Assessing susceptibility to distraction along the vocal processing hierarchy

Muhl, Constanze; Bestelmeyer, Patricia

Quarterly Journal of Experimental Psychology

DOI:
10.1177/1747021818807183

Published: 01/07/2019

Peer reviewed version

Assessing susceptibility to distraction along the vocal processing hierarchy

Short title: *Voice perception ability tests*

Constanze Mühl[a*] & Patricia E.G. Bestelmeyer[a]

[a]School of Psychology, Bangor University, LL57 2AS, UK

[*]Correspondence:
Constanze Mühl
School of Psychology
Bangor University
Brigantia Building
Penrallt Road
Bangor, Gwynedd
LL57 2AS
UK
E-mail: constanze.muhl@bangor.ac.uk
Phone: +44 (0)1248 383254

Abstract

Recent models of voice perception propose a hierarchy of steps leading from a more general, “low-level” acoustic analysis of the voice signal to a voice-specific, “higher-level” analysis. We aimed to engage two of these stages: First, a more general detection task in which voices had to be identified amidst environmental sounds, and, second, a more voice-specific task requiring a same/different decision about unfamiliar speaker pairs (Bangor Voice Matching Test, BVMT). We explored how vulnerable voice recognition is to interfering distractor voices, and whether performance on the aforementioned tasks could predict resistance against such interference. Additionally, we manipulated the similarity of distractor voices to explore the impact of distractor similarity on recognition accuracy. We found moderate correlations between voice detection ability and resistance to distraction (r = .44), and BVMT and resistance to distraction (r = .57). A hierarchical regression revealed both tasks as significant predictors of the ability to tolerate distractors (R^2 = .36). The first stage of the regression (BVMT as sole predictor) already explained 32% of the variance. Descriptively, the “higher-level” BVMT was a better predictor (β = .47) than the more general detection task (β = .25), although further analysis revealed no significant difference between both beta weights. Furthermore, distractor similarity did not affect performance on the distractor task. Overall, our findings suggest the possibility to target specific stages of the voice perception process. This could help explore different stages of voice perception and their contributions to specific auditory abilities, possibly also in forensic and clinical settings.

Keywords: voice perception, voice detection, voice recognition

Successful social interaction relies on our capacity to extract relevant information from our surroundings and the people with whom we are interacting. While there is an extensive amount of research into the perception of such cues from faces, the perception of these cues from voices has been neglected until recently (Blank, Wieland, & von Kriegstein, 2014; Gainotti, 2014). Theoretical models of voice perception closely follow those already established for face perception but have received little empirical evaluation. Belin and colleagues suggest a voice perception model adapted from Bruce and Young's (1986) model of familiar face perception (Belin, Fecteau, & Bédard, 2004). This voice perception model proposes that after an initial low-level analysis of the voice signal, a number of different independent modules are responsible for the analysis of vocal speech, vocal affect, and speaker identity information, before additional semantic knowledge about a person is accessed through the activation of Person Identity Nodes (Belin et al., 2004; Campanella & Belin, 2007). This proposal suggests that the independent levels and modules can be investigated separately.

An alternative model by Kreiman and Sidtis (2013) suggests that the recognition process for voices relies simultaneously on the Gestalt perception of the whole (pattern recognition) and the analysis of specific auditory cues within the voice (feature analysis). The degree to which both are engaged depends on the familiarity of the voices. Recognition of unfamiliar voices calls for the extraction of features more than for an overall pattern recognition, possibly also involving comparison to a known "average" voice, and is more stimulus-driven. Familiar voice recognition is more top-down in that it relies heavily on the overall voice pattern, with only voice-identity specific features becoming salient throughout recognition. As such, recognising an unfamiliar voice is a question of discriminating and matching two voice signals, and is therefore often described as the ability of voice discrimination. Recognising a familiar voice, in contrast, is the recognition of an overall vocal pattern specific to a single person. The term "voice recognition" therefore often applies to the recognition of voice identity for familiar speakers in particular (see also van Lancker &

Kreiman, 1987). Furthermore, a recent neuroimaging study with lesion patients (Roswandowitz, Kappes, Obrig, & von Kriegstein, 2018) has also found that different brain structures are involved in the perception of newly-learnt unfamiliar vs. familiar voices, which supports this distinction.

Although Kreiman and Sidtis' model does not indicate independent feature-specific modules (e.g. for vocal affect perception) like Belin and colleagues' model does, it nevertheless posits the involvement of several distinct brain regions. Tasks related to voice perception therefore recruit the distributed areas that are relevant for solving a specific task. Findings of distributed time scales, for example in vocal affect perception (Iredale, Rushby, McDonald, Dimoska-Di Marco, Swift, 2013; see also model for vocal affect processing by Schirmer & Kotz, 2006, and Bestelmeyer et al., 2014), suggests that voice perception involves hierarchical stages. According to these, earlier stages represent more general analyses, and in the case of unfamiliar voices possibly also more stimulus-driven analyses, before voices are processed in a more abstract, integrative manner (e.g. Warren, Jennings, & Griffiths, 2005; Schirmer & Kotz, 2006).

The need for research on this topic, and indeed support for the existence of different independent voice perception modules, becomes more apparent when surveying the diversity of clinical symptoms reported for individuals with phonagnosia, or an impairment in voice perception. For example, an extensive study of patients with brain lesions revealed that while most patients with voice recognition deficits (in this case the recognition of famous familiar voices) were still able to discriminate between two different unfamiliar voices, one of the patients showed an impairment in both (Neuner & Schweinberger, 2000). However, in this sample no further tests were reported to see whether other domains of voice perception like the perception of gender or affect were selectively impaired as well. In recent years, cases of individuals with developmental phonagnosia have emerged. To assess the extent of their voice recognition deficits, these individuals often complete a number of voice perception tests that target specific voice perception abilities. Usually, only certain functions of voice perception are impaired (e.g. identity perception), while others like gender perception remain intact (see also the first reported case of developmental phonagnosia in Garrido et al., 2009). Both acquired and developmental voice perception deficits underline the need for a more in-depth assessment of possible singular processing stages in order to establish the range of functions that can be selectively impaired.

Apart from clinical contexts and the focus on general perception mechanisms, voice identity perception has also received attention in non-clinical contexts, particularly in the field of forensic psychology. As Kreiman and Sidtis (2013) point out, recognising an unfamiliar person by voice alone is not a task we often encounter in natural settings, yet witnesses to a crime might only be exposed to a perpetrator's voice. The reliability of witness testimony therefore depends on a witness's ability to extract identity information from a typically unfamiliar voice (i.e. process and compare the features of that voice to a stored representation of average voices) and store this information for the newly heard voice. Then, at a later point, the witness needs to distinguish the initial target voice from other unfamiliar voices (all of which require the same processing steps), and match it to its correct target at a later voice line-up. In terms of Belin and colleagues' more general model of possible distinct modules, this forensic line-up task requires structural encoding of the perpetrator's voice beyond just low-level auditory processing. Ideally, identity-specific features of the target voices also have to be accessible at a later time point to allow for correct identification of the perpetrator. This process is, of course, prone to error (Legge, Grosmann, & Pieper, 1984; Yarmey, 1995), and studies on it are often tailored to match specific criminal cases, making connections to existing, more general voice perception literature difficult (Kreiman & Sidtis, 2013).

Despite the ecological validity of such voice line-up tasks, more controlled, lab-based experiments are necessary. A recent study by Stevenage and colleagues (2013) explored the detrimental impact of interference on speaker perception. Listeners heard an unfamiliar speaker articulating a single sentence. In a fixed 16 s interval, participants then heard either nothing, or two or four distractor voices. This was followed by a test voice. Participants had to decide whether this test voice was identical to the initial target voice or not. Accuracy on this task was reduced as soon as any distractor voice was introduced. The detrimental effect distractors had on overall task performance occurred both when the distracting voices were similar (as defined by same speaker sex as target voice) or different (opposite speaker sex).

Our aim for the current study was, on the one hand, to test two potentially separate abilities that occur at different stages of voice perception. On the other hand, we also wanted to explore their impact on a third, complex auditory task that has been used previously and in more ecologically valid contexts. The aforementioned potentially separate abilities are first, the ability to detect voices as a discrete class of sound objects (voice detection ability), and, second, the ability to determine whether two utterances were spoken by the same speaker or not (voice matching ability). To investigate whether both are suitable to determine the accuracy on a more complex auditory task, we chose a distractor task examining how vulnerable or susceptible someone is to the interference of a distracting voice. This third task follows the example of voice perception tasks common in forensic contexts (same/different decisions about a voice that one had previously been exposed to, following interfering information). However, for the current study this takes place within a lab-based environment, allowing for stricter control of voice variables. For this reason, we also wanted to revisit the issue of distractor similarity, i.e. whether distractors that are either similar or different from the initial target voice affect the accuracy of one's same/different decision.

Voices are arguably the most salient sound in our environment. Although there is some debate about the timescale of this development, several studies have reported that infants already show preferential brain activation patterns for vocal sounds within the first twelve months after birth (e.g. Blasi et al., 2011; Grossman, 2011; Cheng, Lee, Chen, Wang, & Decety, 2012). Additionally, lesions studies have shown that voices are processed independently of other object sounds (Peretz et al., 1994; Neuner & Schweinberger, 2000).

As such, the detection of voices should be part of the earlier processing stream of vocal sounds (as described in Belin and colleagues' model). In our study we aimed to measure participants' ability to detect voices in an ongoing stream of vocal and non-vocal sounds. This task was inspired by a visual detection task for faces to investigate an individual with severe face recognition impairments (prosopagnosia; Duchaine, Yovel, Butterworth, & Nakayama, 2006). Our task was adapted to address the inherent differences between the visual domain (faces) and the analysis of auditory information as it unfolds over time. While Duchaine and colleagues embedded their target stimuli (faces) in a noisy background, we chose an ongoing stream of auditory, undistorted stimuli.

To examine a later module of voice perception, we included the Bangor Voice Matching Test (BVMT; Mühl, Sheil, Jarutytė, & Bestelmeyer, 2017). This task involves listening to two different utterances and then deciding whether these stem from the same or different speakers. It thereby requires the extraction of identity information from a voice before making a same/different judgment. Belin et al.'s (2004) model proposes that voice identity cues are processed after the structural configuration of a voice has been extracted. In contrast, Kreiman and Sidtis' (2013) model proposes that for this particular task, participants have to extract the features of both unfamiliar voices and then compare these to a template of an average voice.

Both the voice detection task and the BVMT will be examined in conjunction with the performance on a third task, a voice distractor task. Here, participants have to make an old/new judgment following initial exposure to a target voice. Crucially, a distractor voice is introduced between hearing the first target voice and the same/different judgment needed for the second target voice. We propose that the complexity of this distractor task should require both of the processing stages we aim to tap into using the detection task and the BVMT. The voice detection task depends on an earlier perception stage in which the signal is processed as a vocal (as opposed to a non-vocal) sound. The BVMT, on the other hand, requires a more complex analysis of the vocal signal. In fact, we assume that the BVMT and the distractor task require the extraction of the same kind of vocal cues (voice identity information/feature-based processing and comparison to an average voice). This reflects the proposed succession of voice perception modules in Belin and colleagues' model (2004). We therefore predict that both the voice detection task and the BVMT should correlate with the distractor task as they all rely on the analysis of a sound as a vocal object, but that the correlation with the BVMT should be higher. In order to complete the distractor task accurately, both an intact ability to detect voices and an intact ability to extract identity cues from voices are necessary. We therefore also expect that performance in the voice detection task and in the BVMT will both be predictors for the performance in the distractor task. However, given the proposed similar, later processing stages necessary for the BVMT and distractor task, we assume that the BVMT will be a better predictor.

Finally, we plan to revisit the issue of distractor similarity as initially explored by Stevenage and colleagues (2013). They chose an arguably lenient criterion for their manipulation of vocal similarity as it was solely based on speaker sex. A more fine-tuned approach to voice similarity (relative proximity vs. relative distance in voice space) will determine whether we classify distractors as similar or different. It has been proposed that we perceive different voice identities by comparing them to a prototypical, average voice (Latinus & Belin, 2011; Lavner, Rosenhouse, & Gath, 2001). Specifically, the existence of a two-dimensional voice space based on two acoustic parameters (fundamental frequency, F0, and first formant frequency, f1) has been suggested. Different vocal identities are located within this voice space according to their vocal characteristics. The closer two voices are within this voice space, the more likely it is that they are judged to belong to the same person (Baumann & Belin, 2010). Therefore, our prediction is that the closer a distractor voice is in terms of physical voice distance (i.e. the more similar it is in its physical characteristics to a given target voice), the more distracting it will be. We chose this particular design, including the similarity manipulation, to incorporate both the concept of voice recognition after interfering information (as in previous forensic studies), and the increased control over the nature of the distracting information afforded by the lab-based conditions.

Method

Participants

The sample consisted of 100 native-English speakers (25 male; M_{age} = 21.2, SD_{age} = 6.5) who took part in exchange for course credit. All participants reported normal hearing. Written informed consent was obtained from all participants. The study was approved by the Ethics committee of the School of Psychology at Bangor University.

Stimuli and Materials

Voice recordings for both the Bangor Voice Matching Test and the distractor task consisted of non-sense syllables (different combinations of vowels and consonants like 'aga' or 'hed') spoken by young female and male British-English native speakers. Sounds were recorded in a sound attenuated booth using Audacity (16-bit, 44.1 kHz sampling rate, mono). All speakers were between 18 – 28 years of age. All test stimuli were root-mean square normalised and edited in Cool Edit Pro to start with onset of phonation and end with the offset of phonation (mean duration = .51s; S.D. = .11). For each speaker gender, the distance between each individual speaker and every other speaker was calculated using Pythagoras theorem. This distance was defined as the distance in a two-dimensional voice space between F0 and F1 (see Baumann & Belin, 2010). The smaller this distance, the more similar the speakers are perceived to sound (Baumann & Belin, 2010). For a more detailed explanation of this concept, see Figure S4 in the supplementary online material (SOM). Further detail on the audio recordings as well as selection of voice pairs is provided in the stimulus details described in Mühl et al., 2017.

Voice Detection Task

For this task, a total of 144 high quality sounds were chosen from a number of different sources, including the Multimodal Stimulus Set (Schneider, Engel, & Debener, 2008). Sounds belonged to one of three categories: (1) human vocalisations like laughter or singing (72 sounds; 32 male, 32 female, 8 children's voices), (2) inanimate environmental sounds like telephone ringing (36 sounds), or (3) animate environmental sounds like a cat meowing (36 sounds). Each stimulus was edited to include a 10 ms ramp up and down at its start and end, respectively, using Cool Edit Pro, version 2.00 to avoid clipping. Sounds were then RMS normalised using Matlab (R2013a). To ensure sufficient task difficulty, several pilot versions of the detection task were run with differing stimulus lengths between 75 ms and 250 ms. To avoid ceiling or floor effects we decided on a stimulus duration of 150 ms which revealed an average performance of 77.36% during pilot testing (n = 8).

In the main part of the experiment, participants listened to the 144 sounds described above. These sounds were either presented to the right or left ear, to follow the structure of the face detection task used in Duchaine et al. (2006) where an intact face, presented within an array of detached facial features, had to be spotted either on the left or the right side of the picture. Ear assignments of sounds were counterbalanced across participants. Participants had to indicate via keypress in which ear a human sound appeared ('x' for left ear, 'm' for right ear). No response was necessary for the environmental sounds. Participants had 2 seconds to react before the next sound was presented. During stimulus presentation, participants saw a fixation cross centred on the screen as well as a reminder of the key assignments in the upper half of the screen. Test duration was roughly 7 minutes.

Bangor Voice Matching Test

The Bangor Voice Matching Test is a computerised voice matching test in which participants make a same/different identity decision after hearing 2 different syllables per trial. Syllables were either articulated by the same speaker (40 trials) or by two different speakers (another 40 trials; for further details on item selection for the Bangor Voice Matching Test see Mühl et al., 2017). Speaker sex was balanced, with half of the trials presenting male or female speakers, respectively. Instructions were given on the screen and testing was self-paced. For each trial, participants saw two red speaker icons on the screen and, below them, two response boxes, one for same and one for different speakers. Clicking on the speaker icons led to the audio for each item being played. Responses were then given by clicking on either of the response boxes. Participants could listen to each item multiple times if they wished. Between trials, participants saw a centred fixation cross for 800 ms. On average, completion of the BVMT took less than 10 minutes.

Distractor Task

For the distractor task, each trial consisted of 3 voices: a first target voice (T1) followed by a distractor voice (D) which, in turn, was followed by a second target voice (T2). Voices were separated by a 0.8 s interval. Speaker sex throughout each trial was consistent with 32 trials presenting male speakers and 32 trials presenting female speakers (64 trials in total). For half of the items for each speaker block (male/female), T1 and T2 were the same speaker. For the other half, T1 and T2 speaker identity differed. These formed the same/different items. For all of those items, T1-D combinations represented the voice pairs mentioned above. Items were formed in such a way that T1-D distances were either small (< .020), representing similar speakers, or large (between .204 and .936), representing speakers that were not similar and thus more easily distinguishable. This was done to allow for an analysis of whether the similarity of a distractor D influences the recognisability of a target voice T1. Half of the ‘same’ items and half of the ‘different’ items presented small T1-D distances. For all different items, similarity between T1 and T2 was also balanced so that half of the ‘different items’ consisted of similar T1 and T2. Similarity between distractor voices and Target 2 voices (D-T2 similarity) could not be fully balanced due to the limited number of voice pairings available, and were therefore not considered in our predictions. Nevertheless, we tried to keep the distribution of D-T2 distances comparable for male and female trials with 13 small and 19 larger D-T2 distances each. All syllables uttered within an item were different (e.g. aba – hed – ubu, and not aba – hed - aba), and T2 syllable type (consonant-vowel-consonant or vowel-consonant-vowel) either matched only T1 syllable type (13 items), D syllable type (13 items), both T1 and D (18 items), or was different to T1 and D (20 items).

Independent t-tests between the female and male voices that were used in the distractor task revealed no significant difference between the mean T1-D distance overall ($t[62] = -.068$, $p = .946$). Additionally, there was no significant difference between either similar T1-D voice pairings for female and male speakers, $t(30) = -.681$, $p = .541$, or different T1-D voice pairings for female and male speakers, $t(30) = -.087$, $p = .931$. The same was the case when considering the D-T2 similarities instead (all $p > .602$).

Participants' task was to listen to the three voices per trial, and then decide whether the first and the third speaker were the same or not. Decisions were made using the 'f' and 'j' key for same or different voices (key assignment counterbalanced across participants). The next trial started following a button press. During stimulus presentation, participants saw a fixation cross in the centre of the screen. After the third voice (T2) had been played, the key assignment was displayed on the upper half of the screen. Completion of this task took about 20 minutes.

Procedure

All tasks were implemented in Psychtoolbox-3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007) for Matlab (R2013a). Stimuli were presented via Beyerdynamic DT770 Pro headphones (250 Ω). Up to 2 participants were tested at the same time. The order of the three tasks was randomised across all participants. After being given general information about the nature of the experiments, participants filled in a consent form before starting the tasks. Each task was introduced by the experimenter, and both spoken and written instructions were provided. Both voice detection task and distractor task included practice blocks (8 trials/4 trials, respectively). Stimuli presented in those practice trials were not used in the main parts of the experiments. Moreover, participants were encouraged to ask questions in case of uncertainty about a task. After completion of all three tasks, participants were debriefed and given contact details in case of further questions.

Data analysis and design

Data was analysed using Matlab (R2013a) and SPSS (version 22). Performance in detection and distractor tasks were calculated as sensitivity A', using signal detection theory, to control for possible response bias in tasks that require detection of a signal within noise. Accuracy in percentage correct, where reported, were calculated based on the corrected hit and miss rates for detection and distractor task. These calculations followed the steps proposed in Stanislaw and Todorov (1999) for use in SPSS packages (see equation SE1 in the SOM). Only valid trials with reaction times over 250ms were included. Bivariate Pearson's correlations were used to determine the relationship between all three tasks. Following that, a hierarchical linear regression analysis was performed to understand whether the general ability for voice matching (BVMT score) and performance in the detection task predicted the performance in the distractor task. Finally, paired t-tests on the overall percentage correct in the distractor task were used to determine whether the similarity of distractor voices influences the similarity decision for T1 and T2.

Two participants were identified as outliers for their performance on the distractor task (studentised residuals ±3 SDs), and excluded from subsequent analysis to meet the assumptions for the regression analysis. Sample size for both the hierarchical linear regression and the t-tests was N = 98. Inclusion of both outliers did not affect conclusions.

Supplementary Figures S2 and S3 further illustrate the standardised residuals of the regression analysis.

Results

Descriptive statistics (% correct) and correlation coefficients (Pearson's r) for all three tasks can be found in Table 1. Both the performance in the BVMT and in the voice detection task correlated moderately to highly with participants' ability to resist distraction in the distractor task. The correlation between BVMT and distractor task was greater than between voice detection and distractor task. Fisher's z-transformation showed a trend in the expected direction for the first correlation (BVMT with distractor task) to be higher than the latter (detection task with distractor task), $p = .073$ (1-tailed; Lee & Preacher, 2013).

Table 1

Descriptive statistics (% correct), and bivariate correlations (Pearson's r) for percentage correct in BVMT, and A' measures for voice detection task and distractor task

	M	*SD*	Correlation with Detection Task	Correlation with Distractor Task
BVMT	85.14	7.13	.399**	.570**
Detection	87.31	5.11	-	.437**
Distractor	77.67	7.78	-	-

Note. $N = 98$. BVMT = Bangor Voice Matching Test. *M* is mean, *SD* is standard deviation. **$p < .001$.

A two-stage hierarchical multiple regression analysis was calculated to predict the overall accuracy score (A') in the distractor task based on performance on the BVMT (BVMT score; voice-specific, "high-level" voice perception task) and on performance on the voice detection task (A'; more general, "low-level" voice perception task). At stage one, performance on the voice matching task (BVMT score) served as a significant predictor for accuracy in the distractor task, $F(1,96) = 46.30$, $p < .001$, adjusted $R^2 = .318$. The addition of performance on a "low-level" voice perception task (A' of voice detection task) to the prediction of how vulnerable voice matching is to distraction (stage two) lead to a statistically significant increase in R^2 (change statistics: $F[1,95] = 7.91$, $p = .006$). In the full model, both BVMT score and A' of the voice detection task are significant predictors of performance on the distractor task, $F(2,95) = 28.77$, $p < .001$, adjusted $R^2 = .364$. To test whether the BVMT score was a significantly better predictor than performance in the detection task, we estimated the 95% confidence intervals for both standardised beta weights (calculated after z-transformation of all variables) following bias corrected bootstrap (10000 iterations). Confidence intervals overlapped by more than 50%, suggesting that the difference between both predictors ($\Delta\beta = .223$) is not significant, and that the BVMT score was not a statistically significant better predictor of resilience against distraction. Table 2 gives full details of each regression stage, and Figure 1 illustrates both predictors. Supplementary Figure S1 shows the relationship between both predictors.

Table 2

Hierarchical Multiple Regression Predicting Performance on distractor task from BVMT score and voice detection task (A')

	Accuracy in Distractor Task			
	Stage 1		Stage 2	
Variable	B	β	B	β
Constant	0.394**		-.043	
BVMT	0.007**	0.570	.006**	.471
Detection			.556*	.248

Note. N = 98. *B* is unstandardised coefficients, *β* is standardised coefficients after z-scoring of variables. * $p < .05$, ** $p < .001$.

[Insert Figure 1 here]

Finally, paired t-tests did not reveal a difference in accuracy between trials in which T1 voice and the distractor voice were similar vs. different, neither in overall percentage correct, $t(97) = 1.31$, $p = .195$, nor in reaction times, $t(97) = .70$, $p = .484$.

Discussion

The experiment was designed to engage two different stages of the voice perception hierarchy through a more general voice detection task and a more voice-specific, "higher-level" voice matching task (BVMT), and investigate how both relate to the ability to tolerate interference from distractor voices (distractor task). As predicted, task performance on the BVMT correlated more highly with resilience against distraction than performance on the voice detection task. Nevertheless, both correlations were of medium to high strength (voice detection: r = .44, BVMT: r = .57). A hierarchical regression analysis further explored these relationships and revealed that both voice detection and voice matching task (BVMT) are significant predictors of the ability to resist distraction in a voice line-up task (distractor task). Including the voice detection task as an additional predictor in the model led to a significant change of variance explained, and although BVMT performance was descriptively a better predictor than detection task performance, further analysis revealed that the difference between both predictors was not significant. In terms of variance explained, though, BVMT performance alone accounted for 31.8% of the variance (stage 1), whereas the inclusion of detection task performance led to 36.4% of the variance explained in the full model. We suggest that this is due to both voice matching (BVMT) and voice discrimination in the distractor task occurring at later processing stages along the voice perception pathway whereas detecting a human voice in an array of sounds represents an earlier voice perception task.

Face perception research has tried to explore the different processing stages in face recognition and their interactions systematically (e.g. Bate & Bennetts, 2015; Calder & Young, 2005). One possible approach is to thoroughly assess the range of deficits in individuals with known impairments in face perception. Developmental prosopagnosia, a deficit to recognise faces since childhood, has been reported in a number of case studies (e.g. de Haan, 1999; Duchaine et al., 2006), and several possible explanations for these deficits, including non-face specific theories, have been suggested (e.g. Farah, 1990; Moscovitch, Winocur, & Behrmann, 1997). Duchaine and colleagues (2006) give a thorough account of these competing alternative explanations. They also tested these alternatives against each other by having an individual (Edward) with developmental prosopagnosia complete a vast array of face and object perception tasks, and comparing his performance to that of suitable control groups. While most face perception tasks were indeed impaired (e.g. recognition of famous faces, recognition of gender or affect in faces), Edward showed normal scores in a face detection task. Duchaine and colleagues therefore concluded that

Edward's deficits must arise at some point after the initial, low-level processing of faces as a distinct category of stimuli, namely at the stage of structural encoding (as defined by Bruce & Young, 1986). This would explain Edward's ability to correctly detect faces while the analyses of more complex facial cues (e.g. facial affect, face identity) are disrupted. Given the highly similar proposed structure of face and voice perception (Belin et al., 2004; Campanella & Belin, 2007), this supports our interpretation of voice detection being one of the earliest processing stages in the voice perception pathway.

One limitation of our findings lies in the different characteristics of each task. Of all correlations, the ones with the voice detection task were the smallest, while BVMT and distractor task showed the highest correlation. This could be due to the differences in structure between all three tasks. Arguably, the nature of the stimuli as well as the memory demands of the voice detection task (rapid presentation of human vocalisations/animate and inanimate environmental sounds) differed to those of both BVMT and the distractor task (judgment of two/three vocalisations per trial without time limits). The variances introduced by each specific method could therefore partly drive the strength of the correlations reported here. Similarly, the fact that the BVMT showed a higher correlation with the distractor task, and explained more variance in the regression model than the detection task, could lie in the similarity of stimuli used for both tasks (BVMT and distractor task). Both employ short non-speech syllables for which speakers have to be matched. However, task demands still differ considerably. Each trial in the distractor task consisted of three voices, one played shortly after the other (interval between each voice: 0.8 s). Instructions then called for a same/different decision regarding the first and the third voice. The BVMT, on the other hand, is a task in which participants can replay the two voices per trial as often as they like before making their same/different decision. As such, memory demands and time constraints of both BVMT and distractor task differ considerably. In addition to that, the strength of the correlation between BVMT and distractor task was only moderate to high (.57), suggesting that both tasks are sufficiently different and engage overlapping but still specific abilities. In order to fully address these issues in future research, an additional assessment of auditory memory, as well as the inclusion of pre-ratings on all stimuli used (both in terms of physical characteristics like F0, but also perceptual attributes like distinctiveness of sounds) could prove helpful. Additionally, introducing a time limit on the completion of the BVMT (e.g. time constraints on each trial) might help making both predictor tasks more comparable in future studies, and therefore eliminate some of the variance introduced by mere task differences.

Distractor voices were controlled in a way that half of them showed high similarity to the first target voice (T1) while the other half were markedly different. Surprisingly, we did not find an effect of distractor similarity on target identification, neither in the overall performance (percentage correct) nor in the reaction time data. This is in line with the findings of Stevenage and colleagues (2013) who tested the resilience to distraction in both face and voice perception and found that voice perception is more susceptible to distraction, regardless of whether the distractor is similar or not. It is worth noting, though, that the similarity manipulation in that study only matched speaker sex for target and distractor voices (e.g. similar distractors being female speakers for female targets and different distractors being male speakers for female targets). Stevenage and colleagues argued that voice recognition was vulnerable in itself due to the relative weakness of voice perception pathways. As our design used a more stringent approach to what constitutes as a similar distractor (smaller distance in voice space) rather than just speaker sex, our findings support the notion of voice recognition pathways being vulnerable in general.

Alternatively, Kreiman and Sidtis (2013) present evidence that voice identification in line-up situations are always dependent on the specific listeners as well. They suggest that listeners differ widely in respect to which specific voice features are attended to during voice perception. It is possible that our similarity manipulations based on physical difference cannot suitably account for all possible voice features that were used by the participants in our particular sample. If that is the case, it could also explain our null-result for the impact of distractor similarity. For further discussion of our findings regarding distractor similarity, see supplementary text ST1.

Research into the vulnerability of voice perception and, indeed, the robustness of voice identity representation over time, has mainly occurred in forensic contexts to ascertain the credibility of earwitness testimony. A number of studies have tried to identify factors that determine the reliability of earwitness accounts, including the duration and variability of the voice sample, the number of voices that need to be identified, whether the target's face was visible or not, and how much time has passed between initial exposure to a voice and subsequent identification of a target from a line-up (e.g. Clifford, 1980; Cook & Wilding, 1997; Cook & Wilding, 2001; Legge et al., 1984; Yarmey, 1995). Our study differs from these classical designs by only presenting very short voice samples without speech content and an almost immediate same/different decision following voice exposure. While this design is not suitable to use in forensic voice line-up situations, our findings can still contribute to our insight into voice perception in general. This is relevant for our understanding of the neural mechanisms underlying human voice perception on the one hand, but can ultimately also lead to a better application of such findings in a more ecologically relevant setting. For example, it has been proposed that a certain percentage of the population are super recognisers for faces, that is, they are extremely good at using facial identity cues to recognise a person (Bobak, Bennetts, Parris, Jansari, & Bate, 2016; Russell, Duchaine, & Nakayama, 2009). Indeed, a special unit of UK police officers has been formed in which such super-recognisers are employed to identify individuals in particularly demanding identification tasks (Robertson, Noyes, Dowsett, Jenkins, & Burton, 2016). An equivalent for such super-recognisers but for voices seems feasible. Having a better understanding of how voice recognition at all its different stages works could therefore help in identifying such voice super-recognisers.

The heightened interest in developmental impairments in voice perception (Roswandowitz et al., 2014; Shilowich & Biederman, 2016) as well as recent research into the more general question of individual differences in voice perception (Aglieri et al., 2016; Mühl et al., 2017) underline the need for a better understanding of how we perceive people by their voices. We propose that a more systematic approach to identifying and probing possible distinct processes in the voice perception pathway will not only help our theoretical understanding of voice perception, but will ultimately also impact its application in clinical and, possibly, forensic settings.

Funding

Constanze Mühl is a Ph.D. student funded by the School of Psychology at Bangor University.

References

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2016). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods*. Advance Online Publication. doi:10.3758/s13428-015-0689-6

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, *74*, 110-120.

Bate, S., & Bennetts, R. (2015). The independence of expression and identity in face-processing: evidence from neuropsychological case studies. *Frontiers in Psychology*, *6:770*. doi:10.3758/BF03207704

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. doi:10.1016/j.tics.2004.01.008

Bestelmeyer, P., Maurage, P., Rouger, J., Latinus, M., & Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *Journal of Neuroscience, 34*(24), 8098-8105.

Blank, H., Wieland, N., & von Kriegstein, K. (2014). Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neuroscience and Biobehavioral Reviews*, *47*, 717–734. doi:10.1016/j.neubiorev.2014.10.022

Blasi, A., Mercure, E., Lloyd-Fox, S., Thomson, A., Brammer, M., Sauter, D., Deeley, Q., Barker, G. J., Renvall, V., Deoni, S., Gasston, D., Williams, S. C. R., Johnson, M. H., Simmons, A., & Murphy, D. G. M. (2011). Early specialization for voice and emotion processing in the infant brain. *Current Biology, 21*(14), 1220-1224.

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex*, *82*, 48-62.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 443-436.

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

Calder, A. J, & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, *6*(8), 641-651. doi:10.1038/nrn1724

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535-543.

Cheng, Y., Lee, S-Y, Chen, H-Y., Wang, P-Y., & Decety, J. (2012). Voice and emotion processing in the human neonatal brain. *Journal of Cognitive Neuroscience, 24*(6), 1411-1419.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability*. Law and Human Behavior*, *4*(4), 373-394.

Cook, S., & Wilding, J. (1997). Earwitness Testimony 2: Voices, faces and context. *Applied Cognitive Psychology*, *11*, 527-541.

Cook, S., & Wilding, J. (2001). Earwitness testimony: Effects of exposure and attention on the Face Overshadowing Effect. *British Journal of Psychology*, *92*(4), 617-629. doi:10.1348/000712601162374

De Haan, E. H. F. (1999). A familial factor in the development of face recognition deficits. *Journal of Clinical and Experimental Neuropsychology*, *21*(3), 312-315. doi:10.1076/jcen.21.3.312.917

Duchaine, B. C., Yovel, G., Butterworth, E. J., & Nakayama, K. (2006). Prosopagnosia as an impairment to face-specific mechanisms: Elimination of the alternative hypotheses in a developmental case. *Cognitive Neuropsychology*, *23*(5), 714-747. doi:10.1080/02643290500441296

Farah, M. J. (1990). *Visual agnosia*. Cambridge, MA: MIT Press.

Gainotti, G. (2014). Familiar people recognition disorders: An introductory review. *Frontiers in Bioscience*, *6*, 58-64.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L, Sauter, D., Hanley, J. R., Schweinberger, S. R., Warren, J. D., & Duchaine, B. (2009). Developmental phonagnosia: A selective deficit of vocal identity recognition. *Neuropsychologia*, *47*, 123-131. doi:10.1016/j.neuropsychologia.2008.08.003

Grossman, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience, 28*, 219-236.

Iredale, J., M., Rushby, J. A., McDonald, S., Dimoska-Di Marco, A., & Swift, J. (2013). Emotion in voice matters: Neural correlates of emotional prosody perception. *International Journal of Psychophysiology*, *89*, 483-490.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception 36 ECVP Abstract Supplement.*

Kreiman, J., & Sidtis, D. (2013). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Malden, MA: Wiley-Blackwell.

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*. doi: 10.3389/fpsyg.2011.00175

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology, 4*, 63-74.

Lee, I. A., & Preacher, K. J. (2013, September). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from http://quantpsy.org

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 298-303.

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9*, 555–604.

Mühl, C., Sheil, O., Jarutytė, L., & Bestelmeyer, P. (2017). The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. *Behavior Research Methods*. Retrieved from https://doi.org/10.3758/s13428-017-0985-4

Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairment in the recognition of faces, voices, and personal names. *Brain and Cognition*. *44*(3), 342-366. doi:10.1006/brcg.1999.1196

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., & Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain, 117*, 1283-1301.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by Metropolitan Police super-recognisers. *PLoS ONE*, *11*(2): e0150036. doi:10.1371/journal.pone.0150036

Rossi-Katz, J., & Arehart, K. H. (2009). Message and talker identification in older adults: Effects of task, distinctiveness of the talkers‘ voices, and meaningfulness of the competing message. *Journal of Speech, Language, and Hearing Research*, *52*(2), 435-453.

Roswandowitz, C., Mathias, S. R., Hintz, F., Kreitewolf, J., Schelinski, S., & von Kriegstein, K. (2014). Two cases of selective developmental voice-recognition impairments. *Current Biology*, *24*(19), 2348–2353. doi:10.1016/j.cub.2014.08.048

Roswandowitz, C., Kappes, C., Obrig, H., & von Kriegstein, K. (2018). Obligatory and facultative brain regions for voice-identity recognition. Brain, 141, 234-247. doi:10.1093/brain/awx313

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252-257. doi:10.3758/PBR.16.2.252

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10*(1), 24-30.

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology*, *55*(2), 121-132. doi:10.1027/1618-3169.55.2.121

Shilowich, B. E., & Biederman, I. (2016). An estimate of the prevalence of developmental phonagnosia. *Brain & Language*, *159*, 84-91.

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

Stevenage, S. V., Neil, G. J., Barlow, J., Dyson, A., Eaton-Brown, C., & Parsons, B. (2013). The effect of distraction on face and voice recognition. *Psychological Research*, 77, 167-175. doi:10.1007/s00426-012-0450-z

van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829-834.

Warren, J. D., Jennings, A. R., Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *Neuroimage, 24*(4), 1052-1057.

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law*, *1*(4), 792-816.

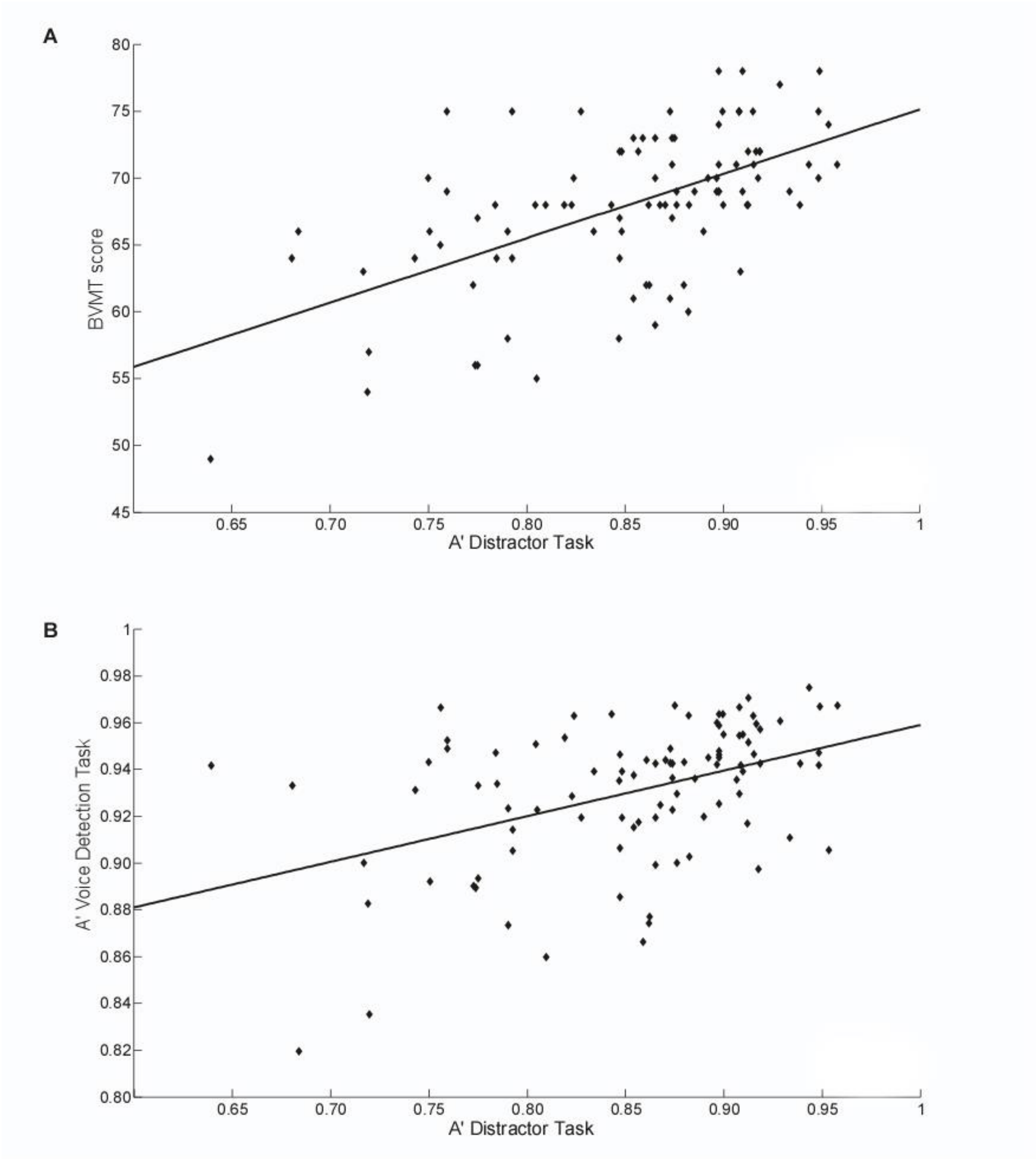


Figure 1. Relationship between (**A**) performance on BVMT (score) and distractor task (A') and (**B**) performance on voice detection task (A') and distractor task (A'). Lines represent linear regression fits to data points.

Acknowledgements

We would like to thank Dr. Till Schneider and his team for providing us with the Multimost materials which made up part of the stimulus set used in this study. The Multimodal Stimulus Set was developed by T. R. Schneider, S. Debener and A. K. Engel at the Dept. of Neurophysiology, University Medial Center Hamburg-Eppendorf, Germany.

Supplemental Online Material

SE1. Equation for A' calculation for SPSS from Stanislaw & Todorov (1999). H denotes hit rate, F denotes false alarm rate:

$$A' = 0.5 * \left(\frac{abs(H - F)}{H - F}\right) * \frac{(H - F)^2 + abs(H - F)}{4 * MAX(H, F) - 4 * H * F}$$

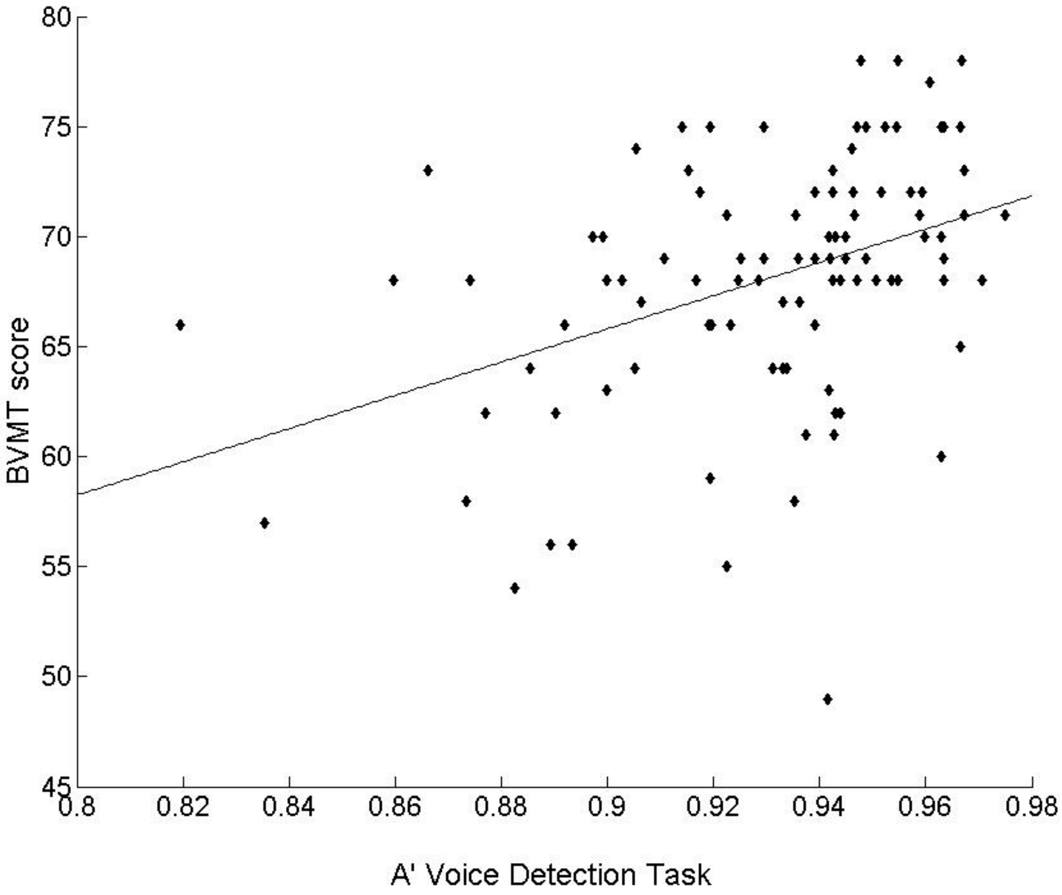


Figure S1. Relationship between performance on both independent variables (BMVT score and A’ in distractor task; *r* = .40). Line represents linear regression fit to data points.

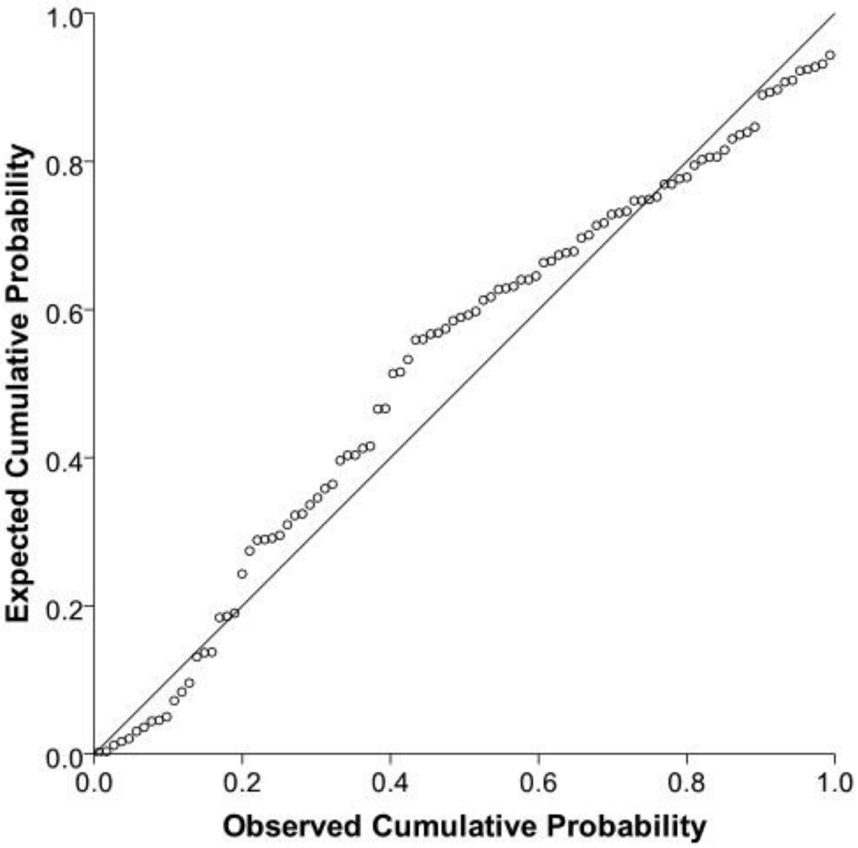
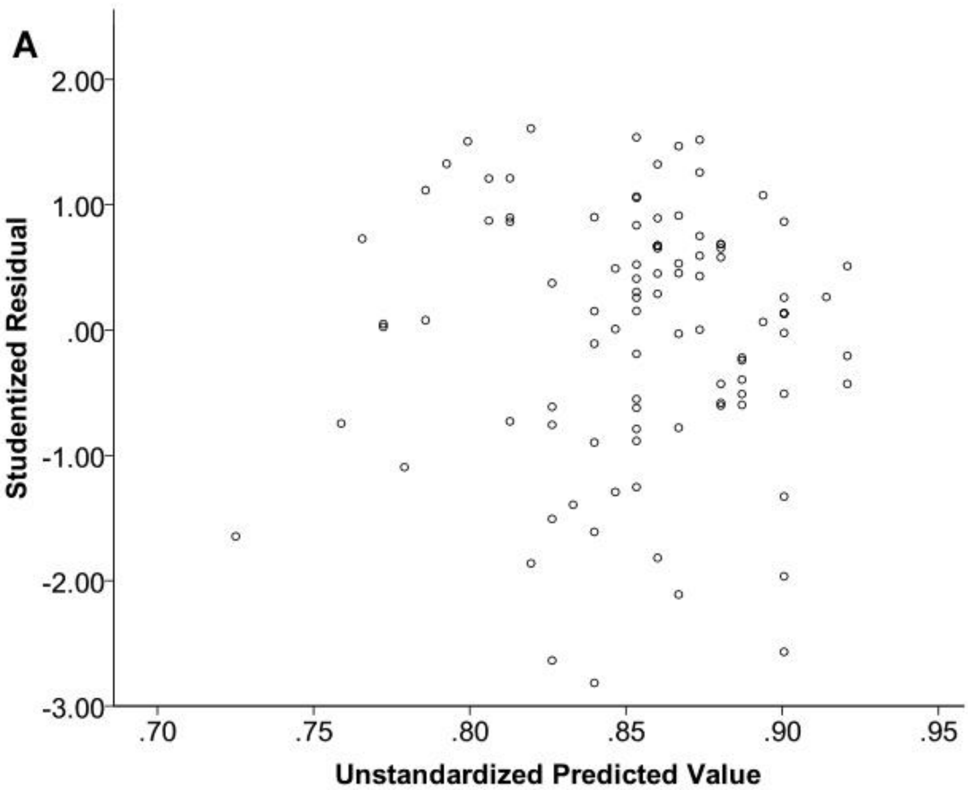


Figure S2. P-P-Plot for regressions standardised residual.



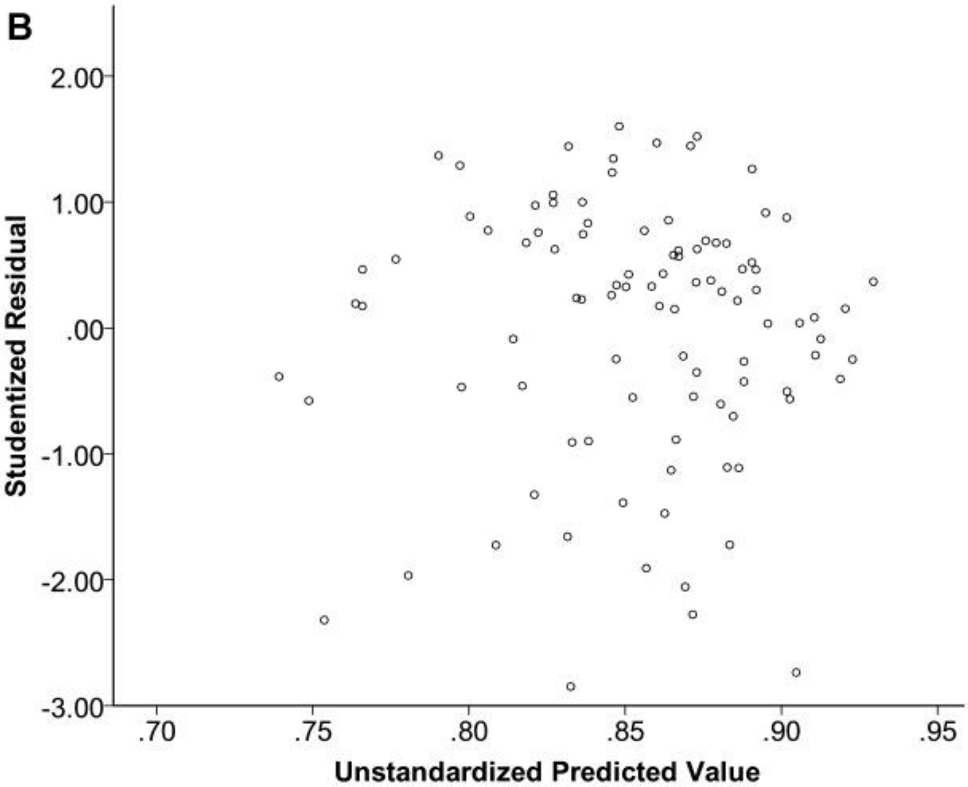


Figure S3. Distribution of standardised residuals against unstandardized predicted values for stage 1 of the model (**A**; BVMT score as sole predictor) and the full model (**B**; BVMT score and A' of detection task).

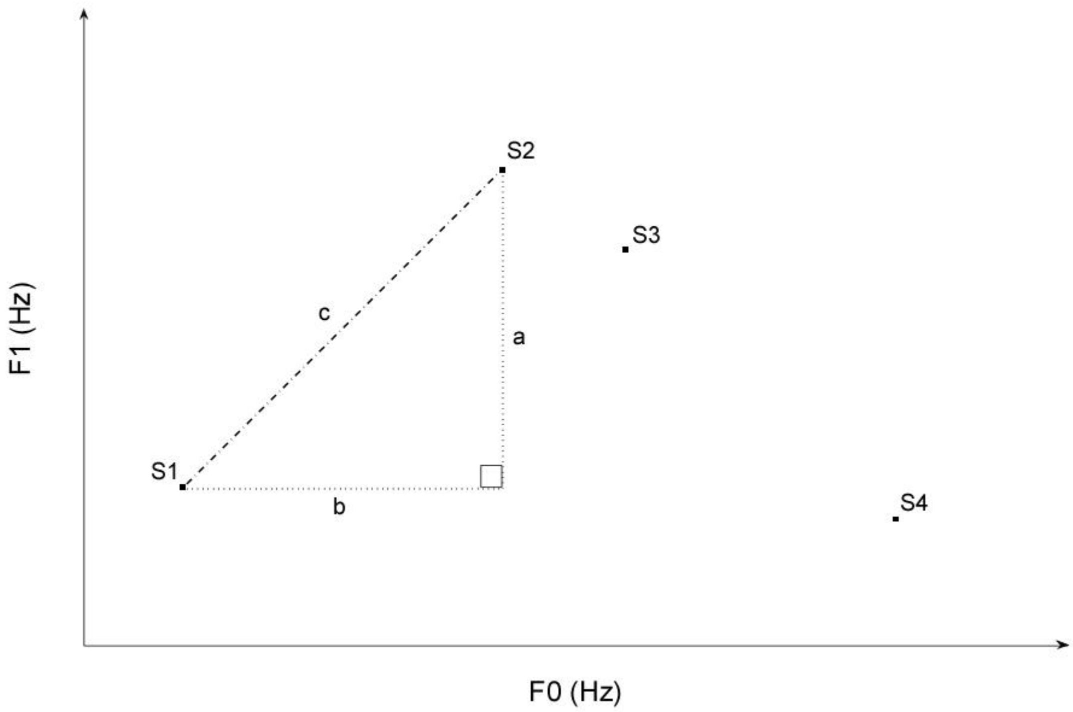


Figure S4. Schematic representation of voice space. Individual speakers (S1 to S4) are illustrated within a 2-dimensional voice space (Baumann & Belin, 2010), according to their fundamental frequency (F0) and their first formant frequency (F1). Voices that are close to each other (e.g. S2 and S3) sound more similar than those further apart (e.g. S2 and S4). Physical difference between S1 and S2 (alternating dashed line, hypotenuse c) is calculated using the Pythagoras theorem, given a right triangle with legs a and b (simple dashed lines), $c = \sqrt{a^2 + b^2}$.

ST1. Accuracy in distractor task based on similarity between distractor and second target voice (T2)

Additional post-hoc analyses of accuracy for trials with similar vs. different distractor and T2 voice pairings showed a significant difference in mean percentage correct, $t(97) = -2.53$, $p = .013$, with a higher accuracy for trials in which physical D-T2 distance was greater ($M = 78.72\%$, $SD = 8.73$) compared to smaller D-T2 distances ($M = 76.26\%$, $SD = 10.90$). However, this difference did not reach significance in the reaction time data ($t[97] = 1.79$, $p = .077$). Our post-hoc analysis therefore revealed a significantly higher accuracy if the distractor voice was markedly different to the T2 voice.

While this is in line with our initial prediction for the impact of distractor similarity, we are cautious to interpret this finding. Unlike for the T1-D pairings, the number of similar/different D-T2 pairings was not equal due to the limited availability of suitable voice pairings. Consequently, as stated before, our predictions only considered the effect a distractor voice could have for the accuracy of identifying a previously heard target voice (T1). This issue needs to be revisited in future studies where the distractor similarity for both target voices, T1 and T2, can be controlled more stringently (given a larger pool of initial voice pairings).

Further indication of an effect of distractor similarity comes from research into changes of our ability to identify speakers from different age ranges. Rossi-Katz and Arehart (2009) manipulated distinctiveness of distractor voices via speaker sex, and investigated its effect on the accuracies of (a) identifying a target message, that is, speech content, and (b) identifying a target speaker identity. Both manipulations were tested in a group of young adults (23 – 25 years of age) as well as in a group of older adults (> 65 years of age). While the target message task profited from increased speaker distinctiveness (albeit to a lower extent in the older group), target identification did not. Young adults showed high speaker identification accuracy regardless of distractor distinctiveness whereas older adults showed a decline of speaker identification accuracy for more distinct distractors (meaningful speech condition). The null effect of distractor similarity/differences in Stevenage and colleagues' study (2013) as well as in ours might therefore be due to the nature of the samples used (young adults), and further investigation into different samples seems necessary.