prime (*d´*). As detailed by Stanislaw and Todorov (1999; Green & Swets, 1966), *d´* provides a bias free measure of participants’ ability to distinguish between signals and distracters. Here, a value of 0 indicates an inability to distinguish signals from distracters and, for each participant and sound, a value of 1.19 indicates perfect performance (i.e., all signals correctly identified as signals and all distracters correctly identified as distracters). These data were submitted to a within participant analysis of variance (ANOVA), which revealed a statistically significant effect of stimulus conditions. By convention, an alpha level of .05 was used to infer statistical significance, and confidence intervals are reported in line with recommendations made by Cumming (2013). Post-hoc comparisons were then conducted using Tukey HSD tests to explore differences in performance, as assessed in terms of *d´,* between stimulus conditions. Further, correlational analysis failed to show any statistically significant relations between *d´,* and musical experience. IBM SPSS version 21 was used to analyse the data.

Results

In the first instance, sensitivity in terms of *d´* — defined as $d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$ — was calculated for each auditory signal and each participant separately. Following the procedures as detailed by Stanislow and Todorov (1999; Macmillian & Kaplan, 1985) rates of 0 were replaced with $0.5 / n$ and rates of 1 were replaced with $(n – 0.5)/n$, where n was the number of signal or distracter trials as appropriate. Figure 1 provides a summary representation of these data.

The data summarized in Figure 1 were submitted to a repeated measures ANOVA with two within-participant factors (stimulus type; sex of speaker [male, female]) and one between-participants factor (listener [male, female]). This analysis revealed a main effect of stimulus type, Wilks' Lambda = .24, F (11, 33) = 9.41, $p < .001$, partial $\eta^2 = .76$. Tukey HSD tests were conducted to examine the main effect of stimulus type in detail. These analyses showed that the two three-syllable words and the two two-syllable words were recognised better than all other stimulus types, namely one-syllable words, sighing "Aaa" sound, pondering "Hmm" sound, hocking "Ahem" sound, breathing, scream, and laughter (mean d' respectively, .934, 95% CI[.776, 1.092]; .708, CI[.475, .942]; .816, CI[.623, 1.009]; and .705, CI[.507, .903]; vs. .448, CI[.258, .639]; .484, CI[.314, .654]; .453, CI [.249, .656]; .373, CI[.217, .528]; .018, CI[-.123, .159]; .342, CI[.137, .548]; $p < .05$). In addition, these analyses showed that non-verbal screams, breathing and for male voices also laughter sounds were recognised worse than all other auditory stimulus types ($p < .05$). The main effect of the sex of the speaker failed to reveal a statistically significant F-value, Wilks' Lambda = .99, F (1, 43) = .081, p = .778, partial $\eta^2 = .002$, as did the main effect of sex of the listener F (1, 43) = 3.758, p = .06, partial $\eta^2 = .08$, interaction between sex of the speaker and sex of the listener, Wilks' Lambda = .996, F (1, 43) = .189, p = .666, partial $\eta^2 = .004$, interaction between stimulus type and sex of the listener, Wilks' Lambda = .665, F (11, 33) = 1.158, p = .150, partial $\eta^2 = .345$, and three way interaction between sex of the speaker, stimulus type and sex of the listener Wilks' Lambda = .799, F (11, 33) = .757, p = .678, partial $\eta^2 = .201$.

Figure 1 about here

Relations between voice recognition performance in terms of d', musical experience, and auditory imagery were assessed by way of Pearson product-moment correlation coefficients. These analyses failed to show a statistically significant correlation between the number of years each participant claimed to play an instrument and voice recognition performance, r = .062, n = 45, p = .686, and failed to show a statistically significance correlation between the number of years each participants reported to have sung in a choir and voice recognition performance, r = -.034, n = 45, p = .826. In similar vein, relations between scores on the clarity of auditory imagery scale and voice recognition performance showed a small negative, but not statistically significant relation, r = -.121, n = 45, p = .430.

Discussion

This study investigated voice recognition abilities from oral linguistic and non-linguistic sounds by means of a binary choice experiment. Participants were tested to assess their ability to recognise correctly a previously unfamiliar voice from vocal sounds when two other voices acted as distracters. The findings support the first hypothesis; there was a difference in recognition performance between one, two, and three-syllable words, where two and three-syllable words increased voice recognition performance. This can be compared to Roebuck and Wilding's (1993) study where the number of vowels increased identification of previously heard voices. On these grounds the present study conforms to the view that increasing the amount of vocal information, on which participants can base their judgement, facilitates voice recognition. This is consistent with Amino and Arais's (2007) findings where voice familiarity facilitated voice recognition and their proposition that individual voice characteristics facilitate familiarity and consequently correct identification. Similarly, participants in Nygaard et al.'s (1994) study improved their performance of recognising one-syllable words over nine days training. Participants in the current study did not receive such extensive repetition of words yet still performed well on the recognition task with two and three-syllable words. The indication is, therefore, that additional vocal information such as speaker attributes is provided with more syllables, despite little training, and this is enough to facilitate voice perception and subsequently voice recognition.

The second hypothesis was likewise supported. Voice recognition performance did change significantly with stimulus type where two and three-syllable words were the easiest to recognise and a reasonable effect size of stimulus type was obtained. Participants were sensitive to all non-linguistic sounds with the exception of screams, and for male voices breathing and laughter. This is contrary to Armony et al.'s (2007) findings where participants showed superior ability in emotionally laden sounds like laughing and crying. It is also contrary to Yarmey's (2004) findings where participants showed superior performance for laughter sounds. Similarly, Phillippon et al. (2013) found that performance was better when participants were exposed to speech and laughter than either alone. In the current study, participants were not exposed to laughter in the familiarization session but only during the subsequent test phase; the first time participants heard the target voice utter the non-linguistic sounds were in the recognition task. Consequently, superior performance with the linguistic as compared to non-linguistic sounds may have arisen simply because participants heard the linguistic, but not the non-linguistic sounds, at the beginning of each session when the speaker read the poem and freely described their home. In this respect, participants poor

performance with the vocal sounds of screams, and for male voices breathing and laughter, which failed to reach conventional levels of statistical significance in terms of unbiased recognition ability, suggests that voice familiarity may not increase voice recognition abilities on all levels. In light of evidence (Philippon, Cherryman, Vrij, & Bull, 2008; Yarmey, 2004) that voice familiarity may not facilitate voice recognition on the basis of short screams, breathing, and laughter Yarmey (2004) and Philippon et al. (2008) suggest that the justice system should not rely solely on earwitness testimonies as such reliance could result in unjust outcomes.

This study agrees with Philippon et al. (2013) who highlight that research and findings on non-linguistic sounds can offer valuable information to the criminal justice system that can be used when estimating the validity of earwitness identification inclusion in line-ups. Blatchford and Foulkes's (2006) research is likewise valuable and suggest caution in regards to external validity as it highlights that even friends may not correctly identify another friend's voice. Blatchford and Foulkes recorded their speakers in a soundproof studio and failed to obtain statistically significant evidence to suggest that participants could recognise their friend's voice from short sound clips. The present study, lends tentative support to the view that the justice system should be reasonably sceptical about earwitness testimonies, especially when those testimonies are based on the recognition of speakers from short non-linguistic cues.

The third hypothesis was not supported; no statistically significant sex differences or interactions were found in voice recognition abilities. This is in line with Cook and Wilding (1997). However, it is in contrast to Roebuck and Wilding's (1993) findings, and in contrast to the same sex interaction of Cook and Wilding (1997). Moreover, the present study failed to show statistically relations between musical ability and voice recognition abilities. More specifically, years sung in a choir and years played an instrument did not relate statistically to voice recognition abilities. A possible explanation for the outcome of both the third hypothesis concerning sex differences and the fourth hypothesis regarding musicians and non-musicians could be that there was not sufficient variability between participants. Restriction of range was an issue, as all participants were from a student population and of similar ages. To overcome this, future research should aim to include men and women with a greater age range and from nonstudent populations in order to observe any sex differences in the results. It may also be that for the fourth hypothesis, self-reported level of musical experience was not entirely accurate, because participants may have inflated their self-reported musical experience for socially desirable reasons. Future research may benefit from testing

participants' level of musical ability rather than merely collecting self-reported information.

Limitations and Implications

The current study has contributed to the gap in the research literature regarding information about where the limits for voice recognition lie. The current study measured what it intended to measure, non-linguistic and linguistic voice recognition abilities. It is unlikely that any threats to internal validity such as practice effect or fatigue were present due to the use of two different target voices and the short number of trials in the experiment. Neither is it likely that there are any extraneous threats to the reliability of the study as all participants were assigned to the same condition with the same instructions.

As highlighted by Perrachione and Wong (2007), speech and voice perception has societal consequences beyond that of a increased understanding of the human auditory cortex. Successful and pragmatic speaker identification systems is an objective for electrical and computer engineering due to its applicability to security and intelligence matters (Perrachione & Wong, 2007). At this date the accuracy of speaker identifications systems are not ideal and it has been suggested that adding phonetic variation that exist between talkers may be advantageous (Li & Espy-Wislon, 2004; Perrachione & Wong, 2007). To obtain this phonetic information linguistic knowledge is required (Perrachione & Wong, 2007). Such information is in addition beneficial in the field of forensic voice identification, as demonstrated by Perrachione and Wong. Likewise, employment decisions in the fields of forensics and crime prevention where voice identification plays an important role may need to consider the importance of language proficiency (Perrachione & Wong, 2007). The ability to recognise voices on the basis of non-linguistic sounds has barely been examined in earwitness research. However, the results of the present study provide some initial clues about exactly what kind of vocal cues can facilitate speaker recognition and ultimately contribute to better evaluation of earwitness testimonies and improved speaker identification performance.

Future Directions

In order to fully understand voice recognition, it is of interest to investigate where and when a voice becomes familiar, how much voice information is required and how the process of becoming familiar with a voice operates. Future research would benefit by replicating the current research with a more diverse sample, especially if the objective is to investigate developmental aspects of voice recognition. For instance inclusion of participants of various ages, i.e., young, adults, and elderly, would enable detailed examination of whether age

influences any of the variables while difference in hearing ability and impairments could be controlled statistically.

Sheffert and Olson (2004) demonstrated that novel words by familiar speakers are likely to be retrieved from long-term episodic memory, and suggest that effects of familiarity is robust over time. Therefore, future research may benefit by retesting participants on a second day or even a week to see to what extent linguistic and non-linguistic voice information has transferred to long-term memory.

In line with previous research showing differences in voice recognition abilities depending on language and dialect of both the speaker and listener (Foulkes & Barron, 2000; Goggin et al., 1991; Perrachione & Wong, 2007; Wong et al., 2004), it is of interest to investigate whether recognition and discrimination are two independent abilities. Future research could compare monolingual and bilingual individuals' voice recognition abilities in a language that they do not comprehend, to examine the role of voice characteristics with and without semantic word understanding.

Conclusion

In conclusion, the current study investigated voice recognition on the basis of oral linguistic and non-linguistic sounds. This research contributes to a greater understanding of human perception and memory by demonstrating that an effect of stimulus type, such that the number of syllables in words increases voice recognition abilities. Similarly, it was found that participants are sensitive to the oral non-linguistic sounds of, sighing "Aaa" sounds, pondering "Hmm" sounds, hocking "Ahem" sounds, and also breathing and laughter with the exception of male voices, and this supports the notion that voice familiarity increases voice recognition. The present research has important implication for understanding human perception – how well individuals can recognise other individuals based on their voice. More digitalised ways of operating our daily lives and our society are developing and may depend more on various voice operating and voice identification systems, and as a result, additional voice recognition research is warranted. An extended understanding of voice recognition abilities and the precise perceptual and cognitive mechanisms involved can enhance the development of speech technologies in various fields including automatic speaker recognition systems and assist court cases that rely on auditory evidence to prevent incorrect speaker identifications and unjust court outcomes.

References

Amino, K., & Arai, T. (2007). Effects of stimulus contents and speaker familiarity on perceptual speaker identification. *Acoustical Science & Technology, 28*(2), 128-130. doi: 10.1250/ast.28.128

Amino, K., Sugawara, T., & Arai, T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical Science & Technology, 27*(4), 2006. doi: 10.1250/ast.27.233

Armony, J. L., Chochol, C., Fecteau, S., & Belin, P. (2007). Laugh (or cry) and you will be remembered. *Psychological Science 18*(12), 1027-1029 doi: 10.1111/j.1467-9280.2007.02019.x

Blatchford, H., & Foulkes, P. (2006). Identi!cation of voices in shouting. *The Journal of Speech, Language, and the Law, 13*(2), 241-254. doi: 10.1558/ijsll.2006.13.2.241

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433-436. doi: 10.1163/156856897X00357

Chartrand, J. P., Peretz, I., & Belin, P. (2008). Auditory recognition expertise and domain specificity. *Brain Research, 1220,* 191-198. doi: 10.1016/j.brainres.2008.01.014

Cohen, M. A., Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). Auditory and visual memory in musicians and nonmusicians. *Pyschonomic Bulletin & Review, 18,* 586-591. doi: 10.3758/s13423-011-0074-0

Cook, S., & Wilding, J. (1997). Earwitness testimony: never mind the variety, hear the length. *Applied Cognitive Psychololgy, 11,* 95-111. doi: 10.1002/(SICI)1099-0720(199704)11:2<95::AID-ACP429>3.0.CO;2-O

Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the face overshadowing effect. *British Journal of Psychology, 92,* 617-629. doi: 10.1348/000712601162374

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. doi: 10.1177/0956797613504966

Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics, 7*(2), 180-198.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*(5), 448-458. doi: 10.3758/BF03199567

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(1), 152-162. doi: 10.1037/0278-7393.17.1.152

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York, NY: John Wiley and Sons.

Hollien, H. (2002). *Forensic voice identification.* San Diego, CA: Academic Press.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers *The Journal of the Acoustic Society of America, 87,* 820-857. doi: /10.1121/1.398894

Kriegsteina, V. K., Egera, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*(1), 48-55. doi: 10.1016/S0926-6410(03)00079-X

Lancker, D. V., Cummings, J., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia, A dissociation between familiar and unfamiliar voices. *Cortex, 24*(2), 195-209. doi: 10.1016/S0010-9452(88)80029-7

Lancker, D. V., & Kreiman, J. (1987). Unfamiliar voice discrimination and familiar voice recognition are independent and unordered abilities. *Neuropsychologia, 25,* 829-834. doi: 10.1121/1.2023449

Li, G., & Espy-Wislon, C. (2004). A novel dynamic acoustical model for speaker verification *The Journal of the Acoustical Society of America, 115,* 2428. doi: 10.1121/1.4781439

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition, 24*(4), 708-715. doi: 10.3758/BF03211391

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98,* 185-199. doi: apa.org/journals/bul/98/1/185

Miller, J. (1996). The sampling distribution of d'. *Perception & Psychophysics, 58*(1), 65-72. doi: 10.3758/BF03205476

Mullennix, J. W., Johnson, K. A., Topcu-Durgun, M., & Farnsworth, L. M. (1995). The perceptual representation of voice gender. *Journal of the Acoustical Society of America, 98*(6), 3080-3095. doi: 10.1121/1.413832

Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition, 44,* 342-366. doi: 10.1006/brcg.1999.1196

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355-376. doi: 10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*(1), 42-46. doi: 10.1111/j.1467-9280.1994.tb00612.x

Oxenham, A. J., Fligor, B. J., Mason, C. R., & Kidd, G. J. (2003). Informational masking and musical training. *The Journal of the Acoustical Society of America, 114*, 1543-1549. doi: 10.1121/1.1598197

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(2), 309-328. doi: 10.1037/0278-7393.19.2.309

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musicians enhancement for speech-in-noise. *Ear and Hearing, 30*(6), 653-661. doi: 10.1097/AUD.0b013e3181b412e9

Pelli, D. G. (1997). The video toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442. doi: 10.1163/156856897X00366

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia, 45*, 1899-1910. doi: 10.1016/j.neuropsychologia.2006.11.015

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology, 21*, 539-550. doi: 10.1002/acp.1296

Philippon, A. C., Cherryman, J., Vrij, A., & Bull, R. (2008). Why is my voice so easily recognized in identity parades? Influence of first impression on voice identification. *Psychiatry, Psychology and Law, 1*, 70-77. doi: 10.1080/13218710701873999

Philippon, A. C., Randall, L. M., & Cherryman, J. (2013). The impact of laughter in earwitness identification performance. *Psychiatry, Psychology and Law, 20*(6), 887-898. doi: 10.1080/13218719.2013.768194

Roebuck, R., & Wilding, J. (1993). Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology, 7*, 475-481. doi: 10.1002/acp.2350070603

Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language, 35*(2), 665-685. doi: 10.1006/jmla.1995.1030

Sheffert, S. M., & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics, 66*(2), 352-362. doi: 10.3758/BF03194884

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave and reversed speech sample. *Journal of Experimental Psychology: Human Perception and Performance, 28*(6), 1447-1469. doi: 10.1.1.103.27

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149. doi: 10.3758/BF03207704

Willander, J., & Baraldi, S. (2010). Development of a new clarity of auditory imagery scale. *Behavior Research Methods, 42*(3), 785-790. doi: 10.3758/BRM.42.3.785

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory & Cognition, 38*(2), 163-175. doi: 10.3758/MC.38.2.163

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*(7), 1-13. doi: 10.1016/j.neuropsychologia.2006.11.015

Yarmey, A. D. (1991). Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Science Society, 31*(4), 421-428. doi: 10.1016/S0015-7368(91)73183-6

Yarmey, A. D. (2004). Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations. *International Journal of Speech, Language and the Law, 11*(2), 268-277.

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology, 15*, 283-299. doi: 10.1002/acp.702

Author note

Louise Owetz, Department of Psychology, Lund University, Sweden.

The research reported here was supported by the Department of Psychology, Lund University. I thank Geoffrey R. Patching for discussion and supervision of the work. In addition, I specially thank Jesper Alvarsson and Artin Arshamian for technical support. Last but not least I am thankful to Holly Knapton and Lisa Espinosa for encouragement and support throughout.

Correspondence regarding this work may be addressed to Louise Owetz at the Department of Psychology, Lund University, Sweden. E-mail: louiseowetz@hotmail.com.

Figure 1

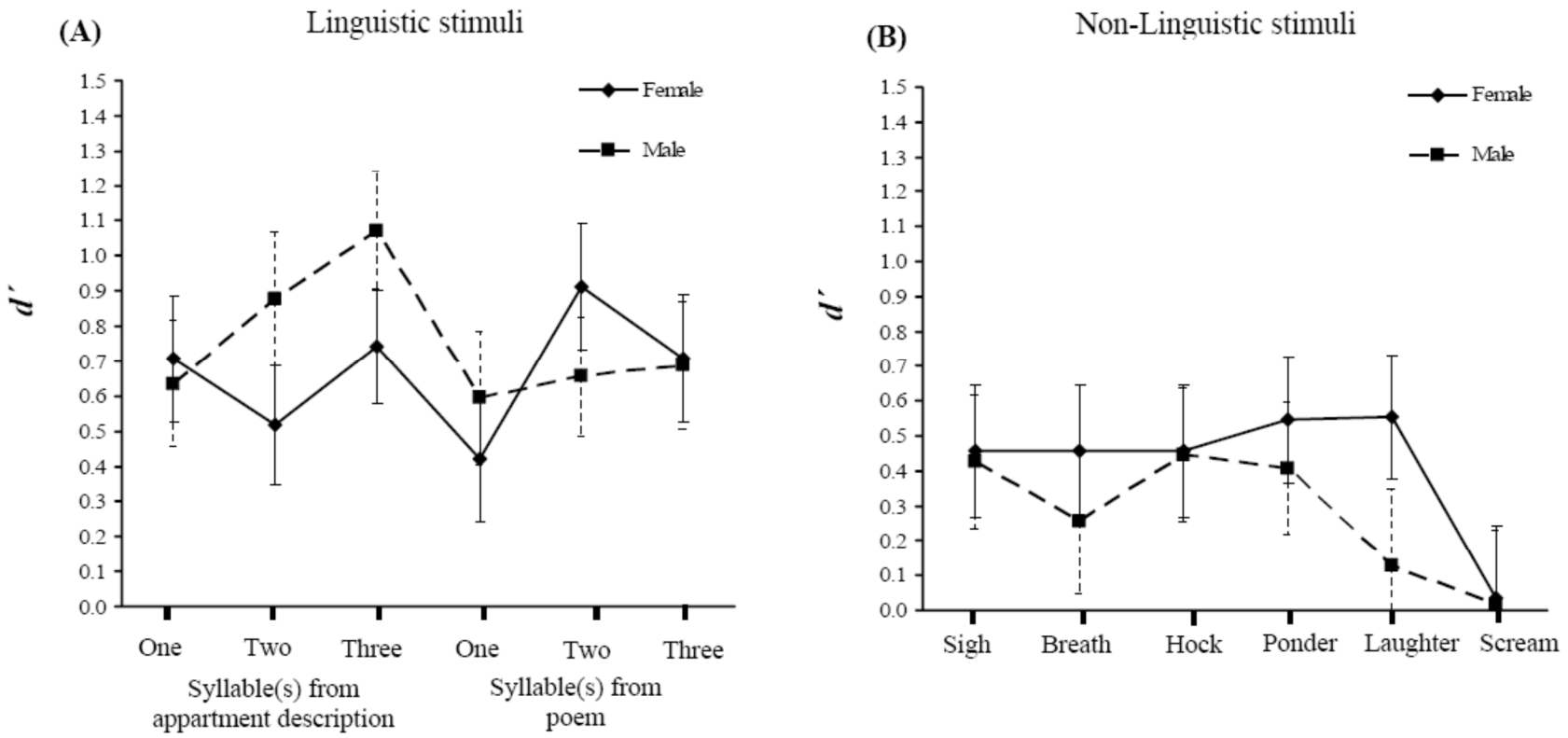

Figure 1. Mean of *d'* for male and female voices for (A) each of the 6 linguistic sounds and (B) each of the 6 non-linguistic sounds. Error bars indicate the 95% confidence intervals following the procedures described by Miller (1997), using Gourevitch and Galanter's (1967) variance approximation formula.