The effects of high variability training on voice identity learning

Nadine Lavan[1,2], Sarah Knight[1], Valerie Hazan[1] and Carolyn McGettigan[1,2]

[1] Department of Speech, Hearing and Phonetic Sciences, University College London

[2] Department of Psychology, Royal Holloway, University of London

Correspondence to:

Nadine Lavan, Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, United Kingdom.

E-mail: n.lavan@ucl.ac.uk

or

Carolyn McGettigan, Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, United Kingdom.

E-mail: c.mcgettigan@ucl.ac.uk

Acknowledgements: This work was supported by a Research Leadership Award from the Leverhulme Trust (RL-2016-013) awarded to Carolyn McGettigan

Abstract

The effects of high variability training during voice identity learning

High variability training has been shown to benefit the learning of new face identities. In three experiments, we investigated whether this is also the case for voice identity learning. In Experiment 1a, we contrasted high variability training sets – which included stimuli extracted from a number of different recording sessions, speaking environments and speaking style – with low variability stimulus sets that only included a single speaking style (read speech) extracted from one recording session (see Ritchie & Burton, 2017 for faces). Listeners were tested on an old/new recognition task using read sentences (i.e. test materials fully overlapped with the low variability training stimuli) and we found a high variability *dis*advantage. In Experiment 1b, listeners were trained in a similar way, however, now there was no overlap in speaking style or recording session between training sets and test stimuli. Here, we found a high variability advantage. In Experiment 2, variability was manipulated in terms of the number of unique items as opposed to number of unique speaking styles. Here, we contrasted the high variability training sets used in Experiment 1a with low variability training sets that included the same breadth of styles, but fewer unique items; instead, individual items were repeated (see Murphy, Ipser, Gaigg & Cook, 2015 for faces). We found only weak evidence for a high variability advantage, which could be explained by stimulus-specific effects. We propose that high variability advantages may be particularly pronounced when listeners are required to generalise from trained stimuli to different-sounding, previously unheard stimuli. We discuss these findings in the context of mechanisms thought to underpin advantages for high variability training.

Keywords: voice identity; person perception; high variability training; voice learning

Introduction

The effects of high variability training during voice identity learning

Within-person variability is one of the defining features of the human voice. Speakers constantly change the acoustic and perceptual properties of their voices to convey information about their emotional states, intentions or social relationships and similarly adapt their speech to suit different speaking environments and audiences. Such within-person variability has been shown to be challenging when attempting to accurately perceive voice identity (Lavan, Burton, Scott & McGettigan, 2018). Studies report decreased performance when making judgements of identity across stimuli that include within-person variability – especially when the voices are not familiar to listeners. These effects have been shown across non-verbal vocalisations (laughter versus vowels; Lavan, Scott & McGettigan, 2016), across speaking styles and background noise (Smith, Baguley, Robson, Dunn & Stacey, 2018), across languages (Wester, 2012), across sung versus spoken words (Peynircioğlu, Rabinovitz, & Repice, 2017) and across different pitches in sung vowels (Erickson & Phillips, 2018). Recently it has been suggested that these effects may arise because unfamiliar listeners tend to misperceive within-person variability as between-person variability. In the absence of a person-specific representation of a voice, unfamiliar listeners perceive multiple variable examples of a single person's voice as having been produced by multiple speakers, thus failing to "tell people together". Familiar listeners, on the other hand, tend to outperform unfamiliar listeners in identity perception tasks: they can access person-specific representations, enabling them to perceive within-person variability appropriately and to thus succeed in "telling people together" (Lavan, Burston & Garrido, 2018 Lavan, Burston, Ladwa, Merriman, Knight & McGettigan, 2019; Lavan, Merriman, Ladwa, Burston, Knight & McGettigan, 2018).

The effects of high variability training during voice identity learning

While within-person variability poses challenges that can be overcome through familiarity with a voice, it may not at all times be detrimental to voice identity perception. It has been shown in the face perception literature that learning faces through high variability exposure may actually be advantageous. When directly comparing high versus low variability training, a number of studies have produced empirical evidence in support of such an advantage: Murphy, Ipser, Gaigg and Cook (2015) exposed participants to a number of 6x8 arrays of ambient images (i.e. static photographs including natural variability) each including 6 exemplars of 8 individuals. In a between-subjects design, the low variability condition repeated the same 6x8 grid for all the 16 learning trials (6 exemplars per identity x 16 repetitions). In the high variability condition, each 6x8 grid included 6 novel exemplars of the same 8 identities (96 exemplars, no repetitions). In a subsequent old/new recognition task, listeners were more accurate if they learned the identities through the high variability training compared to participants who learned the identities through the low variability training. In a within-subjects design, Ritchie and Burton (2017) trained participants to associate names with 10 identities, with half of the identities being learned through low variability exemplars and the other half through high variability exemplars (all static photographs). Here, variability was manipulated by using ambient images (see above) in the high variability condition. In the low variability condition, stills from a single video were used, reducing variability in, for example, lighting, exposure, camera and hairstyle. In two experiments, which used either a speeded naming task based on novel ambient images or a face matching task at test, performance for identities learned based on high variability training exceeded that for identities learned based on low variability training. In a third experiment, the authors show that even when the test images were previously unseen stills from the

same video used during the low variability training – thus giving a clear advantage for identities learned through low variability exposure – performance for high versus low variability was still matched. Finally, Baker, Laurence and Mondloch (2017) report an advantage of high variability training for children: for the high variability conditions, a target identity was shown to participants based on videos recorded across a number of days in different locations. For the low variability training conditions, videos of the target identity were recorded in a single recording session at a single location. At test, children then identified the target identity based on ambient images (in contrast to a number of distractor identities). Note, however, that no high variability advantage was found for adults in their study.

There are debates as to how this advantage may arise: it has been proposed that high variability exposure enables viewers to detect (and abstract) the reliable features of a face, while discounting transient features (e.g. Burton, Jenkins, Hancock & White, 2005; Jenkins & Burton, 2011). More recently, a new view was proposed in which the within-person variability itself is considered to be an essential cue when learning a new identity. Based on principal component analyses of the visual properties of naturally-varying images of faces, Burton, Kramer, Ritchie and Jenkins (2016) have shown that the within-person variability for each face is idiosyncratic, opening up the possibility that variability may not simply be noise that is to be abstracted away to detect reliable and robust cues to identity. Instead, the variability may be an informative signal that could in principle be used by humans during identity perception. Experiencing the range and nature of variability may thus be a fundamental aspect of learning new identities (see Ritchie & Burton, 2017 for a discussion). Since recent work on voice identity perception has highlighted

similarities between faces and voices in how within-person variability affects identity perception, we tested whether high variability training also confers an advantage for voice identity learning in three experiments. We note that variability can be conceptualised in many ways: the type of variability can be manipulated (e.g. phonemic variability vs. variability in specific acoustic features vs. variability in broader characteristics such as speaking styles), as can the ways in which high and low variability are empirically contrasted. As a first step, we opted to implement the contrast of high vs low variability in two different ways, following the studies of face learning by Ritchie and Burton (2017; Experiment 1a and 1b) and Murphy et al. (2015; Experiment 2; see *Methods*; illustrated in Figure 1). All experiments were preregistered via the Open Science Framework (https://osf.io/7xvjw/).

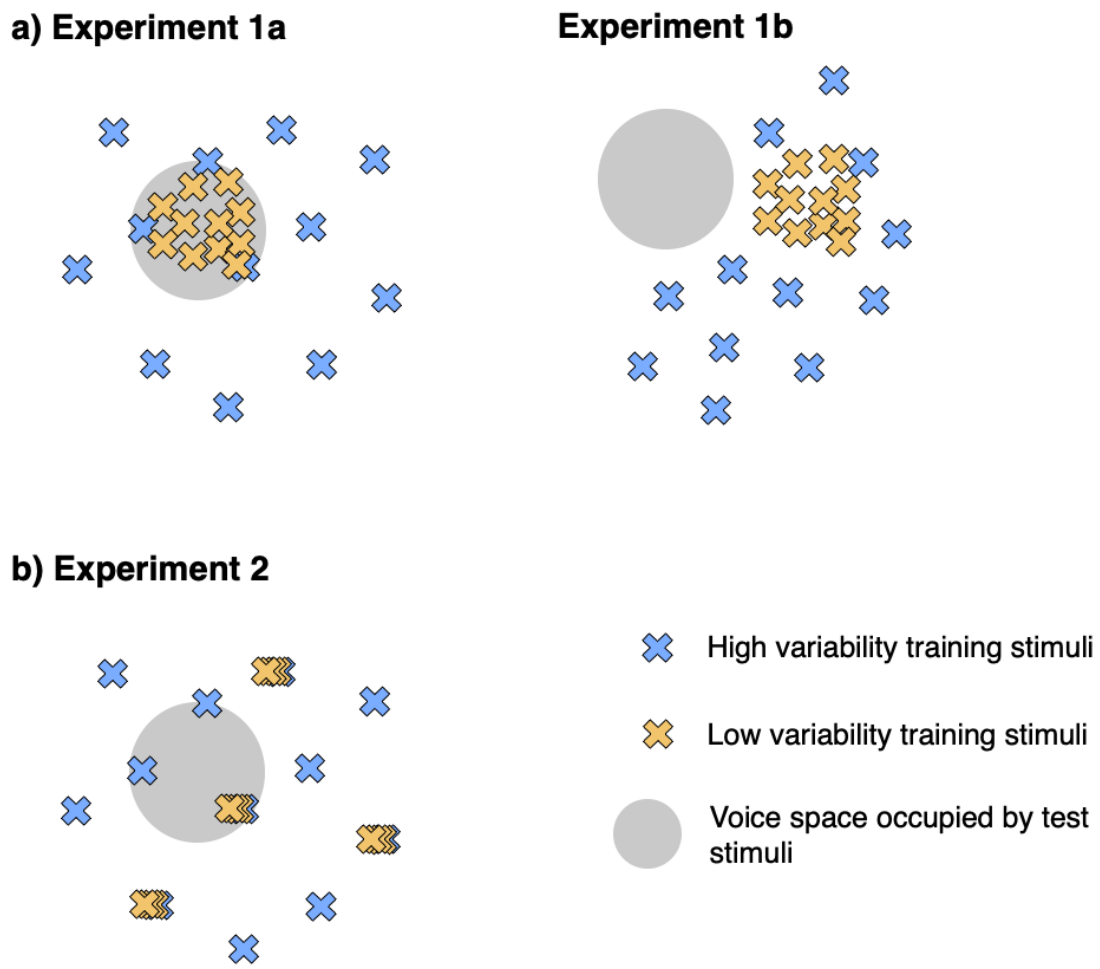

Figure 1 Illustration of the high versus low variability manipulations for Experiments 1a and 1b and Experiment 2. Each illustration depicts a notional "voice space", within which the X symbols indicate training items. a) In Experiment 1a, high variability training sets cover a wider range of the voice space than the low variability training set, where the latter overlaps with the test items (i.e. both are comprised of read sentences). In Experiment 1b, high variability training sets cover a wider range of the voice space than the low variability training set (multiple speaking styles vs one speaking style only). However, test and training sets never overlap in this experiment. b) In Experiment 2, the range of voice space covered by high and low variability training is notionally more comparable through use of the same range of speaking styles. However greater repetition of individual items in the low variability training limits the variation in exposure for each speaking style used. As in Experiments 1a and 1b, the test items are read sentences.

Experiment 1a

In Experiment 1a, high variability training sets included items from a number of different recording sessions, speaking styles and speaking situations (see Methods). Low variability training sets included only one speaking style (read sentences) with exemplars being extracted from within a single recording session. Following the training, listeners were tested on a voice recognition task (old/new judgements)

based on read sentences produced by the 4 learned voice identities and 4 distractor identities. We predicted that if high variability training is indeed advantageous when learning novel voice identities, performance on an old/new recognition task should be either higher or the same for voice identities trained via high variability compared to identities trained via low variability. In the current study, the acoustic space occupied by the test items overlap to a greater degree with the low variability training items than the higher variability training items: the test items and low variability training items were both recorded during the same session and both include the same speaking style. Thus, the study is biased towards finding better performance for identities learned via the low variability training sets. As a result, similar performance at test in this experiment is interpreted as evidence for a high variability advantage, since it implies that training was efficient enough to overcome the potential initial disadvantage (see Ritchie & Burton, 2017, Experiment 1B). Better performance for voice identities learned through low variability training would, however, indicate that there is no meaningful advantage (and possibly a disadvantage) for high variability training.

Methods

Participants

122 participants were tested online using Gorilla (gorilla.sc/about; Anwyl-Irvine, Massonié, Flitton, Kirkham & Evershed, 2018). Participants were recruited via Prolific (prolific.ac) and were reimbursed for their time. All participants were aged between 18 and 40 years, were native speakers of English, had no reported hearing difficulties and had an approval rate over 90% on Prolific. The study was approved by the ethics committee of the Department of Speech, Hearing and Phonetic

Sciences at University College London. All participants were provided with an information sheet and completed a consent form before the start of the study. None of the participants had taken part in any of the pilot studies associated with the current experiment. Two participants were excluded from this data set: 1 participant failed to give the correct response for more than 20% of vigilance trials (see Methods); another participant's overall performance was more than 3 standard deviations below the mean performance of the sample. All participants performed significantly better than chance (± 95% confidence intervals) for the last 16 trials of Training 2, which was a final exclusion criterion. The final participant sample thus included 120 participants (mean age: 28.5 years, SD = 6.1 years; 62 female).

Stimuli

Stimuli were extracted from the LUCID corpus (Baker & Hazan, 2011). This corpus includes recordings of 40 native speakers of Southern British English recorded across 5 recording sessions. Each session features speech produced in a different speaking style. In Session 1 sex-matched pairs of friends were recorded completing a DIAPIX task (an interactive 'spot the difference' task) to elicit spontaneous, conversational speech. In Session 2, the same pairs completed additional DIAPIX tasks, with one person's voice now being noise-vocoded. This creates adverse communication conditions through the degradation of the speech signal, leading to modulations in speech production and the adoption of a clear speaking style to increase the intelligibility of the speech (e.g. Hazan & Baker, 2010). In Session 3, participants completed further DIAPIX tasks but were either paired with a conversation partner who was a stranger while speech was presented in multi-talker babble, or a low-proficiency non-native speaker of English who was also a stranger.

In Session 4, speakers read a number of sentences and recorded semi-spontaneous speech elicited via a picture naming task (“I can see a [OBJECT]”; “The verb is to [VERB]”). In Session 5 listeners recorded the same materials and tasks as in Session 4 but now produced the speech as if they were talking to someone who is hearing impaired, leading to exaggerated, clear speech (e.g. Hazan & Baker, 2010).

From this corpus, we extracted stimulus sets from 8 female speakers of Southern British English (age range = 20-27 years). Of these 8 speakers, 4 were selected to be trained identities and the remaining 4 were used as distractor identities at test (sets counterbalanced across participants). The number of training stimuli was determined via pilot testing, which indicated a desirable level of overall performance, avoiding ceiling or floor effects. Training stimulus sets included 24 unique items. All items lasted between 1 and 4 seconds in duration. Low variability training sets included 24 unique read sentences selected from the recordings from Session 4 of the LUCID corpus. No sentence was repeated across voice identities. High variability training sets included 24 unique items extracted from across the 5 sessions, thus covering a range of speaking styles, speaking situations, and recording times, increasing the degree of within-person variability included. Specifically, 6 items were extracted from the dialogue recorded during Session 1, 6 items were extracted from Session 2, 4 from Session 3, 4 items from Session 4 (2 items of the picture naming task, 2 from read sentences). Items were selected based to include meaningful utterances (e.g. “A blue can and a crisp packet”, “Yours definitely aren’t bags?”) and based on their total duration (see above). Finally, 5 items from Session 5 (2 items from the picture naming task, 2 read sentences, e.g. “The beach stall sold bats and balls”). The test stimulus sets included 12 read sentences from Session 4 – these

sentences were distinct from the items that listeners had been exposed to during training (see Figure 1). Note however that, although the exemplars are distinct, the speaking style and recording session for the low variability stimulus sets fully overlaps with the test stimulus sets (both are read speech).

Procedure

Listeners first completed a headphone screening (Woods, Siegel, Traer & McDermott, 2016) before completing two brief training phases (Training 1 and Training 2). We opted for this particular structure for the training phase based on its efficacy in previous studies of voice identity learning (Latinus & Belin, 2012; Lavan, Knight & McGettigan, 2019). For Training 1, participants were presented with the 24 items associated with each of the 4 training identities, with a name to be associated with that voice identity presented on the screen (e.g. “This is Beth”). For two of those identities, the high variability training sets were used, and for the remaining two identities, listeners were presented with the low variability training lets. The assignment of identities to high and low variability training was counterbalanced across participants: Specifically, half the participants learned one set of 4 identities, the half learned the other set of 4 identities. Within these subgroups, the assignment of high and low variability was counterbalanced again. The presentations during this phase were blocked by identity, with each block including 12 items. The order of the 8 resulting blocks (4 identities x 2 blocks) was randomised across participants. Participants were instructed to listen attentively and to try to memorise the different voices and their names. No responses were collected during this training phase. For Training 2, participants were presented with the same items as in Training 1 (24 high variability items x 2 identities + 24 low variability items x 2 identities = 96 trials in

total) in randomised order and were asked to complete a 4-way forced choice recognition task (“Is this Anna, Beth, Clara or Debbie?”) with audio-visual feedback on whether their response was correct or not. If a response was incorrect, listeners were shown the correct answer in writing on the screen. Both learning phases were self-timed and lasted on average around 15 minutes in total. Performance for the final 20% of trials of Training 2 was used as an index to track whether listeners had learned to recognise the two identities. These data showed listeners were able to correctly identify the 4 voice identities with high accuracy towards the end of the learning phase (mean accuracy = 85.0%, SD = 13.4%; chance level = 25%).

After this learning phase participants completed the test phase, which consisted of an old/new judgement task (“Was the voice you just heard an old voice or a new voice?” Response options: “Old voice”, “New voice”). Listeners were presented with 12 sentences produced by the 4 learned identities plus 12 sentences from the remaining 4 distractor identities. Participants additionally completed 10 vigilance trials: here listeners were asked to follow the instructions of a computer-generated male voice to either respond with “old voice” or “new voice” (e.g. “Please click on ‘New voice’”). The task was self-paced and took participants around 7 minutes to complete.

Results

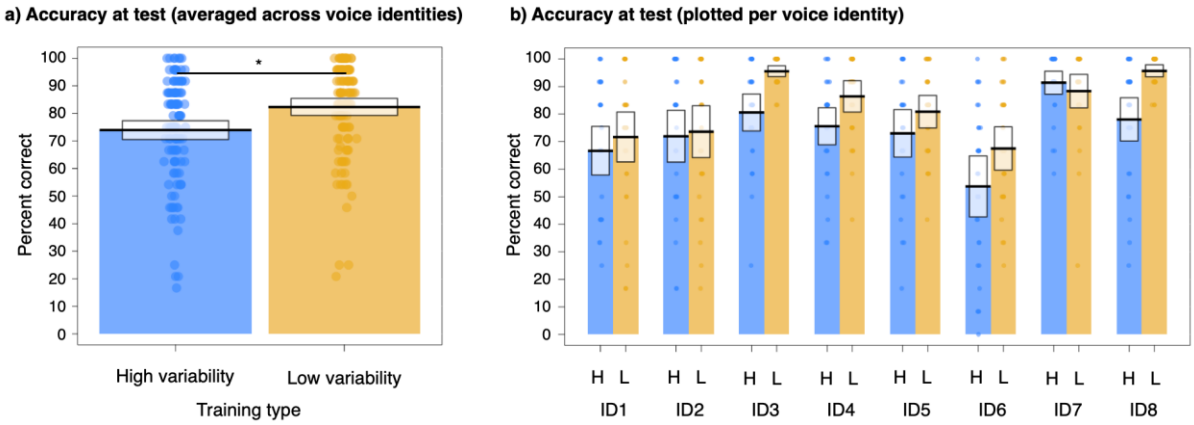

Figure 2 Summary of old/new recognition performance for learned identities in Experiment 1a. a) Accuracy for high versus low variability training is plotted averaged across all identities. b) Accuracy for high (H) versus low (L) variability training broken down by identity. Due to the counterbalancing of trained and distractor identities and assignments of high vs low variability stimuli to the trained voices, each individual bar shows the data of 30 participants. Boxes show the 95% confidence intervals, dots indicate the mean accuracy per participant.

Confirmatory analyses

To assess the effect of high variability training on accuracy for the learned identities (i.e. not taking the data from the distractor identities into account), we ran a binomial generalised linear mixed model (GLMM) using *lme4* (Bates, Maechler, Bolker & Walker, 2014) in the *R* environment (R Core Team, 2013). Training type (high vs low variability) was defined as a fixed effect. Participant, speaker, stimulus as well as the counterbalancing were entered as random effects (Barr, Levy, Scheepers & Tily, 2013). Statistical significance was established via likelihood ratio tests contrasting the full model including the fixed effect plus the random effects with a null model, i.e. a model that did not include the fixed effect. These models confirmed that the type of training had an effect on accuracy (coefficient of -0.66, *SE* = .07) and the comparison of the full and null model was significant ($\chi^2[1] = 81.16$, $p < .001$).

Accuracy was thus higher for identities learned through low variability training (82% correct vs 74% correct; Figure 2a).

Exploratory analyses

We further explored whether this effect differed across the individual voice identities. Figure 2b indicated that this trend can be observed for 7 out of the 8 identities. We ran another GLMM, similar to the one described above. For this GLMM, we, however, included both speaker and training as fixed effects, as well as the interaction between speaker and training. Statistical significance was again established via likelihood ratio tests contrasting the full model including all fixed effects plus the random effects with a reduced model that did not include the fixed effect of interest (i.e. speaker x training). These models confirmed that there was a significant interaction between speaker and training: the comparison of the full and null model was significant ($\chi^2[13] = 81.29$, $p < .001$). Thus, while this effect is present in numerical terms for most of the identities, the difference in performance introduced by high and low variability training varies from identity to identity.

Discussion

In the context of this study, high variability training was not advantageous compared to low variability training. Indeed, performance for identities learned through high variability training was worse than for identities learned through low variability training. This result may not be surprising, given the experimental design: since the test and low variability training sets overlapped to a greater degree - in terms of their speaking style and having been extracted from the same recording session - than the test and high variability training sets, it could be argued that the odds are stacked

against finding an advantage for high variability training. The result is, however, still different to findings obtained in a face learning study (Experiment 1B in Burton & Ritchie, 2017,) using a comparable design: test performance for faces was matched across high and low variability training regimes, indicating that high variability training did not *cost* the participants.

Experiment 1b

In Experiment 1b, we constructed high and low variability training sets that do not overlap in speaking style and recording session with the test stimuli. To successfully perform the old/new recognition task, listeners were therefore required to generalise from training stimuli to a previously unheard speaking style/variability (see Figure 1a on the right). Since there is no overlap for either the low or the high variability training stimuli with the test stimuli, we predicted that if high variability training is advantageous when learning novel voice identities, performance on an old/new recognition task should be higher for voice identities trained via high variability compared to identities trained via low variability.

Methods

Participants

162 participants were recruited via Prolific and tested online. Recruitment strategies and exclusion criteria were identical to the ones listed for Experiment 1a. There was no overlap in participants between Experiment 1a, Experiment 1b and Experiment 2 (see below) and no participants in Experiment 1b had taken part in any of the pilot studies associated with this set of studies. One participants were excluded based on their responses to the vigilance trials and one participant was excluded since they

did not show any learning during training. All the remaining participants' performances on the main task fell within 3 SDs of the mean performance. The final participant sample thus included 160 participants (mean age: 28.4 years, SD = 6.2 years; 91 female).

Stimuli

Stimuli were sourced from the LUCID corpus (Baker & Hazan, 2011) and the same identities were included in the task as in Experiment 1a. We used the same test stimuli as in Experiment 1a (12 read sentences per identity extracted from Session 4 of the LUCID corpus). To avoid any overlap in speaking styles between test and training, we created new high and low variability training sets. For the high variability condition, we used 4 of the available remaining speaking styles: conversational speech (Session 1; "conversational speech"), conversational speech in adverse communication conditions (Session 2; "conversational speech (clear)"), picture naming "as if talking to a hearing-impaired person" (Session 5, "picture naming (clear)") and sentence reading "as if talking to a hearing-impaired person (Session 5; "sentence reading (clear)"). Each of these speaking styles was represented by 6 stimuli, resulting in a total of 24 stimuli per identity. We did not include materials from Session 3 since some of the to-be-learned voices completed a spot the difference task while talking to a stranger in adverse speaking conditions, while the others spoke a to low-proficiency non-native speaker, so we would have been unable to create a uniform set of low variability training stimuli based on these recordings.

For the low variability training sets, we extracted sets of 24 stimuli from each of the 4 speaking styles included in the high variability training sets. These speaking styles

differ in any number of characteristics, from each other and from the training sets: some speaking styles can be considered relatively spontaneous speech (conversational speech and conversational speech (clear)), others are non-spontaneous (picture naming (clear) and sentence reading (clear)). Some are produced in “clear” speech style to enhance intelligibility for potential listeners (picture naming (clear) and sentence reading (clear), conversational speech (clear)), others are produced without trying to enhance intelligibility (e.g. conversational speech). We therefore predicted that the type of speaking style used in the low variability training set may affect performance. For this reason, we tested subsamples of participants (40 per training type, sample size determined *a priori* via a power analyses) on each of the four low variability training conditions (high variability training stimuli remained the same for all of these subsamples). Since the effect of high versus low variability training was present in 7 out of the 8 identities at were trained across participants in Experiment 1a, we opted to not counterbalance the two sets of identities used for training and distractor voices at test in this study. As in Experiment 1a, participants learned 4 new identities each, 2 based on high variability training, the other 2 on the of types of low variability training. The assignment of which identities were trained for the high versus low variability training was counterbalanced across participants.

Procedure

Procedure was the same as described in Experiment 1a.

Results

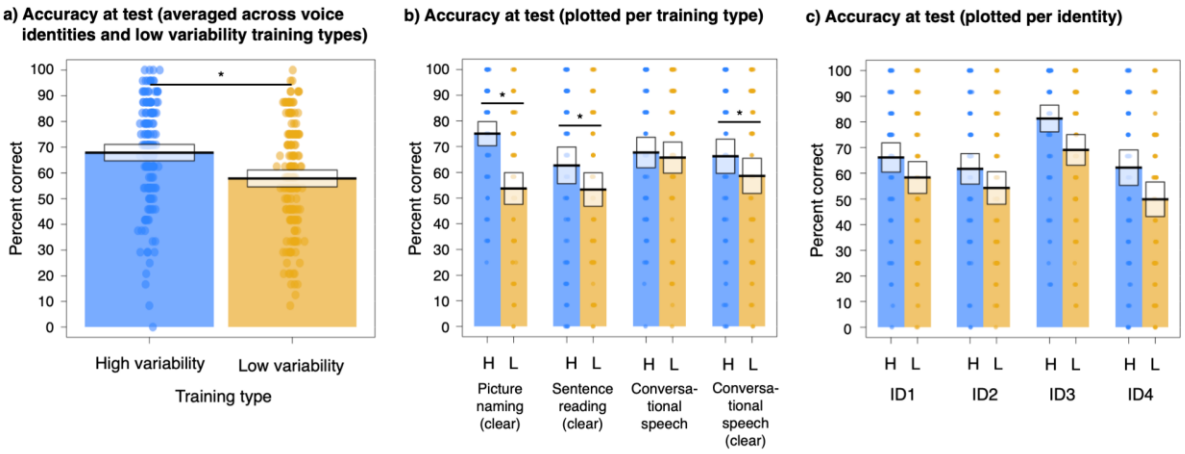

Figure 3 Summary of old/new recognition performance for learned identities in Experiment 1b. a) Accuracy for high versus low variability training is plotted averaged across all identities and all versions of the low variability training. b) Accuracy for high versus low variability training is plotted averaged across all identities by the different low variability training types. c) Accuracy for high (H) versus low (L) variability training is plotted averaged across the different low variability training types by the different identities. Due to the counterbalancing of low variability training types and assignments of high vs low variability stimuli to the trained voices, each individual bar in panel b) and c) shows the data of 20 participants. Boxes show the 95% confidence intervals, dots indicate the mean accuracy per participant.

Confirmatory analyses

As in Experiment 1a, we assessed the effect of high variability training on accuracy for the learned identities (i.e. not taking the data from the distractor identities into account). We again ran a binomial generalised linear mixed model (GLMM) using *lme4* (Bates, Maechler, Bolker & Walker, 2014) in the *R* environment (R Core Team, 2013). This model was almost identical to the one used in Experiment 1a, where training type (high vs low variability) was defined as a fixed effect. The counterbalancing (of assignments of identities to high or low variability training sets) was entered as a fixed effect in this experiment since it only had two levels in this study. Participant, speaker and stimulus were entered as random effects. In the current model, we additionally added type of low variability training as a random effect. Type of training had an effect on accuracy (coefficient of -.51, *SE* = .05) and the comparison of the full and null model was significant ($\chi^2[1] = 96.356$, $p < .001$).

Accuracy was thus higher for identities learned through high variability training (68% correct vs 58% correct; Figure 3a). High variability training was advantageous for all 4 identities (Figure 3c).

Exploratory analyses

We further explored whether this effect differed across the different low variability training conditions. For this purpose, we ran 4 GLMMs. For each GLMM we included only one subgroup of the participant sample, split by the type of low variability training they received (N = 40 per group). The models again included training type and counterbalancing as fixed effects and participant, speaker and stimulus as random effects. These models confirmed that there was a significant effect of training for 3 of the 4 low variability training types (all coefficients < -.36, all $\chi^2[1] > 12.26$, all ps < .001), but not for conversational speech (Session 1) (coefficient of -.11, SE = .11, $\chi^2[1] = 1.26$, p = .263; see Figure 3b).

Discussion

In Experiment 1b, we tested whether high variability advantages may arise when listeners are required to generalise from learned stimuli to stimuli located in a previously unheard location of the within-person voice space of an identity. Specifically, we used an experimental design in which test items did not overlap in speaking style and recording session with the items that listeners had heard during training. In Experiment 1a, where test and training stimuli overlapped in speaking style and time of recording – and most clearly so for low variability training conditions – we found a high variability disadvantage. In contrast, in Experiment 1b, we now

find a high-variability advantage. This advantage was apparent for all of the different identities and relative to 3 out of the 4 different training conditions.

The only subsample of participants that showed no advantage for high variability training were those whose experienced low variability training from conversational speech. This is the only speaking style that was not produced in adverse communication conditions (i.e. that would elicit clear speech from speakers to increase intelligibility): all other conditions either required speakers to imagine communication with a hearing-impaired person or indeed to communicate in real time with a conversation partner struggling to understand their speech (due to their speech being noise-vocoded; Hazan & Baker, 2010). These conditions thus elicited speech intended to increase intelligibility. By contrast, the conversational speech task did not impose any adverse communication conditions – and, crucially, neither did the read sentences used during test. We would thus speculate that the similarity between training and test items for these conditions may have been higher, enhancing test performance. Specifically, listeners may have needed to bridge less of a gap between test and training items, diminishing the high-variability training advantage (see Experiment 1a).

Experiment 2

In Experiment 2, we took a different approach to implementing a contrast between high and low variability training. In this study, both stimulus sets include naturally varying items. We retained the high variability stimulus sets from Experiment 1a: In these sets, all items are unique. For the low variability stimulus sets, variability was limited through presenting participants with only subsets of items from the high

variability sets, which were repeated during training to match the overall exposure to high variability (see Murphy et al., 2015 for faces). Variability here is therefore manipulated based on the assumption that each item includes novel variability, and that repeating items limits listeners' exposure to this novel variability. In contrast to the manipulations of high versus low variability in Experiment 1a and 1b, the type of variability remains constant in Experiment 2 (high variability stimulus sets, spanning different speaking style and recording sessions) but the *amount* of variability is reduced. If high variability training is advantageous for identity learning, we should see better performance for voice identities that have been learned through high variability training. If performance is matched across training types, or worse for identities learned via high variability exposure, we can conclude that there is no clear advantage for high variability training.

Methods

Participants

66 participants were recruited from Prolific.ac based on the same criteria as in Experiment 1a and 1b and were paid for their time. The study was approved by the departmental ethics committee at Speech, Hearing and Phonetic Sciences at University College London. All participants were provided with an information sheet and completed a consent form before the start of the study. We intended to test a sample of 60 participants following a power analysis using the *simr* package in *R* (Green & McLeod, 2016). This power analysis indicated that this sample size would be adequate to detect an effect of a similar size to the one detected in Experiment 1a for the main contrast of low vs high variability training. There was no overlap in participants between Experiment 1a, Experiment 1b and Experiment 2 and no

participants in Experiment 2 had taken part in any of the pilot studies associated with this set of studies. Six participants were excluded from this sample: 2 participants failed to give the correct response for more than 20% of vigilance trials, another participant's overall performance was more than 3 standard deviations below the mean performance of the sample and 3 participants did not perform significantly better than chance (± 95% confidence intervals) for the last 20% trials of Training 2. The final participant sample thus included 60 participants (mean age: 27.05 years, SD = 6.2 years; 38 female).

Stimuli

Stimulus sets for both the high variability training sets and the test stimulus set were the same as in Experiment 1a. Low variability training sets were created from subsets of the items used in the high variability training sets. Thus, low variability training sets also include items from across the 5 sessions. However, only one item per speaking style was included, and repeated during training to match the relative exposure to the different speaking styles present in the high variability training sets (see Figure 1): 1 item from Session 1 (6 repetitions), 1 item from Session 2 (6 repetitions), 1 from Session 3 (4 repetitions), 2 items from Session 4 (1 item from the picture naming task, 1 read sentence; 2 repetitions each) and 2 from Session 5 (1 item from the picture naming task, 1 read sentence; 2 repetitions each).

Procedure

The procedure was identical to that reported for Experiment 1, with only the items for the low variability training sets differing between experiments.

Results

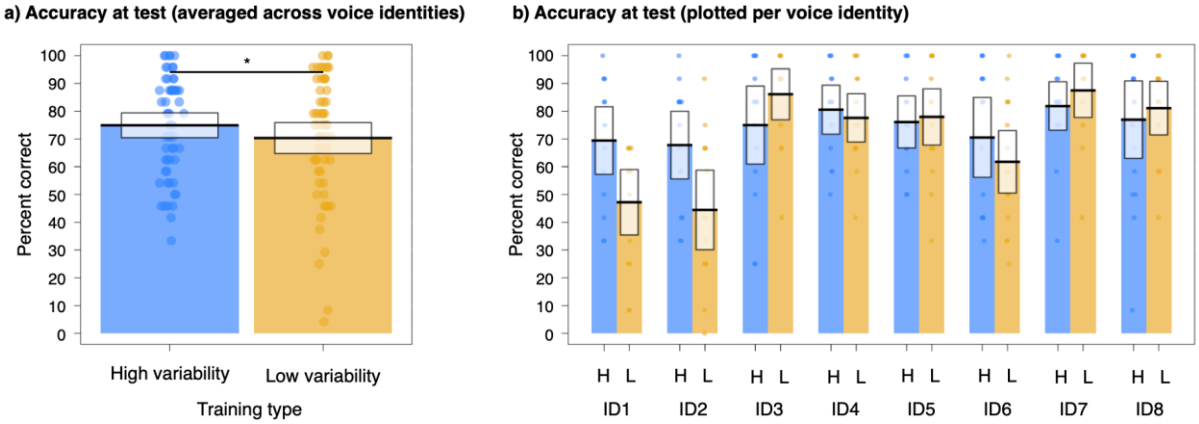

Figure 4 Summary of old/new recognition performance for learned identities in Experiment 2. a) Accuracy for high versus low variability training plotted averaged across all identities. b) Accuracy for high (H) versus low (L) variability training broken down by identity. Due to the counterbalancing of trained and distractor identities and assignments of high vs low variability stimuli to the trained voices, each individual bar shows the data of 15 participants. Boxes show the 95% confidence intervals, dots indicate the mean accuracy per participant.

Confirmatory analyses

To assess the effect of high variability training on accuracy for the learned identities (i.e. not taking the data from the distractor identities into account), we ran a GLMM that was identical to the one reported for Experiment 1a: Training type (high vs low variability) was defined as a fixed effect. Participant, speaker, stimulus as well as counterbalancing were entered as random effects. This confirmed that the type of training had an effect on accuracy (coefficient of 0.18, *SE* = .09) and that the comparison of the full and null model was significant (χ^2[1] = 3.93, *p* = .047). Accuracy was thus higher for identities learned through high variability training (75% correct vs 70% correct; Figure 4a). This is the opposite of what we found in Experiment 1a. We note that the overall accuracy for the high variability condition is very similar across the two experiments using the same high variability training and test stimuli (74% in Experiment 1a and 75% in Experiment 2).

Exploratory analyses

In line with the analyses reported in Experiment 1a, we further explored whether this effect differed across the individual voice identities. Figure 4b indicates that the trend showing a high variability advantage can be observed for only 4 out of the 8 identities. There appears to be a large advantage for high variability training for 2 identities (ID1 and ID2), which is likely to drive the result toward indicating an overall advantage for high variability training. As in Experiment 1a, we therefore ran a GLMM including speaker and training as fixed effects alongside the interaction between speaker and training. This model shows that there was again a significant interaction between speaker and training and the comparison of the full and null model was significant (χ^2[13] = 47.35, p < .001). This interaction again reflects that the effect differs across the different identities. This underlines the observation that the overall high variability advantage is driven by the two identities. Despite the presence of a significant speaker x training interaction in Experiment 1a, there was consistency in the direction of the effect for 7 out of the 8 speakers – the lack of consistency in Experiment 2 suggests much greater speaker effects that should temper the interpretation of the overall high-variability training advantage.

Discussion

Experiment 2 used a different definition for the high vs low variability comparison: in contrast to the definition employed in Experiment 1a and 1b, both high- and low variability training sets included items from across different speaking styles (see Figure 1). While the high variability training set still included 24 unique items, the low variability training set included only 6 unique items, repeated a number of times to match the relative exposure to the different speaking styles in the high variability set.

We found an overall advantage for high variability training. However, when looking at performance for the individual voice identities, the high variability training advantage (defined here as performance for identities trained through high variability exposure numerically exceeding performance for identities trained through low variability exposure) is only apparent for 4 out of the 8 voice identities. Indeed, the overall effect appears to be mainly driven by 2 identities (ID1 and ID2). For both of these identities, the low variability conditions stand out as having the worst performance out of all of the speakers (at chance level [50%]). Performance for the high variability conditions was overall well matched to performance for the same identities in Experiment 1a and 1b (~60-70% across all experiments). We inspected the accuracy for these two identities for all trials of Training 2 to further explore the origin of these results (NB feedback was provided during this training, so the overall accuracy cannot be meaningfully interpreted): Listeners' accuracy for ID1 and ID2 was 57% and 58% respectively. While these are the lowest scores across all 8 identities numerically, they do not stand out in comparison to accuracy for the remaining voices (62%-79%). This therefore indicates that listeners were able to learn the voices in the context of the training.

We propose that the small number of items included in the low variability training set resulted in poor representations of these speakers' voices (in relation to the test sets). Due to the relatively sparse training materials (only 6 unique items), listeners were apparently unable to compensate for these poor representations as they may have done for the high variability training sets including a larger number of unique items. If this is the case, this finding does not so much show a high variability advantage but a cost for this particular type of low variability training: representations

formed based on items that provide only sparse coverage of the speaker's within-person voice space may be fragile. Distances between the different sounds may be too large for listeners to adequately interpolate between them. Alternatively, the small number of unique items being repeated several times across training may have led listeners to form mostly exemplar-based representations, offering little scope for generalisation. While this did not affect performance for 6 out of the 8 tested identities, it appears that listeners could not successfully generalise from training to test for those 2 identities. These poor representations thus lead to a low variability training *disadvantage,* where some representations were not functional enough to allow listeners to recognise new recordings from the corresponding voice identities above chance level. Why this was the case for only these two identities remains unclear. However, we note that this effect is likely a stimulus set effect and not an effect of the two specific voice identities: accuracy for these two voice identities was comparable to the other 6 identities for both the low and high variability conditions in Experiment 1a and 1b. Overall, we conclude that Experiment 2 does not offer convincing evidence for a high variability training advantage.

General Discussion

In the current set of experiments, we tested whether high variability training can be advantageous in comparison to low variability training when learning new voice identities. Across three experiments, we find mixed evidence. In Experiment 1a, where low variability training stimuli and test stimuli fully overlapped in terms of speaking style and time of recording, we found a disadvantage for high variability training. In Experiment 1b, where none of the training stimuli overlapped with test stimuli, we found a high variability advantage. In Experiment 2, we implemented a

different interpretation of high versus low variability training by contrasting training sets of unique items in the high variability conditions with training sets comprising a smaller number of repeated stimuli in the low variability condition. Through, in effect, manipulating the amount of variability listeners experience in this experiment, we found a small advantage for high variability training; however, this appear to have been the result of stimulus set-specific effects.

Taken together, Experiment 1a and 1b suggest that high variability training may not be advantageous when training and test stimuli are similar to each other and thus do not require listeners to generalise to previously unexposed locations of a person's within-person voice space. Indeed, if the similarity between training and test stimuli is very high, low variability training may become advantageous (see Experiment 1a): in these cases, listeners get much exposure to the "right" kind of variability, which then provided a good match at test. However, high variability exposure during learning may be advantageous for listeners in the context of adaptively generalising from previously experienced to previous unexposed types of variability in a voice. Here, listeners may be better able to generalise because they already expect the voice to vary as a result of having been exposed to a wide range of the voice's potential variability. These findings and interpretation align with findings in the visual domain showing that the degree of category variability affects whether ambiguous or critical exemplars falling between two categories are perceived as part of one category or the other (e.g. Sakamoto, Love & Jones, 2006; Stewart & Chater, 2002). Alternatively, it could also be argued that listeners exposed to high variability training have more successfully abstracted the diagnostic, stable features of the voices due to greater exposure to the within-person variability. Whether the effects reported

here occur based on decision making criteria or based on the nature of the representation formed cannot be answered within our study but this question warrants further research.

Two of our experiments did not show conclusive evidence for high variability advantages in voice identity learning. This stands somewhat in contrast to evidence from the face perception literature, where consistent high variability advantages are reported. Divergent findings between face and voice processing are of interest in the context of a literature that has mainly stressed the parallels between the two modalities (Campanella & Belin, 2007; Yovel & Belin, 2013). While we aimed to design our study to be broadly comparable to studies reporting high variability training advantages for face identity perception, there are nonetheless many differences between our experiments and face perception studies that may explain why we may not have found a consistent advantage for high variability training across all experiments. Aside from differences in the stimuli and the specific study designs, such divergent findings may originate from general differences in the nature of face and voice perception and/or the time course of learning in these two modalities. For example, better performance on identity perception tasks for faces compared to voices has been widely reported (e.g. Barsics, 2014; Stevenage & Neil, 2014). It could therefore be the case that the differences in findings across modalities arise from the differential difficulty of processing (and learning) information from voices in the context of identity perception. If we assume that all mechanisms underpinning voice and face identity processing are similar but that voice learning is more challenging, the learning and test phases of otherwise similar face- and voice-based tasks may tap into different stages of identity learning. It has been shown that

within-person variability has detrimental on voice identity perception when dealing with unfamiliar voices (e.g. Lavan et al., 2018a). As noted above, the lack of a clear effect for Experiment 2 could there be explained by listeners struggling to cope with the within-person variability when trying to form a robust, abstracted representation due to the limited duration of exposure and/or number of unique items they were presented with during training (see for example the chance-level performance for ID1 and ID2 in Experiment 2). For faces, where identity learning may be generally faster, a broadly similar level of exposure as used in our experiments may have been sufficient to form a stable enough representation to overcome the detrimental effects of variability (e.g. Jenkins et al., 2011). Once the detrimental effects of within-person variability have been at least partially overcome by establishing such an initial representation, exposure to (high) variability may then become helpful in forming more robust representations. This line of argument may tie in with reports from the phonetic training literature (where listeners are trained to discriminate/recognise linguistic sounds): here, studies report that the high variability advantage is present for reasonably successful learners whereas less successful learners were disadvantaged by high variability training (Perrachione, Lee, Ha & Wong, 2011; Sadakata & McQueen, 2014). Thus, it could be the case that if the task is relatively easy for participants, they could cope with more variability and even benefit from it, whereas if the task is more difficult (e.g. faces versus voices, successful vs unsuccessful learners, easy-to-learn voices vs hard-to-learn voices) high variability may not confer advantages or may indeed be costly.

We note that we opted to use speech stimuli because these are most representative of what listeners experience of human voices on a daily basis and manipulated the

amount of type of variability in our experiments via changes in the number of speaking styles, unique items and recording times included within the different training conditions. Voices, however, vary in any number of ways that were not manipulated in the current study (e.g. Lavan et al., 2018 for a review of variability in the voice). It is possible that the kind of variability we contrasted within our stimuli may not be the most relevant type of variability used during voice identity learning, such that effects high variability training may in fact not be reliably detected with our stimuli. For example, items in both of our training sets included full meaningful utterances. In the context of voice learning, a substantial amount of information is encoded in such stimuli, with listeners being able to sample a large number of phonemes from each voice identity in *both* high and low variability training conditions. It has previously been proposed that the accuracy of recognition of familiar voices is closely related to stimulus content as indexed by the number of phonemes and that this is possibly more influential than simply increasing stimulus duration (e.g. Bricker & Pruzansky, 1966). Future work defining high versus low variability training based on different types and sources of variability present in voices will be necessary to build up a more comprehensive account of its effect on voice identity learning and perception.

The question whether high variability training is advantageous during learning has been asked in a number of different contexts in the auditory domain: studies have explored the effects of high variability training when learning to perceive new phonetic contrasts (e.g. /r/ and /l/ in Japanese listeners; Lively, Logan & Pisoni, 1993), learning the meaning of new words in a foreign language (Barcroft & Sommers, 2005, Sommer & Barcroft, 2006, 2007) and for the learning of native

dialects (Clopper & Pisoni, 2004). Voice (and similarly face) identity learning is thus only a small part of the expansive literature. These other literatures may provide further insights into how to further probe this question as there a number of ways in which high variability training could be advantageous during voice identity learning that the current set of studies has not yet addressed. In general, only little is known so far about how representations of individual identities are formed from variable signals (but see Lavan et al., 2019). Variable signals are however exactly what we encounter in everyday life, when we learn to recognise a new person. We therefore believe that this warrants further work looking at the nature of within-person variability in voices and how this variability affects and interacts with voice identity perception and learning on different timescales.

References

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2018). Gorillas in our Midst: Gorilla. sc. Behavior Research Methods.

Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition*, *161*, 19-30.

Baker, R., & Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods*, *43*(3), 761-770.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387-414.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Barsics, C. G. (2014). Person recognition ss easier from faces than from voices. *Psychologica Belgica*, *54*(3), 244-254.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version*, *1*(7), 1-23.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441-1449.

Burton, A. M., Jenkins, R., Hancock, P. J., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive psychology*, *51*(3), 256-284.

Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223.

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

Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *journal of phonetics*, *32*(1), 111-140.

Erickson, M. L., & Phillips, P. (2018). Can Listeners Hear Who Is Singing? The Role of Listener Experience in Singer Discrimination Across Pitch. *Journal of Voice*. [Epub ahead of print].

Green P. & MacLeod C.J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498.

Hazan, V., & Baker, R. (2010). Does reading clearly produce the same acoustic-phonetic modifications as spontaneous speech in a clear speaking style?. In *DiSS-LPSS Joint Workshop 2010*, p 7-10.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1671-1683.

Latinus, M., & Belin, P. (2012). Perceptual auditory aftereffects on voice identity using brief vowel stimuli. *PLoS One*, *7*(7), e41384.

Lavan N., Burston L.F.K., & Garrido L. (2018). How many voices did you hear? Natural variability disrupts identity perception in unfamiliar listeners. *British Journal of Psychology.* [Epub ahead of print].

Lavan N., Burston, L.F.K., Merriman, S.E., Ladwa P., Knight, S. & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology.*

Lavan N., Burton A.M., Scott S.K., & McGettigan C. (2018). Flexible voices: identity perception from variable vocal signals. *Psychonomic Bulletin & Review*. [Epub ahead of print].

Lavan, N., Knight, S., & McGettigan, C. (2019). *Listeners form average-based representations of individual voice identities even when they have never heard the average. Nature Communications.*

Lavan, N., Merriman, S. E., Ladwa, P., Burston, L., Knight, S., & McGettigan, C. (2018d). “Please sort these sounds into 2 identities”: Effects of task instructions on performance in voice sorting studies. *PsyArXiv*.

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, *145*(12), 1604-1614.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the acoustical society of America*, *94*(3), 1242-1255.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 577-581.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461-472.

Peynircioğlu, Z. F., Rabinovitz, B. E., & Repice, J. (2017). Matching Speaking to Singing Voices and the Influence of Content. *Journal of Voice*, *31*(2), 256-e13.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*(5), 897-905.

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in psychology*, *5*, 1318.

Sakamoto, Y., Love, B. C., & Jones, M. (2006). Tracking variability in learning: Contrasting statistical and similarity-based accounts. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society. Vancouver, Canada: Cognitive Science Society*.

Sommers, M. S., & Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, *119*(4), 2406-2416.

Sommers, M. S., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics*, *28*(2), 231-249.

Stacey, P., Dunn, A., Smith, H., Robson, J., & Baguley, T. (2018). Forensic voice discrimination: The effect of speech type and background noise on performance. Applied Cognitive Psychology. [Epub ahead of print].

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica*, *54*(3), 266-281.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 893-807.

Team, R. C. (2013). R: A language and environment for statistical computing.

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781-790.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064-2072.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in cognitive sciences*, *17*(6), 263-271.