UNIVERSITY OF CALIFORNIA
Los Angeles

Speaking Style Variability in Speaker Discrimination by Humans and Machines

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

Amber Afshan

2022

ABSTRACT OF THE DISSERTATION

Speaking Style Variability in Speaker Discrimination by Humans and Machines

by

Amber Afshan
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2022
Professor Abeer A. Alwan, Chair

A speaker's voice constantly varies in everyday situations, such as when talking to a friend, reading aloud, talking to pets, or narrating a happy incident. These changes in speaking style affect human and machine abilities to distinguish speakers based on their voice. This dissertation studies the effects of speaking style variability on speaker discrimination performance by humans and machines.

We compare human speaker discrimination performance for read speech versus casual conversations. Listeners perform better when stimuli are style-matched, particularly in read speech – read speech trials. They perform the worst in style-mismatched conditions. Moderate style variability affects the "same speaker" task more than the "different speaker" task. The speakers who are "easy" or "hard" to "tell together" are not the same as those who are "easy" or "hard" to "tell apart." Analysis of acoustic variability suggests that listeners find it easier to "tell speakers together" when they rely on speaker-specific idiosyncrasies and that they "tell speakers apart"

based on their relative positions within a shared acoustic space.

The effects of style variability on automatic speaker verification (ASV) systems are systematically analyzed using the UCLA Speaker Variability database, which comprises multiple speaking styles per speaker. The performance is better when enrollment and test utterances are of the same style, but it substantially degrades when styles are mismatched. We hypothesize that between-frame entropy can capture style-related spectral and temporal variations. We propose an entropy-based variable frame rate (VFR) technique to address style variability in two different approaches: data augmentation and self-attentive conditioning. Both approaches improve performance in style-mismatch scenarios and are comparable in performance.

Furthermore, humans and machines seem to employ different approaches to speaker discrimination. In an attempt to improve ASV performance in the presence of style variability, insights learnt from the human speaker perception experiments are used to design a training loss function, referred to as "C_{llr}CE loss". C_{llr}CE loss focuses on both speaker-specific idiosyncrasies and relative acoustic distances between the speakers to train the ASV system. This loss function improves ASV performance in case of style variability, especially in the case of moderate style variations from conversational speech.

The dissertation of Amber Afshan is approved.

Sudhakar Pamarti

Jonathan Chau-Yan Kao

Jody E. Kreiman

Abeer A. Alwan, Committee Chair

University of California, Los Angeles

2022

To my family ...

for their love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation has been possible due to the endless support from many people. First and foremost, I would like to express my most earnest appreciation to my doctoral advisor, Professor Abeer Alwan. She has been a wonderful mentor and provided me with gracious support. She has been a constant source of inspiration. She encouraged me to tackle challenging problems and think critically about my work. Under her supervision, I was able to develop as a better researcher.

I am immensely grateful to Professor Jody Kreiman for her advice and guidance. She provided me with invaluable knowledge that has been instrumental in shaping my research on human speaker perception. I would also like to thank the members of my Ph.D. committee: Professor Jonathan Kao and Professor Sudhakar Pamarti, whom I have gotten to know and learn from throughout my many years at UCLA.

I am immensely thankful to Dr. Alan McCree for his guidance and for providing insight into my research. I also would like to thank Professor Patricia Keating for her support and advice, sharing her unparalleled knowledge in linguistics. I also want to express my gratitude to Professor Jonathan Flint for his guidance towards my research and an opportunity to work on depression research.

I had amazing mentors during my internships Dr. Kshitiz Kumar, Dr. Jian Wu, Dr. Martin Graciarena, Mahesh Kumar Nandwana, Dr. Andreas Tsiartas, Dr. Diego Castán, and Dr. Gang Chen. Their guidance enriched my training as a researcher.

My heartfelt gratitude for my undergraduate mentors, Professor Prasanta Kumar Ghosh and Professor Deepu Vijayasenan, who introduced me to speech processing and guided me to graduate school. A special thanks to Professor Ghosh for encour-

aging me to get involved in research and for helping me find my bearings initially.

I am grateful to my amazing labmates at SPAPL. I have had numerous fun, thought-provoking, and inspiring conversations with them that helped me keep up with the latest research and gave me ideas whenever I was stuck. Many thanks to Jom, Jinxi, and Soo for guiding me in my initial days. I would also like to thank Cynthia, Usha, Gary, Rohit, Kailun, Vijay, AJ, Morgan, Ruchao, Jinhan, Yunzheng, Huanhua, and Ben. Special thanks to Soo, Jinxi, Vijay, Cynthia, and Ruchao for their collaborations without which this work would not have been possible.

I cannot express enough gratitude to all the administrative staff: Deeona Columbia, Ryo Arreola, Julio Romero, Ylena Requena, Jose Cano, Vanessa Ramirez, and Mandy Smith for their gracious help and support throughout my time at UCLA.

I am immensely fortunate to have amazing friends who made Los Angeles a home-away-from-home from their company, celebrating small and big wins and support through day-to-day struggles. I am truly blessed to have great friends who were one call away whenever I needed them. They provided constant support, encouragement, and a hearing ear for all my breakdowns, for which I cannot thank them enough.

I am ever grateful to my mother Zareena Perveen and my father Syed Zabiulla without whom I would not be the person I am today. Finally, I would like to thank my sister Nishath Afza who has been my biggest cheerleader. Her profound belief in my abilities lifted me even when I was frustrated by my failures. I will always be grateful to my family for all their love, encouragement, and support through both challenging and joyful occasions during graduate school.

My Ph.D. experience would not have been the same without all these people. Thank you all for being there through the ups and downs of my journey.

VITA

2014	B.Tech., Electronics and Communication Engineering, National Institute of Technology, Karnataka.
2015–present	Graduate Student Researcher, Speech Processing and Auditory Perception Laboratory, UCLA.
2016–2021	Teaching Assistant, Electrical and Computer Engineering, UCLA.
2016	Interim Intern, Qualcomm.
2017	M.S., Electrical Engineering, UCLA.
2017, 2018	Student Associate, SRI International.
2020	Research Intern, Microsoft Corporation.

SELECTED PUBLICATIONS

A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination for "easy" versus "hard" voices in style-matched and -mismatched speech," The Journal of the Acoustical Society of America, 151(2):1393–1403, 2022.

A. Afshan, K. Kumar, and J. Wu, "Sequence-level Confidence Classifier for ASR

Utterance Accuracy and Application to Acoustic Models," in Interspeech, 2021.

R. Fan, **A. Afshan**, and A. Alwan, "Bi-APC: Bidirectional Autoregressive Predictive Coding for Unsupervised Pre-Training and its Application to Children's ASR," ICASSP, 2021.

A. Afshan, J. Guo, S. J. Park, V. Ravi, A. McCree, and A. Alwan, "Variable frame rate-based data augmentation to handle speaking-style variability for automatic speaker verification," in Interspeech, 2020.

A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination in humans and machines: Effects of speaking style variability," in Interspeech, 2020.

V. Ravi, R. Fan, **A. Afshan**, H. Lu, and A. Alwan, "Exploring the Use of an Unsupervised Autoregressive Model as a Shared Encoder for Text-Dependent Speaker Verification," in Interspeech, 2020.

S. J. Park, **A. Afshan**, J. Kreiman, G. Yeung, and A. Alwan, "Target and Non-Target Speaker Discrimination by Humans and Machines," in ICASSP, 2019.

A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of Voice Quality Features in Detecting Depression," in Interspeech, 2018.

CHAPTER 1

Introduction

1.1 Motivation

The manner in which a speaker says an utterance can change unintentionally from one scenario to another, for example, due to social context (e.g., talking to a friend versus public speaking) or emotional or physiological state; or it can change intentionally, for example, to express irony or in an attempt to hide one's identity [KS11]. These variations introduce within-speaker variability that could affect speaker discrimination abilities of both humans and machines.

Within-speaker variability can be further categorized into extrinsic variability and intrinsic variability. *Extrinsic variability* is associated with factors not directly related to the speaker's behavior (e.g., recording conditions, channel types, and environmental noise). On the other hand, *intrinsic variability* is related to the speaker's conscious and/or unconscious behavior that can influence speech signal production. It could be variations due to vocal effort, speaking styles, speaking rate, loudness, emotional state, or physical status. In this dissertation, we focus on the effects of speaking style variability on speaker discrimination performance by humans and machines.

1.2 Speaker Discrimination by Humans

Within-speaker variability strongly impacts the perception of unfamiliar voices [LBG19, LBL19]. For example, the effects of speaking style variability on speaker identification accuracy have been studied extensively, particularly in the forensic literature [SY80, BF06, GHK19, GKH15]. One study [Jes08] showed that style variability confuses ear witnesses hearing a suspect shouting versus reading aloud during a voice lineup. In non-forensic work, human performance has shown to suffer when style changes from read to pet-directed speech, which is characterized by exaggerated prosody [PYV18]. Differences in style were extreme in both examples (shouting and pet-directed speech).

Two recent studies have provided insights into the ways in which listeners deal with moderate speaker variability. The authors in [SBR19] compared style-matched read speech trials with read versus spontaneous speech trials, and found that listeners were more accurate and confident in style-matched trials compared to style-mismatched ones. However, their experiments included style-matched trials only from read speech, leaving open the question as to whether the perception of a particular style is more robust than another in identifying a speaker. A second study [STN21] addressed this limitation by including style-matched spontaneous speech as well. They found that performance on style-matched trials exceeded that for mismatched trials, with performance on style-matched read speech trials better than that for style-matched spontaneous speech. Their results also revealed a significant bias toward "same speaker" over "different speaker" responses.

Although these studies show that acoustic variability confuses listeners, they leave open the important questions of why and how this occurs. Neither study quantified

the extent of acoustic variability within- and across-speakers and styles, nor did they examine the relationship between acoustic variability and how well listeners performed in "same speaker" versus "different speaker" tasks.

Evidence from voice sorting tasks indicates that humans do vary their perceptual strategies when "telling people together" (i.e., assessing within-speaker variability in voice) versus "telling people apart" (i.e., assessing between-speaker variability in voice) [LBG19, JML20]. However, we do not know how or why listeners vary their perceptual strategies in trials where speakers are the same (a "same speaker" task) versus trials in which speakers are different (a "different speaker" task). In "same speaker" trials, differences between stimuli reflect within-talker acoustic variability, while in "different speaker" trials differences largely reflect between-speaker variability. The nature and extent of differences in listener performance in these two trial types should follow from differences in the nature and extent of these two kinds of variability. Thus, three major questions arise: (i) How does human speaker discrimination performance vary with speaking style?; (ii) Is there a difference in how speaking style variations affect "same speaker" versus "different speaker" tasks?; and (iii) How does human speaker perception relate to the nature and extent of acoustic variability that occurs within- versus between-speakers?

Recent studies [LKK19, LK19] showed that the most important principal components (typically 2-3) describing acoustic variability for individual speakers were shared by all the speakers, but the majority of the principal components (typically 5) were idiosyncratic. Moreover, individual speakers' acoustic spaces (within-speaker variability) and spaces for whole populations of speakers (between-speaker variability) shared a similar structure. This shared structure was mainly computed over higher-frequency harmonic ($H_4^* - H_{2k}^*, H_{2k}^* - H_{5k}$), inharmonic energy in the voice

and over formant dispersion in read speech. In conversational speech, the structure corresponded to variability in source spectral shape, spectral noise, F0, and in higher formant frequencies. However, little is known about the relationships among within- and between-speaker acoustic variability and listener performance, particularly in the context of differences in speaking style. In this study we examined these relationships by asking listeners to discriminate among speakers with moderate speaking style variations. Listener performance for individual speakers was interpreted with respect to the speakers' acoustic spaces, with separate analyses for "same speaker" and "different speaker" trials.

We hypothesize that speaking style variability would have a large effect on performance in the case of unfamiliar speaker discrimination, because the "same speaker" task largely relies on within-speaker variability. Moreover, casual conversations have a higher degree of variation in comparison to read speech [LBS19], suggesting that the "same speaker" task may be more difficult for conversational speech. Performance on "different speaker" tasks theoretically relies on the relative positions of voices in a shared acoustic structure (between-speaker variability). Previous research [Laa92] has shown that there are inconsistencies between listeners when classifying read and conversational speech, indicating that the moderate differences between these styles have minor perceptual effects, and suggesting that moderate speaking style variations result in small within-speaker variability. Based on [Laa92] and the studies reviewed earlier, we hypothesize that moderate speaking style variability would have a smaller effect on speaker discrimination performance for "different speaker" trials as they primarily rely on between-speaker variability.

1.3 Speaker Discrimination by Machines

Automatic speaker verification (ASV) refers to the task of "enrolling" speakers with one or more utterances each and verifying the "test" utterance against the enrolled speakers. Automatic speaker discrimination (ASD) by machines is a special case of ASV systems where "enrollment" is done using only one utterance from the speaker. We use ASV when referring to machine experiments in general, and ASD when human and machine performances are compared.

ASV systems generally assume that the variability across speakers is greater than the variability within speaker. However, this is not always the case. There is some overlap between variability across speakers and variability within-speakers. When the acoustic properties of an individual's speech differ between the enrollment and test utterances, ASV system performance generally degrades [SRG14]. These within-speaker variations can be from extrinsic or intrinsic variability. There has been considerable progress in studying the effects of extrinsic variability on ASV performance [GYM16, SS20, RFA20, ZZW21, FMB22]. While some studies showed that ASV performance also degraded due to intrinsic variabilities–vocal effort, speaking styles, speaking rate, loudness, emotional state and physical status [SGB08, SKS09, CX12].

The focus of this dissertation is the effect of speaking style variability which is another source of intrinsic variability that can make acoustic characteristics considerably different within a speaker. Only a limited number of studies have investigated the effects of style variability on ASV performance. *Style factors* are shown to be present in widely-used speaker representations [WK19] such as i-vectors [DKD11] and x-vectors [SGS18]. ASV performance degradation due to style mismatch between the

enrollment and test utterances were systematically analyzed in [PSK16, PYV18]. To alleviate the degradation due to style variabilities, some studies proposed the use of a joint factor analysis framework [SKS09, CX12]. In [ZRH18], curriculum-learning based transfer learning was done using neutral/physical stressed as well as read and spontaneous speech to compensate for style mismatches during testing. Note that the compensation techniques proposed in these studies require a variety of speaking styles per speaker to train the systems, i.e., the training data includes all the styles occurring in the test utterances [ZRH18]. However, one might not always have prior knowledge of the speaking style of the test utterances.

Variations in speech due to speaking style can be broadly classified into rhythmic variations, including speech rate, long pauses, changes in the duration of individual sounds, boundary articulation, and prosodic variations. However, the latter is directly associated with speaker identity and disentangling prosody from speech will result in performance degradation in speaker verification tasks. Hence, we focus on addressing the effects of rhythmic variations between styles. Acoustical differences between read and conversational speech include different speaking rates and inconsistent pauses between words. There are also variations in the number and type of phonological phenomena observed. For example, vowels are modified or reduced in conversational speech, and word-final plosive bursts are not released while it is not the case in read speech [PDB86]. Similar differences are observed across other speaking styles as well [Esk93]. Hence, in this work we focus on developing ASV systems that are robust to speaking style variations.

1.3.1 Data Augmentation

One can expect that including various speaking styles in the training data might improve the speaking-style robustness of the system. However, corpora with sufficient numbers of speakers speaking with different styles are not available. A widely-used approach to address insufficient data to train different conditions in ASV is *data augmentation* using artificially generated data. Augmentation strategies include adding variations of noise, reverberation [HMB16, SGS18], collecting additional domain-specific data [ZRH18], and synthesizing data [RMG19]. Yet, for style variability, artificially synthesizing speaking styles is not yet reliable enough to be applied [WSX17, WK19]. Hence, we need an alternative approach to generating style variants for data augmentation.

1.3.2 Self-attention

ASV systems generally use pooling to obtain a fixed-dimension representation from variable-length utterances. In [VLM14], the pooling was performed at the last hidden layer. More recent works [SGP17, SGS18] used a statistics pooling layer to calculate the mean and standard deviation of the utterance resulting in a fixed-dimension representation assuming each frame to be equally important. However, we know that not all frames are equally important in conveying speaker or content information [ZA00]. To address this issue few works [ZCZ17, ZKS18, OKS18] have proposed using self-attention in the pooling layer and have observed performance improvements in ASV tasks. Recently [WOL18] decoupled attention weights extracted from an x-vector system and used it in combination with an i-vector system and showed performance improvements. The results confirm that attention weights can better represent the

relative importance of each frame irrespective of the underlying embeddings. To learn weights so that the embeddings are style-robust, the attention network needs information to address style effects.

Thus, we experiment with two different approaches to address the effects of style variability: (i) data augmentation, and (ii) self-attention conditioning. We expect these ASV systems to be more robust to style variations than their traditional counterparts.

1.4 Loss functions to train ASV systems

Automatic speaker verification is an open-set problem, i.e., the test speakers are unavailable to the ASV system during training. ASV is, hence, a metric learning problem that needs to map speakers to a discriminative embedding space.

However, most of the work on speaker verification has focused training with identification objectives (for example, cross-entropy loss [KSH12, GBC16], also referred to as softmax loss). Identification loss functions learn linearly separable embeddings by focusing on maximizing inter-speaker distances. However, identification objectives do not minimize intra-speaker distances (i.e., increasing embedding similarity). Hence, the resulting embeddings do not have adequate discriminative properties. So the networks trained on identification objectives are often combined with different backends to build an ASV system. One such backend is PLDA [KSO13] (probabilistic linear discriminant analysis) which is used to obtain scores on the verification task.

Addressing the drawbacks of identification loss in ASV systems, one work used

Angular softmax [LGO18] loss. Angular softmax uses cosine similarity as the logit input to the softmax layer. Additive margin variants of Angular softmax such as AM-Softmax [WWZ18, WLL18] and AAM-Softmax [DGX19] use a cosine margin penalty on the target logit. These techniques though effective, have been proven sensitive to the value of scale and margin in the loss function, making them challenging for use in ASV systems.

As an alternative to identification objectives, metric learning approaches that focus on minimizing intra-speaker distances have been used. Metric learning approaches such as contrastive loss [CX12] and triplet loss [SKP15] have been used in ASV tasks with some success [ZKH18, CWM18]. However, these approaches require careful selection of triplet pairs i.e. anchor, positive and negative pairs, resulting in longer training cycles. Apart from the high computational cost, these losses do not consider the performance measures (such as equal error rate (EER) and detection cost function (DCF)) used in the final evaluation of the speaker verification task.

It has been shown that considering metric related to final evaluation improves ASV performance at least in text dependent ASV systems by using aAUC [MMO20], aDCF [MMR19] and C_{llr} [MMO21] objectives. The C_{llr} loss, in particular, provides performance improvements without the need for triplet pairs and provides computational cost similar to that of identification objectives such as cross-entropy loss. However, [MMO21] used C_{llr} in a text dependent speaker verification task and its efficacy has not been evaluated in a text independent case.

Given that the best training objective would be the one that maximizes inter-speaker distances and minimizes intra-speaker distances, objective designs need to consider style variations. Style variations play a critical role in determining inter and

intra-speaker variabilities. However, none of the above objectives regard the effects of style variations. Therefore, there is a need for a training objective (loss function) that focuses on style variations. In this research, we experiment on developing such an objective function.

1.5 Comparison between Humans and Machines

Speaking style variations are prevalent in everyday life, changing as we move from talking to a friend to reading aloud, from public speaking to talking to an infant. Regardless of these variations, humans are often able to recognize a familiar voice after hearing it for a few seconds [WM12]. Previous research suggests that for humans, recognizing familiar talkers entails matching a sample to stored voice templates, whereas recognizing unfamiliar talkers is a much more involved process requiring acoustic feature comparisons [VK87]. In this dissertation, we are interested in comparing human abilities to distinguish between unfamiliar voices from short duration ($\sim$3 s), text-independent utterances, in the presence of moderate style variability against state-of-the-art automatic speaker verification systems.

Recent studies [PYV18, PAK19] showed that humans consistently outperformed machines in both style-matched and -mismatched conditions when discriminating speakers from samples of read versus pet-directed speech (characterized by exaggerated prosody), although style variations resulted in worse performance for both humans and machines. Forensic literature includes comparisons between human and machine performances. The authors in [HHF17] found that forensic experts were able to resolve speaker pairs incorrectly classified by ASV systems. In these examples, differences in style were extreme, and little is known about how moderate

variations in style, for example between read and conversational speech, affect the relative performance of humans versus machines in speaker discrimination.

Evidence from voice sorting tasks indicates that humans vary their perceptual strategies when "telling people together" versus "telling people apart" [LBG19, JML20], while machines apply the same classification approach in target and non-target trials [PAK19]. Given that humans and machines seem to employ different approaches to speaker discrimination, this suggests that machines can adopt strategies from humans, and humans might do better with machine assistance in certain situations.

Therefore, we focus on learning from human speaker perceptual strategies to develop ASV algorithms, in particular, training loss functions. We employed an unfamiliar *speaker discrimination* task in which the listener decides if two samples are from the same speaker or not in presence of moderate style variability (i.e, between read and conversational speech). We hypothesized that such algorithms might improve ASV system performance for conversational styles.

1.6 Dissertation Outline

The rest of this dissertation is organized as follows:

Chapter 2 introduces the databases used in the experiments reported in the dissertation. The chapter also describes the features used.

Chapter 3 begins by laying out the perception experiments used to study human speaker discrimination performance in presence of moderate style variability. It also analyzes the results of the experiments and attempts to answer questions on the

effects of style variability on human performance and human approaches to speaker discrimination.

Chapter 4 outlines the two different style-robust methods proposed in this dissertation: a data augmentation method, and a self-attention conditioning method. It also reports comparisons between the two methods and on approaches to choosing the appropriate style-robust method for different applications.

Chapter 5 is concerned developing ASV algorithm inspired by human perception to improve performance, especially in style-mismatched conditions.

Chapter 6 concludes the dissertation with a summary of the key results, possible applications of the work and suggestions for future work.

CHAPTER 2

Databases and features

Four databases were used in this dissertation. The UCLA speaker variability database being the key database used for style variability analysis. For extracting representations from speech, we used two different feature sets: Mel-frequency cepstral coefficients (MFCCs) and voice quality features (VQual).

2.1 Databases

2.1.1 The UCLA Speaker Variability Database

In order to systematically study both within- and between-speaker variability, a multi-speaker speech database including multiple speech tasks per speaker is needed. Hence, the UCLA Speaker Variability Database [KKA19, KPK15, KKA21] was employed. The database is available from the Linguistic Data Consortium (LDC).[1] It incorporates commonly-occurring variations in voice deriving from phonetic content, speaking style, and affect conditions. This database includes speech from 101 female and 101 male speakers, recorded with a 1/2" Brüel & Kjær microphone in a sound-attenuated booth at a sampling rate of 22kHz. Each speaker recorded 12 recordings

[1] https://doi.org/10.35111/c5gk-6j49

over three sessions, providing a total of 2424 recordings in the database.

This work used **read** sentences representing scripted speaking style (≈ 75 sec per speaker); giving **instructions** representing unscripted clear monologue style (≈ 30 sec per speaker); **narrating** a recent neutral, happy, or annoying conversation representing unscripted affective speech (≈ 30 sec each affect per speaker); speaker's side of the conversation on a call with a familiar person representing unscripted **conversational** style (60–120 sec per speaker); and talking to pets in a video i.e., **pet-directed** speech, characterized by exaggerated prosody (60–120 sec per speaker).

2.1.2 The Speakers in the Wild Database (SITW)

SITW has 2,883 recordings from 117 male and 63 female speakers divided into 6,445 utterances sampled at 16 kHz. SITW consists of both single- and multiple-speaker audio with segment labels for person of interest (POI) in enrollment utterances. Enrollment utterances include **core** (single POI) and **assist** (multiple speakers with segmentation labels for POI) and test utterances include **core** (single POI) and **multi** (multiple speakers with no segmentation labels for POI). This dissertation uses SITW in order to gain insights into the effects of some of the proposed approaches on a large-scale database as SITW includes some speakers employing multiple speaking styles.

2.1.3 NIST SRE and Switchboard databases

The NIST Speaker Recognition Evaluation (SRE) 04, 05, 06, 08 and 10 databases [PM04, PML06, MG09] and the Switchboard II corpus, phase 2 [GWC99] were used in this

work, mainly for training the ASV systems. These databases provide more than 3,000 hours of speech samples from 3,408 female and 1,832 male talkers, sampled at 8 kHz.

The SRE and Switchboard databases offer many recordings from a large number of speakers with multiple speech tasks. However, there are certain drawbacks associated with using them for analyses of the effects of style variability. First, they do not provide multiple speech tasks per speaker under controlled recording environments. Second, they do not provide metadata regarding speaking style.

2.1.4 VoxCeleb Database

The Voxceleb2 dataset [CNZ18] consists of speech from YouTube videos of 3,682 male and 2,313 female speakers and includes 1,092,009 utterances with a sampling rate of 16 kHz. We use the *DEV* set from the Voxceleb2 dataset for training the ASV systems. The main disadvantage of using VoxCeleb2 for testing is that it comprises interview-style speech only and does not include different styles for each speaker. Hence, we believe that it does not provide a good representation of the test case scenario targeted in this work.

2.2 Features

2.2.1 Voice Quality Features

Voice Quality (VQual) feature selection was motivated by a psychoacoustic model of voice quality [GSG16, KLG21]. The set comprised F_0: fundamental frequency, F_1, F_2, F_3, F_4: the first four formants, CPP: cepstral peak prominence [HCE94],

and the amplitude differences between the first (H_1^*), second (H_2^*), and fourth (H_4^*) harmonics, and the harmonics nearest 2 kHz (H_{2k}^*) and 5 kHz (H_{5k}), denoted as $H_1^* - H_2^*$, $H_2^* - H_4^*$, $H_4^* - H_{2k}^*$, and $H_{2k}^* - H_{5k}$. These measures quantified the harmonic source spectral shape. Harmonic values marked with '*' were corrected for the influence of formants on harmonic amplitudes [HC99, IA04]. Following Lee et. al. [LKK19, LK19], we also included FD: formant dispersion (calculated as the average difference in the frequency between each adjacent pair of formants), *energy* (a measure of amplitude given by root-mean-square energy calculated over five pitch pulses), and the ratio of amplitudes of SHR: subharmonics to harmonics [Sun02, Her21] as a measure of period doubling, for a total of 13 features for every analysis frame. We used VoiceSauce [SKV11] to extract the VQual features. These measures have also shown to be useful in detecting affect [PAC18], depression [AGP18, RWF22] and sleepiness [RPA19] in speech.

2.2.2 Mel Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCCs) are one of the widely used feature sets for speech representation [DM80]. MFCCs represent the overall spectral envelope of the speech signal, and are closely related to the phonetic information in speech at the frame level. A key feature of MFCCs is that they use a perceptual frequency scale, referred to as the Mel scale.

A standard MFCC extractor process the speech segments with a window size of 25 ms, a window shift of 10 ms, and a pre-emphasis filter with coefficient 0.97. This dissertation, however, will not only use the standard extractor but also a few of its variants.

CHAPTER 3

Speaker discrimination performance by humans

In this chapter, we compare human speaker discrimination performance for read speech versus casual conversations, and explore differences between unfamiliar voices that are "easy" versus "hard" to "tell together" versus "tell apart." Thirty listeners were asked whether pairs of short style-matched or -mismatched, text-independent utterances represented the same or different speakers. Perceptual experiments were conducted with unfamiliar voices, using short duration ($\sim$3 s) utterances. The main focus was to study the effects of moderate variations in speaking style (i.e., between read and conversational speech) on human speaker recognition and discrimination performance. We aim to answer these major questions: (i) How does human speaker discrimination performance vary with speaking style?; (ii) Is there a difference in how speaking style variations affect "same speaker" versus "different speaker" tasks?; and (iii) How does human speaker perception relate to the nature and extent of acoustic variability that occurs within- versus between-speakers? This work was published in [AKA20, AKA22].

3.1 Methods

3.1.1 Perceptual speaker discrimination

3.1.1.1 Stimuli

Voice samples from 40 female speakers (also used in [PYV18, PAK19]) were drawn from the UCLA Speaker Variability Database [KKA19, KPK15, KKA21]. Forty speakers were studied to balance concerns about testing duration versus sampling considerations, and to provide continuity with our previous perception experiments using this dataset ([PYV18, PAK19]). Samples were restricted to female speakers to avoid any gender-dependent cues, and because females produced clearer contrasts between speaking styles than male speakers did (as judged by the authors). All speakers were self-reported native speakers of American English (confirmed *post hoc* by two linguists). Two sets of voice samples were selected for each speaker. The first (clear read speech) included five phonetically-rich Harvard sentences [IEE69], read twice in random order. The second (casual conversational speech) consisted of the speakers' side of a 2-minute telephone conversation with a family member or friend. The recordings were post-processed to remove any long preceding or trailing silences and all non-speech vocalizations (laughing, giggling, sighing). Six $\sim$3 sec clips were taken from each recording. Selections were carefully made to ensure that semantic cues would not bias responses. For instance, stimuli were chosen from different topics in the conversation. All chosen stimuli were recorded on the same day.

3.1.1.2 Listeners and listening task

All experimental procedures were approved by the UCLA Institutional Review Board. Thirty normal-hearing listeners including 24 native speakers of English (22 female, 8 male) participated in this experiment. An additional 6 speakers (3 were native speakers of Spanish, 2 of Mandarin, and 1 of Hindi) were also tested, but were later deleted from the data set because preliminary analyses suggested effects of native language on listener performance. There were not enough data to explore these effects in detail. However, in Section 3.2.1 we present the overall speaker discrimination performance of non-native speakers. The sample size was determined such that there are 12-15 listeners per subset of voices, as will be explained later.

Each listener undertook three kinds of comparisons, in random order. In one they heard two different read sentences; in another they compared two different clips excerpted from a conversation; and in the third, they compared one read sentence and one conversational sentence. Equal numbers of "same speaker" and "different speaker" trials were included for each of these three trial types, resulting in six different kinds of trials per experiment. Care was taken to make sure that a listener never heard the same stimulus twice. As only five different sentences had been recorded in the case of read speech, we randomly chose a second recording of one of the five sentences to repeat for the sixth trial.

Listeners were tested individually in a sound-attenuated booth. Stimulus pairs were played in random order over Etymotic insert earphones (model ER-1) at a constant comfortable listening level. To minimize fatigue, listeners heard one of two subsets of speakers (15 listeners per subset). Each subset included 24 speakers selected at random from the pool of 40, for a total of 144 trials per listener (6 trial

types x 24 speakers). On "different speaker" trials, two speakers were paired at random, such that each was compared with every other speaker an equal number of times.

Each listener heard the stimulus pairs in a unique random order and was asked (i) "Did the two voices represent the same speaker or two different speakers?, and (ii) "How confident are you in your response on a scale of 0 to 5 (0 = wild guess and 5 = very confident)?" Pairs of stimuli could be heard twice, once in each presentation order (AB/BA). Listeners were not aware of the number of speakers included in the experiment. They were encouraged to complete the experiments at their own pace, taking breaks as necessary. Testing time averaged about 45 minutes.

3.1.2 Evaluation metric

3.1.2.1 Calculation of scores

Same/different responses were combined with confidence ratings to create an unfolded similarity score for each stimulus pair. Confidence ratings (0 to 5) were multiplied by the decision (different = -1 and same = 1) to provide continuous scores ranging from -5 (highly confident that the voices are different) to 5 (highly confident that the voices are the same). This ensured that the similarity score reflected listeners' confidence as well as their same/different decisions.

Similarity scores were used to calculate calibrated log-likelihood ratios (LLRs), denoted as L. LLRs are used in this work instead of similarity scores by themselves, as they provide reliable probabilistic interpretations of the comparisons of the two hypotheses ("same speaker" or "different speaker"). Thus, LLRs provide a single identification score that can be meaningfully interpreted. In other words, calibrated

log-likelihood ratios provide numerical representations of listeners' degree of support for either hypothesis in each trial. This allows us to measure not only the listeners' discriminating power, but also the strength of the trials evaluated by them [RFG11]. Moreover, the calibrated LLRs are needed to obtain the log-likelihood-ratio cost function in Section 3.1.2.2, which, unlike the standard measures, is application-independent. This provides a universal probabilistic interpretation in the analysis. A calibration system based on a standard logistic regression solution [BD11a] was used to estimate the LLRs by optimizing the following mapping:

$$L_t = a + bs_t \tag{3.1}$$

where L_t is the calibrated output log-likelihood-ratio for trial t and s_t is the similarity score for trial t. Offset parameter a and the weight b are optimized with logistic regression [Bru10].

3.1.2.2 Analysis of performance errors

Speaker discrimination performance was evaluated in terms of equal error rates (EER) and the log-likelihood-ratio cost function (C_{llr}) [VB07]. While the EER is a widely-used measure, it does not measure ability to set good decision thresholds. Hence, C_{llr}, an application-independent measure for evaluating soft decisions, was also used. It can be interpreted as a measure that is inversely related to information. The lower the C_{llr}, the more the average information per trial (in bits) increases. In

[VB07] a closed-form solution for C_{llr} is provided:

$$C_{\text{llr}}(L_t) = \frac{1}{2}\left(\sum_{t\in\text{same}} \frac{\log_2(1+e^{-L_t})}{N_{\text{same}}} + \sum_{t\in\text{diff}} \frac{\log_2(1+e^{L_t})}{N_{\text{diff}}}\right) \tag{3.2}$$

where L_t is the log-likelihood-ratio for trial t , 'same' is a set of N_{same} "same speaker" trials and 'diff' is a set of N_{diff} "different speaker" trials. These two normalized terms represent the costs for "same speaker" (first term) and "different speaker" (second term) trials. We will refer to the first term as $C_{\text{llr}}^{\text{same}}$ and the second term as $C_{\text{llr}}^{\text{diff}}$.

We used the Bosaris toolkit [BD11b] to perform calibration and to calculate the evaluation measures. As data were limited, the calibration parameters were trained on and applied to the same set of scores.

3.1.2.3 Speaker-level analysis

This section describes speaker-level measures. The log-likelihood ratio L_t, which represents listeners' scalar responses to each given trial, was obtained for each trial t, as outlined in Section 3.1.2.1. To compare the scores for "same speaker" and "different speaker" trials involving each speaker, L^{same} for "same speaker" and L^{diff} for "different speaker" trials were calculated separately. L^{same} for a speaker was obtained by averaging the L_t values over the "same speaker" trials that included that particular speaker. It measures within-speaker variability across the stimuli as perceived by the listeners: a large L^{same} means small perceived within-speaker variability (i.e., these "same speaker" trials are easy). L^{diff} for a given speaker was calculated by averaging the L_t values over the "different speaker" trials that included a given speaker; it represents between-speaker variability across the stimuli as perceived by the listeners.

A large L^{diff} value indicates that the speaker has small perceived between-speaker variability, making it difficult for listeners to distinguish her from others.

A speaker-level aggregation of the log-likelihood-ratio cost function (C_{llr}; section 3.1.2.2) was also computed by calculating the mean across listeners over all the trials that included that particular speaker. The speaker-level C_{llr} represents the confidence listeners had when identifying that speaker. Speaker-level C_{llr} values for "same speaker" trials ($C_{\text{llr}}^{\text{same}}$) and "different speaker" trials ($C_{\text{llr}}^{\text{diff}}$) were also computed by calculating the average of their respective C_{llr}s across listeners.

For speaker-level analysis, trials were combined across conditions due to the limited number of trials per speaker (9 trials x number of listeners). Although collapsing conditions in this way precludes examination of factors other than speaker, these analyses focus primarily on the main effect of differences among speakers, so we felt adding power to tests of this main effect outweighed other considerations. Note that the system-level values are denoted by using a ($'$), i.e, C'_{llr}, $C'^{\text{same}}_{\text{llr}}$, $C'^{\text{diff}}_{\text{llr}}$, L', L'^{same} and L'^{diff}.

3.1.3 Speaker acoustic variability

3.1.3.1 Feature extraction and data processing

Voice quality features from utterances of vowels and approximants (i.e., /l/, /r/, /w/) in the stimuli were used as acoustic measures. There were a total of 13 features (F_0: fundamental frequency, F_1, F_2, F_3, F_4: the first four formants, CPP: cepstral peak prominence, $H_1^* - H_2^*$, $H_2^* - H_4^*$, $H_4^* - H_{2k}^*$, and $H_{2k}^* - H_{5k}$: the amplitude differences of the harmonics, FD: formant dispersion, SHR: subharmonics to harmonics

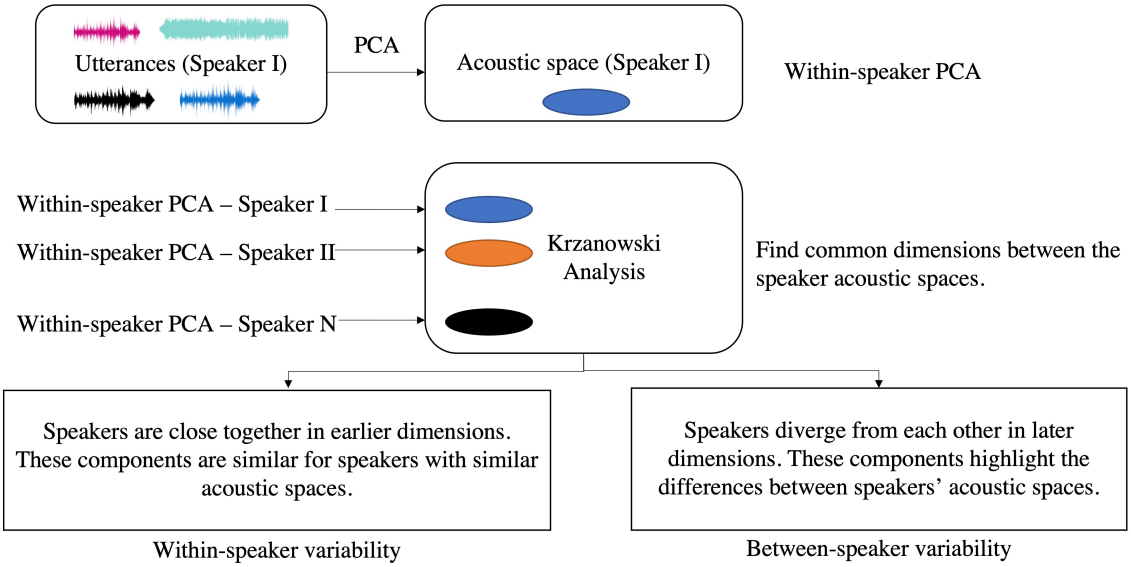


Figure 3.1: Block diagram representing the analysis of variability in speaker acoustic spaces using principal component analysis (PCA) and Krzanowski analysis.

ratio and Energy) for every analysis frame. Frames with missing or unrealistic values[1] were removed, after which features were normalized with reference to global maxima and minima, for a range across speakers of 0-1. We then calculated the moving average and the moving coefficient of variation (moving CoV = $\frac{\text{moving standard deviation}}{\text{moving average}}$) over a 25 ms window (commonly used in speech feature extraction, equivalent to 5 observations) for each of the 13 features. This resulted in a total of 26 acoustic features (13 moving averages and 13 moving CoVs). These 26 features were used for subsequent analysis.

3.1.3.2 Principal component analysis

Figure 3.1 represents the block diagram of the speaker variability analysis. Following Lee et al. [LKK19, LK19], we applied principal component analysis (PCA) to characterize acoustic variability in the voices of individual speakers. Utterances from each speaker were used to calculate the within-speaker PCA representing that individual's acoustic space. We retained only the principal components with eigenvalues greater than one so that each represented an interpretable amount of variance in the data [Kai60].

The analytical approach proposed by [Krz79] was used to compare PCA spaces, to avoid reliance on subjective criteria associated with visual examination. In this approach, let g be the number of speakers being compared with n_t observations for the t^{th} speaker ($t = 1, 2, \ldots, g$), with the same set of p variables measured for each speaker. Let us assume that for each speaker, k_t principal components represent that speaker's acoustic variability. Next, let b be an arbitrary vector in the original p-dimensional data-space and let δ_t be the angle between b and the vector most parallel to it in the space generated by the k_t principal components of speaker t ($t = 1, 2, \ldots, g$). We represent the loadings using the matrix L_t where the element $l_{ij}^{(t)}$ represents the loading of the j^{th} variable on the i^{th} principal component of the t^{th} speaker. Then the value of b that minimizes $V = \sum_{t=1}^{g} \cos \delta_t{}^2$ is given by the eigenvector b_1, corresponding to the largest eigenvalue μ_1 of $H = \sum_{t=1}^{g} L_t' L_t$.

The eigenvector b_2, corresponding to the second largest eigenvalue of H, satisfies the criterion for the next largest value of V and is orthogonal to b_1. When k_t

[1]For example, impossible zero values; measurements inconsistent between different techniques such as F0 estimated by praat, snack, straight; values with NaNs; etc.

different components have been obtained for the t^{th} speaker ($t = 1, 2, \ldots, g$) and $k = \min(k_1, k_2, \ldots, k_g)$, then only a k-dimensional comparison will be useful. Any further dimension will be orthogonal to at least one of the speaker spaces. Using this transformation thus allows us to compare different principal component subspaces, because the eigenvalues μ_i (alternatively, the minimum angles $\cos^{-1}(\mu_i)^{\frac{1}{2}}$) can provide a measure of the extent to which the subspaces differ, and the eigenvectors b_i can describe the nature of their similarities or differences. The smaller the angles between the subspaces, the higher the similarity. Algorithm 1 provides a psuedocode of the Krzanowski analysis implementation for one set of speakers.

Algorithm 1: Krwazonski analysis for set with g speakers

```
k ← min(k_1, ..., k_g)                    // k-dimensional comparison
for speaker t in set do
    L_t ← normalized loadings of speaker t
    H ← H + L_t' L_t
V ← Eigenvectors(H)  /* Loadings of the directions closest to the
 speakers in the set */
for variable j in set of p variables do
    b ← V_j          // Eigenvector corresponding to variable j
    for speaker t in set do
        c ← b' * L_t' * L_t * b
        δ_{j,t} ← arccos √c   // Angle between speaker t and direction j
```

Krzanowski analysis was performed over the within-speaker PCAs for all the speakers in a set, to obtain the dimensions common to speaker acoustic spaces. The earlier (lower) dimensions represent the components that are similar for speak-

Table 3.1: Speaker discrimination performance in terms of equal error rates (EER, %) and log-likelihood-ratio cost function for combined (C'_{llr}), "same speaker" trials ($C'^{\text{same}}_{\text{llr}}$), and "different speaker" trials ($C'^{\text{diff}}_{\text{llr}}$). The better (lower cost) value for "same speaker" versus "different speaker" trials in each condition is underlined. All reported comparisons are statistically significant.

read – read				**conversation – conversation**				**read – conversation**			
EER %	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$	**EER %**	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$	**EER %**	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$
6.96	0.264	<u>0.210</u>	0.318	15.12	0.529	<u>0.501</u>	0.557	20.68	0.691	<u>0.690</u>	0.692

ers with similar acoustic spaces, and which are thus suitable for comparing within-speaker variability. The speakers diverge from each other in later (higher) dimensions. Hence, the components in these dimensions highlight the differences between speakers' acoustic spaces, making them appropriate indices of between-speaker variability.

3.2 Results

Table 3.1 shows speaker discrimination performance for the three speaking-style conditions (read speech – read speech, conversation – conversation, and read speech – conversation). Statistical significance was evaluated using a two-sample Kolmogorov-Smirnov (KS) test [Smi48]. The statistical significance is reported in terms of p:the statistical significance, h: (0)accept/(1)reject null hypothesis, D: the KS-test statistic and N: degrees of freedom . All reported comparisons are statistically significant. EER values in this table indicate that listeners performed best when voice samples were style-matched read speech (EER = 6.96%). Performance decreased for conversation – conversation trials (EER = 15.12%; $p = 0.035, D = 0.059, N = 2304$), even though these were also style-matched. This decrease in performance is likely due to

Table 3.2: Speaker discrimination performance of non-native listeners in terms of equal error rates (EER, %) and log-likelihood-ratio cost function for combined (C'_{llr}), "same speaker" trials ($C'^{\text{same}}_{\text{llr}}$), and "different speaker" trials ($C'^{\text{diff}}_{\text{llr}}$). The better (lower cost) value for "same speaker" versus "different speaker" trials in each condition is underlined.

read – read				**conversation – conversation**				**read – conversation**			
EER %	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$	**EER %**	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$	**EER %**	C'_{llr}	$C'^{\text{same}}_{\text{llr}}$	$C'^{\text{diff}}_{\text{llr}}$
12.39	0.4292	<u>0.3836</u>	0.4748	23.22	0.7026	<u>0.6667</u>	0.7385	31.46	0.8723	0.8730	<u>0.8716</u>

additional variability in casual conversations (formal/informal, happy/sad/angry/neutral, etc.; [LBS19]). The style-mismatched read speech – conversation trials resulted in performance that was significantly worse than in either style-matched condition (read speech: $p = 4.95 \times 10^{-14}, D = 0.164, N = 2304$; conversation: $p = 4.61 \times 10^{-7}, D = 0.115, N = 2304$).

A comparison of the log-likelihood-ratio cost functions (see Section 3.1.2.2), $C'^{\text{same}}_{\text{llr}}$, and $C'^{\text{diff}}_{\text{llr}}$ values in Table 3.1 indicates that "same speaker" trials were easier than "different speaker" trials in all conditions (read speech – read speech: $p = 1.4 \times 10^{-101}, D = 0.88, N = 1152$; conversation – conversation: $p = 4.76 \times 10^{-73}, D = 0.72, N = 1152$; read speech – conversation: $p = 9.07 \times 10^{-74}, D = 0.59, N = 1152$). Differences in difficulty between the two tasks depended on speaking style, with style-matched read speech – read speech trials showing the best performance overall (0.210 and 0.318 for $C'^{\text{same}}_{\text{llr}}$ and $C'^{\text{diff}}_{\text{llr}}$, respectively), and the most difference between the same and different speaker tasks.

3.2.1 Speaker discrimination performance of non-native listeners

Table 3.2 shows speaker discrimination performance of non-native listeners for the three speaking-style conditions (read speech – read speech, conversation – conversation, and read speech – conversation). EER values in this table indicate that similar to natives, non-natives performed best when voice samples were style-matched read speech (EER = 12.39%). Performance decreased for conversation – conversation trials (EER = 23.22%; $p = 0.0.82, D = 0.052, N = 576$) but it was not statistically significant. The style-mismatched read speech – conversation trials resulted in performance that was significantly worse than in either style-matched condition (read speech: $p = 6.194 \times 10^{-7}, D = 0.226, N = 576$; conversation: $p = 3.702 \times 10^{-6}, D = 0.212, N = 576$). A comparison of the log-likelihood-ratio cost functions (see Section 3.1.2.2), $C'^{\text{same}}_{\text{llr}}$, and $C'^{\text{diff}}_{\text{llr}}$ values in Table 3.2 indicates that non-native listeners found "different speaker" trials easier in the read speech – conversation condition and "same speaker" trials easier in the other two conditions. However, as mentioned in Section 3.1.1.2 there were not enough data to explore the perception strategies of non-natives in detail. In subsequent analyses, we only use scores from native listeners.

3.2.2 Speaker-level log-likelihood-ratio analysis

Figure 3.2 compares the distribution kernel density plots overlaid onto histograms of speaker-level log-likelihood-ratios (see Section 3.1.2.3) for "same speaker" (L^{same}) and "different speaker" (L^{diff}) trials for the three style conditions. Recall that the positive end of this scale represents highly confident "same" responses, and the negative end represents highly confident "different" responses. The means of L^{same} and L^{diff} are

shifted towards correct responses in the read speech – read speech conditions. This increased separation of "same speaker" and "different speaker" trial distributions indicates that discrimination was easier and resulted in better performance in the read speech – read speech condition compared to the other two conditions. For example, compare discrimination performance from the distributions in Figures 3.2(a) (read speech – read speech) and 3.2(c) (read speech – conversational speech) (EERs = 6.96% versus 20.68%, respectively). The read speech – read speech trials resulted in an L^{same} distribution with small variance ($variance = 0.05$) confined to the positive response region. This was not the case with L^{diff} ($variance = 0.57$), indicating that listeners were more confident when classifying "same speaker" pairs than "different speaker" pairs.

In comparison, in the conversation – conversation condition [Figure 3.2(b)] variance in the L^{same} distribution increased ($variance = 0.34$), and this distribution overlapped with the L^{diff} distribution ($variance = 0.70$), suggesting that listeners' confidence decreased overall with a change in style from read to conversational speech. Finally, in the read speech – conversation condition [Figure 3.2(c)] the variance in the L^{same} distribution increased further ($variance = 0.75$), while it decreased slightly in the L^{diff} distribution ($variance = 0.55$). This overall pattern suggests that style affected the listeners' confidence in "same speaker" tasks, but not in "different speaker" tasks.

Given the multimodal shape of the distributions for conversation and style-mismatched tasks, the findings in terms of variances of LLRs were helpful. We evaluated the differences between the distributions across styles. The speaker-level log-likelihood-ratios for "same speaker" (L^{same}) tasks for the three style conditions differed significantly from one another, with means of $1.8915, 1.4131$ and 0.9475

for style-matched read speech, style-matched conversation and style-mismatched tasks, respectively (read speech – read speech versus conversation – conversation: $h = 1, p = 3.57 \times 10^{-5}, D = 0.45, N = 80$, read speech – read speech versus read speech – conversation: $h = 1, p = 7.34 \times 10^{-9}, D = 0.68, N = 80$, and conversation – conversation versus read speech – conversation: $h = 1, p = 0.04, D = 0.30, N = 80$). In contrast, the speaker-level LLRs for "different speaker" (L^{diff}) tasks for the three style conditions did not differ significantly (means $= -1.4891, -1.3433$ and -1.2165 for style-matched read speech, style-matched conversation and style-mismatched tasks, respectively; read speech – read speech versus conversation – conversation: $h = 0, p = 0.14, D = 0.25, N = 80$, read speech – read speech versus read speech – conversation: $h = 0, p = 0.08, D = 0.28, N = 80$, and conversation – conversation versus read speech – conversation: $h = 0, p = 0.72, D = 0.15, N = 80$). This result is consistent with our hypothesis that the effect of speaking style-variability is greater in "same speaker" tasks than in "different speaker" tasks.

3.2.3 Speaker-level log-likelihood-ratio cost analysis

Recall that the speaker-level log-likelihood-ratio cost function, C_{llr}, denotes the overall speaker information available when the listener is performing speaker discrimination. It is calculated by averaging the values for "same speaker" trials ($C_{\text{llr}}^{\text{same}}$) and "different speaker" trials ($C_{\text{llr}}^{\text{diff}}$) for a given speaker. A higher C_{llr} indicates less information available to the listener for the speaker discrimination task, hence more difficulty. For "same" and "different" trials, speaker-level C_{llr} values from LLR scores were used to group speakers into three subsets. The correlation between the speaker-level $C_{\text{llr}}^{\text{same}}$ and $C_{\text{llr}}^{\text{diff}}$ is weak ($r = -0.0892$), hence our preference for the

categorical approach used here, versus treating difficulty as a continuous variable. We classified the thirteen speakers with the lowest C_{llr} values ("same speaker" task: mean = 0.251; range = 0.169-0.361, "different-speaker" task: mean = 0.243; range = 0.127-0.367) into an "easy" subset and the thirteen speakers with the highest C_{llr} values ("same speaker" task: mean = 0.964; range = 0.669-1.434, "different-speaker" task: mean = 1.076; range = 0.741-1.582) as "hard" (difficult to distinguish speakers). The remaining fourteen speakers were referred to as "average" ("same speaker" task: mean = 0.508; range = 0.368-0.647, "different-speaker" task: mean = 0.538; range = 0.370-0.708).

The joint distribution of speakers across the three subsets for the "same speaker" versus "different speaker" tasks is shown in Figure 3.3. An entry $count_{m_i,h_j}$ denotes the number of speakers from subset i of the "different speaker" task overlapping with subset j of the "same speaker" task. For example, in the first column, 6 samples were "easy" in both the "same speaker" and "different speaker" tasks, whereas 2 samples that were "easy" in the "same speaker" task were "hard" in the "different speaker" task. More observations fall off diagonal (speakers are not equally "easy" to "tell together" and "tell apart") than on diagonal (the tasks are equally "easy" for that speaker), consistent with findings that humans rely on different information when performing the two tasks [LBG19, JML20].

3.2.4 Variability in the speaker acoustic spaces

Because the acoustic signal is the input to human perceptual processes, examination of acoustic variability may provide insight into the perceptual strategies listeners use when performing "same speaker" and "different speaker" tasks. To address this, we

used PCA to generate principal component subspaces and applied Krzanowski analysis (Section 3.1.3.2) to compare acoustic variability for speakers who were "easy", "average", or "hard" to discriminate. As noted above, Krzanowski analysis provides a means of quantifying the similarity of the acoustic spaces for different talkers, by generating loadings of the directions in the acoustic spaces that are closest to the PCs for the speakers in each subset.

Figures 3.4 and 3.5 show each orthogonal direction as a separate subplot. "Same speaker" trials are shown in Figure 3.4 and "different speaker" trials are shown in Figure 3.5. The angles listed at the top of each subplot are inversely related to the similarity between all speakers in the set. For ease of comparison, each subplot shows the same dimension for all three subsets of speakers ("easy", "average", and "hard" to discriminate). Speaking styles are combined, however, because speaker C_{llr} values were calculated across all conditions. Plots are additionally restricted to the absolute values of the top three contributing factor loadings to focus attention on the most important contributors to similarity and differences in the acoustic space. Finally, we restricted the number of orthogonal directions to a dimension of $k = 7$, which is the minimum number of principal components extracted per speaker.

3.2.4.1 "Same speaker" task

As Figure 3.4 shows, "easy," "average," and "hard" speakers were acoustically similar along the first two dimensions (as indicated by small mean angular separations), but they increasingly diverged after this, with the maximum variation along the 7th dimension. The mean angular separations quantify the extent to which the dimensions represent the similarity between speakers for each subset. Within-speaker

variations can be compared along dimension 1, which is associated with CoVs of F_1, F_2, and FD for all speakers. F_2 and FD contribute to separating voices on the second dimension; $H_4^* - H_{2k}^*$ also contributed for "easy" speakers, and F_1 for the "average" and "hard" speakers.

On the other hand, examination of dimension 7 shows that different features underlie acoustic differences for each group of speakers, with mean angular separations of $33.35°, 28.11°$, and $29.97°$ for "easy," "average," and "hard" speakers, respectively. For "easy" speakers, this dimension is related to *Energy*, *Energy* CoV, and F_2 CoV, suggesting that differences between speakers in these factors have little effect on listeners' ability to tell voices together. For "average" speakers, this dimension is related to CoVs of F_0, F_2, and FD, while voices that were hardest to tell together varied along F_3, *Energy*, and its CoV. Dimensions 3-6 explained a mixture of similarities and differences, with some speakers closer to each other along those dimensions and others farther apart.

3.2.4.2 "Different speaker" task

Figure 3.5 compares the principal components describing acoustic variability for speakers classified as "easy," "average," or "hard" to "tell apart" in "different speaker" trials. The coefficients of variation (CoVs) for F_1 and FD contributed to separating voices based on their within-speaker variability on the first dimension for all three groups; F_2 CoV also contributed for "easy" and "hard" speakers, while CPP CoV contributed for "average" speakers. The second dimension is related primarily to moving averages of F_2 and FD. Telling voices apart in the "easy" and "hard" subsets also depended on F_1. Similarity for "average" speakers was also related to

$H_4^* - H_{2k}^*$, and "average" speakers were more similar to one another along the second dimension (7.83°) compared to "easy" and "hard" subsets (12.83° and 9.43°, respectively). These results suggest that the means of formant frequencies contribute little to making voices "easy" or "hard" to distinguish in a "different speaker" task. Dimension 7 describes the majority of the between-speaker variability across subsets, with mean angular separations of $28.38°, 31.49°$, and 29.75° for "easy," "average," and "hard" speakers, respectively. Speakers who are "easy" to tell apart differed from each other primarily in F_0 CoV, followed by CoVs of $H_4^* - H_{2k}^*$ and $H_{2k}^* - H_{5k}$. In comparison, "average" speakers varied almost equally in terms of F_0, F_2 CoV, and FD CoV, while most variation in "hard" speakers was attributable to F_0, followed by smaller contributions from $H_1^* - H_2^*$ and $Energy$. In other words, for the "different speaker" task, discrimination is the easiest for talkers whose speech acoustics are mainly separated by the three CoVs (F_0 CoV, $H_4^* - H_{2k}^*$ CoV and $H_{2k}^* - H_{5k}CoV$), is less easy for "average" talkers whose acoustics differed mainly in mean F_0 and two formant-variable CoVs (extent of variability in relation to the average), and is the hardest for talkers whose speech is distinguished only by moving averages.

3.3 Discussion

In this chapter we examined the effects of moderate speaking style variations (read speech versus casual conversations) and of within- versus between-speaker acoustic variability on human speaker discrimination performance. The stimuli comprised short text-independent utterances from speakers who were not familiar to the listeners.

The first objective of this work was to identify the effects of speaking style vari-

ations on human speaker discrimination performance. Listeners performed better in style-matched cases (EER = 6.96% when both stimuli were read sentences and EER = 15.12% when both stimuli were excerpts from conversations) than in the style-mismatched case (EER = 20.68%). Moderate speaking style variations affected speaker discrimination performance when stimuli were style-mismatched and also when they were style-matched i.e., read speech – read speech trials were easier than conversation – conversation trials.

In comparison to our previous findings based on read and pet-directed speech from the same speakers [PYV18] the performance gap between the style-matched and style-mismatched conditions appears to depend at least partly on the extent of the mismatch (moderate in the present study and extreme in the previous study). For example, the EER in [PYV18] for the style-matched read speech – read speech condition was 19.02%, while for the style-mismatched condition it was 39.23%, versus 6.96% and 20.68%, respectively, in the present study. Note that, the sampling rate was higher in the present study than in our previous work (22 kHz, versus 8 kHz in [PYV18]).

Another objective of this research was to determine the differences in how speaking style variations affect "same speaker" and "different speaker" trials. The speaker-level log-likelihood ratio distribution (see Figure 3.2) skewed heavily toward the positive region with small variance in "same speaker" trials, indicating that listeners were more accurate and more confident in the "same speaker" trials. Confidence on these trials was highest for read speech – read speech and worst for read speech – conversation; this pattern did not occur for "different speaker" trials. The changes in listeners' confidence in "same speaker" trials seem to follow the same pattern as did overall performance. Listeners were highly confident for the style-matched

read speech trials, but confidence decreased substantially for the other two conditions. Taken together, these results are consistent with our hypotheses that the "same speaker" task largely relies on within-speaker variability, and that moderate style variations impact human performance. However, no such confidence differences arose from style variability in the "different speaker" trials. This suggests that between-speaker variability in the "different speaker" task has greater influence on human performance compared to the effects of moderate speaking style variability.

We also found that which voices listeners judged most accurately depended not only on the voices but also on the task: "telling speakers together" was easier for some voices, while "telling speakers apart" was easier for others. This suggests that listeners rely on different acoustic information when performing these two tasks. In the "same speaker" task, the "easy" speakers varied widely along F_2 CoV, $Energy$, and its CoV, while "average" speakers varied the most along CoVs of F_0, F_2, and FD. Finally, "hard" speakers varied mainly along F_3, $Energy$, and its CoV in this task. The features that made the "same speaker" task easier (F_2 CoV, $Energy$, and its CoV) were the ones that appeared in later dimensions (dimension 3 or higher), i.e., the ones that contributed to speaker idiosyncrasies in Lee et al.'s [LKK19, LK19] acoustic voice space model. This further suggests that listeners rely on speaker idiosyncrasies for the "same speaker" task. Note that in this task, CoVs of formants (F_2 CoV for "easy" speakers and CoVs of F_2 and FD for "average" speakers) played a critical role in assisting listeners in "telling speakers together." Forensic studies [McD04] argue that formant frequency variations have relevant speaker identification information as they are determined not only by the shape and size of the vocal tract but also by the speaker's style of configuring articulators for speech.

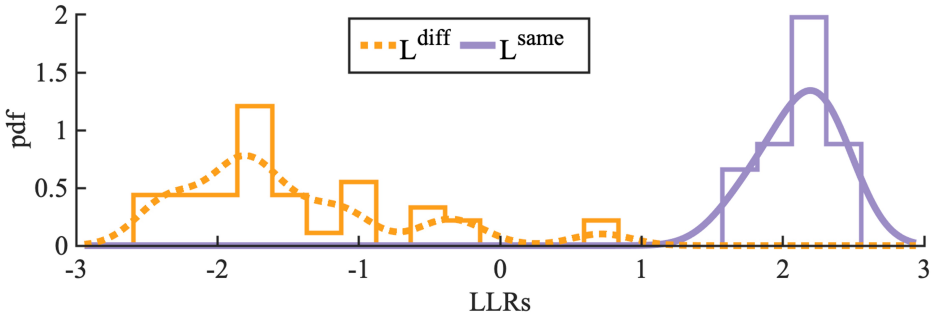
In the "different speaker" task, "easy" speakers differed in the CoVs of ampli-

tude differences of the higher harmonics ($H_4^* - H_{2k}^*, H_{2k}^* - H_{5k}$) and F_0. These were some of the variability features that described the shared acoustic structure across speakers in Lee et al. [LKK19, LK19]. These results provide further evidence in support of our hypothesis that the distance along the shared acoustic structure is critical for speaker discrimination in the "different speaker" task. Voices that differed in moving average of acoustic properties, including a combination of mean F_0, lower harmonic amplitudes ($H_1^* - H_2^*$), and energy, were difficult for listeners to distinguish. Moreover, average voices were distinguished by both moving average and variability (CoVs) features, implying that variations between speakers along moving average of acoustic properties could be insufficient for listeners to tell them apart, while variations along feature CoVs assisted listeners in this task. In general, the measures characterizing hard-to-distinguish voices are known to be important for speaker characterization (e.g., fundamental frequency and $H_1^* - H_2^*$ correlate with perceived breathiness [WJ03]), but challenges arise in this task given that it involves female-only comparisons. In a female-only comparison, there are smaller variations in F_0 and smaller influence of nasality on $H_1^* - H_2^*$ [Sim12].

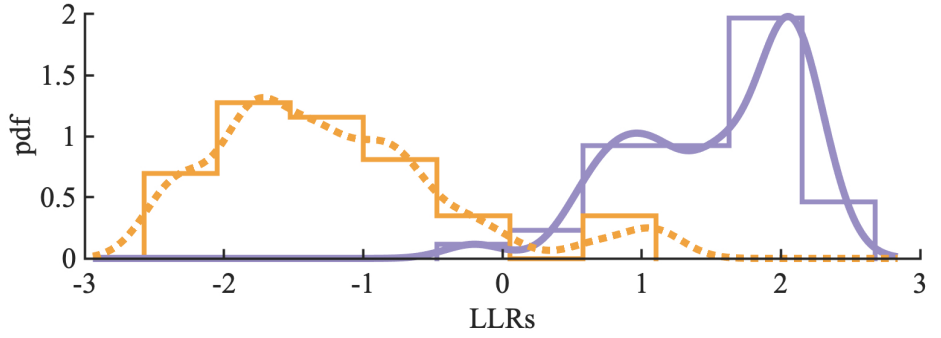
In summary, it seems that listeners find it easier to "tell speakers together" using speaker-specific idiosyncrasies, i.e., we can best explain the performance on the "same speaker" task by the nature and extent of within-speaker variability. In contrast, listeners "tell speakers apart" based on differences in features (alternatively, relative positions) within a shared acoustic structure rather than speaker-specific features. This implies that "telling speakers apart" relies more on the nature and extent of between-speaker variability as the differences here are across acoustic features representing shared variability. Therefore, it should be possible to perform acoustic-based predictions of which voices will be "easy" or "hard" to "tell apart" using the

relative positions in the shared acoustic structure. However, similar acoustic-based predictions about "telling together" different samples of a speaker's voice might be challenging, as this would require finding the speaker-specific idiosyncrasies.

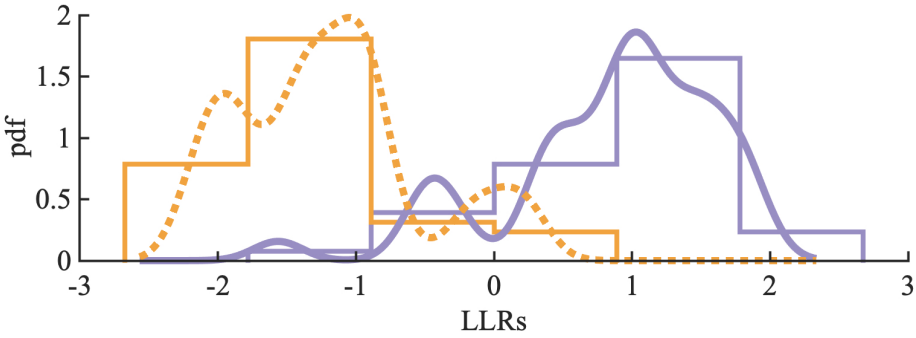
One limitation of this work must be noted. The perception experiments used a homogenous panel of listeners (22 female out of 30 listeners with an age range of 17-21 years old). Hence, these findings may not fully generalize to other populations. The results presented nevertheless provide a means of investigating the question of the effects of moderate style-variability on speaker discrimination performance. In the future, a heterogeneous population will be used for the listeners' panel.



(a) read speech – read speech



(b) conversational speech – conversational speech



(c) read speech – conversational speech

Figure 3.2: Distributions as kernel density plots overlaid onto histograms of speaker-level log-likelihood-ratios (LLRs) for "same speaker" (L^{same}) and "different speaker" (L^{diff}) trials represented as probability density functions. L^{same} and L^{diff} are denoted with solid ('–') and dotted ('..') lines, respectively.

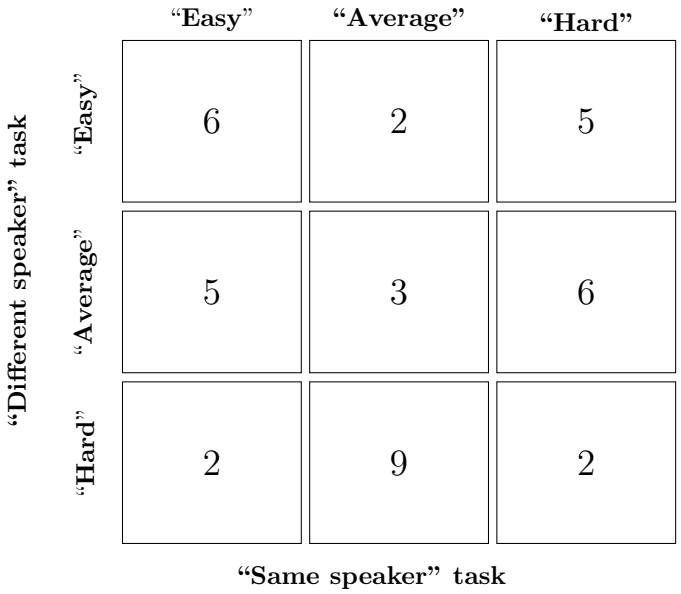


Figure 3.3: The number of speakers that were "easy" versus "average" or "hard," as indexed by overall accuracy, for "different speaker" versus "same speaker" tasks. Columns show the number of speakers who were easy, average, or hard to "tell together" on the "same speaker" trials, while rows show how difficult the same voices were to "tell apart" on the "different speaker" trials.

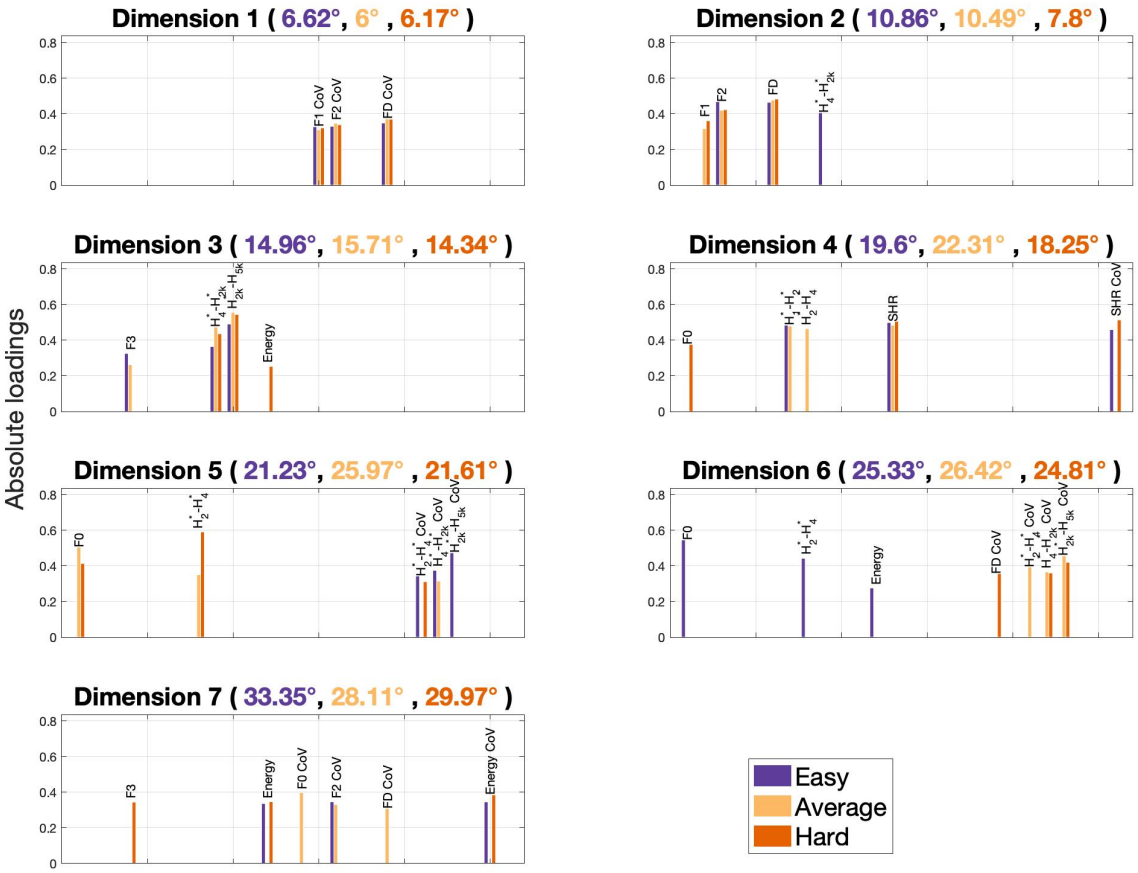


Figure 3.4: For the "same speaker" task, the absolute loadings/coefficients of the directions that are closest to the principal components of all speakers in a subset versus acoustic features. The mean angular separation between groups and each direction is shown above each subplot. The features are represented along the x-axis. F_0: fundamental frequency, F_1, F_2, F_3, F_4: the first four formants, CPP: cepstral peak prominence, $H_1^* - H_2^*$, $H_2^* - H_4^*$, $H_4^* - H_{2k}^*$, and $H_{2k}^* - H_{5k}$: the amplitude differences of the harmonics, FD: formant dispersion, SHR: subharmonics to harmonics ratio, CoV: coefficient of variation.

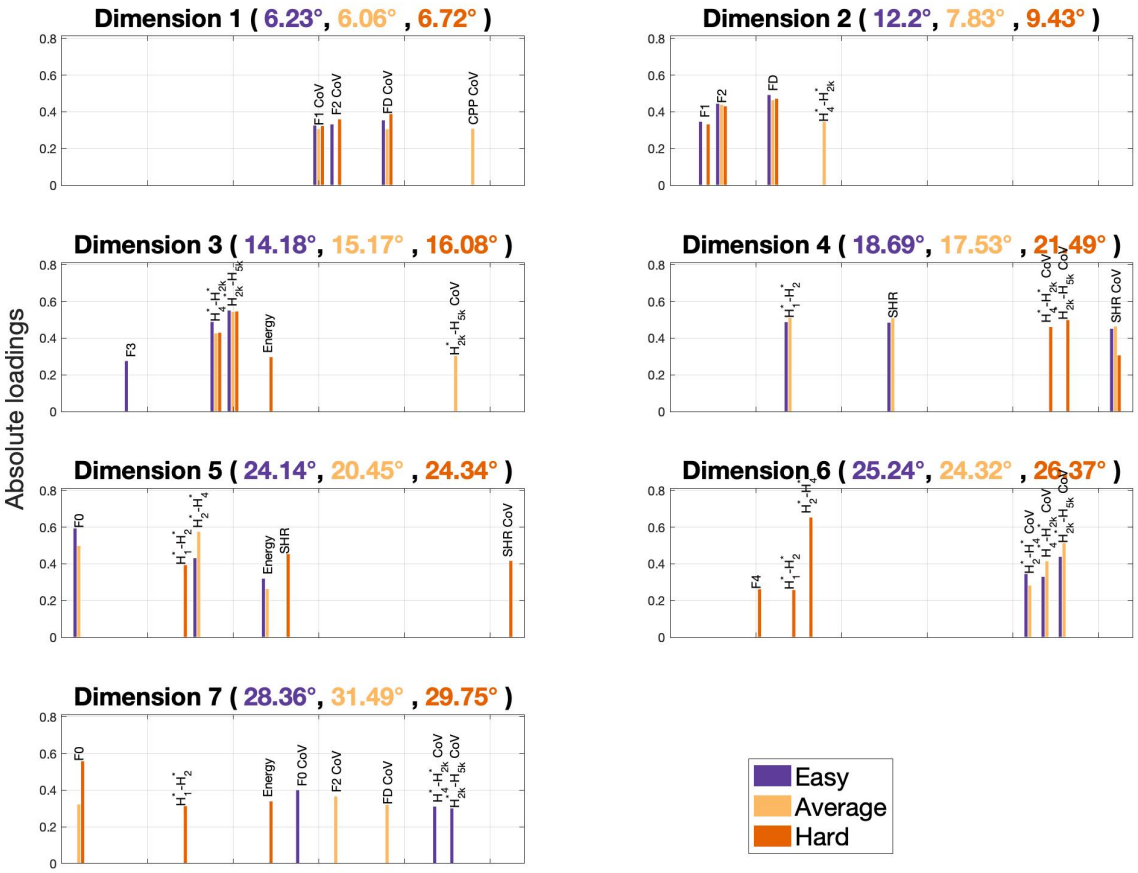


Figure 3.5: For the "different speaker" task, the absolute loadings/coefficients of the directions that are closest to the principal components of all speakers in a subset versus acoustic features. The mean angular separation between groups and each direction is shown above each subplot. The features are represented along the x-axis. F_0: fundamental frequency, F_1, F_2, F_3, F_4: the first four formants, CPP: cepstral peak prominence, $H_1^* - H_2^*$, $H_2^* - H_4^*$, $H_4^* - H_{2k}^*$, and $H_{2k}^* - H_{5k}$: the amplitude differences of the harmonics, FD: formant dispersion, SHR: subharmonics to harmonics ratio, CoV: coefficient of variation.

CHAPTER 4

Style-robust speaker verification systems

In this chapter, the effects of speaking-style variability on automatic speaker verification were investigated using the UCLA Speaker Variability database which comprises multiple speaking styles per speaker. The performance was better when enrollment and test utterances were of the same style, but it decreased substantially when styles were mismatched between enrollment and test utterances. This chapter focuses on reducing the effect of style mismatch when multiple styles per speaker are not available for training the speaker verification systems.

We hypothesize that entropy can capture acoustic variability due to style variations. Hence, an entropy-based variable frame rate (VFR) technique was proposed. Using the VFR technique two approaches for style-robust speaker verification systems are developed: (i) artificially generating style-variant representations for PLDA adaptation, and (ii) extracting speaker embeddings which are robust to speaking style variations using attention-based conditioning. Parts of this chapter were published in [AGP20].

4.1 Variable frame rate-based data augmentation

4.1.1 Method

The Kaldi [PGB11] SRE16 recipe was used to develop an x-vector/PLDA ASV system [SGS18]. The input acoustic features were 23-dimensional mel-frequency cepstral coefficients (MFCCs) with a frame length of 25 ms and a frame shift of 10 ms, which were mean normalized over a sliding window of up to 3 secs. Standard extrinsic data augmentation (as in the recipe) was applied on the training data for both x-vector and PLDA.

A widely-used strategy to attenuate within-speaker variability is to train the PLDA with data for the conditions of variability from each speaker [GMS14, GZE12]. Although this strategy has been mainly used for external sources of variability (e.g, noise, channel, etc.) [SGS18, GZE12], it could be also applied to deal with speaking style variability. However, a sufficient amount of data is not available in the UCLA database to train a robust PLDA in this manner. Therefore, a PLDA model was trained with the training data and the in-domain adaptation (using the version provided in Kaldi) was performed with the UCLA database. The experimental configurations for adaptation will be described in Section 4.1.2.2.

In cases when multiple speaking styles per talker are not available in the training dataset, a method to artificially generate speaking style-variant representations for augmentation is required. We propose to use the entropy-based variable frame rate to generate such variants. The differences across speaking styles can be broadly categorized into rhythmic variations, including speech rate, long pauses, changes in the duration of individual sounds, boundary articulation, and prosodic variations.

However, the latter is directly associated with speaker identity and disentangling prosody from speech will result in performance degradation. Hence, in this work, we focus on addressing the effects of rhythmic variations between styles. Specifically, we propose to generate style-variant speaker representations by applying the entropy-based variable frame rate approach [YZA04].

4.1.1.1 Entropy Computation

Consider a random variable $\nu \in \mathcal{R}^K$ where $p(\nu)$, the probability distribution function (PDF) of ν is a K-dimensional Gaussian. Let μ and Σ be the mean and covariance matrix of the random variable. The entropy can be calculated as:

$$\begin{aligned} H(\nu) &= -\int p(\nu) \ln p(\nu) d\nu \\ &= -\int p(\nu) \left[-\frac{1}{2}(\nu-\mu)^T \Sigma^{-1} (\nu-\mu) - \ln |2\pi\Sigma|^{\frac{1}{2}} \right] d\nu \\ &= \frac{K}{2} + \frac{1}{2} \ln |2\pi\Sigma| \end{aligned} \tag{4.1}$$

To facilitate faster computation and to avoid an ill-posed problem when the random variable's covariance matrix is not full rank, the following approximation is used to calculate the entropy [YZA04]:

$$H(\nu) \approx K \ln \sqrt{2\pi} + \ln \operatorname{Tr} \Sigma \tag{4.2}$$

4.1.1.2 Implementation

The variable frame rate approach dynamically changes the frame rate based on between-frame entropy using the steps shown in Figure 4.1. First, a signal is win-

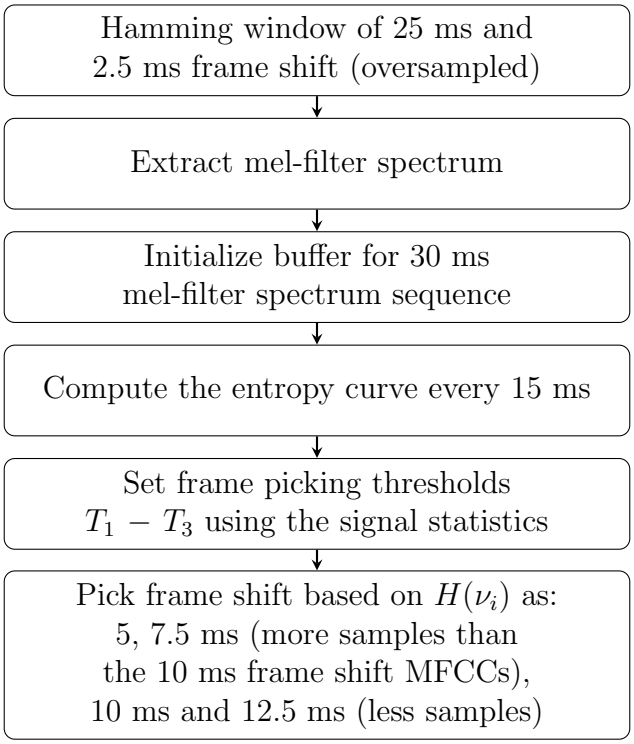


Figure 4.1: Overview of the entropy-based variable frame rate approach.

dowed using 25 ms Hamming window by first sampling with frame shift of 2.5 ms, a much lower value than the widely-used 10 ms frame shift. With these densely sampled, or "oversampled" frames, varying the frame rate becomes a simple task of retaining frames selectively. Mel-filter spectra are then computed. The frames spanning a duration of 30 ms are then used to calculate the entropy curve using the local entropy every 15 ms. VFR was carried out by comparing the signal's entropy to certain thresholds in order to calculate the frame picking rate in the extraction of MFCCs. Using the entropy curve of the speech signal $H(\nu_i)$, $i = 1, ..., N$, the frame-picking thresholds T_1, T_2, T_3 are set as in Equation 4.3.

$$\begin{cases} T_1 &= \omega_1 M_{max} + (1-\omega_1) M_{med} \\ T_2 &= (1-\omega_2) M_{max} + \omega_2 M_{med} \\ T_3 &= (1-\omega_3) M_{med} + \omega_3 M_{min}, \end{cases} \tag{4.3}$$

where ω_1, ω_2, and ω_3 are weighting parameters of values 0.7, 0.8, and 0.5, respectively. M_{max}, M_{med}, and M_{min}, are the maximum, median, and minimum of the entropy curve, respectively. The values of the weighting parameters were chosen empirically. In this implementation, the x-vector extractor is trained using a frame shift of 10 ms. Hence, frame rates of 5 ms ($H(\nu_i) \geq T_1$) and 7.5 ms ($T_1 > H(\nu_i) \geq T_2$) are used to obtain more frames from the regions where the signal has rapid changes of information. A 10 ms frame shift is used when entropy is close to average ($T_2 > H(\nu_i) \geq T_3$). Whereas the frame rate is 12.5 ms ($T_3 > H(\nu_i)$) when the signal has low information gain, so that we obtain lesser frames from the region.

Recall that we aimed at reducing the effects of some of the key differences across speaking styles such as speech rate, long pauses, changes in the duration of individual sounds, and boundary articulation. These variations can be captured by between-frame entropy. For instance, fast speech rate, short pause, or incomplete word final stops (/b/, /t/) can lead to a rapid change of information in spectral characteristics between frames resulting in a high between-frame entropy. On the other hand, a decrease in speech rate, long pauses or an increase in the duration of an individual sound could result in a low between-frame entropy. Hence, we hypothesize that extracting features by changing between-frame entropy could, in-turn, result in generating different style variants and hence compensate for the effects of spectral and temporal variations from style variability through augmentation.

Based on this hypothesis, VFR was used to generate style-variant utterance representations. This approach is expected to be more robust than varying the speaking rate of the entire utterance because the variations within an utterance and within speaking style are not always uniform and can vary based on speaker characteristics, context of the conversation, emotion, interlocutor, etc [LBS19].

4.1.2 Experimental Setup

4.1.2.1 Database Statistics

A randomly selected subset of 50 female and 50 male speakers from the UCLA database was set aside as the "development set". The remaining subset of 50 female and 50 male speakers was used as the "evaluation set". The evaluation set was further split into "enrollment" and "test" set.

In order to analyze the effect of style variability on system performance, the effect of phonetic variability across utterances needs to be negligible. Based on studies [HSH13] reporting that 30-sec utterances cover enough phonetic variability to capture speaker-specific information, 30-sec long speech samples were used both for enrollment and test utterances. Table 4.1 shows the number of speech samples from the UCLA database used in this experiment. Note that at least 1 min of speech is required per speaker to generate style-matched enrollment – test utterance pairs. Because the majority of speakers in the UCLA database did not have enough speech in the narrative and pet-directed speaking styles, style-matched conditions for those styles were omitted. This resulted in 23 different evaluation combinations. All possible trials were generated for the five styles, which resulted in more non-target trials than target trials. The recordings from the UCLA database were downsampled

to 8 kHz during the ASV experiments to match the sampling rate of NIST SRE and Switchboard databases that were used for training the ASV systems.

Table 4.1: Number of utterances distributed across each set used in VFR data augmentation for the UCLA database.

Style	**read**	**instructions**	**narrative**	**conversation**	**pet-directed**
Development	196	100	36	184	19
Enroll	102	104	35^{+}	99	16^{+}
Test	101	104	35^{+}	88	16^{+}

$^{+}$ Same enroll and test utterances.

4.1.2.2 PLDA Adaptation Configurations

The PLDA trained on SRE and Switchboard data is adapted using the development set from the UCLA database. Recall that the major focus here is data augmentation using VFR for PLDA adaptation. Hence, we designed five different adaptation configurations to experimentally analyze the advantages of the proposed technique:

Baseline: In-domain data with a single speaking style, the same as that of the enrollment set, is used (development set size X).

Extrinsic augmentation: Extrinsic variability is added using artificial data augmentation (development set size $5X$). The implementation here is similar to the one in x-vector training [SGS18], but we use all the extrinsic variants and not a subset. We add music, noise and babble from the MUSAN corpus [SCP15] and reverb by convolving with simulated room impulse responses [KPP17].

VFR: Entropy-based VFR is applied to the development data (development set size X). This generates style-variant development set.

[Proposed] VFR augmentation: Both the original representations of the development data and their style-variant counterparts, obtained by performing VFR, were used (development set size $2X$)[1].

Multi-style: Multiple speaking styles from the in-domain data were used (development set size $4X$).

In the baseline, extrinsic augmentation, VFR, and VFR augmentation configurations, the speaking style used in the development set matched that of the enrollment utterances. For instance, when enrolling with *read* and testing with other styles, the development set for PLDA adaptation contained only *read* sentences. In contrast, all styles in the development set were used in the multi-style configuration.

The baseline configuration was used to assess the effects of speaking style variability on ASV performance, as well as to establish baseline performance to be compared with the other configurations.

The extrinsic augmentation configuration represents standard techniques [SCP15, KPP17] that increase the amount of data, and it was used to understand how the proposed VFR data augmentation performs in comparison. The VFR configuration was used to analyze the effectiveness of style-variants with the VFR approach and also to assess if style-variants alone would be enough to compensate for style variability. Note that the multi-style configuration is the best-case scenario, but it is not realistic to assume that one can obtain all speaking styles for each speaker.

[1]We experimented with VFR augmentation of size $3X$, $5X$, and $7X$ and did not find a significant improvement over $2X$.

4.1.3 Results and Discussion

System performance in terms of the EER for the UCLA database is shown in Table 4.2. Statistical significance was verified using McNemar's test [McN47]. Unless mentioned explicitly, all performance differences reported in this section are significant with $p < 0.05$.

In the baseline, a style-mismatch between enrollment and test utterances consistently degraded ASV performance compared to their style-matched task. For instance, when enrolled with conversational speech, the style-matched task (conversation – conversation) had an EER of 0.57%. The performance degraded for style-mismatched tasks resulting in EERs of 3.03%, 3.24%, 2.96%, and 22.12% for conversation – read, conversation – instructions, conversation – narrative, and conversation – pet-directed pairs, respectively.

The second configuration of extrinsic augmentation performed slightly better than the baseline in 9/23 tasks, especially for pet-directed speech which had fewer utterances for adaptation and hence, the increase in the amount of data from augmentation could explain the improvement. On the other hand, the extrinsic augmentation performed worse than the baseline in 6/23 tasks. Interestingly, these were the tasks with reading or conversational speech as the development set. These styles had more utterances than others. The standard augmentation techniques used in the *extrinsic augmentation* setup merely increased the amount of data and might not have been sufficient to address style-variability.

VFR was better than the baseline in 10/23 tasks, the same in 6/23 tasks, and worse in 7/23 tasks. This inconsistency in performance gains between the two setups may be due to: (i) the style variant from VFR only partially addressed style variabil-

Table 4.2: Performance in terms of EER (%) on the UCLA database. In the baseline, extrinsic augmentation, VFR, and VFR augmentation configurations, the speaking style used in the development set matched that of the enrollment utterances. All styles in the development set were used in the multi-style configuration. The best result in each condition with improvements over other configurations is boldfaced. If denoted by a '*' the difference from the baseline is not statistically significant.

	Development and Enroll style	Test style				
		read	**instructions**	**conversation**	**narrative**	**pet-directed**
Baseline	**read**	0.98	1.91	**2.25**	2.20	15.87
	instructions	2.89	0.09	4.67	1.82	23.53
	conversation	3.03	3.24	0.57	2.96	22.12
	narrative	0.63	0.91	1.09	NA	11.76
	pet-directed	18.75	18.24	**10.00**	14.57	NA
Extrinsic aug.	**read**	0.98	1.91	3.37	1.89	**12.50**
	instructions	3.85	0.22	**3.30**	1.52	23.53
	conversation	4.04	3.92	1.14	2.70	18.75
	narrative	0.63	0.91	1.09	NA	11.76
	pet-directed	**12.50**	17.65	**10.00**	13.73	NA
VFR	**read**	0.98	1.47	3.37	1.89	18.75
	instructions	**1.92**	0.08*	**3.30**	**1.22**	23.53
	conversation	3.03	4.90	1.14	**2.27**	18.75
	narrative	**0.48**	0.91	1.09	NA	11.76
	pet-directed	**12.50**	17.65	13.33	15.69	NA
VFR aug.	**read**	0.98	1.91	2.62	1.29	**12.50**
	instructions	**1.92**	**0.07***	**3.30**	1.52	23.53
	conversation	2.69	3.27*	**0.38**	**2.27**	18.75
	narrative	0.63	0.91	**0.55**	NA	11.76
	pet-directed	**12.50**	15.20	14.44	**12.64**	NA
	Enroll style	**Test style**				
		read	**instructions**	**conversation**	**narrative**	**pet-directed**
Multi-style	**read**	0.98	**0.95**	**2.25**	**1.26**	**12.50**
	instructions	**1.92**	0.22	4.40	1.52	**18.82**
	conversation	**2.02**	**2.94**	1.14	**2.27**	**12.50**
	narrative	0.63	0.91	0.73	NA	11.76
	pet-directed	**12.50**	**13.24**	13.33	15.59	NA

ity (ii) the VFR variant was only applied to development data and not to enrollment and test data. We did not apply VFR to enroll and test utterances because it would result in the loss of speaker-specific information.

The proposed approach of entropy-based VFR augmentation performed better

than the baseline in 13/23 tasks. The most notable improvement was seen when the testing was on pet-directed speech (read – pet-directed and conversation – pet-directed) which is often characterized by exaggerated prosody. However, for two tasks, read – conversation and pet-directed – conversation, the proposed approach did not improve the results compared to the baseline.

When compared to VFR, the proposed approach showed significant improvement in 10/23 tasks. The performances were the same in 9/23 tasks. There was a degradation in performance of the proposed approach for 4/23 tasks.

The proposed approach was better than extrinsic augmentation in 11/23 tasks and the same in 11/23 tasks. The proposed approach was generally better even if it used less data than extrinsic augmentation. This result verifies the hypothesis that VFR, in fact, improved the ASV performance by providing style-variant utterance representations and not by simply increasing the number of samples seen by the PLDA classifier. However, in the pet-directed – conversation task the proposed approach was worse than using extrinsic augmentation.

The multi-style configuration had more style information available in the development set as compared to the proposed approach (VFR aug.), still, their performances were comparable. Their performances were the same in 9/23 tasks, 5/23 tasks VFR aug. was better, and multi-style was better in 9/23 tasks. These findings support the hypothesis that VFR methods can be used as a data-augmentation technique when multi-style data are limited. One of the tasks where VFR aug. was better than multi-style was a style-matched task of conversation – conversation. There are probably variations within a speaking style that could be compensated by the style-variant augmentation approach.

4.2 Attention-based conditioning

Section 4.1 addressed the issue of speaking style variability by using an entropy-based variable frame rate (VFR) technique to perform data augmentation. Entropy inherently captures spectral and temporal variations across different styles [AGP20, RPA19]. Thus, by applying VFR, style-variant speaker representations were obtained. However, that work involved data augmentation on the PLDA backend. In contrast, this section focuses on the embedding extractor instead of the PLDA backend to make the ASV system robust to speaking style variations.

Typically, speaker embedding extraction includes training a deep neural network for speaker classification and using the bottleneck features as speaker representations. Such a network has a pooling layer to transform frame-level features to utterance-level by calculating statistics over the frames, weighing them equally. On the other hand, self-attentive embeddings perform weighted pooling such that the weights correspond to the importance of the frames in a speaker classification task. This work aims to extract speaker embeddings robust to style variations. Entropy can capture acoustic variability due to such style variations. Hence, an entropy-based variable frame rate output is proposed as an external conditioning vector for the self-attention layer to provide the network with information to compensate style effects.

4.2.1 Method

The proposed method includes self-attentive statistical pooling with an entropy-based VFR conditioning for style-robust speaker verification. This approach uses an x-vector/PLDA framework.

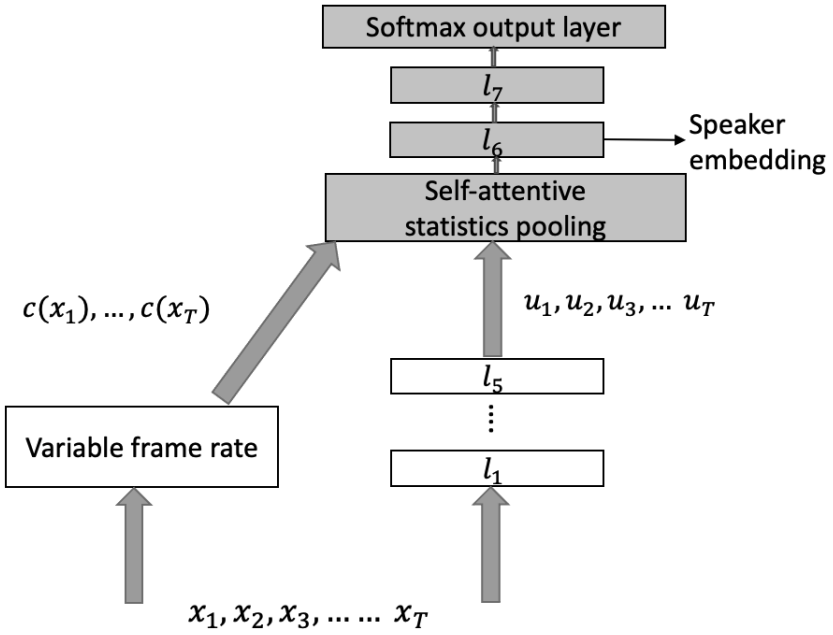


Figure 4.2: Self-attentive statistics pooling with VFR conditioning.

The inputs to the x-vector/PLDA system [SGS18] are 30-dimensional mel-frequency cepstral coefficients (MFCCs) using a 25 ms frame length and a 10 ms frame shift. The MFCCs are mean normalized over a sliding window of up to 3 secs. Extrinsic data augmentation of noise and reverberation [SGS18] was applied to the training data.

4.2.1.1 Network architecture

The network architecture of the proposed method is shown in Figure 4.2. It builds upon the network structure from x-vectors. Layers l_1 to l_5 operate at the frame-level, with a small temporal context centerd at the current frame t.

l_1 operates on frames $(t-2)$ to $(t+2)$, followed by l_2 which operates on the output l_1 at time steps $\{t-2, t, t+2\}$ and finally l_3 operates on the output of l_2 at

time steps $\{t-3, t, t+3\}$. Layers l_4 and l_5 do not add temporal contexts, resulting in a total temporal context of fifteen frames. The pooling layer uses self-attention with conditioning vector, $c(\mathbf{x}_t)$ providing weighted statistics. The output of pooling is propagated to the fully connected layers l_6 and l_7 and to the softmax output layer. The network is trained to classify speakers using cross-entropy. ReLUs are used as non-linearities. The output of the affine component of l_6 is used as speaker embedding and sent to the PLDA backend.

4.2.1.2 Self-attentive pooling

As described earlier, self-attentive pooling learns weights to maximize the speaker classification performance during training, resulting in better speaker representations. Let the input to the pooling layer from hidden layer l_5 at frame t be $\mathbf{u}_t$. Self-attention [CDL16] calculates attention scores α_t for each frame providing us the weighted average ($\tilde{\mu}$) and the weighted standard deviation ($\tilde{\sigma}$) of $\mathbf{u}_t$:

$$\alpha_t = \text{softmax}(\mathbf{W}_2^T f(\mathbf{W}_1 \mathbf{u}_t + \mathbf{b}_1) + b_2) \tag{4.4}$$

$$\tilde{\mu} = \sum_{t=1}^{T} \alpha_t \mathbf{u}_t \tag{4.5}$$

$$\tilde{\sigma} = \sqrt{\sum_{t=1}^{T} \alpha_t \mathbf{u}_t \odot \mathbf{u}_t - \tilde{\mu} \odot \tilde{\mu}} \tag{4.6}$$

where $\odot$ is the Hadamard product; $\mathbf{W}_2$ and $\mathbf{W}_1$ are the weight matrices for the attention layer and $\mathbf{b}_1$ and b_2 are biases for attention; $f(.)$ is a non-linear activation function, a sigmoid in this case.

4.2.1.3 External conditioning: Variable frame rate

The entropy-based VFR discussed in Section 4.1.1.1 is used to extract a conditioning vector for the self-attention network. Using the output from the thresholding stage in Figure 4.1, we create a vector $z(\mathbf{x})$ composed of 1's and 0's where 1 indicates that the frame is to be picked and 0 indicate that the frame is to be skipped. We compare the entropy, $H(\nu_i)$ with thresholds from Equation 4.3 and pick every r^{th} frame from $z(\mathbf{x})$ where r is a multiple of the 2.5 ms frame shift:

$$r = \begin{cases} 2, & \text{if } H(\nu_i) \geq T_1 \\ 3, & \text{if } T_1 > H(\nu_i) \geq T_2 \\ 4, & \text{if } T_2 > H(\nu_i) \geq T_3 \\ 5, & \text{if } T_3 > H(\nu_i) \end{cases} \tag{4.7}$$

The number of frames set to 1 is proportional to entropy. When the entropy is high, more frames are selected, and fewer frames are selected when the entropy is low. Thus, equalizing the entropy across the utterance. This "oversampled" (4 times that of MFCCs) vector $z(\mathbf{x})$ is reduced by calculating a sum over every 4 frames to obtain the conditioning vector $c(\mathbf{x})$,

$$c(\mathbf{x}_{4i}) = \sum_{j=1}^{4} z(\mathbf{x}_{4i+j}) \ , \quad i = 0, 1, \ldots, \frac{N}{4} \tag{4.8}$$

Recall, that we focus on compensating for the rhythmic variations between styles i.e, speech rate, long pauses, changes in the duration of individual sounds and boundary articulation. As discussed in Section 4.1.1.2, these variations can be captured by

between-frame entropy. Hence, we hypothesize that an entropy-based conditioning vector may implicitly represent spectral and temporal variations in style and thereby provide self-attention with information to compensate for style effects.

4.2.1.4 Conditional Attention

As mentioned earlier, VFR is used as a conditioning vector for the speaker embeddings, $c(\mathbf{x}_t)$ in self-attentive pooling by updating $f(\mathbf{u}_t)$ in Equation 4.4 with $f(\mathbf{u}_t, c(\mathbf{x}_t))$. There are multiple possibilities for adding the conditioning vector and three such methods are explored: concatenation, gating, and affine transformation [MBP19].

Conditioning by concatenation: The conditioning vector is concatenated with the output of l_5, adding extra dimensions to $\mathbf{u}_t$. These new dimensions carry information about the signal's entropy. $||$ indicates concatenation, $\mathbf{W}_c$ is the weight matrix, and $\mathbf{b}_c$ is the bias vector. Hence, updating self-attention as:

$$f_c(\mathbf{u}_t, c(\mathbf{x}_t)) = tanh(\mathbf{W}_c[\mathbf{u}_t||c(\mathbf{x}_t)] + \mathbf{b}_c) \tag{4.9}$$

Conditioning by gating: A gating mechanism is used to learn a feature mask from $c(\mathbf{x}_t)$ and apply it to $\mathbf{u}_t$ the output of the hidden layer (l_5) before pooling, $\mathbf{u}_t$, $t = 1, \ldots, T$. A sigmoid is used for the mask to generate values between 0 and 1. As the VFR output conditions gating, frames are selected for pooling based on signal entropy. $\mathbf{W}_g$ is the weight matrix and $\mathbf{b}_g$ is the bias vector. Hence, updating

self-attention as:

$$f_g(\mathbf{u}_t, c(\mathbf{x}_t)) = \sigma(\mathbf{W}_g(c(\mathbf{x}_t) + \mathbf{b}_g) \odot \mathbf{u}_t \tag{4.10}$$

Conditioning using affine transformation: An affine transformation is applied on the hidden layer (l_5) output, $\mathbf{u}_t$ by using the conditional vector to calculate scaling $\gamma(.)$ and shifting $\beta(.)$. $\mathbf{W}_\gamma, \mathbf{W}_\beta$ are the weight matrices and $\mathbf{b}_\gamma, \mathbf{b}_\beta$ are the bias vectors for the transformation. Hence, updating self-attention as:

$$f_a(\mathbf{u}_t, c(\mathbf{x}_t)) = \gamma(c(\mathbf{x}_t)) \odot \mathbf{u}_t + \beta(c(\mathbf{x}_t)) \tag{4.11}$$

$$\gamma(\mathbf{x}) = \mathbf{W}_\gamma \mathbf{x} + \mathbf{b}_\gamma, \quad \beta(\mathbf{x}) = \mathbf{W}_\beta \mathbf{x} + \mathbf{b}_\beta \tag{4.12}$$

Two additional methods are studied: (A) concatenation in combination with gating, and (B) concatenation in combination with affine. Gating is a special case of affine transformation with shifting factor, $\beta = 0$ and scaling factor, $\gamma \in [0, 1]$. Hence, those two methods are not combined.

4.2.2 Experimental Setup

As is commonly used, the embedding extractor had 512 nodes in each of l_1 to l_4, l_5 had 1500 nodes, while l_6 and l_7 had 512 nodes. The self-attention layer had 500 hidden nodes. The experiments were setup using Pytorch [PGM19] and Kaldi [PGB11]. The optimizer was Adam [KB17] with a batch size of 128 and trained for 100 epochs. We train the embedding extractor using the Voxceleb2 dev set.

4.2.2.1 Database Statistics

To effectively evaluate style-robustness, we require negligible effect from phonetic variability. As discussed in Section 4.1.2.1, we use 30 sec long speech samples to cover enough phonetic variability and capture speaker idiosyncratic information. Hence, 1,838 30 sec segments are extracted and the statistics of the subset used are in Table 4.1. We require a minimum of 1 min of speech per speaker to generate style-matched trials. However, the majority of the speakers had less than 1 min of speech for pet-directed speech and affect-matched narrative case. So the style-matched tasks for those two styles were omitted. Thus resulting in 23 tasks (5 styles in matched and mismatched conditions except for the style-matched pet-directed and narrative cases). To match the sampling rate of the other databases used in this set of experiments, the data were downsampled to 16 kHz.

Table 4.3: Number of utterances distributed across each set used in VFR conditioning of the UCLA database.

Style	**read**	**instructions**	**narrative**	**conversation**	**pet-directed**
Enroll	200	204	625^{+}	197	35^{+}
Test	199	204	625^{+}	174	35^{+}

$^{+}$ Same enroll and test utterances.

4.2.3 Results and Discussion

4.2.3.1 UCLA SVD Evaluation

Table 4.4 shows the equal error rate (EER) over all style trial combinations for the UCLA SVD data. The best result in each condition with improvement over the baseline is boldfaced. If denoted by a '*' it is not statistically significant improvement over

Table 4.4: Performance using the UCLA database (EER %). The best result in each condition with improvement over the baseline is boldfaced. If denoted by a '*' it is not a statistically significant improvement over the baseline. Combined A (concatenation with gating) and Combined B (concatenation with affine).

Enroll	Test	x-vector	VFR weights	Self-attention	Concatenation	Gating	Affine	CombinedA	CombinedB
read	read	0.50	1.00	0.50	0.50	0.50	0.50	0.50	0.50
	instructions	0.49	2.44	0.49	0.49	0.49	0.49	0.49	0.49
	conversation	2.86	6.86	**2.29**	**2.29**	2.86	**2.29**	**2.29**	2.86
	narrative	0.80	2.55	0.80	0.80	0.80	**0.64***	0.80	**0.64***
	pet-directed	17.14	22.86	17.14	**14.29**	17.14	**14.29**	**14.29**	**14.29**
instructions	read	1.47	3.43	1.47	**0.98**	1.47	1.47	**0.98**	1.47
	instructions	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
	conversation	2.79	6.70	2.79	2.79	3.35	3.35	2.79	**2.24**
	narrative	1.23	2.61	1.08	0.92	0.92	0.92	**0.77**	0.92
	pet-directed	18.92	24.32	16.22	16.22	16.22	16.22	**13.51**	16.22
conversation	read	2.03	5.08	**1.52**	**1.52**	2.03	**1.52**	**1.52**	**1.52**
	instructions	2.97	4.95	**2.48***	**2.48***	2.97	**2.48***	**2.48***	**2.48***
	conversation	0.57	1.72	0.57	0.57	0.57	0.57	0.57	0.57
	narrative	1.94	5.98	**1.78**	1.94	2.59	2.10	1.94	2.10
	pet-directed	20.00	22.86	20.00	20.00	**17.14**	**17.14**	**17.14**	**17.14**
narrative	read	0.48	1.76	0.32*	0.32*	0.32*	0.32*	0.32*	**0.16**
	instructions	0.46	1.08	0.46	0.46	0.46	0.46	0.46	0.46
	conversation	1.46	4.56	1.64	1.83	1.64	1.28	**1.10**	**1.10**
	pet-directed	18.58	26.55	18.58	**13.27**	18.58	15.93	**13.27**	16.81
pet-directed	read	14.29	20.00	14.29	**11.43**	14.29	14.29	14.29	14.29
	instructions	18.92	27.03	18.92	18.92	16.22	16.22	**13.51**	16.22
	conversation	21.21	24.24	21.21	21.21	**18.18**	21.21	**18.18**	21.21
	narrative	19.47	28.32	17.70	15.93	19.47	17.70	**14.16**	17.70

the baseline. Statistical signficance ($p < 0.05$) was evaluated using the McNemar's test [McN47]. The results of an x-vector baseline is compared with self-attention and the five conditioning methods.

The baseline x-vector performs better for style-matched conditions than style-mismatched ones. For instance, in conditions where the data is enrolled with conversational speech, the style-matched condition results in an EER of 0.57%. However, style-mismatched conditions have EERs of 2.03%, 2.97%, 1.94% and 20% for read, instructions, narrative, and pet-directed speech, respectively.

To evaluate the need for a self-attention layer, another model is trained with VFR as weights for statistical pooling. However, as seen in Table 4.4, performance degrades when using this approach (VFR weights). Thus, VFR by itself may not be sufficient to provide meaningful weights for each frame.

Self-attentive speaker embeddings provide a statistically significant improvement over the x-vector baseline in 6/23 tasks and only degrades in the narrative–conversation task. These improvements are due to representations with better speaker discrimination capabilities, in agreement with the results in [ZKS18, OKS18].

Compared to the x-vector performance, among the proposed approaches of VFR conditioning, Combined A (concatenation with gating) results in statistically significant improvements in 12/23 tasks, while Combined B (concatenation with affine transformation) results in statistically significant improvements in 6/23 tasks. Among the three individual VFR conditioning methods, the concatenation results in statistically significant improvements in 8/23 tasks, gating in 5/23 tasks, and finally affine transformation in 10/23 tasks. Gating is a special case of affine transformation, and individually gating performs worse than affine, but when combined with con-

Table 4.5: Performance using the SITW evaluation set (EER %). The best performance in each condition is boldfaced and is a statistically significant improvement over the baseline. Combined A (concatenation with gating) and Combined B (concatenation with affine).

Model	Core-Core	Core-Multi	Assist-Core	Assist-Multi
x-vector	3.66	5.87	5.47	6.9
VFR weights	8.17	10.67	10.17	11.82
Self-attention	3.91	6.09	5.51	6.64
Concatenation	3.86	5.83	5.35	**6.35**
Gating	4.32	6.64	6.25	7.61
Affine	3.91	5.95	5.41	6.58
Combined A	3.69	**5.81**	**5.26**	6.54
Combined B	3.91	6.17	5.83	7.16

catenation, it performs better overall. Moreover, the best performing method, VFR conditioning by concatenation with gating, provided significant improvement over the self-attentive embeddings in 10/23 tasks and only degraded in conversation–narrative tasks. In general, VFR conditioning has shown improvement in the case of style-matched and -mismatched tasks. The most gain was in the case of pet-directed speech. These results support the hypothesis that including the VFR conditioning vector in self-attention facilitates the speaker representations to be robust to speaking style variations.

4.2.3.2 SITW Evaluation

SITW evaluation results are in Table 4.5. The best performance in each case is boldfaced and is a statistically significant improvement over the baseline. Conditioning provides improvements over x-vectors in Core-Multi, Assist-Core, and Assist-Multi. For the Core-Multi and Assist-Core cases, the best performing method is conditioning

using concatenation with gating. However, in Assist-Multi, conditioning with concatenation performs the best. Interestingly, the proposed methods, although aimed at addressing style-robustness, provide performance improvements in multi-speaker scenarios.

4.3 Comparison between VFR augmentation and self-attention conditioning using VFR

To determine the appropriate style-robust method for different applications, in this section, we compare the VFR augmentation with the self-attention conditioning approaches.

4.3.1 Database Statistics

To match the experimental setups for both experiments, we use the x-vector setup from Section 4.2. We then use the same subset of 50 female and 50 male speakers from the UCLA database that was set aside in Section 4.1.2.1 as the "development set". The remaining subset of 50 female and 50 male speakers was used as the "evaluation set". The evaluation set was further split into "enrollment" and "test" set.

4.3.2 Results and Discussion

Table 4.4 compares VFR augmentation with the best performing VFR conditioning method which is referred to as combined A (concatenation with gating). The results are in terms of equal error rate (EER) over all style trial combinations for the UCLA

Table 4.6: Performance using the UCLA database (EER %) with VFR augmentation and conditioning. The better result in each condition is boldfaced. All reported differences are statistically significant. Combined A (concatenation with gating).

Enroll	Test	VFR augmentation	VFR conditioning combined A
read	read	0.98	0.98
	instructions	0.95	0.95
	conversation	3.37	**2.25**
	narrative	0.94	0.94
	pet-directed	12.50	12.50
instructions	Read	1.92	1.92
	instructions	0.86	0.86
	conversation	2.20	2.20
	narrative	1.22	1.22
	pet-directed	11.76	11.76
conversation	Read	**1.01**	2.02
	instructions	4.90	**3.92**
	conversation	1.14	**0.38**
	narrative	**1.62**	2.59
	pet-directed	12.50	12.50
narrative	Read	**0.32**	0.63
	instructions	0.91	0.91
	conversation	**0.73**	1.09
	pet-directed	11.76	11.76
pet-directed	Read	12.50	12.50
	instructions	11.76	11.76
	conversation	13.33	13.33
	narrative	17.65	**11.76**

SVD data. The better result in each condition is boldfaced. All reported differences are statistically significant. Statistical significance ($p < 0.05$) was evaluated using the McNemar's test [McN47].

The VFR augmentation and VFR conditioning approaches are comparable in performance. VFR augmentation performs better than conditioning in 4/23 tasks, while VFR conditioning performs better than augmentation in 4/23 tasks. The performances are comparable in the remaining 15/23 tasks.

Given that VFR augmentation does not require any front-end model changes and

only changes the backend, it is a faster and cost-effective approach to address style variability. Hence, it could be used in applications that are time-sensitive and/or require cost-effective updates.

On the other hand, VFR conditioning is easily applicable with other embedding extractors with pooling layer (for instance, ResNet architectures [ZZW21]). Thus, this approach can be used when embedding extractors other than x-vectors are being used. Moreover, the performance of VFR conditioning is more consistent in the case of exaggerated prosody tasks and hence could be the choice for experiments with high prosodic variations.

4.4 Chapter Summary

Speaking style variations degrade the performance of ASV systems. We hypothesize that signal entropy can capture style-related spectral and temporal variations. Hence, we proposed to use entropy-based VFR in two different ways to develop style-robust ASV approaches. First, we used VFR to perform data augmentation when multiple styles were not available to perform an in-domain adaptation of the PLDA classifier. The ASV performance showed significant improvement in the presence of a speaking-style mismatch by addressing performance degradation using VFR data augmentation. The performance of the proposed approach was comparable to the best-case scenario of having multiple styles available for PLDA augmentation. Second, VFR was used to condition self-attentive speaker embeddings providing style-robust representations. The best conditioning approach, concatenation with gating, results in statistically significant improvements over the x-vector baseline in 12/23 tasks and outperforms self-attention in 10/23 tasks. In addition, performance improve-

ments on multi-speaker scenarios in SITW evaluation due to the proposed approach were observed. The simplicity of the proposed method allows for extension to other embedding extractors that utilize a pooling layer. Finally, VFR augmentation and conditioning approaches are comparable in performance. Hence, the requirements of the application (for example, time-sensitivity, low-cost, generalization, data characteristics etc.) will be the key determining factors to chose among the two methods. Augmentation is useful for time-sensitive and/or low-cost applications, while, the conditioning approach can be used for other embedding extractors (for example, ResNet architectures [ZZW21]) or in applications with high prosodic variations.

CHAPTER 5

Can we learn from human speaker perception strategies to improve ASV?

In Chapter 1, we reviewed literature that showed differences in human and machine approaches to speaker discrimination. Subsequent chapters illustrated the effects of speaking style variability on human and machine speaker discrimination abilities. In this chapter, we develop ASV algorithms that are inspired by human speaker perception strategies in an effort to improve ASV performance in the presence of style variability.

5.1 Method

5.1.1 Human speaker perception model

Chapter 3 modeled human speaker perception for moderate style variability between read and conversational speech. Our results showed that listeners find it easier to "tell speakers together" using speaker-specific idiosyncrasies, while listeners "tell speakers apart" based on relative positions within a shared acoustic structure rather than speaker-specific features.

This section aims to incorporate this model in the training loss function. Thus,

we need a loss function that focuses on speaker-specific idiosyncrasies for the "target speaker" task while using acoustic distances between speakers for the "non-target speaker" task.

5.1.2 Baseline models

An x-vector/PLDA system [SGS18] is the baseline. Additionally, the best performing VFR conditioning network from Chapter 4: concatenation with gating, referred to as "combined A" is also used.

The inputs to the embedding extractor are 30-dimensional mel-frequency cepstral coefficients (MFCCs) using a 25 ms frame length and a 10 ms frame shift. The MFCCs are mean normalized over a sliding window of up to 3 secs. Extrinsic data augmentation of noise and reverberation [SGS18] was applied to the training data.

5.1.3 Loss Functions

5.1.3.1 Cross-Entropy (CE) Loss

A widely-used loss function for training ASV systems, including the x-vector system is the cross-entropy loss. This function calculates loss for a multi-class classification problem. CE loss can be calculated as,

$$L_{CE} = -\frac{1}{m}\sum_{i=0}^{m}\log\frac{e^{(\mathbf{W}_{\mathbf{y}_i}^T.\mathbf{x}_i+\mathbf{b}_{\mathbf{y}_i})}}{\sum_{j=0}^{N}e^{(\mathbf{W}_j^T.\mathbf{x}_j+\mathbf{b}_j)}} \tag{5.1}$$

where the $\mathbf{x}_i$ is the i^{th} training sample, $\mathbf{y}_i$ is the ground truth speaker label of the i^{th} training sample, $i \in \{1, \ldots, m\}$, where m is the total number of training samples.

$\mathbf{W}$ indicates the weight matrix, $\mathbf{b}$ is the bias vector. $\mathbf{W}_j$ and $\mathbf{W}_{\mathbf{y}_i}$ are the j^{th} and $\mathbf{y}_i^{\text{th}}$ columns of $\mathbf{W}$, respectively. The CE loss is calculated for a total of N speakers.

The CE loss aims at maximizing inter-speaker distances. However, it does not minimize intra-speaker distances. By maximizing inter-speaker distances, the extracted embeddings are linearly separable. On the other hand, for the embeddings to include desirable discriminative features, the loss should also minimize intra-speaker distances. The embeddings trained on CE loss–maximizing inter-speaker distances– are equivalent to the human approach of focusing on relative positions within a shared acoustic structure to "tell speakers apart". To minimize intra-speaker distances and implement other aspects of human perception strategies, we need a loss function that focuses on speaker-specific idiosyncrasies.

5.1.3.2 C_{llr} Loss

In order to focus on speaker-specific idiosyncrasies without increasing the length of the training cycles, we chose log-likelihood-ratio cost function (C_{llr}) [VB07] as a loss function for training the embedding extractor, referred to as "C_{llr} loss".

C_{llr} is an application independent measure for evaluating soft decisions in ASV performance. There is a closed-form solution for C_{llr} [VB07] that provides the C_{llr} loss function as follows:

$$C_{\text{llr}}(\theta) = \frac{1}{2}\left(\frac{C_{tar}(\theta)}{N_{\text{tar}}} + \frac{C_{non}(\theta)}{N_{\text{non}}}\right) \tag{5.2}$$

$$C_{tar}(\theta) = \sum_{i \in \text{tar}} \log_2(1 + e^{-s_\theta(\mathbf{x}_i, \mathbf{y}_i)}) \tag{5.3}$$

$$C_{non}(\theta) = \sum_{i \in \text{non}} \log_2(1 + e^{s_\theta(\mathbf{x}_i, \mathbf{y}_i)}) \tag{5.4}$$

where θ represents the model parameters, $s_\theta(\mathbf{x}_i, \mathbf{y}_i)$ is the score from the last layer of the embedding extractor for speaker $\mathbf{y}_i$ from input $\mathbf{x}_i$, 'tar' is a set of target speakers and 'non' is a set of non-target speakers. The two terms in Equation 5.2 represent the costs for N_{tar} "target" ($C_{tar}(\theta)$) and N_{non} "non-target" speakers ($C_{non}(\theta)$).

C_{llr} can be interpreted as a measure that is inversely related to information. The lower the C_{llr}, the more the average information per trial (in bits) increases. Optimization is performed with the objective to minimize C_{llr} loss. C_{llr} loss is calculated for each minibatch by considering the output of the last linear layer as scores and using the class labels to define target and non-target speakers.

Thus, C_{llr} loss minimizes intra-speaker distances by focusing on speaker-specific idiosyncrasies. This is similar to the human approach to "tell speakers together".

5.1.3.3 Proposed method: C_{llr}CE loss

Inspired by human speaker-discrimination strategies, we propose to use the combination of cross-entropy loss and C_{llr} loss for training ASV systems. In other words, we aim to combine the approach for "telling speaker apart" and "telling speakers together" in the loss function to focus on maximizing "inter-speaker" distances and minimizing "intra-speaker" distances. We thus use a combined loss function referred to as "C_{llr}CE loss",

$$C_{\text{llr}}CE(\theta) = \frac{1}{2}\left(C_{\text{llr}}(\theta) + L_{CE}\right) \tag{5.5}$$

Given that the perception model used to derive this loss function is based on human speaker discrimination strategies in the presence of moderate style variability,

i.e, between read and conversational speech, we hypothesize that this loss function will provide the most improvement in conversational speech tasks.

5.2 Experimental Setup

The experimental setup for training and testing the embedding extractor is same as in Section 4.2.2. Adam [KB17] optimization was used with a batch size of 128 and trained for 100 epochs. The embedding extractor was trained on the Voxceleb2 dev set.

The UCLA SVD data and SITW evaluation set were used for testing. The statistics of the UCLA database used is provided in Section 4.2.2.1. The proposed method was tested on a total of 23 style conditions from the UCLA SVD data. Further it was tested on four evaluation conditions of the SITW corpus.

5.3 Results and Discussion

5.3.1 UCLA SVD Evaluation

Table 5.1 compares the performance (EER %) of different loss functions on the UCLA database. The loss functions used are cross-entropy loss (CE), C_{llr} loss, and C_{llr}CE loss. These loss functions are used to train the x-vector system and the best performing VFR conditioning: combined A (concatenation with gating). Statistical significance was verified using McNemar's test [McN47]. Unless mentioned explicitly, all performance differences reported in this section are significant with $p < 0.05$.

The C_{llr} loss function by itself does not provide an improvement over the widely-

used CE loss function in both x-vector and VFR conditioning architectures. However, when combined with the CE loss, the performance improves in both architectures.

In the x-vector setup, the C_{llr}CE loss provides the best performance in 9/23 conditions. While in the VFR conditioning setup C_{llr}CE loss results in the best performance in 11/23 conditions.

When compared against their CE counterparts, we can see that the conditions where the C_{llr}CE loss provides improvements are the ones that include conversation style. In the VFR conditioning setup, the improvements by using the C_{llr}CE loss are in read – conversation, instructions – conversation, conversation – narrative, and narrative–read conditions. While with the x-vector setup, these conditions include read – conversation, instructions – narrative, conversation – read, conversation – instructions, conversation – narrative, conversation – pet-directed, pet-directed – conversation, narrative – read, and narrative – pet-directed. In this case, the improvements are with conditions involving conversation and narrative styles (closest to the conversation style).

The results agree with our hypothesis that using the perception model in training the ASV system will improve performance in the conversation style. Interestingly, it also improved performance in some cases of the narrative style, which might resemble conversational speech.

5.3.2 SITW Evaluation

Table 5.2 presents the performance on the SITW evaluation set using different loss functions. The loss functions used are cross-entropy loss (CE), C_{llr} loss, and C_{llr}CE loss. These loss functions are used to train the x-vector system and the best per-

forming VFR conditioning: combined A (concatenation with gating). The best performance in each condition is boldfaced and is a statistically significant improvement over the baseline. The results are reported in terms of EER (%) and also $\text{minDCF}_{0.01}$ values. minDCF evaluates scores produced by the classifier, i.e., soft decisions and is a calibration independent measure.

The results show that the best performing system in terms of $\text{minDCF}_{0.01}$ values is the one with combination loss in VFR conditioning setup. However, EER values of the C_{llr}CE loss in the VFR conditioning setup are slightly worse than the CE loss counterpart for assist-core and assist-multi evaluations. Overall, the proposed loss function with the VFR conditioning setup results in the best performance on the SITW evaluation. Since SITW involves mainly conversational speech, this result agrees with our hypothesis that the new loss function improves ASV system performance for conversational styles.

In summary, Tables 5.1 and 5.2 show that the combined C_{llr}CE loss improves ASV performance, especially in conversational style tasks. Thus, implying that these two loss functions are complementary where C_{llr} loss focuses on speaker-specific idiosyncrasies and CE loss focuses on the relative distance between speakers in a shared acoustic space.

Table 5.1: Performance using the UCLA database (EER %) with different loss functions. The loss functions are used to train the x-vector system and the best performing VFR conditioning: combined A (concatenation with gating). The best result in each condition is boldfaced. If denoted by a '*' it is not a statistically significant improvement over the baseline.

Loss		**CE**		C_{llr}		C_{llr}**CE**	
Enroll	**Test**	**x-vector (Baseline)**	**VFR conditioning**	**x-vector**	**VFR conditioning**	**x-vector**	**VFR conditioning**
read	read	0.50	0.50	0.50	0.50	0.50	0.50
	instructions	0.49	0.49	0.49	0.49	0.49	0.49
	conversation	2.86	2.29	2.29	2.29	2.29	**1.71**
	narrative	**0.80**	**0.80**	1.12	0.96	**0.80**	**0.80**
	pet-directed	17.14	**14.29**	17.14	17.14	17.14	17.14
instructions	read	1.47	**0.98**	1.47	1.47	1.47	1.47
	instructions	0.45	0.45	0.45	0.45	0.45	0.45
	conversation	2.79	2.79	3.35	3.35	2.79	**2.24**
	narrative	1.23	**0.77**	1.08	1.08	0.92	**0.77**
	pet-directed	18.92	13.51	16.22	13.51	16.22	16.22
conversation	read	2.03	**1.52**	2.54	2.03	**1.52**	**1.52**
	instructions	2.97	**2.48***	2.97	**2.48***	**2.48***	**2.48***
	conversation	0.57	0.57	0.57	0.57	0.57	0.57
	narrative	**1.94**	**1.94**	2.26	2.26	**1.94**	2.10
	pet-directed	20.00	**17.14**	**17.14**	**17.14**	**17.14**	**17.14**
narrative	read	0.48	0.32*	0.32*	0.32*	**0.16**	**0.16**
	instructions	0.46	0.46	0.46	0.46	0.46	0.46
	conversation	1.46	1.10	1.83	1.64	1.10	**0.73**
	pet-directed	18.58	**13.27**	17.70	20.35	**13.27**	14.16
pet-directed	read	**14.29**	**14.29**	17.14	20.00	**14.29**	**14.29**
	instructions	18.92	**13.51**	16.22	16.22	16.22	16.22
	conversation	21.21	**18.18**	21.21	21.21	**18.18**	21.21
	narrative	19.47	**14.16**	18.58	20.35	15.93	15.04

Table 5.2: Performance using the SITW evaluation set. The loss functions are applied to the x-vector system and the best performing VFR conditioning: combined A (concatenation with gating). The best performance in each condition is boldfaced and is a statistically significant improvement over the baseline.

		Core-Core		**Core-Multi**		**Assist-Core**		**Assist-Multi**	
Loss	**Model**	**EER %**	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**
CE	x-vector (Baseline)	3.66	0.382	5.87	0.4629	5.47	0.4041	6.90	0.4512
	VFR conditioning	3.69	0.3989	5.81	0.4740	**5.26**	0.4027	**6.54**	0.4651
C_{llr}	x-vector	4.13	0.4153	6.46	0.4940	6.24	0.4376	7.57	0.4824
	VFR conditioning	4.29	0.4009	6.65	0.4821	6.28	0.4337	7.68	0.4776
C_{llr}CE	x-vector	3.77	0.3654	5.88	0.4394	5.70	0.3833	6.74	0.429
	VFR conditioning	**3.47**	**0.3346**	**5.73**	**0.4178**	5.36	**0.3738**	6.73	**0.4191**

CHAPTER 6

Conclusion

This dissertation examines the speaker discrimination abilities of humans and machines in the presence of speaking style variability.

6.1 Summary

Chapter 2 presented several databases and features that were used in the experiments reported in this dissertation. In particular, the UCLA SVD database with multiple styles per speaker is an important database to systematically examine the effects of style variability. Other publicly available databases used in this dissertation include the Speakers in the Wild Database, the NIST SRE and Switchboard databases, and the Voxceleb database. The features used were the voice quality feature set motivated by a psychoacoustic model of voice quality [GSG16, KLG21] and Mel-frequency cepstral coefficients.

Chapter 3 examined the effects of speaking style variations (read speech versus conversational speech) on human speaker discrimination accuracy. The results showed that the difficulty of the discrimination task changed with style: the style-matched read speech – read speech condition was easiest, followed by conversation – conversation. The style-mismatched condition resulted in the worst performance.

Moderate speaking style variability affects the "same speaker" task more than the "different speaker" task. The same speakers were not "easy" or "hard" to distinguish in the "same speaker" and "different speaker" tasks. Analysis of acoustic variability suggested that listeners found it easier to "tell speakers together" when they rely on speaker-specific idiosyncrasies and that they "tell speakers apart" based on their relative positions within a shared acoustic space. The chapter contributes to our understanding of the relationship between human speaker perception and the nature and extent of acoustic variability.

Chapter 4 focuses on automatic speaker verification systems when there is speaking style variability. We analyzed automatic speaker verification (ASV) system performance in both style-matched and style-mismatched conditions. Style-mismatched conditions resulted in performance degradation. To develop style-robust ASV systems, we hypothesized that signal entropy could capture style-related spectral and temporal variations. Thus, two approaches using entropy-based variable frame rate (VFR) were proposed. The first approach focused on performing data augmentation when multiple styles per speaker are not available to train the PLDA classifier. The second approach extracted style-robust embeddings using a self-attentive embedding extractor with VFR conditioning. We evaluated five different conditioning approaches. The best conditioning approach, concatenation with gating, results in statistically significant improvements over the x-vector baseline in 12/23 tasks. Finally, both augmentation and conditioning approaches improved performance in the style-mismatched conditions and were comparable in performance. The augmentation approach is more suitable for time-sensitive and low-cost applications. On the other hand, the conditioning approach can be used in high prosodic variation conditions, as well as when applying these approaches to other embedding extractors (for

example, ResNet architectures [ZZW21].

Finally, in Chapter 5, we introduce a loss function (C_{llr}CE) that is inspired by human perception strategies. The C_{llr} loss by itself focuses on speaker-specific idiosyncrasies to "tell speakers together", while the CE loss focuses on relative acoustic distances between the speakers to "tell speakers apart". The perception model was derived from moderate style variability between read and conversational speech. Hence, the C_{llr}CE loss provides the most improvement in conditions with conversational speech.

6.2 Future Work

The perception experiments used a homogeneous set of listeners (22 female out of 30 listeners with an age range of 17-21 years old). It would be interesting to repeat the perception experiments with a heterogeneous set of listeners to evaluate the generalizability of the proposed human speaker perception model.

Preliminary analyses suggested some effects of native language on listener performance. However, the subject pool did not include sufficient listeners who were non-native speakers of English to perform further analyses. A future perception study focusing on the effects of native language on listener performance would provide further insights into human speaker perception.

The VFR conditioning approaches only focus on using a TDNN/x-vector-based embedding extractor. Given the generalizability of the proposed approach, it would be interesting to evaluate the performance in other embedding extractors (for example, ResNet).

The perception study only focused on moderate style variability. Since the proposed ASV approach based on human perception strategies provided improvements only in conversational speech conditions, a perception study of higher variability in speaking styles might aid in developing more robust ASV approaches.

REFERENCES

[AGP18] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, and Abeer Alwan. "Effectiveness of Voice Quality Features in Detecting Depression." In *Interspeech*, pp. 1676–1680, Hyderabad, India, 2018.

[AGP20] Amber Afshan, Jinxi Guo, Soo Jin Park, Vijay Ravi, Alan McCree, and Abeer Alwan. "Variable frame rate-based data augmentation to handle speaking-style variability for automatic speaker verification." In *Interspeech*, pp. 4318–4322, Shanghai, China, 2020.

[AKA20] Amber Afshan, Jody Kreiman, and Abeer Alwan. "Speaker discrimination in humans and machines: Effects of speaking style variability." In *Proceedings of Interspeech*, pp. 3136–3140, Shanghai, China, 2020.

[AKA22] Amber Afshan, Jody Kreiman, and Abeer Alwan. "Speaker discrimination performance for "easy" versus "hard" voices in style-matched and -mismatched speech." *The Journal of the Acoustical Society of America*, **151**(2):1393–1403, 2022.

[BD11a] Niko Brummer and Edward De Villiers. "The BOSARIS toolkit: Theory, algorithms and code for surviving the new DCF." In *NIST SRE Analysis Workshop*, Atlanta, GA, 2011.

[BD11b] Niko Brummer and Edward De Villiers. "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing." *Documentation of BOSARIS toolkit*, **24**, 2011.

[BF06] Helen Blatchford and Paul Foulkes. "Identification of voices in shouting." *International Journal of Speech, Language and the Law*, **13**(2):241–254, August 2006.

[Bru10] Niko Brummer. *Measuring, refining and calibrating speaker and language information extracted from speech.* Ph.D. dissertation, University of Stellenbosch, South Africa, December 2010.

[CDL16] Jianpeng Cheng, Li Dong, and Mirella Lapata. "Long Short-Term Memory-Networks for Machine Reading." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 551–561, Austin, Texas, November 2016. Association for Computational Linguistics.

[CNZ18] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition." *Interspeech 2018*, pp. 1086–1090, September 2018. arXiv: 1806.05622.

[CWM18] F. A. Rezaur Rahman Chowdhury, Quan Wang, Ignacio Lopez Moreno, and Li Wan. "Attention-Based Models for Text-Dependent Speaker Verification." *arXiv:1710.10470 [cs, eess, stat]*, January 2018. arXiv: 1710.10470.

[CX12] Sheng Chen and Mingxing Xu. "Compensation of Intrinsic Variability with Factor Analysis Modeling for Robust Speaker Verification." In *INTERSPEECH*, 2012.

[DGX19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[DKD11] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. "Front-end factor analysis for speaker verification." *IEEE Transactions on Audio, Speech and Language Processing*, **19**(4):788–798, 2011. ISBN: 1558-7916.

[DM80] Steven Davis and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." *IEEE transactions on acoustics, speech, and signal processing*, **28**(4):357–366, 1980.

[Esk93] Maxine Eskenazi. "Trends in speaking styles research." In *Third European Conference on Speech Communication and Technology*, 1993.

[FMB22] Luciana Ferrer, Mitchell McLaren, and Niko Brümmer. "A speaker verification backend with robust performance across conditions." *Computer Speech & Language*, **71**:101258, January 2022.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[GHK19] Rosa González Hautamäki, Ville Hautamäki, and Tomi Kinnunen. "On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise." *J. Acoust. Soc. Am.*, **146**(1):693–704, July 2019.

[GKH15] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, and Anne-Maria Laukkanen. "Automatic versus human speaker verification: The case of voice mimicry." *Speech Communication*, **72**:13–31, September 2015.

[GMS14] Daniel Garcia-Romero, Alan McCree, Stephen Shum, Niko Brummer, and Carlos Vaquero. "Unsupervised domain adaptation for i-vector speaker recognition." *Odyssey*, p. 5, 2014.

[GSG16] Marc Garellek, Robin Samlan, Bruce R. Gerratt, and Jody Kreiman. "Modeling the voice source in terms of spectral slopes." *J. Acoust. Soc. Am.*, **139**(3):1404–1410, 2016.

[GWC99] D Graff, K Walker, and A Canavan. "Switchboard-2 phase ii." *LDC 99S79–http://www. ldc. upenn. edu/Catalog*, 1999.

[GYM16] Jinxi Guo, Gary Yeung, Deepak Muralidharan, Harish Arsikere, Amber Afshan, and Abeer Alwan. "Speaker Verification Using Short Utterances with DNN-Based Estimation of Subglottal Acoustic Features." In *INTERSPEECH*, pp. 2219–2222, 2016.

[GZE12] Daniel Garcia-Romero, Xinhui Zhou, and Carol Y Espy-Wilson. "Multi-condition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition." In *ICASSP*, pp. 4257–4260. IEEE, 2012.

[HC99] H M Hanson and E S Chuang. "Glottal characteristics of male speakers: acoustic correlates and comparison with female data." *J. Acoust. Soc. Am.*, **106**(2):1064–77, 1999.

[HCE94] James Hillenbrand, Ronald A. Cleveland, and Robert L. Erickson. "Acoustic correlates of breathy vocal quality." *Journal of Speech Language and Hearing Research*, **37**(4):769–778, 1994.

[Her21] Christian T. Herbst. "Performance evaluation of subharmonic-to-harmonic ratio (SHR) computation." *Journal of Voice*, **35**(3):365–375, May 2021.

[HHF17] Vincent Hughes, Philip Harrison, Paul Foulkes, Peter French, Colleen Kavanagh, and Eugenia San Segundo. "Mapping Across Feature Spaces in Forensic Voice Comparison: The Contribution of Auditory-Based Voice

Quality to (Semi-)Automatic System Testing." In *Interspeech*, pp. 3892–3896, Stockholm, Sweden, 2017.

[HMB16] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. "End-to-end text-dependent speaker verification." In *ICASSP*, pp. 5115–5119. IEEE, 2016.

[HSH13] Taufiq Hasan, Rahim Saeidi, John HL Hansen, and David A Van Leeuwen. "Duration mismatch compensation for i-vector based speaker recognition systems." In *ICASSP*, pp. 7663–7667. IEEE, 2013.

[IA04] M. Iseli and A. Alwan. "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. I–669–I–672, Montreal, Canada, May 2004.

[IEE69] IEEE Subcommittee on Subjective Measurements. "IEEE recommended practice for speech quality measurements." *IEEE Transactions on Audio and Electroacoustics*, **17**(3):225–246, 1969.

[Jes08] Michael Jessen. "Forensic phonetics." *Language and Linguistics Compass*, **2**(4):671–711, 2008.

[JML20] Justine Johnson, Carolyn McGettigan, and Nadine Lavan. "Comparing unfamiliar voice and face identity perception using identity sorting tasks." *Quarterly Journal of Experimental Psychology*, **73**(10):1537–1545, October 2020.

[Kai60] Henry F. Kaiser. "The application of electronic computers to factor analysis." *Educational and Psychological Measurement*, **20**(1):141–151, April 1960.

[KB17] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization." *arXiv:1412.6980 [cs]*, January 2017. arXiv: 1412.6980.

[KKA19] Patricia Keating, Jody Kreiman, and Abeer Alwan. "A new speech database for within- and between-speaker variability." In *Proceedings of the ICPhS XIX*, volume 126, pp. 736—-739, Melbourne, Australia, 2019.

[KKA21] Patricia Keating, Jody Kreiman, Abeer Alwan, Adam Chong, and Yoonjeong Lee. "UCLA Speaker Variability Database." 2021. (Last viewed July 20, 2021).

[KLG21] Jody Kreiman, Yoonjeong Lee, Marc Garellek, Robin Samlan, and Bruce R. Gerratt. "Validating a psychoacoustic model of voice quality." *J. Acoust. Soc. Am.*, **149**(1):457–465, January 2021.

[KPK15] Jody Kreiman, Soo Jin Park, Patricia A. Keating, and Abeer Alwan. "The Relationship Between Acoustic and Perceived Intraspeaker Variability in Voice Quality." In *Proceedings of Interspeech*, pp. 2357–2360, Dresden, Germany, 2015.

[KPP17] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. "A study on data augmentation of reverberant speech for robust speech recognition." In *ICASSP*, pp. 5220–5224. IEEE, 2017.

[Krz79] W. J. Krzanowski. "Between-groups comparison of principal components." *Journal of the American Statistical Association*, **74**(367):703–707, 1979.

[KS11] Jody Kreiman and Diana Sidtis. *Foundations of voice studies: An interdisciplinary approach to voice production and perception.* WileyBlackwell, Walden, MA, 2011. pp. 245-246.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[KSO13] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Md. Jahangir Alam, and Pierre Dumouchel. "PLDA for Speaker Verification with Utterances of Arbitrary Duration." In *Proc. ICASSP*, pp. 7649–7653, 2013. ISSN: 15206149.

[Laa92] Gitta PM Laan. "Perceptual differences between spontaneous and read aloud speech." In *Proceedings of the Institute of Phonetic Sciences Amsterdam*, volume 16, pp. 65–79, 1992.

[LBG19] Nadine Lavan, Luke F. K. Burston, and Lúcia Garrido. "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices." *British Journal of Psychology*, **110**(3):576–593, 2019.

[LBL19] Nadine Lavan, Luke FK Burston, Paayal Ladwa, Siobhan E Merriman, Sarah Knight, and Carolyn McGettigan. "Breaking voice identity perception: Expressive voices are more confusable for listeners." *Quarterly Journal of Experimental Psychology*, **72**(9):2240–2248, September 2019.

[LBS19] Nadine Lavan, A. Mike Burton, Sophie K. Scott, and Carolyn McGettigan. "Flexible voices: Identity perception from variable vocal signals." *Psychonomic Bulletin & Review*, **26**(1):90–102, February 2019.

[LGO18] Yutian Li, Feng Gao, Zhijian Ou, and Jiasong Sun. "Angular Softmax Loss for End-to-end Speaker Verification." In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 190–194, November 2018.

[LK19] Yoonjeong Lee and Jody Kreiman. "Within- and between-speaker acoustic variability: Spontaneous versus read speech." *J. Acoust. Soc. Am.*, **146**(4):3011–3011, October 2019.

[LKK19] Yoonjeong Lee, Patricia Keating, and Jody Kreiman. "Acoustic voice variation within and between speakers." *J. Acoust. Soc. Am.*, **146**(3):1568–1579, September 2019.

[MBP19] Katerina Margatina, Christos Baziotis, and Alexandros Potamianos. "Attention-based Conditioning Methods for External Knowledge Integration." *arXiv:1906.03674 [cs, stat]*, June 2019. arXiv: 1906.03674.

[McD04] Kirsty McDougall. "Speaker-specific formant dynamics: An experiment on Australian English /aI/." *International Journal of Speech Language and the Law*, **11**(1):103–130, March 2004.

[McN47] Quinn McNemar. "Note on the sampling error of the difference between correlated proportions or percentages." *Psychometrika*, **12**(2):153–157, 1947. Publisher: Springer.

[MG09] Alvin F. Martin and Craig S. Greenberg. "NIST 2008 Speaker Recognition Evaluation: Performance across Telephone and Room Microphone Channels." In *Proc. Interspeech*, pp. 2579–2582, Brighton, UK, 2009. ISSN: 19909772.

[MMO20] Victoria Mingote, Antonio Miguel, Alfonso Ortega, and Eduardo Lleida. "Optimization of the area under the ROC curve using neural network

supervectors for text-dependent speaker verification." *Computer Speech & Language*, **63**:101078, September 2020.

[MMO21] Victoria Mingote, Antonio Miguel, Alfonso Ortega, and Eduardo Lleida. "Log-Likelihood-Ratio Cost Function as Objective Loss for Speaker Verification Systems." In *Interspeech 2021*, pp. 2361–2365, Brno, Czech Republic, August 2021. ISCA.

[MMR19] Victoria Mingote, Antonio Miguel, Dayana Ribas, Alfonso Ortega, and Eduardo Lleida. "Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems." In *Interspeech 2019*, pp. 2903–2907. ISCA, September 2019.

[OKS18] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding." *Interspeech 2018*, pp. 2252–2256, September 2018. arXiv: 1803.10963.

[PAC18] Soo Jin Park, Amber Afshan, Zhi Ming Chua, and Abeer Alwan. "Using Voice Quality Supervectors for Affect Identification." In *INTERSPEECH*, pp. 157–161, 2018.

[PAK19] Soo Jin Park, Amber Afshan, Jody Kreiman, Gary Yeung, and Abeer Alwan. "Target and non-target speaker discrimination by humans and machines." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 6326–6330, Brighton, United Kingdom, May 2019.

[PDB86] Michael A Picheny, Nathaniel I Durlach, and Louis D Braida. "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech." *JSLHR*, **29**(4):434–446, 1986.

[PGB11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, and others. "The Kaldi speech recognition toolkit." Technical report, IEEE Signal Processing Society, 2011.

[PGM19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy,

Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." *arXiv:1912.01703 [cs, stat]*, December 2019. arXiv: 1912.01703.

[PM04] Mark Przybocki and Alvin Martin. "NIST Speaker Recognition Evaluation Chronicles." In *Proc. Odyssey*, pp. 12–22, 2004.

[PML06] Mark Przybocki, Alvin Martin, and Audrey Le. "NIST Speaker Recognition Evaluation Chronicles - Part 2." In *Proc. Odyssey*, pp. 1–6, 2006.

[PSK16] Soo Jin Park, Caroline Sigouin, Jody Kreiman, Patricia A. Keating, Jinxi Guo, Gary Yeung, Fang-Yu Kuo, and Abeer Alwan. "Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition." In *Interspeech*, 2016.

[PYV18] Soo Jin Park, Gary Yeung, Neda Vesselinova, Jody Kreiman, Patricia A. Keating, and Abeer Alwan. "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles." *J. Acoust. Soc. Am.*, **144**(1):375–386, 2018.

[RFA20] Vijay Ravi, Ruchao Fan, Amber Afshan, Huanhua Lu, and Abeer Alwan. "Exploring the Use of an Unsupervised Autoregressive Model as a Shared Encoder for Text-Dependent Speaker Verification." In *Interspeech*, Shanghai, China, 2020.

[RFG11] Daniel Ramos, Javier Franco-Pedroso, and Joaquin Gonzalez-Rodriguez. "Calibration and weight of the evidence by human listeners. the ATVS-UAM submission to NIST human-aided speaker recognition 2010." In *Proceedings of ICASSP*, pp. 5908–5911, Prague, Czech Republic, 2011.

[RMG19] Esther Rituerto-González, Alba Mínguez-Sánchez, Ascensión Gallardo-Antolín, and Carmen Peláez-Moreno. "Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence." *Applied Sciences*, **9**(11):2298, 2019.

[RPA19] Vijay Ravi, Soo Jin Park, Amber Afshan, and Abeer Alwan. "Voice Quality and Between-Frame Entropy for Sleepiness Estimation." In *Interspeech*, pp. 2408–2412, Graz, Austria, 2019.

[RWF22] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. "FrAUG: A Frame Rate Based Data Augmentation Method for Depression Detection from Speech Signals." In *Proceedings of ICASSP (in press)*, Singapore, 2022.

[SBR19] Harriet MJ Smith, Thom S Baguley, Jeremy Robson, Andrew K Dunn, and Paula C Stacey. "Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance." *Applied Cognitive Psychology*, **33**(2):272–287, 2019.

[SCP15] David Snyder, Guoguo Chen, and Daniel Povey. "MUSAN: A Music, Speech, and Noise Corpus." *arXiv:1510.08484 [cs]*, October 2015.

[SGB08] Elizabeth Shriberg, Martin Graciarena, Harry Bratt, Andreas Kathol, Sachin S Kajarekar, Huda Jameel, Colleen Richey, and Fred Goodman. "Effects of vocal effort and speaking style on text-independent speaker verification." In *9th Annual Conf. of the Intl. Speech Communication Association*, 2008.

[SGP17] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. "Deep Neural Network Embeddings for Text-Independent Speaker Verification." In *Interspeech*, pp. 999–1003, 2017.

[SGS18] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. "X-vectors: Robust dnn embeddings for speaker recognition." In *ICASSP*, 2018.

[Sim12] Adrian P. Simpson. "The first and second harmonics should not be used to measure breathiness in male and female voices." *Journal of Phonetics*, **40**(3):477–490, May 2012.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering." In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, Boston, MA, USA, June 2015. IEEE.

[SKS09] Elizabeth Shriberg, Sachin Kajarekar, and Nicolas Scheffer. "Does session variability compensation in speaker recognition model intrinsic variation under mismatched conditions?" In *INTERSPEECH*, 2009.

[SKV11] Yen-Liang Shue, Patricia A. Keating, Chad Vicenik, and Kristine Yu. "VoiceSauce: A program for voice analysis." In *Proceedings of the ICPhS XVII*, volume 126, pp. 1846–1849, Hong Kong, 2011.

[Smi48] N. Smirnov. "Table for estimating the goodness of fit of empirical distributions." *The Annals of Mathematical Statistics*, **19**(2):279–281, 1948.

[SRG14] Stephen H Shum, Douglas A Reynolds, Daniel Garcia-Romero, and Alan McCree. "Unsupervised clustering approaches for domain adaptation in speaker recognition systems." *Odyssey*, 2014.

[SS20] Susanta Kumar Sarangi and Goutam Saha. "Improved Speech-Signal Based Frequency Warping Scale for Cepstral Feature in Robust Speaker Verification System." *Journal of Signal Processing Systems*, March 2020.

[STN21] Sarah V Stevenage, Rebecca Tomlin, Greg J Neil, and Ashley E Symons. "May I Speak Freely? The Difficulty in Vocal Identity Processing Across Free and Scripted Speech." *Journal of Nonverbal Behavior*, **45**(1):149–163, 2021.

[Sun02] X. Sun. "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. I–333–I–336, May 2002.

[SY80] H. Saslove and A. D. Yarmey. "Long-term auditory memory: speaker identification." *The Journal of Applied Psychology*, **65**(1):111–116, February 1980.

[VB07] David A Van Leeuwen and Niko Brummer. "An introduction to application-independent evaluation of speaker recognition systems." In *Speaker Classification I: Fundamentals, Features, and Methods*, pp. 330–353. Springer, Berlin, Heidelberg, 2007.

[VK87] Diana Van Lancker and Jody Kreiman. "Voice discrimination and recognition are separate abilities." *Neuropsychologia*, **25**(5):829–834, 1987. ISBN: 0028-3932.

[VLM14] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. "Deep neural networks for small footprint text-dependent speaker verification." In *ICASSP*, pp. 4052–4056, May 2014. ISSN: 2379-190X.

[WJ03] Ratree Wayland and Allard Jongman. "Acoustic correlates of breathy and clear vowels: the case of Khmer." *Journal of Phonetics*, **31**(2):181–201, April 2003.

[WK19] Jennifer Williams and Simon King. "Disentangling Style Factors from Speaker Representations." *Interspeech*, pp. 3945–3949, 2019.

[WLL18] Feng Wang, Weiyang Liu, Haijun Liu, and Jian Cheng. "Additive Margin Softmax for Face Verification." *IEEE Signal Processing Letters*, **25**(7):926–930, July 2018. arXiv: 1801.05599.

[WM12] Stanley J Wenndt and Ronald L Mitchell. "Machine recognition vs human recognition of voices." In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4245–4248. IEEE, 2012.

[WOL18] Q. Wang, K. Okabe, K. A. Lee, H. Yamamoto, and T. Koshinaka. "Attention Mechanism in Speaker Recognition: What Does it Learn in Deep Speaker Embedding?" In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1052–1059, December 2018.

[WSX17] Yuxuan Wang, R. J. Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A. Saurous. "Uncovering Latent Style Factors for Expressive Speech Synthesis." *arXiv:1711.00520 [cs]*, November 2017.

[WWZ18] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. "CosFace: Large Margin Cosine Loss for Deep Face Recognition." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, Salt Lake City, UT, June 2018. IEEE.

[YZA04] Hong You, Qifeng Zhu, and Abeer Alwan. "Entropy-based variable frame rate analysis of speech signals and its application to ASR." In *ICASSP*, volume 1, pp. I–549. IEEE, 2004.

[ZA00] Qifeng Zhu and Abeer Alwan. "On the use of variable frame rate analysis in speech recognition." In *ICASSP*, volume 3, pp. 1783–1786. IEEE, 2000.

[ZCZ17] Shi-Xiong Zhang, Zhuo Chen, Yong Zhao, Jinyu Li, and Yifan Gong. "End-to-End Attention based Text-Dependent Speaker Verification." In *arXiv:1701.00562 [cs, stat]*, January 2017. arXiv: 1701.00562.

[ZKH18] Chunlei Zhang, Kazuhito Koishida, and John H. L. Hansen. "Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings." *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(9):1633–1644, September 2018. Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[ZKS18] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey. "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification." In *Interspeech 2018*, pp. 3573–3577. ISCA, September 2018.

[ZRH18] Chunlei Zhang, Shivesh Ranjan, and John HL Hansen. "An Analysis of Transfer Learning for Domain Mismatched Text-independent Speaker Verification." In *Odyssey*, pp. 181–186, 2018.

[ZZW21] Tianyan Zhou, Yong Zhao, and Jian Wu. "ResNeXt and Res2Net Structures for Speaker Verification." In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 301–307, January 2021.