University of Glasgow School of Psychology

ID NUMBER: 2304168

DATE: 05.04.2021

WORD COUNT: 8295

PSYCH4007P: L4H Dissertation

2020-21

Cue or Noise: A Computational Analysis of Within-Speaker Variability and its Role in Voice Identity Perception

Acknowledgements

First of all, I would like to thank Dr Philip McAleer for supervising my project. Without your encouragement, I might not have pursued the risky but more interesting path and this dissertation would not be the same today.

To Gaby Mahrholz for sharing her work with me and making this project possible in the first place.

To my friends and my partner for their relentless support throughout this pandemic and this project.

And finally, to my parents, without whom I would not be here today. Ich danke euch für die Wurzeln und die Flügel und liebe euch über alles.

Contents

Acknowledgements 1

Contents 2

Abstract 4

Cue or Noise: A Computational Analysis of Within-Speaker Variability and its Role in Voice Identity Perception 5

Between-Speaker Variability 5

Within-Speaker Variability 8

Machine-Learning Approaches to Voice Identification 11

The Present Study 13

Methods 15

Ethics Statement 15

Open Science Statement 15

Participants 15

Speakers 15

Materials 16

Voice Stimuli 16

Acoustical Features 16

Model 17

Procedure 19

Training and Testing Models 19

Data Analysis 21

Results..21

Exploratory Analysis 24

Discussion 27

References...36

Abstract

Research on voice identity perception has extensively studied between-speaker variability in an attempt to understand how listeners tell people apart. Within-speaker variability, the way the same voice varies across situations, has mostly been eliminated in these studies, effectively treating it as a form of random noise that impairs voice identity perception. However, recent studies focusing on the ability to tell people together suggest that within-speaker variability is speaker-specific and plays an equally important role in identifying others. This study attempted to quantitatively verify these claims by training and testing a machine-learning model on variable and non-variable voice samples, examining emerging differences in identification accuracy based on differences in their training and test sets. Our results revealed that the complexity of variable voices makes them more difficult to identify compared to non-variable voices. However, the model was also able to extract speaker-specific information from variable voice samples that improved identification accuracy for variable voices. These findings show that within-speaker variability provides important cues for the identification of naturally varying voices, highlighting the need for further research on its role in voice identity perception, as well as demonstrating how the use of machine-learning models can benefit future psychological research.

Cue or Noise: A Computational Analysis of Within-Speaker Variability and its Role in Voice Identity Perception

The ability to recognise and identify other individuals is a critical function for human social interaction, which relies on a complex interplay between perceptual inputs, memory and semantic knowledge (Barton & Corrow, 2016). Cognitive models attempting to explain this mechanism suggest that identity perception is facilitated through a parallel processing approach that incorporates three major sources of information: facial features, names and vocal cues (Gainotti, 2014). The latter of these components, vocal cues, can convey large amounts of information about a speaker, informing others of characteristics such as their age (Goy et al., 2016; Ma & Wu, 2020), gender (Bishop & Keating, 2012; Whiteside, 1998), weight (Bruckert et al., 2006; Souza et al., 2018) and personality (Belin et al., 2017; McAleer et al., 2014).

Our ability to process these vocal cues and use them to identify others is developed early on. Fetuses, for instance, start attending to sounds of speech in gestation and newborns can already distinguish between different individuals (Bahrick et al., 2003) and recognise their parents' voices (DeCasper & Fifer, 1980; DeCasper & Prescott, 1984). By the time we reach adulthood, we can identify several individuals from only a few syllables of speech (Pollack et al., 1954; Schweinberger et al., 1997) or even obscured speech such as whisper (Smith et al., 2017). This advanced ability to distinguish and identify individuals by their voice can be divided into two major functions (Lavan et al., 2019a): telling speakers apart by focusing on between-speaker variability and telling people together in different instances by focusing on within-speaker variability. However, it currently remains unclear how these two functions interact with each other, with the majority of research studying them as separate processes.

Between-Speaker Variability

Our ability to tell people apart is based on between-speaker variations in voice production, which is primarily defined by two major elements, the vocal folds and the vocal tract (Fant, 1971). The vibrating vocal folds in the larynx act as a sound source while the resonances produced by the vocal tract act as a filter. The dimensions of the vocal folds determine the fundamental frequency (F0) of a voice and therefore its pitch (Ghazanfar & Rendall, 2008). Female vocal folds are shorter and lighter than male vocal folds, therefore vibrating at roughly double the frequency (200-240Hz) compared to those of the average male (100-120Hz). Similarly, the size and structure of the vocal tract influence its resonance characteristics, leading to the reinforcement of specific frequencies referred to as formant frequencies (Latinus & Belin, 2011a). Both fundamental frequency and formant frequencies play a crucial role in processes that rely on vocal cues such as gender categorisation (Gelfer & Bennett, 2013; Neisi et al., 2019; Pernet & Belin, 2012). However, fundamental frequency also differs between individuals of the same gender and provides a unique source of information that enables us to identify an individual both via spoken language and non-verbal vocalisations (Pisanski et al., 2020). Similarly, negative formant dynamics, referring to the speed at which a formant decreases from its peak frequency to its adjacent minimum, vary from individual to individual, with the negative dynamics of the first formant accounting for 70% of perceived between-speaker variability (He et al., 2019). However, beyond fundamental frequency and formant frequencies, individual voices also vary due to behavioural differences in speech tempo and rhythm (Dellwo et al., 2015) and other voice quality characteristics (Podesva & Callier, 2015). This demonstrates that our rather intuitive ability to tell people apart is based on a complex interplay of both simple acoustical features and speech characteristics.

Beyond uncovering potential acoustical features that facilitate our ability to identify voices, research has also explored ways to formalise the mechanisms by which these features are encoded as voice identities. In an attempt to create a model that describes how individual voice identities are encoded, Baumann and Berlin (2008) asked participants to make similarity judgments for a range of different male and female voices uttering three different vowels. A multidimensional analysis revealed that both male and female voices could each be represented in a two-dimensional perceptual space, most accurately defined by the fundamental frequency (F0) and the first formant frequency (F1) of a given voice. Using a similar model, Latinus et al. (2013) found that voices are coded in relation to a sex prototype. MRI scans showed that neural activity correlated with the perceptual distance between a voice identity and the corresponding sex prototype. However, their chosen perceptual space was not only defined by fundamental frequency and formant dispersion but also by harmony to noise ratio (HNR), which captures the amount of added noise in a voice and its perceived roughness (Awan & Frenkel, 1994). While their model represents a good approximation, Latinus et al. (2013) admit that the true voice space is likely to include a much larger number of complex dimensions that correspond to various acoustical features. Beyond HNR research established other objective voice quality features such as jitter, the variation in frequency around the fundamental frequency, and shimmer, the variation in amplitude (Farrus et al., 2007). While all three of these features appear to contribute to differentiating between voice identities (Kreiman et al., 1992), listeners do not seem to be able to accurately perceive these features, rather using them inconsistently across different voice identities (Mathias & Kriegstein, 2014). Listeners appear to use a different combination of acoustical features for each voice identity (Lavner et al., 2000), making it impractical to define a specific set of acoustical features that facilitates voice identity perception. Instead, there is a large number of acoustical features that can influence voice identity perception (Kreiman & Sidtis, 2011).

This more nuanced perspective that voice identity is based on more than just a few acoustical features is supported by research going beyond objective acoustical measures. Perrachione et al. (2009) reviewed studies examining voice identity perception across languages and found that voice identification is significantly impaired if a listener does not understand the speaker’s language. In a follow-up study, Perrachione et al. (2011) found that dyslexic participants performed worse on a speaker identification task than controls. These findings show that voice identification also relies on higher-level features encoded in language, a theory supported by Kamide (2012) who observed that listeners associate the structural speech preferences of a speaker with their identity. An experiment by Kreiman and Emmorey (1985) shows how these cues are used in practice. They presented participants with voice samples from celebrities before sorting them into familiar and unfamiliar identities for each participant. Participants then were tested on samples played in reverse, a process that eliminates most idiosyncratic features of speech. Their results showed that this process rendered familiar voices unrecognisable whereas unfamiliar voices were still recognised normally by participants. Together with the findings on between-speaker discrimination discussed above, this observation suggests that the identification of familiar voices is more reliant on idiosyncratic phonetic and articulatory information while unfamiliar voices are primarily identified based on voice quality features such as fundamental frequency and formants (Schweinberger et al., 2014).

Within-Speaker Variability

Research on between-speaker variability has produced a significant number of findings supporting the notion that beyond core acoustical features such as fundamental frequency and formant frequencies, voice identity perception also heavily relies on the specific manner in which an individual speaks and adapts their speech in different situations (Kreiman et al., 2015). Most studies on voice identity perception have largely ignored these individual nuances, removing as much within-speaker variability from stimuli as possible in an attempt to emphasise between-speaker variability (Lavan, 2019). As a result, research on voices has generally limited its scope to the ability to tell people apart. However, as outlined earlier, voice identity perception is also based on our ability to tell people together in different contexts. To fill this gap, Lavan et al. (2019b) presented participants with naturally varying voice samples from popular TV characters and asked participants to sort the samples into perceived identities. Results showed that listeners who were unfamiliar with the TV show reported more perceived identities than the sample contained, pointing towards a systematic failure in telling people together. Consequently, a lack of research on within-speaker variability ignores not only a natural component of voices but also the function that might be impaired the most by this variability.

In practice, the acoustical features of a single voice can greatly vary from one situation to another, for example when trying to convey different emotions (Patel et al., 2011) or dominance (Cheng et al., 2016), addressing babies or animals (Burnham, 2002) or in non-verbal vocalisations such as laughter or screams (Pisanski et al., 2020). Since studies show that naturally varying voices are more difficult to identify compared to voices without this variability (Lavan et al., 2019b; Lavan et al., 2019c), it seems intuitive to assume that within-speaker variability is just an added layer of random noise that lacks informational value. However, research on face identity perception has produced findings that suggest otherwise. Experiments in face research found within-person variability to be a crucial component of the face recognition process (Burton, 2013; Burton et al., 2016). Although high within-person variability impairs our ability to consistently identify faces (Jenkins et al., 2011), follow-up studies revealed that the within-person variability observed across different pictures of the same face is highly person-specific and cannot be generalised to the entire population. As such, within-person variability could be a source of additional information that is valuable for face identity perception (Murphy et al., 2015; Ritchie & Burton, 2017).

Since face and voice perception are parallel processes in current identity perception models (Gainotti, 2014), within-speaker variability likely plays a similar role in voice identity perception. Two recent studies by Lavan et al. (2019d, 2019e) attempted to examine the impact of within-speaker variability in voice identification tasks. In the first study, Lavan et al. (2019d) created three unique voice identities and introduced a controlled form of within-speaker variability for each identity by modifying the glottal pulse rate and vocal tract length, which correspond to fundamental frequency and formant frequencies, respectively (Smith & Patterson, 2005). These modified samples were distributed in clusters so that the original voice identities were at the relative centres of these clusters. All samples were mapped onto a two-dimensional perceptual space defined by the glottal pulse rate and vocal tract length, similar to the norm-coding based perceptual space suggested by prior voice identity perception research (Baumann & Belin, 2008; Latinus et al., 2013; Latinus & Belin, 2011b). Results showed that even when participants were trained only on the modified voice samples, participants were still most accurate in identifying the original voice samples at the centre of the identity's cluster, even though participants were never presented with these original voice samples. This implies that the variable voice samples helped participants form an average-based identity for each speaker that corresponded to the non-variable original sample. These findings match those from face research (Burton et al., 2016; Murphy et al., 2015; Ritchie & Burton, 2017), suggesting that within-speaker variability is useful for building robust and unique voice identity perceptions.

In a follow-up study that also examined the impact of within-speaker variability, Lavan et al. (2019e) performed several experiments in which they trained participants on either low-variability voice samples consisting of recordings of read speech or high-variability voice samples consisting of recordings from various speech styles. Participants were then asked to identify voice samples of read speech as a product of either a new or an old voice. Results from the first experiment showed that high-variability training led to a statistically significant decrease in identification accuracy compared to low-variability training. However, Lavan et al. (2019e) pointed out that due to an overlap in speech style between the low-variability training set and the test set the difference in accuracy could be unrelated to the degree of the variability itself. To confirm this assumption, in the second experiment any overlap in speech style between the training sets and the test set was eliminated, with the same test set being used as in the first experiment. Under these new conditions, high-variability training now significantly increased identification accuracy compared to low-variability training. Lavan et al. (2019e) suggested that high-variability training could provide important information specifically when individuals generalise from previously heard speech to novel speech styles, overall indicating that within-speaker variability was, in fact, informative for voice identity perception under certain circumstances.

Considering findings from face research (Burton et al., 2016; Murphy et al., 2015; Ritchie & Burton, 2017) and voice research (Lavan et al., 2019d; Lavan et al., 2019e) there is support for the theory that within-person variability is beneficial to identity perception. In voice identity perception specifically, within-speaker variability appears to be unique to each individual, providing additional information instead of just being a source of random noise. Lavan et al. (2019a) suggested that one potential avenue to quantitatively test this hypothesis could be the use of computational models to create voiceprints that formally capture the uniqueness of within-speaker variability. Previous attempts in this area failed to establish robust models due to a lack of computing resources and appropriate technologies (Hollien, 2002). However, recent advances in computational methods render this approach potentially viable today.

Machine-Learning Approaches to Voice Identification

One of the greatest advances in computational methods over the last decade has been the widespread emergence and application of sophisticated machine-learning models. While these models have been used in areas such as clinical medicine (Obermeyer & Emanuel, 2016), psychiatry (Orrù et al., 2012; Vieira et al., 2017) and forensics (Pace et al., 2019), psychological research has so far made little use of them, favouring more traditional statistical measures such as p-values and effect sizes (Orrù et al., 2020). To understand how machine-learning models can benefit research on within-speaker variability and voice identity perception, it is crucial to understand how they differ from the traditional statistical approaches used in psychological research.

Breiman (2001), in a review, characterises the field of statistics as the interplay of three major components: a set of independent variables, a "black box" that represents nature performing an unknown process and a set of dependent variables that emerge as a result of the independent variables passing through the black box. The approach of statistical inference widely used in psychology and other areas of scientific research since the 1940s assumes the black box to be a stochastic model which randomly draws from a probability distribution (Çınlar, 2011). These approaches consequently attempt to fit a data model that best explains this stochastic model through which independent variables from a given sample are transformed into the observed dependent variables. Machine-learning models on the other hand do not try to fit a model that describes the black box and instead attempt to find an exact algorithm that can predict the dependent variables as accurately as possible given the independent variables from the sample. In other words, a machine-learning model learns from the provided data, allowing it to generalise to unseen samples and to make accurate predictions beyond its original training sample.

The ability of machine-learning models to process large and complex sets of independent variables, produce accurate predictions and generalise to new samples has led to their increasing use in voice research, both for speaker identification (Boles & Rad, 2017, Ittichaichareon et al., 2012) and recognition of speaker emotions (Gumelar et al., 2019, Mamyrbayev et al., 2019). Specifically, some studies have examined within-speaker variability using computational models. A recent study by Afshan et al. (2020) presented pairs of voice samples to a machine-learning-based voice identification system and human participants to examine the effects of variable speech on identification accuracy. Their results showed that the model, whilst using a different approach to voice identification compared to human controls, performed similarly to human participants in terms of identification accuracy. Furthermore, observations showed that both the machine-learning system and human participants performed better when presented with pairs of voice samples that were both from the conversational (variable) speech condition, rather than one being read speech (non-variable) and the other one conversational. This implies that in style-matched pairs the similar level of within-speaker variability in both voice samples made it easier to correctly identify whether they belonged to the same voice identity. Park et al. (2016) produced similar findings, showing that a mismatch in speech style reduced identification accuracy when using a similar machine-learning model.

These findings show that machine-learning models can be useful in examining the effects and properties of within-speaker variability, allowing us to make inferences about voice identity perception in humans. Among the large variety of models used in current research, artificial neural networks have emerged as an adaptation of the human brain to a computational model. Mimicking the hierarchical structure of neurons found in humans, these models can directly learn from data and theoretically perform any desired function (Gupta, 2013) and are increasingly utilised in psychological research and neuroscience (Jamshidi et al., 2018; Małgorzata, 2003). As a result, artificial neural networks represent one of the most promising avenues to study within-speaker variability computationally.

The Present Study

Given the focus of prior research on between-speaker variability and the potential usefulness of machine-learning in psychological research suggested by recent studies, this study aimed to expand the line of research by Lavan et al. (2019a) by utilising a machine-learning model, specifically an artificial neural network, to examine whether within-speaker variability is unique to an individual and provides valuable information for voice identity perception. Since machine-learning models perform similarly to humans in voice identification tasks (Afshan et al., 2020; Park et al., 2016), they might enable us to make inferences about the effects of within-speaker variability based on observed identification accuracy, using training and test sets with different speech styles similar to the design employed by Lavan et al. (2019e). As the model's predictions are entirely based on the provided inputs, all variations in identification accuracy on average should occur as a result of the chosen training and test sets. In other words, by controlling and varying speaker variability through the use of different training and test sets, we can gain a better understanding of the impact of within-speaker variability on the voice identification process and its accuracy. To keep results as close to human research as possible we decided to use an artificial neural network to categorise voice samples according to voices identity. By training and testing the model repeatedly on different combinations of training and test sets, we replicated the experimental process of training individual participants under different conditions, allowing us to use statistical analysis to compare overall identification accuracy under different conditions.

Based on prior research, we developed two hypotheses. Firstly, we hypothesised that identification accuracy will be lower for models trained and tested on variable voice samples, compared to models trained and tested on non-variable voice samples. Looking at studies with human participants, variable voices are generally harder to identify as a result of the higher complexity of variable vocal signals (Lavan et al., 2019b; Lavan et al., 2019c). Therefore, the model's identification accuracy should be lower when tested on variable samples, compared to testing on non-variable samples. Secondly, we hypothesised that identification accuracy will be higher for models trained and tested on variable voice samples, compared to models trained on non-variable voice samples and tested on variable voice samples. According to Lavan et al. (2019a), within-speaker variability could be speaker-specific and therefore play an important role in voice identity perception, a theory supported by findings in face (Burton et al., 2016;

Murphy et al., 2015; Ritchie & Burton, 2017) and voice research (Lavan et al., 2019d; Lavan et al., 2019e). If within-variability were to be random or not specific to an individual, training on variable voice samples should not provide a significant increase in accuracy over training on non-variable voice samples.

Methods

Ethics Statement

This study was approved by the University of Glasgow College of Science and Engineering and complies with the BPS Code of Ethics and Conduct. Due to the nature of this project, no participants were directly involved as we only used data from an already existing dataset. The data used was collected by a prior project that received full approval from the University of Glasgow College of Science and Engineering (Ref: 300170044) and also complied with the BPS Code of Ethics and Conduct.

Open Science Statement

All analysis scripts and coding notebooks have been uploaded to the Open Science Framework (Foster & Deardorff, 2017) and are publicly available (https://osf.io/zv3at/?view_only=2b3568d1c27d4f3789b551cdfeddff1e). This serves to facilitate a more transparent research process, allowing for public review of our methodology by other researchers. The voice samples used in this study will be made available at a later date by the project that originally collected them.

Participants

Speakers

The voice samples used in our experiment were collected from 10 male individuals between the age of 18 and 28 (M = 21.7, SD = 3.83) who were recruited for voice recordings by a previous project of the University of Glasgow's School of Psychology (Ref: 300170044).

All participants were native English speakers, with 9 born and raised in Scotland and one raised by Scottish parents. As a result, they all shared a similar accent, predominantly from the Scottish Central Belt area. None of the participants reported any form of hearing or verbal impairment.

Materials

Voice Stimuli

Our vocal data set used to train and test our model consisted of 40 recordings for each of the 10 speakers, for a total of 400 voice samples. Participants were recorded in an anechoic chamber at the University of Glasgow's School of Psychology using Audacity (.wav format, 16-bit mono, 44100 Hz) (Audacity, 2021). Each speaker completed two sets of recordings, reading a set of the same 20 sentences from the Harvard sentences inventory (see List 1 and 2 at https://www.cs.columbia.edu/~hgs/audio/harvard.html) with specific speaking instructions for each set: one in a neutral voice (treated as non-variable) and one as if talking to a family member (treated as variable). Sentences that contained a stutter or other distortions were re-recorded. Each recording set was manually edited by us and cut into 20 individual voice samples. Samples had a mean duration of 3.82 seconds (SD = 1.12).

Acoustical Features

To provide our model with a set of meaningful inputs, we extracted acoustical features from each voice sample using Python v.3.8.5 (Pilgrim & Willison, 2009) and the surfboard library v.0.2.0 (Lenain et al., 2020) using the default settings (specified below when available). Based on prior research, we decided to extract three types of features:

1. Fundamental Frequency (F0) statistics (mean and standard deviation; hop_length_seconds=0.01, method='swipe') that correspond to the vibrations of the vocal folds. Prior research has established that voice identities appear to be formed in relation to a prototype in a two-dimensional perceptual space, with fundamental frequency being one of

these defining dimensions (Baumann & Belin, 2008). In addition, research on voice identification systems suggests that the inclusion of voice quality features such as fundamental frequency can significantly improve identification accuracy (Park et al., 2017), making it a relevant feature from both a psychological and a technical perspective.

2. Formant Frequencies (F1 - F4) that correspond to the resonance frequencies emphasised by the vocal tract. Formant frequencies (either individually or in form of formant dispersion) represent the other primary dimension used in encoding voice identities (Baumann & Belin, 2008), while also being part of the voice quality features that can improve accuracy in voice identification systems (Park et al., 2017).

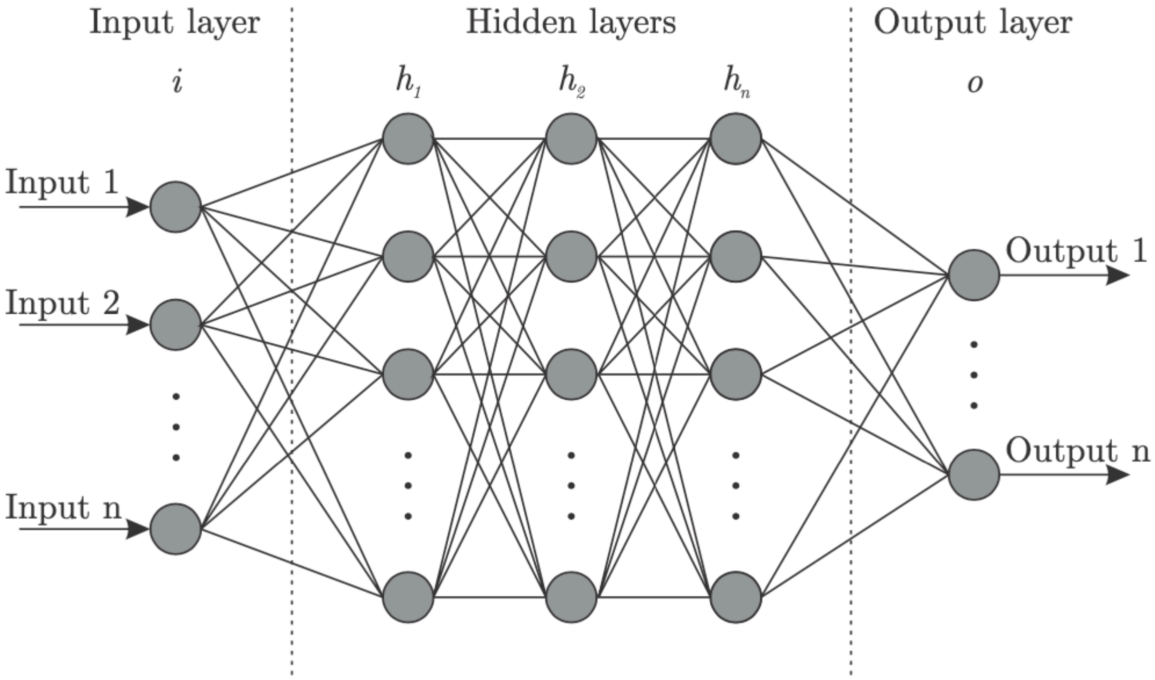
3. Mel Frequency Cepstral Coefficients (MFCCs) (mean and standard deviation; n_mfcc=13, n_fft_seconds=0.04, hop_length_seconds=0.01). This audio transformation technique has been widely used in voice identification systems as it replicates the human peripheral auditory system (Tiwari, 2010). Audio is first filtered through a mel-scale filter that matches the non-linear scale found in human hearing. The resulting mel spectrum is then converted back into the time dimension, providing us with 13 coefficients that provide a good characterisation of the spectral properties of each voice sample (Alim & Rashid, 2018).

Model

The voice identification model used in this study is based on an artificial neural network (see Gupta (2013) for a review). These machine-learning models simulate the structure found in the human brain, with each neuron represented by a simple activation unit that only passes on a signal if it receives a sufficiently strong signal from a set of other units. In a feedforward network like the one used in this study, units are arranged in a hierarchical structure of several layers through which signals travel strictly in one direction (see Figure 1). While units in the same layer cannot directly interact with each other, each unit shares a direct connection with every individual unit in the next layer.

Figure 1

Structural illustration of an artificial neural network (Bre et al., 2017)



The network used in our experiment consists of one input layer with 32 units (corresponding to 32 acoustical features), two hidden layers with 128 units each and one output layer with 10 units (corresponding to 10 possible voice identities). All units in the input layer and hidden layers use a rectified linear activation function (see Figure 2), while output units use a softmax function (see Figure 3), a choice based on prior research on the optimal design of voice identification systems using artificial neural networks (Zeiler et al., 2013).

Figure 2

Rectified linear unit (ReLU) activation function (Bhurtel et al., 2019)

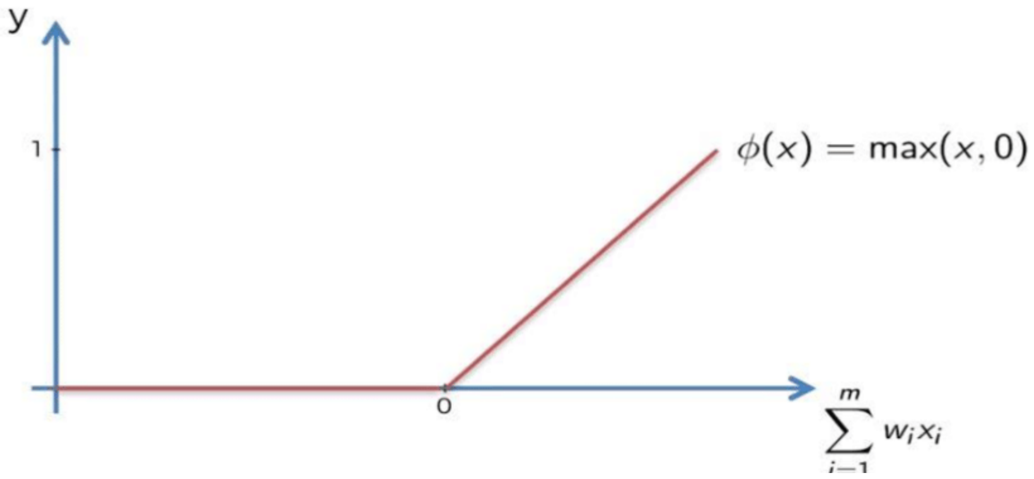


Figure 3

Softmax function (Wang et al., 2018)

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} (i = 1, 2, \ldots, N)$$

Procedure

Training and Testing Models

After importing voice samples and extracting the acoustical features outlined above from each voice sample, the numerical values of features were standardised (or z-scored) with the scikit-learn library (Pedregosa et al., 2011) as this process improves performance in many machine-learning models (Guyon & Elisseeff, 2006). This resulted in 40 feature sets for each identity, with 20 corresponding to variable and 20 to non-variable voice samples. However, as the sentences are the same in both conditions, there are only 20 unique sentences, limiting training and test sets to 10 voice samples each, to avoid any overlap. We created three types of training and test sets: all variable, all non-variable or an equal mix of both. Since each model uses one training set and one test set, there were eight unique conditions a model could be allocated to (see Table 1). Models from each condition were trained iteratively with the TensorFlow v.2.4.0 (Abadi et al., 2016) and Keras v.2.4.3 (Chollet, 2015) libraries. Each model was trained over 75 epochs on a random selection of 10 audio samples per speaker for a total of 100 samples from the specified training condition. The model was then tested on the remaining 100 samples from the specified test condition. After training and testing a model, the identification accuracy of the final epoch was stored in a data frame, providing one data point per model.

Table 1

Outline of model conditions based on the type of voice samples in training and test sets

Model Condition	**Training Set**	**Test Set**
NV-NV	Non-variable	Non-variable
V-V	Variable	Variable
NV-V	Non-variable	Variable
V-NV	Variable	Non-variable
NV-NV+V	Non-variable	Non-variable and variable
V-NV+V	Variable	Non-variable and variable
NV+V-NV	Non-variable and variable	Non-variable
NV+V-V	Non-variable and variable	Variable

Before our experiment, we performed a power analysis using the pwr library (Champely, 2020) to establish how many data points we would need to detect small effect sizes ($g <= 0.2$) at a power of at least 0.8. The analysis suggested a minimum of 310 models per condition. As a result, we decided to train and test 500 models per condition for a total of 4000 models, achieving a power of 0.89.

Data Analysis

Our analysis of identification accuracies was performed using R v.4.0.4 (R Core Team, 2021) and RStudio v.1.4.1106 (RStudio Team, 2021). As all the available data was generated in a controlled manner, no data points were excluded, resulting in a sample of 4000 observations across 8 conditions. We then performed independent t-tests to compare identification accuracies between pairs of two conditions and establish whether differences were statistically significant. To control for an increased risk of a Type I error due to repeated statistical tests, we decided to use a Bonferroni Correction (Napierala, 2012) to adjust our critical cut-off value. Based on this procedure, we divided our initial alpha ($\alpha = 0.05$) by the number of tests we ran ($N = 5$), resulting in a new alpha of $\alpha = 0.01$.

Results

This study aimed to establish whether within-speaker variability is unique to an individual and provides valuable information for voice identity perception. Consequently, we specified two hypotheses that allow us to test this assumption. Firstly, we hypothesised that identification accuracy will be lower for models trained and tested on variable voice samples (V-V), compared to models trained and tested on non-variable voice samples (NV-NV), as the added within-speaker variability in variable voices increases complexity, making identification more difficult. Secondly, we hypothesised that identification accuracy will be higher for models trained and tested on variable voice samples (V-V), compared to models trained on non-variable voice samples and tested on variable voice samples (NV-V), as the variable training set contains speaker-specific information regarding within-speaker variability, improving identification accuracy for variable voices.

Figure 4

Violin- and boxplots of identification accuracy for NV-NV and V-V models

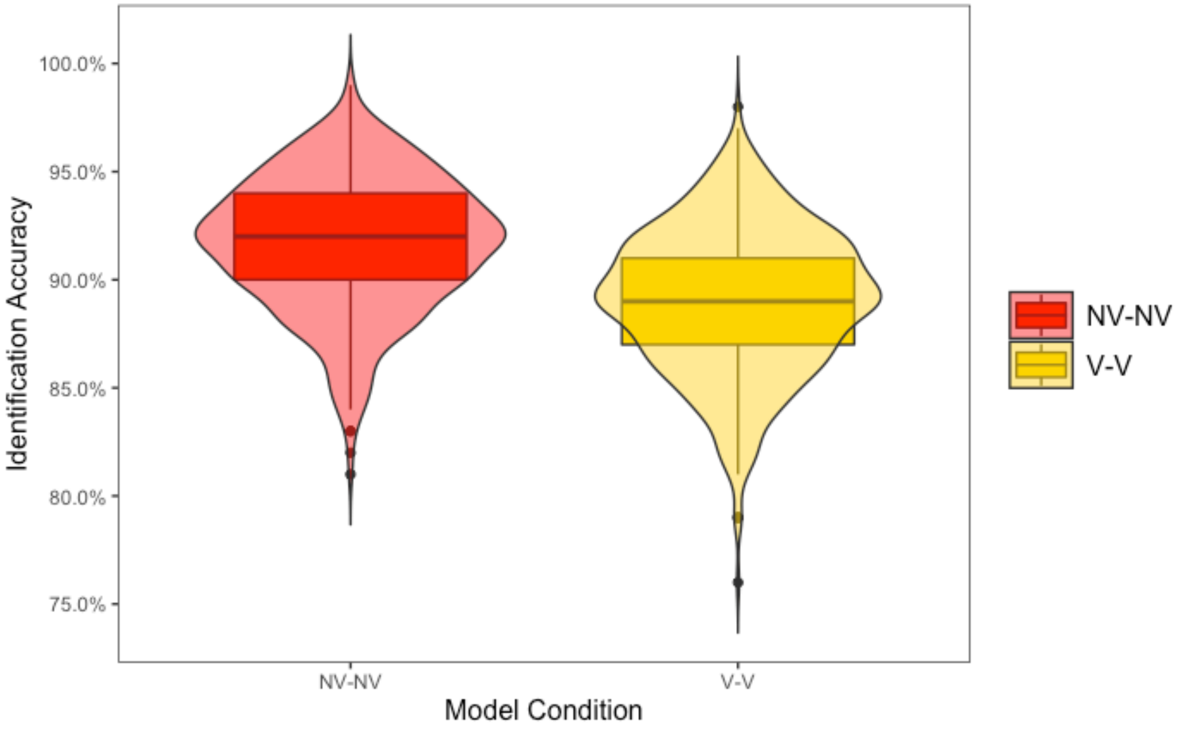


Table 2

Means and standard deviations of identification accuracy for NV-NV and V-V models

Model Condition	**Identification Accuracy**	
	Mean	**Standard Deviation**
NV-NV	91.68%	3.09%
V-V	89.07%	3.37%
Difference	2.61%	

All data relating to our first hypothesis that identification accuracy will be lower for models trained and tested on variable voice samples (V-V), compared to models trained and tested on non-variable voice samples (NV-NV), can be found in Figure 4 and Table 2. Looking at the boxplots it appears that, in line with our hypothesis, V-V models achieved an average identification accuracy (M = 89.07%, SD = 3.37%) lower than that of NV-NV models (M = 91.68%, SD = 3.09%). All assumptions required for a two-sample one-sided Welch's t-test were checked through visual inspection and found to hold. As suggested by our data, the t-test (t(990.63) = 12.74, p < .001, g = 0.81) revealed a statistically significant difference between V-V and NV-NV models, showing that models trained and tested on variable voice samples perform better than models trained and tested on non-variable voice samples. These results match our expectations, as the additional complexity in variable voices should make identification more difficult in contrast to non-variable voices.

Figure 5

Violin- and boxplots of identification accuracy for V-V and NV-V models

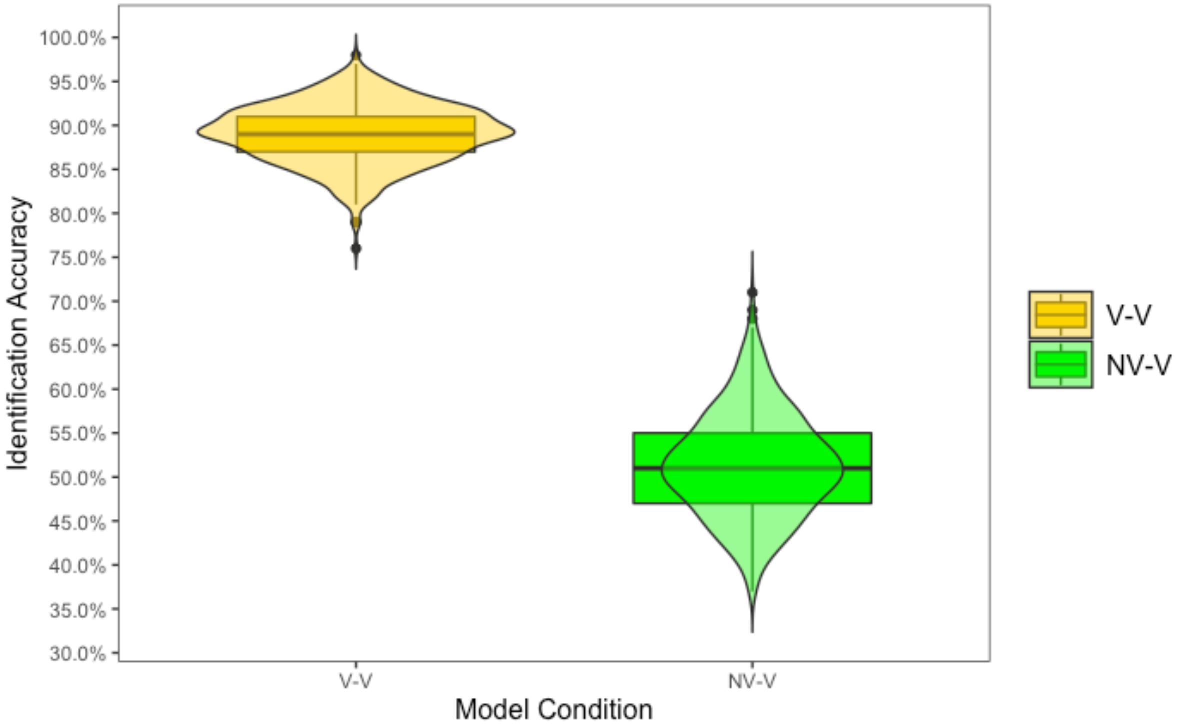


Table 3

Means and standard deviations of identification accuracy for V-V and NV-V models

Model Condition	**Identification Accuracy**	
	Mean	**Standard Deviation**
V-V	89.07%	3.37%
NV-V	51.43%	5.95%
Difference	37.64%	

Regarding our second hypothesis that identification accuracy will be higher for models trained and tested on variable voice samples (V-V), compared to models trained on non-variable voice samples and tested on variable voice samples (NV-V), descriptives and graphs can be found in Figure 5 and Table 3. Looking at the boxplots it appears that, in line with our hypothesis, V-V models achieved an average identification accuracy (M = 89.07%, SD = 3.37%) higher than that of NV-V models (M = 51.43%, SD = 5.95%). All assumptions required for a two-sample one-sided Welch's t-test were checked through visual inspection and found to hold. As suggested by our data, the t-test (t(789.57) = 123.13, p < .001, g = 7.78) revealed a statistically significant difference between V-V and NV-V models, showing that models trained and tested on variable voice samples perform better than models trained on non-variable voice samples and tested on variable voice samples. These results match our expectations, as the variable training set contains speaker-specific information regarding within-speaker variability, improving identification accuracy for variable voices.

Exploratory Analysis

Given the novel approach used in this study and the opportunity to explore additional patterns observed in our models that can inform the current literature, we decided to perform a series of exploratory analyses on our data to supplement the results from our confirmatory analysis. Table 4 and Figure 6 below show the distributions of all individual models, including those using mixed training or test sets (NV+V). For example, NV+V-NV models were trained on a mix of variable and non-variable voice samples and tested on non-variable voice samples. In contrast, NV-NV+V models were trained on non-variable voice samples and tested on an equal mix of variable and non-variable voice samples.

Figure 6

Violin- and boxplots of identification accuracy for all model conditions

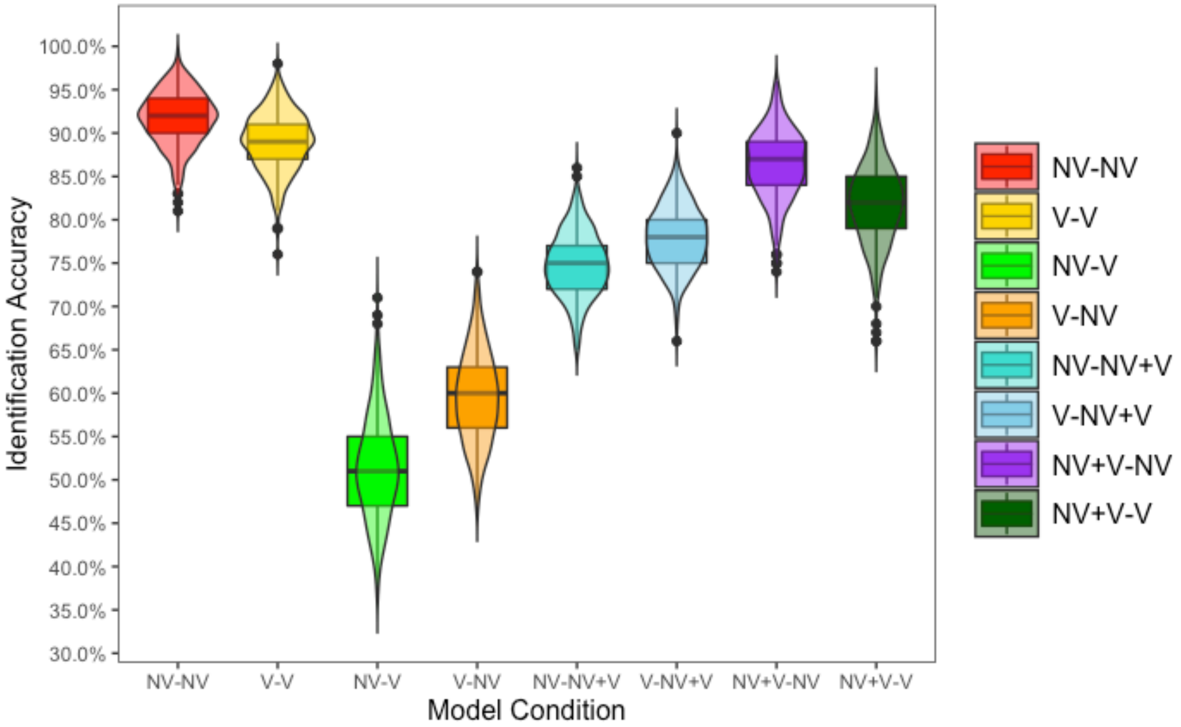


Table 4

Means and standard deviations of identification accuracy for all model conditions

Model Condition	Identification Accuracy	
	Mean	**Standard Deviation**
NV-NV	91.68%	3.09%
V-V	89.07%	3.37%
NV-V	51.43%	5.95%
V-NV	59.57%	5.30%
NV-NV+V	74.85%	3.68%
V-NV+V	78.04%	3.61%
NV+V-NV	86.40%	3.97%
NV+V-V	81.76%	4.57%

Firstly, we examined whether the difference in identification accuracy between models tested on variable or non-variable voice samples, respectively, would also emerge between the models using mixed training sets (NV+V-NV and NV+V-V). Looking at the data in Table 4 and Figure 6, we can observe a difference that matches the pattern we observed between the NV-NV and V-V models we examined for our first hypothesis. In both cases, the models tested on variable samples performed slightly worse than the models tested on non-variable samples, even though the NV+V-NV and NV+V-V models were both trained on similar training sets containing equal shares of non-variable and variable samples. A two-sample two-sided t-test (t(978.60) = 17.14, p < .001, g = 1.08) showed that this observed difference is statistically significant.

Secondly, we examined whether the difference in identification accuracy between models trained on variable or non-variable voice samples respectively would also emerge between the models tested on non-variable test sets (V-NV and NV-NV). Looking at the data in Table 4 and Figure 6, we can observe a similar difference as we observed between the NV-V and V-V models we examined for our second hypothesis, namely that the model trained and tested on the same speech style (NV-NV) performs worse than the model with mismatched speech styles (V-NV). A two-sample two-sided t-test (t(803.30) = 117.04, p < .001, g = 7.40) showed that this difference is statistically significant, with a similar effect size compared to the difference between V-V and NV-V models.

Lastly, we examined whether the difference in identification accuracy between models trained on variable or non-variable voice samples respectively would also emerge between the models tested on mixed test sets (V-NV+V and NV-NV+V). Looking at the data in Table 4 and Figure 6, we can observe a similar pattern as we observed between the NV-V and V-V models we examined for our second hypothesis. While both of the models using mixed test sets perform worse than the models tested exclusively on variable voice samples, the model that was trained on variable samples once again performs better than the one trained on non-variable samples. A two-sample two-sided t-test (t(997.61) = -13.84, p < .001, g = -0.87) showed that this difference is statistically significant, even though the effect size is much smaller compared to the difference between V-V and NV-V models.

Discussion

This study aimed to examine whether within-speaker variability is unique to an individual and provides valuable information for voice identity perception. To assess this theory, we trained and tested an artificial neural network on variable and non-variable voice samples, comparing the resulting identification accuracies across different conditions. Based on prior research, we devised two hypotheses. Firstly, we hypothesised that identification accuracy will be lower for models trained and tested on variable voice samples, compared to models trained and tested on non-variable voice samples. Secondly, we hypothesised that identification accuracy will be higher for models trained and tested on variable voice samples, compared to models trained on non-variable voice samples and tested on variable voice samples. Furthermore, we performed an explanatory analysis of our results, examining any additional patterns that emerged.

Our results show that in line with our first hypothesis, models trained and tested on variable voice samples (V-V) performed worse than models trained and tested on non-variable voice samples (NV-NV). While the absolute difference is statistically significant, it is small at only 2.61%. Furthermore, models trained on variable voice samples (V-V) performed significantly better than models trained on non-variable voice samples (NV-V) when tested on variable voices. The observed absolute difference of 37.64% is not only statistically significant but also large, confirming our second hypothesis.

These findings correspond directly to the expectations set by prior research. In line with studies demonstrating that variable voices are more difficult to identify for humans (Lavan et al., 2019b; Lavan et al., 2019c), our model performed worse when tested on variable voice samples compared to non-variable voice samples. These findings suggest that the presence of higher levels of within-variability adds a layer of complexity to voices that lowers identification accuracy. However, the difference observed in our study is small and both NV-NV and V-V models maintained a high degree of accuracy at around 90%, showing that the added complexity decreases performance but does not distort voices to a large degree. These findings are supported by other studies, even though numerical results vary to a certain extent. For example, Lavan et al. (2019b; 2019c) also reported that high variability impaired participant's ability to accurately distinguish between individuals. This observation was expressed as a difference in the number of reported identities and not in terms of identification accuracy, making it difficult to draw direct comparisons with our results. However, we can directly contrast our findings with those by Afshan et al. (2020), who observed a much larger difference in identification accuracy of 8.16% between variable and non-variable voices in human listeners. Their results also showed that the difference in identification accuracy for a state-of-the-art voice identification system was only 5.52%, implying that computational models suffer from a lower accuracy penalty than human listeners when confronted with variable voices.

In addition, variations in experimental designs and the specific models used are likely to contribute to variations across studies. Afshan et al. (2020) employed a different experimental setup than our study and asked participants and the model to indicate whether pairs of unfamiliar voices belonged to the same voice identity. As this format requires a much more general approach to voice identification than our training-based approach, the added difficulty could cause larger discrepancies. Since our model overall had a higher identification accuracy (NV-NV: 91.68%; V-V: 89.07%) than Afshan et al.'s (2020) (NV-NV: 85.65%; V-V: 80.13%) it seems appropriate that the difference in our study is also smaller than the difference observed in their experiment. As such, while the observed difference in our experiment is not as pronounced as in other human and computational experiments, it is still statistically significant and matches the overall narrative, suggesting that the added complexity in variable voices makes it more difficult for both humans and computational models to accurately identify them.

In regard to our second hypothesis, our results show that models trained on variable voice samples performed significantly better than models trained on non-variable voice samples when tested on variable voices. This pattern matches Lavan et al.'s (2019a) theory that within-speaker variability is speaker-specific and can therefore provide additional information when identifying naturally varying voices. However, while our results fit with the overall findings by other studies, the large difference observed in our experiment (37.64%) stands out in comparison to past research. For example, Park et al. (2016) presented a voice identification system with pairs of voice samples that each consisted of either read, non-variable speech or spontaneous, variable speech. Similar to our findings, Park et al. (2016) reported higher identification accuracy in matching conditions (e.g. V-V) than in mismatched conditions (e.g. NV-V) but the reported difference is significantly smaller at only 2.25%. This much smaller difference is also supported by Afshan et al. (2020) who only found a 1.91% difference when using a voice identification system and a 5.56% difference for human listeners. Therefore, we must acknowledge that our results, while generally supported by prior research, appear exaggerated in contrast to other findings (Afshan et al., 2020; Park et al., 2016).

However, this does not imply that our results are not valid. Instead, as highlighted by Lavan et al. (2019a), we have to consider that results obtained from computational approaches might not directly generalise to results in the human domain, as the underlying mechanisms could be vastly different from each other. The same principle applies when comparing different machine-learning models, as each type of model takes a fundamentally different approach in processing inputs (Dey, 2016). For instance, both Park et al. (2016) and Afshan et al. (2020) used a voice identification system based on probabilistic linear discriminant analysis, an advanced dimensionality reduction technique similar to principal component analysis (Ioffe, 2006). These types of models differ significantly from our artificial neural network approach that mimics the structure of the human brain, processing inputs through a complex network of neurons and combining them in a hierarchical structure to allow for the detection of higher-level features (Gupta, 2013). As a result, it is not feasible to directly compare numerical results between these models. However, as the general objective of both of these models is the accurate identification of voices, we can compare the general implications of their behaviours in regard to within-speaker variability. Both models by Park et al. (2016) and Afshan et al. (2020) were more accurate at identifying pairs of voices when both voices contained similar levels of within-speaker variability. Our results add to this observed pattern, supporting Lavan et al.'s

(2019a) theory that within-speaker variability is speaker-specific and can help identify naturally varying voices.

To add more nuance to our confirmatory results and understand the discussed deviations from other findings, we can examine our exploratory analysis. For instance, we observed that models for which training and test set conditions did not match (NV-V and V-NV) performed significantly worse than those in matched conditions (NV-NV and V-V). This outcome was expected for the NV-V model, which our hypothesis predicted to perform significantly worse than its matched V-V counterpart, as the lack of within-speaker variability in the training set should impair its ability to identify variable voices. However, in the case of the V-NV model, prior research would not predict such a large difference in identification accuracy, as the variable training set should allow it to generalise to the non-variable test set. Lavan et al. (2019d) found that human listeners used within-speaker variability to construct an average for a given voice identity and were more accurate at identifying these averages compared to identifying the variable voice samples. As a result, we would expect our model to build an average from the variable training set and subsequently perform fairly well on the non-variable test set.

The artificial neural network used in our experiment did not seem to process voices in the same way as human listeners and could not build an average representation based on a set of voice samples. The potential reason for this impairment can be found in research on the underlying mechanisms. Since an artificial neural network learns directly from the provided data, its behaviour depends on the specific data set it is trained on (Gupta, 2013). If a data set is biased and does not represent the general population well, the artificial neural network will interpret this bias as meaningful information, impairing its ability to generalise to unseen data (Kim et al., 2019). In other words, an artificial neural network trained on a specific speech style will be inherently biased and fail to generalise well to an unseen speech style. When considering this together with the larger than expected accuracy difference between NV-V and V-V models, as well as the similarly large difference between V-NV and V-V models, it seems that the primary cause of performance degradation is the mismatch in speech style (e.g., read speech and conversational speech) rather than within-speaker variability. Lavan et al. (2019e) encountered a similar phenomenon in their experiments with human listeners, where performance was lower for the high-variability training group when the speech style in the low-variability training set and the test set was the same. After replacing the low-variability training set with a different speech style, this effect disappeared, and the high-variability training group then performed better than the low-variability training group. Since our data set only consisted of two speech styles, read speech and “as if talking to a family member”, an exaggerated form of this phenomenon could be occurring in our experiment as well. This could explain why the observed difference for our second hypothesis is notably larger than those reported by Park et al. (2016) and Afshan et al. (2020), as their probabilistic linear discriminant analysis based voice identification system might not suffer from this issue.

Taken together, these considerations suggest that the observed differences in accuracy between V-V and NV-V models as well as NV-NV and V-NV models are mainly a product of mismatched speech styles. However, we can potentially observe the true difference caused by a lack of exposure to within-speaker variability by looking at the models using mixed test sets (NV-NV+V and V-NV+V). As these models are tested on equal shares of variable and non-variable voice samples, both models should suffer from the same deficit with regard to speech style generalisation as both of their training sets only correspond to half of the test sets, equalising any resulting deficits. In addition, since both models are tested on the same kind of test set we also neutralise the difference observed between models tested on variable or non-variable voice samples. As a result, any persisting difference in identification accuracy should correspond to the lack of exposure to within-speaker variability in the NV-NV+V model. Our results support this assumption, as the difference in identification accuracy is statistically significant and much smaller at only 3.19%, matching the difference observed in models that do not suffer from issues surrounding speech style generalisation (Afshan et al., 2020; Park et al., 2016). This finding adds further support to the theory that within-speaker variability is speaker-specific and improves the identification of naturally varying voices, even when controlling for potential issues in our model. However, due to the exploratory nature of this analysis, this value should not be considered a valid measurement. Rather, it represents an attempt to mitigate the limitations of our model and provide a more appropriate value that matches findings from prior research.

Overall, our results indicate that in line with our hypotheses, the presence of within-speaker variability does make it more difficult to accurately identify voices but also provides speaker-specific information that can improve identification accuracy for naturally varying voices. While our numerical results differ from other studies in terms of magnitude, they match overall findings from both computational (Afshan et al., 2020; Park et al., 2016) and human studies (Lavan et al., 2019b; Lavan et al., 2019c; Lavan et al., 2019e). Furthermore, our explanatory analysis suggests that our model suffers from inherent limitations concerning the generalisation between different speech styles. However, by contrasting models using mixed test sets, this limitation can potentially be circumvented, isolating the effects of within-speaker variability and producing numerical values that match prior findings (Afshan et al., 2020; Park et al., 2016).

It is important to consider our results in light of the unique limitations of our study. Firstly, our use of a computational model severely limits the transferability of our results to the human domain. While our results fit with findings from psychological research (Lavan et al., 2019b; Lavan et al., 2019c; Lavan et al., 2019e) and other computational studies (Afshan et al., 2020; Park et al., 2016) we must acknowledge that the process by which our model identifies voices is unlikely to match the equivalent process in humans. We specifically chose an artificial neural network approach based on its relationship with the human brain (Gupta, 2013), but our model exhibits clear deviations from human voice identity perception, such as a lack of speech style generalisation that can be observed in both humans and other computational models (Afshan et al., 2020; Lavan et al., 2019d; Park et al., 2016). Furthermore, our model is limited by the provided data set. Our study only involved the use of male voice samples, which could limit its validity in regard to female voice identity perception. Since research has established that both male and female voices are encoded in a similar fashion (Baumann & Belin, 2008; Latinus et al., 2013; Latinus & Belin, 2011b), our findings should generally also apply to female voices to a certain extent. However, studies have observed differences in voice identity perception between males and females (Ahrens et al., 2014; Skuk & Schweinberger, 2012), emphasising the importance of examining both groups individually as well. In addition, our data set was limited to two different speech styles, which only represent a small subset of the large range of variations in speech style that we are exposed to on a daily basis (Kreiman et al., 2015). While our study represents a vital addition to the limited research on within-speaker variability, it is primarily a proof-of-concept regarding the usefulness of machine-learning models and demonstrates the need for voice research to re-evaluate the relevance of within-speaker variability in voice identity perception. Future studies should examine a variety of speech styles from both male and female speakers and use different computational models as well as human speakers to gain a better understanding of within-speaker variability, broadening the current one-sided perspective surrounding voice identity perception.

In conclusion, our findings demonstrate that within-speaker variability adds complexity to voices and as a result makes their identification more challenging for a computational model. However, our observations also show that within-speaker variability provides speaker-specific information to the model which improves identification accuracy for naturally varying voices.

Taken together, these results reinforce Lavan et al.’s (2019a) theory that within-speaker variability represents much more than a random form of noise. Instead, it should be considered a central factor in voice identity perception that requires further attention in future research.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., & others. (2016). Tensorflow: A system for large-scale machine learning. *12th ${$USENIX$}$ Symposium on Operating Systems Design and Implementation (${$OSDI$}$ 16)*, 265–283.

Afshan, A., Kreiman, J., & Alwan, A. (2020). Speaker discrimination in humans and machines: Effects of speaking style variability. *ArXiv:2008.03617 [Cs, Eess]*. http://arxiv.org/abs/2008.03617

Ahrens, M.-M., Awwad Shiekh Hasan, B., Giordano, B. L., & Belin, P. (2014). Gender differences in the temporal voice areas. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00228

Alim, S. A., & Rashid, N. K. A. (2018). Some Commonly Used Speech Feature Extraction Algorithms. *From Natural to Artificial Intelligence - Algorithms and Applications*. https://doi.org/10.5772/intechopen.80419

Audacity. (2021). *Audacity*. Audacity ®. https://www.audacityteam.org

Awan, S. N., & Frenkel, M. L. (1994). Improvements in estimating the harmonics-to-noise ratio of the voice. *Journal of Voice*, *8*(3), 255–262. https://doi.org/10.1016/S0892-1997(05)80297-8

Bahrick, L., Lickliter, R., Shuman, M., Batista, L., & Grandez, C. (2003). *Infant discrimination of voices: Predictions from the intersensory redundancy hypothesis*.

Barton, J. J. S., & Corrow, S. L. (2016). RECOGNIZING AND IDENTIFYING PEOPLE: A neuropsychological review. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *75*, 132–150. https://doi.org/10.1016/j.cortex.2015.11.023

Baumann, O., & Belin, P. (2008). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research PRPF*, *74*(1), 110. https://doi.org/10.1007/s00426-008-0185-z

Belin, P., Boehme, B., & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLOS ONE*, *12*(10), e0185651. https://doi.org/10.1371/journal.pone.0185651

Bhurtel, M., Shrestha, J., Lama, N., Bhattarai, S., Uprety, A., & Guragain, M. (2019, November 23). *DEEP LEARNING BASED SEED QUALITY TESTER*.

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America*, *132*(2), 1100–1112. https://doi.org/10.1121/1.4714351

Boles, A., & Rad, P. (2017). Voice biometrics: Deep learning-based voiceprint authentication system. *2017 12th System of Systems Engineering Conference (SoSE)*, 1–6. https://doi.org/10.1109/SYSOSE.2017.7994971

Bre, F., Gimenez, J., & Fachinotti, V. (2017). Prediction of wind pressure coefficients on building surfaces using Artificial Neural Networks. *Energy and Buildings*, *158*. https://doi.org/10.1016/j.enbuild.2017.11.045

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Bruckert, L., Liénard, J.-S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1582), 83–89. https://doi.org/10.1098/rspb.2005.3265

Burnham, D. (2002). What's New, Pussycat? On Talking to Babies and Animals. *Science*, *296*(5572), 1435–1435. https://doi.org/10.1126/science.1069587

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485. https://doi.org/10.1080/17470218.2013.800125

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity From Variation: Representations of Faces Derived From Multiple Instances. *Cognitive Science*, *40*(1), 202–223. https://doi.org/10.1111/cogs.12231

Champely, S. (2020). *pwr: Basic Functions for Power Analysis*. https://CRAN.R-project.org/package=pwr

Chollet, F. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

Çınlar, E. (2011). *Probability and Stochastics*. Springer Science & Business Media.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, *208*(4448), 1174–1176. https://doi.org/10.1126/science.7375928

DeCasper, Anthony J., & Prescott, P. A. (1984). Human newborns' perception of male voices: Preference, discrimination, and reinforcing value. *Developmental Psychobiology*, *17*(5), 481–491. https://doi.org/10.1002/dev.420170506

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, *137*(3), 1513–1528. https://doi.org/10.1121/1.4906837

Dey, A. (2016). Machine learning algorithms: A review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179.

Fant, G. (1971). *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. De Gruyter, Inc. http://ebookcentral.proquest.com/lib/gla/detail.action?docID=3044232

Farrus, M., Hernando, J., & Ejarque, P. (2007). *Jitter and Shimmer Measurements for Speaker Recognition*. 4.

Gainotti, G. (2014). Cognitive models of familiar people recognition and hemispheric asymmetries. *Frontiers in Bioscience (Elite Edition)*, *6*, 148–158. https://doi.org/10.2741/E698

Gelfer, M. P., & Bennett, Q. E. (2013). Speaking Fundamental Frequency and Vowel Formant Frequencies: Effects on Perception of Gender. *Journal of Voice*, *27*(5), 556–566. https://doi.org/10.1016/j.jvoice.2012.11.008

Ghazanfar, A. A., & Rendall, D. (2008). Evolution of human vocal production. *Current Biology*, *18*(11), R457–R460. https://doi.org/10.1016/j.cub.2008.03.030

Goy, H., Kathleen Pichora-Fuller, M., & van Lieshout, P. (2016). Effects of age on speech and voice quality ratings. *The Journal of the Acoustical Society of America*, *139*(4), 1648–1659. https://doi.org/10.1121/1.4945094

Gumelar, A. B., Kurniawan, A., Sooai, A. G., Purnomo, M. H., Yuniarno, E. M., Sugiarto, I., Widodo, A., Kristanto, A. A., & Fahrudin, T. M. (2019). Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks. *2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH)*, 1–8. https://doi.org/10.1109/SeGAH.2019.8882461

Gupta, N. (2013). Artificial Neural Network. *Network and Complex Systems*, *3*(1), 24. https://iiste.org/Journals/index.php/NCS/article/view/6063

Guyon, I., & Elisseeff, A. (2006). An Introduction to Feature Extraction. In I. Guyon, M. Nikravesh, S. Gunn, & L. A. Zadeh (Eds.), *Feature Extraction: Foundations and Applications* (pp. 1–25). Springer. https://doi.org/10.1007/978-3-540-35488-8_1

He, L., Zhang, Y., & Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. *The Journal of the Acoustical Society of America*, *145*(3), EL209–EL214. https://doi.org/10.1121/1.5093450

Hollien, H. F. (2002). *Forensic Voice Identification*. Academic Press.

Ioffe, S. (2006). Probabilistic Linear Discriminant Analysis. In A. Leonardis, H. Bischof, & A. Pinz (Eds.), *Computer Vision – ECCV 2006* (Vol. 3954, pp. 531–542). Springer Berlin Heidelberg. https://doi.org/10.1007/11744085_41

Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012). Speech Recognition using MFCC. *Simulation and Modeling*, 4.

Jamshidi, M. B., Alibeigi, N., Rabbani, N., Oryani, B., & Lalbakhsh, A. (2018). Artificial Neural Networks: A Powerful Tool for Cognitive Science. *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 674–679. https://doi.org/10.1109/IEMCON.2018.8615039

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66–71. https://doi.org/10.1016/j.cognition.2012.03.001

Kim, B., Kim, H., Kim, K., Kim, S., & Kim, J. (2019). *Learning Not to Learn: Training Deep Neural Networks With Biased Data*. 9012–9020. https://openaccess.thecvf.com/content_CVPR_2019/html/Kim_Learning_Not_to_Learn_Training_Deep_Neural_Networks_With_Biased_CVPR_2019_paper.html

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech and Hearing Research*, *35*(3), 512–520. https://doi.org/10.1044/jshr.3503.512

Kreiman, Jody, & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters Part I: Recognition of backward voices. *Journal of Phonetics*, *13*(1), 19–38. https://doi.org/10.1016/S0095-4470(19)30723-5

Kreiman, Jody, Park, S. J., Keating, P. A., & Alwan, A. (2015). *The Relationship Between Acoustic and Perceived Intraspeaker Variability in Voice Quality*. 4.

Kreiman, Jody, & Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. John Wiley & Sons.

Latinus, M., & Belin, P. (2011a). Anti-Voice Adaptation Suggests Prototype-Based Coding of Voice Identity. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00175

Latinus, M., & Belin, P. (2011b). Human voice perception. *Current Biology*, *21*(4), R143–R145. https://doi.org/10.1016/j.cub.2010.12.033

Latinus, M., McAleer, P., Bestelmeyer, P., & Belin, P. (2013). Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology : CB*, *23*. https://doi.org/10.1016/j.cub.2013.04.055

Lavan, N. (2019). Telling people together. *Psychologist*, 48–51. http://search.ebscohost.com/login.aspx?direct=true&db=pbh&AN=139755946&site=ehost-live

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How Many Voices Did You Hear? Natural Variability Disrupts Identity Perception from Unfamiliar Voices. *British Journal of Psychology*, *110*(3), 576–593. http://search.ebscohost.com/login.aspx?direct=true&db=mlf&AN=202016895100&site=ehost-live

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners: *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021819836890

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, *26*(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

Lavan, N., Domone, A., Fisher, B., Kenigzstein, N., Scott, S. K., & McGettigan, C. (2019). Speaker Sex Perception from Spontaneous and Volitional Nonverbal Vocalizations. *Journal of Nonverbal Behavior*, *43*(1), 1–22. https://doi.org/10.1007/s10919-018-0289-0

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition*, *193*, 104026. https://doi.org/10.1016/j.cognition.2019.104026

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*, 2404. https://doi.org/10.1038/s41467-019-10295-w

Lavan, N., & McGettigan, C. (2017). Increased Discriminability of Authenticity from Multimodal Laughter is Driven by Auditory Information: *Quarterly Journal of Experimental Psychology*. http://journals.sagepub.com/doi/10.1080/17470218.2016.1226370

Lavan, N., Mileva, M., Burton, M., Young, A., & McGettigan, C. (2020). *Trait evaluations of faces and voices: Comparing within- and between-person variability* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/3rjc4

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. https://doi.org/10.1037/xge0000223

Lavan, N., Short, B., Wilding, A., & McGettigan, C. (2018). Impoverished encoding of speaker identity in spontaneous laughter. *Evolution and Human Behavior*, *39*(1), 139–145. https://doi.org/10.1016/j.evolhumbehav.2017.11.002

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, *30*(1), 9–26. https://doi.org/10.1016/S0167-6393(99)00028-X

Lenain, R., Weston, J., Shivkumar, A., & Fristed, E. (2020). Surfboard: Audio Feature Extraction for Modern Machine Learning. *ArXiv:2005.08848 [Cs, Eess]*. http://arxiv.org/abs/2005.08848

Ma, E. P.-M., & Wu, M. C.-K. (2020). Age estimation from voice in the Cantonese elderly population: Influence of listener's age and stimulus types. *Speech, Language and Hearing*, *23*(4), 243–249. https://doi.org/10.1080/2050571X.2019.1634348

Małgorzata, S. (2003). Use of artificial neural networks in clinical psychology and psychiatry. *Psychiatria Polska*, *37*(2), 349–357. https://europepmc.org/article/med/12776663

Mamyrbayev, O., Mekebayev, N., Turdalyuly, M., Oshanova, N., Medeni, T. I., & Yessentay, A. (2019). Voice Identification Using Classification Algorithms. *Intelligent System and Computing*. https://doi.org/10.5772/intechopen.88239

Mathias, S., & Kriegstein, K. (2014). How do we recognise who is speaking? *Frontiers in Bioscience (Scholar Edition)*, *6*, 92–109.

McAleer, P., Todorov, A., & Belin, P. (2014). How Do You Say 'Hello'? Personality Impressions from Brief Novel Voices. *PLOS ONE*, *9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

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. https://doi.org/10.1037/xhp0000049

Napierala, M. A. (2012). *What Is the Bonferroni Correction?* 3.

Neisi, L., karimAbadi, F. A., & Shekaramiz, M. (2019). An Investigation of Male and Female Voices: Does Voice Gender Categorization Depend on Pitch? *International Journal of Linguistics, Literature and Translation*, *2*(6), 63–70. https://al-kindipublisher.com/index.php/ijllt/article/view/480

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine*, *375*(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine Learning in Psychometrics and Psychological Research. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02970

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., & Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1140–1152. https://doi.org/10.1016/j.neubiorev.2012.01.004

Pace, G., Orrù, G., Monaro, M., Gnoato, F., Vitaliani, R., Boone, K. B., Gemignani, A., & Sartori, G. (2019). Malingering Detection of Cognitive Impairment With the b Test Is Boosted Using Machine Learning. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01650

Park, S. J., Sigouin, C., Kreiman, J., Keating, P., Guo, J., Yeung, G., Kuo, F.-Y., & Alwan, A. (2016). Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition. *INTERSPEECH*. https://doi.org/10.21437/Interspeech.2016-523

Park, S. J., Yeung, G., Kreiman, J., Keating, P. A., & Alwan, A. (2017). Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems. *Interspeech 2017*, 1522–1526. https://doi.org/10.21437/Interspeech.2017-157

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, *87*(1), 93–98. https://doi.org/10.1016/j.biopsycho.2011.02.010

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pernet, C. R., & Belin, P. (2012). The Role of Pitch and Timbre in Voice Gender Categorization. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00023

Perrachione, T. K., Tufo, S. N. D., & Gabrieli, J. D. E. (2011). Human Voice Recognition Depends on Language Ability. *Science*, *333*(6042), 595–595. https://doi.org/10.1126/science.1207327

Perrachione, T., Pierrehumbert, J., & Wong, P. (2009). Differential Neural Contributions to Native- and Foreign-Language Talker Identification. *Journal of Experimental Psychology. Human Perception and Performance*, *35*, 1950–1960. https://doi.org/10.1037/a0015869

Pilgrim, M., & Willison, S. (2009). *Dive Into Python 3* (Vol. 2). Springer.

Pisanski, K., Raine, J., & Reby, D. (2020). Individual differences in human voice pitch are preserved from speech to screams, roars and pain cries. *Royal Society Open Science*, *7*(2), 191642. https://doi.org/10.1098/rsos.191642

Podesva, R. J., & Callier, P. (2015). Voice Quality and Identity. *Annual Review of Applied Linguistics*, *35*, 173–194. https://doi.org/10.1017/S0267190514000270

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the Identification of Speakers by Voice. *The Journal of the Acoustical Society of America*, *26*(3), 403–406. https://doi.org/10.1121/1.1907349

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology*, *70*(5), 897–905. https://doi.org/10.1080/17470218.2015.1136656

RStudio Team. (2021). *RStudio: Integrated Development Environment for R*. RStudio, Inc. http://www.rstudio.com/

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research: JSLHR*, *40*(2), 453–463. https://doi.org/10.1044/jslhr.4002.453

Schweinberger, Stefan R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *WIREs Cognitive Science*, *5*(1), 15–25. https://doi.org/10.1002/wcs.1261

Skuk, V., & Schweinberger, S. (2012). Gender Differences in Familiar Voice Identification. *Hearing Research*, *296*. https://doi.org/10.1016/j.heares.2012.11.004

Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177–3186. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2346770/

Smith, I., Foulkes, P., Sóskuthy, M., Smith, I., Foulkes, P., & Sóskuthy, M. (2017). Speaker Identification in Whisper. *Letras de Hoje*, *52*(1), 5–14. https://doi.org/10.15448/1984-7726.2017.1.26659

Souza, L. B. R. de, Santos, M. M. dos, Souza, L. B. R. de, & Santos, M. M. dos. (2018). Body mass index and acoustic voice parameters: Is there a relationship?,. *Brazilian*

Journal of Otorhinolaryngology, *84*(4), 410–415. https://doi.org/10.1016/j.bjorl.2017.04.003

Tiwari, V. (2010). *MFCC and its applications in speaker recognition*. 4.

Vieira, S., Pinaya, W. H. L., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience and Biobehavioral Reviews*, *74*(Pt A), 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002

Wang, M., Lu, S., Zhu, D., Lin, J., & Wang, Z. (2018). A High-Speed and Low-Complexity Architecture for Softmax Function in Deep Learning. *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, 223–226. https://doi.org/10.1109/APCCAS.2018.8605654

Whiteside, S. P. (1998). Identification of a speaker's sex: A study of vowels. *Perceptual and Motor Skills*, *86*(2), 579–584. https://doi.org/10.2466/pms.1998.86.2.579

Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., & Hinton, G. E. (2013). On rectified linear units for speech processing. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 3517–3521. https://doi.org/10.1109/ICASSP.2013.6638312