2020

Understanding the Effect of Voice Quality and Accent on Talker Similarity

Anurag Das
Texas A & M University

Guanlong Zhao
Texas A & M University

John Levis
Iowa State University, jlevis@iastate.edu

Evgeny Chukharev-Hudilainen
Iowa State University, evgeny@iastate.edu

Ricardo Gutierrez-Osuna
Texas A & M University

Understanding the Effect of Voice Quality and Accent on Talker Similarity

Abstract

This paper presents a methodology to study the role of nonnative accents on talker recognition by humans. The methodology combines a state-of-the-art accent-conversion system to resynthesize the voice of a speaker with a different accent of her/his own, and a protocol for perceptual listening tests to measure the relative contribution of accent and voice quality on speaker similarity. Using a corpus of non-native and native speakers, we generated accent conversions in two different directions: non-native speakers with native accents, and native speakers with non-native accents. Then, we asked listeners to rate the similarity between 50 pairs of real or synthesized speakers. Using a linear mixed effects model, we find that (for our corpus) the effect of voice quality is five times as large as that of non-native accent, and that the effect goes away when speakers share the same (native) accent. We discuss the potential significance of this work in earwitness identification and sociophonetics.

Keywords

voice quality, accent, voice similarity

Disciplines

Anthropological Linguistics and Sociolinguistics | Computational Linguistics

Comments

This proceeding is published as Das, Anurag, Guanlong Zhao, John Levis, and Evgeny Chukharev-Hudilainen. "Understanding the Effect of Voice Quality and Accent on Talker Similarity." *Proceedings of Interspeech 2020*. [Virtual Conference, October 25-29, 2020.] Pages 1763-1767. DOI: 10.21437/Interspeech.2020-2910. Posted with permission.

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Understanding the Effect of Voice Quality and Accent on Talker Similarity

Anurag Das[1], *Guanlong Zhao*[1], *John Levis*[2], *Evgeny Chukharev-Hudilainen*[2], *Ricardo Gutierrez-Osuna*[1]

[1]Department of Computer Science and Engineering, Texas A&M University, USA
[2]Department of English, Iowa State University, USA

{adas, gzhao, rgutier}@tamu.edu, {jlevis, evgeny}@iastate.edu

Abstract

This paper presents a methodology to study the role of non-native accents on talker recognition by humans. The methodology combines a state-of-the-art accent-conversion system to resynthesize the voice of a speaker with a different accent of her/his own, and a protocol for perceptual listening tests to measure the relative contribution of accent and voice quality on speaker similarity. Using a corpus of non-native and native speakers, we generated accent conversions in two different directions: non-native speakers with native accents, and native speakers with non-native accents. Then, we asked listeners to rate the similarity between 50 pairs of real or synthesized speakers. Using a linear mixed effects model, we find that (for our corpus) the effect of voice quality is five times as large as that of non-native accent, and that the effect goes away when speakers share the same (native) accent. We discuss the potential significance of this work in earwitness identification and sociophonetics.

Index Terms: voice quality, accent, voice similarity

1. Introduction

How do we recognize the voice of non-native speakers, through their non-native accent or their voice quality, or both? What is the relative significance of both attributes? Significant research has been done in talker recognition (by humans), where a large number of acoustic parameters (e.g. F0, spectral slope, formants) have been shown to contribute to voice individuality [1, 2], a robust finding being that the relative significance of these acoustic cues on identity varies from voice to voice [3]. Despite such progress, however, the interaction between voice quality and non-native accents on talker recognition remain open, in part due to the challenges of individually manipulating these two aspects of the speech signal beyond short segments of speech, e.g., vowels [4].

In this work, we propose a new methodology to examine the role of accent and voice quality in talker recognition. Our methodology relies on the use of "accent conversion" techniques [5, 6] to transform utterances from second-language (L2) learners[1] to mimic the pronunciation patterns (i.e., accent) of a native (L1) speaker, and vice versa. Our accent conversion model consists of three basic components: an acoustic model that generates a speaker-independent embedding of an utterance (a posteriorgram, or PPG), a sequence-to-sequence (seq2seq) synthesizer that maps PPGs into Mel-spectrograms, and a vocoder that maps the Mel-spectrogram into a high-quality speech waveform. To perform accent conversion, we build a seq2seq synthesizer and a vocoder on a speech corpus from the L2 learner, then drive both models using PPGs extracted from an L1 utterance. As a result, the output speech waveform has the native speaker's pronunciation (as captured by the input PPG) and the non-native speaker's voice quality (as captured by the seq2seq synthesizer).

We performed accent-conversion experiments on pairs of L2 and L1 speakers from the L2-ARCTIC [7] and ARCTIC corpora [8], respectively. For each speaker pair, we generated accent conversions in both directions: L2 speaker with L1 accent, and L1 speaker with L2 accent. Then, we conducted listening experiments where participants were asked to rate the similarity between all pairs of speakers (the original L1 and L2 speakers, and the two accent conversions.) On our speech dataset, a linear mixed effect model estimated that the contribution of non-native accents is 20% of that associated with the voice quality of the speaker. While these results cannot be generalized to other speaker pairs, they suggest that our methodology is suitable for studying the role of L2 accents in talker recognition.

2. Literature Review

2.1. Talker recognition (by humans)

Hundreds of studies have been conducted over the past 40 years to identify acoustic parameters that correlate with ratings of voice quality [9], *"those characteristics which are present more or less all the time that a person is talking"* [10]. A large number of acoustic features have been identified, including features related to the vocal source (e.g., fundamental frequency, differences between harmonics) and the vocal tract (e.g., formant frequencies and bandwidths) [11], depending on whether the stimuli consists of sustained vowels in isolation or sentences [9]. A perceptual model emerges from these studies, which is that (for unfamiliar voices) listeners do not perceive voice quality as a collection of isolated features but as an overall pattern in a multidimensional "voice space" [12]. This model bears a strong resemblance to a well-established model of facial recognition, according to which faces exist in a multidimensional "face space" [13]. According to this model, faces are encoded as vectors relative to a central, prototypical face (i.e., an "average" face). This suggests that there exists a unified coding strategy for processing faces and voices [14]. Evidence for this hypothesis is the fact that similar phenomena have been observed in the two perceptual spaces; As an example, the "other race effect" explains why recognizing faces from other races (e.g., Caucasians observers of Asian faces) is more difficult than recognizing faces of the same race (e.g., Caucasian observers of Caucasian faces) [15]. Its counterpart in the voice recognition domain is the "language familiarity effect", which explains why listeners can better recognize speakers of their own language (e.g. English listeners recognizing English speakers) than speakers of an unfamiliar language (e.g., English listeners recognizing Mandarin

[1] For compactness, in what follows we will use the terms L2 and L1 instead of non-native and native.

http://dx.doi.org/10.21437/Interspeech.2020-2910

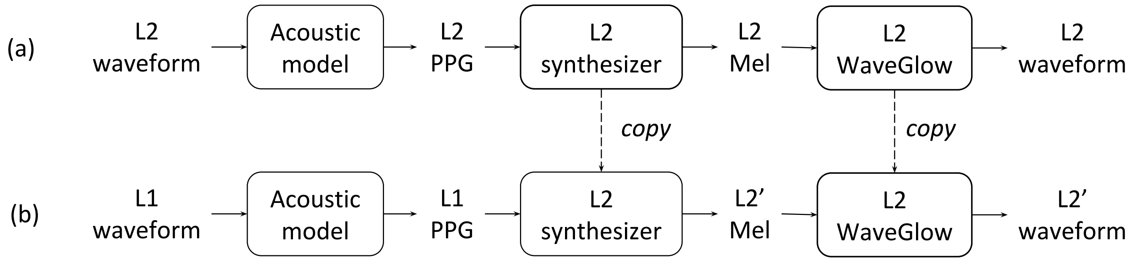

Figure 1: *Pipeline of the accent conversion system during training (a) and testing (b)*

speakers [16]). Two models have been proposed to explain the language familiarity effect [17, 18]: the phonetic familiarity model, which argues that listeners rely on their familiarity with the statistical distribution of phonetic features in their native language, and the linguistic processing model, which argues that listeners also incorporate higher-level information, such as word recognition and comprehension.

2.2. Accent conversion

Accent conversion is related to the more general problem of voice conversion (VC) [19]. In VC, one seeks to transform a source speaker's speech into that of a (known) target speaker. The conversion aims to match the voice characteristics of the target speaker, which include vocal tract configurations, glottal characteristics, pitch range, pronunciation, and speaking rate; ideally, the only information retained from the source speech is its linguistic content, i.e., the words that were uttered. The basic strategy for VC is to collect a parallel corpus for the source and target speaker, and then align the two corpora, e.g. using dynamic time warping. This generates a lookup table with pairs of source and target frames (e.g., Mel Cepstra), which is then used to build a mapping from source frames to target frames. Popular mapping techniques include joint-density Gaussian Mixture Models (GMMs) [20], frequency warping [21, 22], DNNs [23, 24], and sparse coding [25, 26].

In contrast with VC, accent conversion seeks to combine the linguistic content *and* pronunciation characteristics of the source speaker with the voice quality of the target speaker. This is a more challenging problem than VC for two reasons. First, accent conversion lacks ground-truth since there are no recordings of the L2 speaker producing speech with the desired L1 accent. But, more importantly, accent conversion requires decomposing the speech into voice quality and accent, whereas VC does not. The conventional approach used in VC (pairing source and target frames via time alignment) cannot be used in accent conversion, since it would result in a model that maps L1-accented source into L2-accented target speech. Instead, source and target frames have to be paired based on their linguistic content. This may be done by using a speaker-independent acoustic model (e.g. from an ASR system) to estimate the posterior probability that each frame belongs to a set of pre-defined phonetic units (e.g., a phonetic posteriorgram, or PPG [27]). Once a PPG has been computed for each source and target frame in the corpus, the two can be paired in a many-to-many fashion based on the similarity between their respective PPGs [5, 28].

3. Method

Our accent-conversion model is based on the system proposed by Zhao et. al. [28] , which has been shown to produce higher ratings of acoustic quality and naturalness than traditional systems which use conventional vocoders such as STRAIGHT [29] or World [30]. The system consists of three components: an acoustic model (AM) that extracts phonetic posteriorgrams (PPGs) from source utterances, a sequence-to-sequence (seq2seq) synthesizer that maps PPGs to Mel-spectrograms, and a WaveGlow vocoder that synthesizes speech waveforms from Mel-spectrograms. The AM is trained on a large corpus of speech from multiple L1 speakers so that the PPGs are speaker independent, whereas the seq2seq synthesizer and WaveGlow vocoder are trained on speech recordings from the L2 speaker. At test time, we compute a PPG sequence from an L1 utterance, and pass it to the L2 seq2seq synthesizer and WaveGlow vocoder. The output speech signal has the L1 speaker's pronunciation (as captured by the PPG) and the L2 speaker's voice quality (as captured by the seq2seq synthesizer). The overall procedure is illustrated in Figure 1. To perform accent conversion in the reverse direction, the seq2seq synthesizer and WaveGlow vocoder are trained on the L1 speaker, and the PPG sequence is extracted from the L2 utterance.

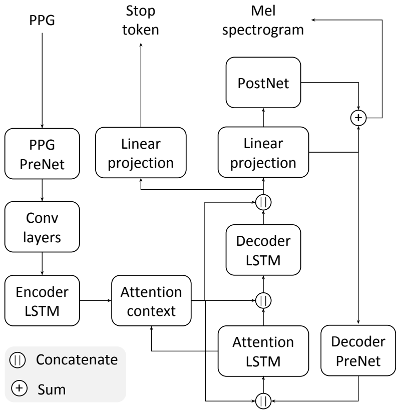

Figure 2: *PPG-to-Mel conversion model*

3.1. Extracting PPGs from acoustic models

The acoustic model takes stacked Mel-frequency coefficients as inputs and outputs senones as class labels. The model is a deep neural network composed of a number of hidden layers, each of which includes a p-norm non-linearity; see section 4.1 for details. The final hidden layer is followed by a softmax layer, which outputs the predicted senones. A detailed description of the AM can be found in [31].

3.2. Seq2seq speech synthesizer (PPG $\rightarrow$ Mel-spectrogram)

We use a modified Tacotron 2 model [32] to convert PPGs to Mel-spectrograms. The original Tacotron 2 model can synthesize natural sounding speech from raw transcripts. It takes as input a one-hot vector of characters, which is then passed to an encoder LSTM. This is followed by an attention network, which summarizes the encoded sequence as a context vector. Finally, a decoder LSTM, generates Mel-spectrograms using a location-sensitive attention mechanism [33]. In order to take PPGs as input, we replace the character-embedding layer with a PPG-embedding network [28]. This PPG-embedding layer consists of two fully connected hidden layers with ReLU nonlinearities. Addition of the PPG embedding layer, transforms the original high dimensional PPG features to low dimensional features. This helps the model converge during training. The PPG to Mel-spectrogram model is shown in Figure 2; more details can be found in [28].

3.3. WaveGlow vocoder (Mel-spectrogram → speech)

To convert the output of the Mel-spectrogram to audio waveform, we use a WaveGlow vocoder [34], a flow-based network [35] that can generate high quality speech from Mel-spectrograms and has low inference time compared to autoregressive models such as WaveNet [36]. It is a generative model that samples from a zero-mean spherical Gaussian with the same number of dimensions as the target, and generates a target distribution by passing through a series of layers. Here, the target distribution is the audio waveform.

4. Results

4.1. Speech corpus

We used a pretrained AM to extract PPGs. The AM had been trained on the Librispeech corpus [37], which contains 960 hours of native English speech, most of which from North America. The AM had five hidden layers and a final softmax layer that produced a 5816-dimensional PPGs. We trained the PPG-to-Mel and WaveGlow models on two L2 speakers from the publicly-available L2-ARCTIC corpus [7]: ABA (male Arabic speaker) and EBVS (male Spanish speaker), and two L1 speakers from the ARCTIC corpus [8]: BDL (male American English) and RMS (male American English). Each speaker in L2-ARCTIC and ARCTIC recorded the same set of 1,132 sentences, or about an hour of speech. For each L2-ARCTIC speaker, we used the first 1,032 sentences for model training, the next 50 sentences for validation, and the remaining 50 sentences for testing. All audio signals were sampled at 16 KHz.

To train the seq2seq models, we used a batch size of 6 and a learning rate of 1×10^{-4}. We trained the model until the validation loss reached a plateau. The WaveGlow models were trained using a batch size of 3 and the learning rate was 1×10^{-4} for 650,000 iterations. We used the same set of parameters for the seq2seq model and the WaveGlow model as those reported in [28]. The AM was trained with Kaldi, and the other models were implemented in PyTorch and trained with the Adam optimizer [38].

4.2. Perceptual listening tests

We performed accent conversions for four pairs of L2/L1 speakers, plus a fifth pair with the L1/L1 speakers, which served as a reference for the linear mixed effects models; see Table 1. For each speaker pair, we then generated accent conversions in two directions: L2 speaker with L1 accent, and L1 speaker with L2 accent. Denoting the first speaker by V1A1 (voice 1, accent 1) and the second speaker by V2A2 (voice 2, accent 2), this led to ten different comparison pairs, as illustrated in Table 2.

We conducted perceptual listening tests on Amazon Mechanical Turk. Following [28], all participants resided in the United States at the time of the recruitment and passed a qualification test where they had to identify several regional dialects in the United States. Only those participants who answered all questions correctly moved on to the listening test. All participants were self-reported native English speakers. Each participant (N=50) rated 50 utterance pairs[2], one from each of the 50 possible comparisons: five speaker pairs (see Table 1) and ten comparisons per speaker pair (see Table 2). The presentation order of the utterance pairs was counterbalanced. Each of the 50 utterances rated by each listener was from one of five possible sentences in ARCTIC. For each comparison, listeners were asked to rate the similarity between the two utterances on a scale of 1 (no similarity) to 7 (excellent similarity).

Speaker pair	**Speaker V1A1**	**Speaker V2A2**
1	ABA*	BDL
2	ABA*	RMS
3	EBVS*	BDL
4	EBVS*	RMS
5	BDL	RMS

Table 1: *Five speaker pairs used in the perceptual listening pairs. * denotes L2*

	V1A1	**V1A2**	**V2A1**	**V2A2**
V1A1	6.93	3.47	1.94	1.52
V1A2		6.84	1.95	2.37
V2A1			6.87	3.95
V2A2				6.92

Table 2: *Average similarity scores between pairs of speakers for the four non-native/native conversions. The term "ViAj" denotes a speaker with voice quality i and accent j.*

Results for L2/L1 pairs are summarized in Table 2. The values on the diagonal elements approach the maximum rating available (7), as one may expect since both utterances in those pairs are from the same condition (e.g., V1A1 vs. V1A1). Likewise, the lowest similarity scores are obtained when the voice and accent of each utterance are mismatched (e.g., V1A1 vs. V2A2, V1A2 vs V2A1), as expected. The more interesting results are those for the conditions where either the voice (e.g., V1A1 vs. V2A1) or the accent (e.g., V1A1 vs. V1A2) are mismatched, but not both. When only the accent is mismatched, the average ratings drop to 3.47-3.95, whereas when only the voice is mismatched, the average rating drops more significantly to 1.94-2.37. Thus, these results indicate that voice quality has a stronger effect than accent on the perceived similarity between speakers. Results for the L1/L1 pair are summarized in Table 3. As with the L2/L1 pairs, the scores on the diagonal elements are close to the maximum, since the voice and accent are both matched. The lowest scores are obtained when both voice and accent are mismatched (2.20-2.46), though these scores are higher than those on the non-native/native pairs (1.52-1.95), a result that can be explained by the fact that accent differences are negligible on the L1/L1 pair.

	V1A1	**V1A2**	**V2A1**	**V2A2**
V1A1	6.96	4.60	2.18	2.46
V1A2		6.78	2.20	2.62
V2A1			6.90	4.50
V2A2				6.96

Table 3: *Average similarity scores between pairs of speakers for the native/native conversions. The term "ViAj" denotes a speaker with voice quality i and accent j.*

[2] Both utterances in each pair were from the same sentence, to make it easier for listeners to attend to pronunciation differences between the two speakers.

4.3. Linear mixed effects model

To further understand the contributions of voice quality and accent, we built a linear mixed effect model (LMM). An LMM explains the variation in a dependent variable using independent variables of interest called fixed effects and the variation not explained by independent variables of interest called random effects. We use voice quality, accent and interaction between voice quality and accent as fixed effects, and sentence, speaker pair and listener as the random effects. An LMM can be represented as:

$$Y = X\beta + Zb + \epsilon \tag{1}$$

where Y represents the similarity scores, X represents the fixed effects of voice quality, accent and interaction between voice quality and accent, Z represents the random effects of sentence, speaker pair and listener, ϵ represents the error and β, b are estimates of the fixed and random effects respectively, learnt by the LMM. Using R notation[3], the LMM was:

$$\begin{aligned} &similarity \sim isSameVoice + isSameAccent + \\ &\quad isSameVoice : isSameAccent + (1|speaker_1) + \\ &\qquad (1|speaker_2) + (1|listener) + (1|sentence) \end{aligned} \tag{2}$$

In a first step, we built a model without fixed effects (model 1), and compared it with a model that included voice as a fixed effect (model 2), and found a statistically significant difference ($p < 0.001$). Then, we compared model 2 against a model that also included accent as a fixed effect (model 3), and again found a statistically significant difference ($p < 0.001$). In a final step, then, we compared model 3 against a model that also included an interaction between both fixed effects, and also found a statistically significant difference between the two models ($p < 0.001$). Results for the final model (4) are shown in Table 4. In the case of the four L2/L1 speaker pairs, the intercept (1.77) can be interpreted as the average similarity between conditions when both voice quality and accent are mismatched (e.g., V1A1 vs. V2A2), which is consistent with the results in Table 2 (1.52-1.95). Matching the voice quality for the two conditions (e.g., V1A1 vs. V1A2) increases the perceived similarity by 1.95 points, whereas matching the accent (e.g., V1A1 vs. V2A1) increases the perceived similarity by 0.42. Accordingly, then, these results suggest that the effect of accent is about 20% that of voice quality. The model also shows a strong interaction effect (2.79) between both fixed effects, which suggests that people rely on a combination of the two factors that act together synergistically and are not simply additive. Results for the L1/L1 pair are shown in Table 4. In this case, the effect of accent disappears (0.05), since the two speakers in the pair have the same accent (general American English). This latter result is not surprising, but it is important since it shows that our methodology is suitable for studying the role of accent/dialect in talker recognition (by humans).

	Estimate Non-native / native	Estimate Native / native
(Intercept)	1.77	2.37
Voice	1.95	2.16
Accent	0.42	0.05
Interaction	2.79	2.37

Table 4: *Results from the mixed effects linear model on non-native / native speaker pairs*

5. Discussion

We have proposed a methodology that may allow researchers to examine the effect of non-native/regional accents on talker recognition. The methodology combines a state-of-the-art accent-conversion system that allows us to resynthesize the voice of a speaker with a different accent, and an experimental protocol for perceptual listening tests that allows us to measure the relative contribution of accent and voice quality. We validated the methodology on a speech corpus with four pairs of L2-L1 speakers and a pair of L1-L1 speakers, for a total of 50 combinations between accent and voice quality. The highest ratings of speaker similarity were obtained when the voice and accent of the two speakers were matched, regardless of whether the speakers were real (i.e., original recordings) or synthesized (i.e., accent conversions), whereas the lowest ratings were obtained when both accent and voice were mismatched. More interestingly, speakers with the same voice quality but different accents were rated as being more similar to each other than speakers with different voice quality but similar accents. Using these results in a linear mixed effects model, we were able to estimate that (for our corpus), the effect of non-native accents is roughly 20% of that of voice quality, and this effect goes away (2%) when performing accent-conversions on L1-L1 pairs with the same accents.

Beyond the voice conversion community, our methodology may be of interest in the field of earwitness identification. Prior studies have shown that differences in accent between a speaker and a listener can lead to degradation in earwitness identification performance [40, 41]. Thus, accent plays a key role that should be considered while verifying earwitness testimony. As an example, using an earwitness lineup experiments, our accent conversion algorithms could be used to examine the effect of accent on memory recall. Our methodology may also be of interest in sociophonetics, where it could be used to study the effect of various non-native accents on social biases, e.g., non-native speakers receive less favorable judgments of credibility, competence and intelligence, and they have fewer employment opportunities, housing options, and access to healthcare [42, 43]. The results also suggest that this approach can identify bias toward differences in L1 social and regional accents [44].

6. Acknowledgements

This work was supported by NSF awards 1619212 and 1623750.

7. References

[1] C. Y. Espy-Wilson, S. Manocha, and S. Vishnubhotla, "A new set of features for text-independent speaker identification," in *Ninth International Conference on Spoken Language Processing*, 2006.

[2] S. J. Park, C. Sigouin, J. Kreiman, P. A. Keating, J. Guo, G. Yeung, F.-Y. Kuo, and A. Alwan, "Speaker identity and voice quality: Modeling human responses and automatic speaker recognition." in *Interspeech*, 2016, pp. 1044–1048.

[3] J. Kreiman and D. Sidtis, "Voices and listeners: Toward a model of voice perception," *Acoustics Today*, vol. 7, no. 4, pp. 7–15, 2011.

[4] M. Latinus and P. Belin, "Anti-voice adaptation suggests prototype-based coding of voice identity," *Frontiers in Psychology*, vol. 2, p. 175, 2011.

[5] G. Zhao and R. Gutierrez-Osuna, "Using phonetic posteriorgram based frame pairing for segmental accent conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1649–1660, 2019.

[3] The linear mixed effects models were trained in R using the lme4 [39] package. Tests for statistical significance were performed using the R anova package.

[6] S. Aryal and R. Gutierrez-Osuna, "Reduction of non-native accents through statistical parametric articulatory synthesis," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. 433–446, 2015.

[7] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native English speech corpus," *in Interspeech*, pp. 2783–2787, 2018.

[8] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004, pp. 223 –224.

[9] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2619–2634, 2009.

[10] D. Abercrombie *et al.*, *Elements of general phonetics*. Edinburgh University Press Edinburgh, 1967, vol. 203.

[11] M. Garellek, R. Samlan, B. R. Gerratt, and J. Kreiman, "Modeling the voice source in terms of spectral slopes," *The Journal of the Acoustical Society of America*, vol. 139, no. 3, pp. 1404–1410, 2016.

[12] Y. Lee, P. Keating, and J. Kreiman, "Acoustic voice variation within and between speakers," *The Journal of the Acoustical Society of America*, vol. 146, no. 3, pp. 1568–1579, 2019.

[13] T. Valentine, "A unified account of the effects of distinctiveness, inversion, and race in face recognition," *The Quarterly Journal of Experimental Psychology Section A*, vol. 43, no. 2, pp. 161–204, 1991.

[14] G. Yovel and P. Belin, "A unified coding strategy for processing faces and voices," *Trends in Cognitive Sciences*, vol. 17, no. 6, pp. 263–271, 2013.

[15] S. G. Young, K. Hugenberg, M. J. Bernstein, and D. F. Sacco, "Perception and motivation in face recognition: A critical review of theories of the cross-race effect," *Personality and Social Psychology Review*, vol. 16, no. 2, pp. 116–142, 2012.

[16] T. K. Perrachione and P. C. Wong, "Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex," *Neuropsychologia*, vol. 45, no. 8, pp. 1899–1910, 2007.

[17] T. K. Perrachione, "Speaker recognition across languages." Oxford University Press, 2017.

[18] S. V. Levi, "Methodological considerations for interpreting the language familiarity effect in talker processing," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 10, no. 2, p. e1483, 2019.

[19] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[20] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[21] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[22] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2011.

[23] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using speaker-dependent conditional restricted boltzmann machine," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[24] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *ICASSP*, 2015, pp. 4869–4873.

[25] C. Liberatore, S. Aryal, Z. Wang, S. Polsley, and R. Gutierrez-Osuna, "Sabr: Sparse, anchor-based representation of the speech signal," in *Interspeech*, 2015, pp. 608–612.

[26] G. Zhao and R. Gutierrez-Osuna, "Exemplar selection methods in voice conversion," in *ICASSP*, 2017, pp. 5525–5529.

[27] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 421–426.

[28] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams," in *Interspeech*, 2019, pp. 2843–2847.

[29] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sadhana*, vol. 36, no. 5, pp. 713–727, 2011.

[30] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[31] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *ICASSP*, 2014, pp. 215–219.

[32] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783.

[33] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *NIPS*, 2015, pp. 577–585.

[34] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *ICASSP*, 2019, pp. 3617–3621.

[35] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *NIPS*, 2018, pp. 10 215–10 224.

[36] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[37] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, G. Grothendieck, P. Green, and M. B. Bolker, "Package 'lme4'," *Convergence*, vol. 12, no. 1, p. 2, 2015.

[40] J. H. Kerstholt, N. J. Jansen, A. G. Van Amelsvoort, and A. Broeders, "Earwitnesses: Effects of accent, retention and telephone," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 20, no. 2, pp. 187–197, 2006.

[41] C. P. Thompson, "A language effect in voice identification," *Applied Cognitive Psychology*, vol. 1, no. 2, pp. 121–131, 1987.

[42] R. Chakraborty, "A short note on accent–bias, social identity and ethnocentrism," *Advances in Language and Literary Studies*, vol. 8, no. 4, pp. 57–64, 2017.

[43] S. Lev-Ari and B. Keysar, "Why don't we believe non-native speakers? the influence of accent on credibility," *Journal of Experimental Social Psychology*, vol. 46, no. 6, pp. 1093–1096, 2010.

[44] J. Baugh, "Linguistic profiling," in *Black Linguistics*. Routledge, 2005, pp. 167–180.