UNIVERSITY OF CALIFORNIA

Los Angeles

Towards Understanding Voice Discrimination Abilities of Humans and Machines

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical and Computer Engineering

by

Soo Jin Park

2019

ABSTRACT OF THE DISSERTATION

Towards Understanding Voice Discrimination Abilities of Humans and Machines

by

Soo Jin Park
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2019
Professor Abeer A. H. Alwan, Chair

An individual's voice can vary dramatically depending on word choice, affect, and other factors. Such intrinsic within-talker variability causes considerable difficulties when distinguishing talkers by their voices, both for humans and machines. For machines, phonetic content variability substantially degrades performance when utterances are short (e.g., < 10 sec). Humans, on the contrary, are less influenced by content variability, and they perform better than machines in such conditions. Hence, understanding which and how acoustic features are related to human responses might provide insights to improve machine performance. Yet, little is known about human and machine voice discrimination ability under various kinds of intrinsic within-talker variabilities.

This dissertation presents studies of voice discrimination abilities of humans and machines under text, affect, and speaking-style variabilities. The main focus is in developing a feature set, based on a psychoacoustic model of voice quality, that can be used to improve machine performance and to find acoustic correlates with human responses. In order to systematically investigate the effects of within- and between-talker variability, a database was developed at UCLA. More than a hundred females and a hundred males were recorded with various speech styles, including sustained vowels, read sentences, affective speech, and pet-directed speech.

Preliminary experiments indicated that the voice quality feature set (VQual1) was promising for predicting human responses, and for improving automatic speaker verification (ASV)

performance which degraded significantly under text, affect and/or speaking-style variabilities. VQual1 was modified to another set (VQual2) to better differentiate talkers, leading to further improvements in short-utterance text-independent ASV tasks. Voice discrimination abilities of humans and machines for very short utterances ($\approx$ 2 sec) under high text and style variability were analyzed using read sentences and pet-directed speech. Humans were more accurate than machines for read sentence pairs, but the performance difference became small for style-mismatched pairs and for perceptually marked talkers. Humans' and machines' decision spaces were weakly correlated, indicating a weak or non-linear relationship between talker representations by humans and machines. However, for different-talker pairs, the VQual2-based system responses were highly correlated with human responses. Results also suggested that machines could supplement human decisions for perceptually marked talkers. Additionally, VQual2 was effective in perceived affect recognition, suggesting another application where voice quality features can contribute to predict human decisions.

The dissertation of Soo Jin Park is approved.

Kung Yao

Alan J. Laub

Jody E. Kreiman

Abeer A. H. Alwan, Committee Chair

University of California, Los Angeles

2019

To my parents

who encouraged me to go on every adventure,

especially this one

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Over the past five years, I have received support and encouragement from a great number of individuals. I would like to express my deepest appreciation to my doctoral adviser, Prof. Abeer Alwan. She taught me fundamentals of conducting scientific research in the speech processing area, and presenting the results with my own voice and words. Not only that, her profound belief in my abilities had lifted me up even when I was frustrated by failures. I am also extremely grateful to Prof. Jody Kreiman. Her comprehensive knowledge in voice production and perception played a crucial role in my research. I must thank the rest of my committee members, Prof. Alan Laub and Prof. Kung Yao, for their wholehearted support and invaluable advice. Thanks should also go to Prof. Patricia Keating for sharing her unparalleled knowledge in linguistics.

My heartfelt gratitude also goes to my labmates. They contributed in various projects related to this dissertation, and they shared invaluable insights with me through frequent and extensive discussions. I appreciate their sense of humor and thoughtfulness as much as their intelligence. Many nights I stayed in the lab could be enjoyable with them. I cannot leave UCLA without mentioning administrative staff members in the department. Life could become much easier with their quick and accurate work. I am truly blessed with wonderful friends. My mental and physical health is absolutely indebted to their constant care. A million thanks to my family members for their unconditional love and enormous support. The completion of this dissertation would not have been possible without these people.

I would also like to extend my sincere gratitude to my former advisers, Dr. Jeung-Yoon Choi and Prof. Hong-Goo Kang. Their expertise and enthusiasm in speech processing attracted me to this area, and their unwavering support and encouragement drove me to explore further.

VITA

2011	B.S. with Highest Honors, Electrical and Electronic Engineering, Yonsei University.
2011–2013	Research Assistant, Digital Signal Processing Lab., Yonsei University.
2013	M.S., Electrical and Electronic Engineering, Yonsei University.
2013–present	Graduate Student Researcher, Speech Processing and Auditory Perception Lab., UCLA.
2014, 2018	Graduate Student Researcher, Head and Neck Surgery Department, Ronald Reagan UCLA Medical Center.
2014–2018	Teaching Assistant, Electrical Engineering Department, UCLA.
2016, 2017	Research Intern, Oben Inc.
2018	Speech Scientist Intern, Gridspace Inc.

PUBLICATIONS

Soo Jin Park, Gary Yeung, Neda Vesselinova, Jody Kreiman, Patricia 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: 375–386, 2018.

Dinesh Chhetri and **Soo Jin Park**, "Interactions of Subglottal Pressure and Neuromuscular Activation on Fundamental Frequency and Intensity," *Laryngoscope*, 126(5):1123–1130, 2016.

Soo Jin Park, Ambr Afshan, Jody Kreiman, Gary Yeung, and Abeer Alwan, "Target and Non-Target Speaker Discrimination by Humans and Machines," in *ICASSP* (accepted)

Soo Jin Park, Amber Afshan, Zhi Ming Chua, and Abeer Alwan, "Using Voice Quality Supervectors for Affect Identification," in *Interspeech*, 2018.

Amber Afshan, Jinxi Guo, **Soo Jin Park**, Vijay Ravi, Jonathan Flint, and Abeer Alwan, "Effectiveness of Voice Quality Features in Detecting Depression," in *Interspeech*, 2018.

Soo Jin Park, Gary Yeung, Jody Kreiman, Patricia Keating, and Abeer Alwan, "Using Voice Quality Features to Improve Short-Utterance Text-Independent Speaker Verification," in *Interspeech*, 2017.

Soo Jin Park, Caroline Sigouin, Jody Kreiman, Patricia 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.

Jody Kreiman, **Soo Jin Park**, Patricia A. Keating and Abeer Alwan, "The Relationship between Acoustic and Perceived Intraspeaker Variability in Voice Quality," in *Interspeech*, 2015.

Gang Chen, **Soo Jin Park**, Jody Kreiman and Abeer Alwan, "Investigating the Effect of F0 and Vocal Intensity on Harmonic Magnitudes: Data from High-Speed Laryngeal Videoendoscopy," in *Interspeech*, 2014.
∗International Speech Communication Association (ISCA) best student paper award finalist

Gang Chen, **Soo Jin Park**, Jody Kreiman and Abeer Alwan, "On Transition between Voice Registers: Data from High-Speed Laryngeal Videoendoscopy," *J. Acoust. Soc. Am.*, 135(4), 2014.

CHAPTER 1

Introduction

1.1 Overview and Motivation

Voice can be viewed as an "auditory face [BFB04]", that allows us to recognize individuals and their emotional status. What characterizes a talker's voice can be related to her/his physical traits and habitual style of speaking. Many researchers have made efforts to reliably retrieve such information. In courts, for example, forensic phoneticians apply linguistic knowledge to identify a criminal from her/his voice, and are also interested in constructing a voice lineup for earwitnesses. The earwitnesses are often asked to identify the voice they heard at the crime scene among the voice lineup. A fair voice lineup should be constructed with foils having voices similar to the suspect so that the suspect does not 'stick out' [NMH11]. An objective measure of perceived voice similarity is desired in that approach. In the signal processing and machine learning areas, researchers are interested in building an acoustic model that can represent a talker. Such a model can be used for automatic speaker verification (ASV), automatic speaker diarization, and many other applications.

The human voice is a performance biometric, unlike widely-used biometrics such as fingerprints or irises. This makes speech signals prone to a large degree of within-talker variability. The within-talker variability falls into two categories: extrinsic variability and intrinsic variability. Intrinsic variability includes variability related to the talker's conscious and/or unconscious behavior that can influence speech signal production, such as phonetic content,

Parts of this chapter were published in [PYV18].

mood, health condition, and speaking style. Extrinsic variability includes variability that may introduce changes in the acoustic properties of speech signals, such as recording conditions, channel types, and environmental noise. Both categories of within-talker variability can lead to considerable difficulties in distinguishing individuals by their voices both for humans and machines.

In this dissertation, the effects of intrinsic within-talker variability on humans and machines in distinguishing voices are analyzed. Of special interest is the role of voice quality (the timbre of voice) in constituting a voice's identity. Little is known about the abilities of humans and machines in detecting identity from voices influenced by various kinds of intrinsic variability, partly because of a lack of database. Recently, a database was developed at the University of California, Los Angeles (UCLA) to represent both within- and between-talker variability and also recording session variability [KPK15, KKA18]. The UCLA Speaker Variability Database includes a large number of talkers (more that 100 females and 100 males), with multiple recording sessions and varying phonetic content, speaking style, and affect conditions per talker, reflecting normal, daily-life variations in voice quality. In particular, we focus on within-talker variability in recording session, phonetic content, affect, and speaking style in the UCLA database.

One of the most basic tasks in distinguishing talkers was employed: deciding whether two speech samples came from a single talker or from two different talkers. This task is referred to as *voice discrimination*. Because humans are known to involve acoustic feature comparison when they discriminate unfamiliar voices, this task is appropriate to learn about acoustic correlates of human responses. This task is also appropriate to analyze machine performances because it is used as a standard evaluation task for ASV systems. By investigating human responses to voice discrimination tasks, we aim at understanding how and to which extent voice quality is related to human responses, and using the knowledge to improve machine performance.

1.2 Acoustic Features Often Used to Represent Voice Identity

Both for humans and machines, some objectively measurable aspects of speech need to be considered to make meaningful comparisons among voices. Generally, these characterizing aspects are referred to as *acoustic features*. An ideal feature would [Ros02, Wol72]: (1) show high between-talker variability and low within-talker variability, (2) be robust against noise and distortion, (3) occur frequently and naturally in speech, (4) be easy to extract and measure from speech signals, (5) be resistant to attempted disguise or mimicry, and (6) not be affected by the talker's health or long-term variations in voice.

1.2.1 Acoustic Theory of Speech Production

As mentioned above, physical traits of an individual's vocal apparatus and the way it is used influence acoustic characteristics of her/his voice. The process of speech production is often modeled by source and a filter [Fan60]. A schematic of the physical speech production system and its model are shown in Figure 1.1. For voiced sounds, the source is the quasi-periodic, harmonic-rich pressure wave produced when the air flow from the lungs is modulated by the vibrating vocal folds. The *fundamental frequency* of the vocal fold vibration is denoted as F_0. In the context of perception, F_0 is strongly related to *pitch*, and these terms are often used interchangeably. Periodicity of the source excitation leads to a discrete spectrum, where each discrete component is placed at integer multiples of F_0. These discrete components are called *harmonics*. The amplitude of the N-th harmonic is denoted as H_N. For example, the harmonic amplitudes at F_0, $2F_0$, and $3F_0$, are denoted as H_1, H_2, and H_3, respectively.

The vocal tract (which extends from the vocal folds to the lips) is modeled as a time-varying filter whose acoustic gain is frequency-dependent due to the resonances produced by the physical geometry. The resonance frequencies of the vocal tract are called *formant frequencies* or simply *formants*. The lowest formant frequency is denoted as F_1, the second lowest one is F_2, and so on. The amplitude of those resonances are called *formant amplitudes*, and denoted as A_1, A_2, and so on.

Over a short duration, the speech production system can be modeled as a linear, time-

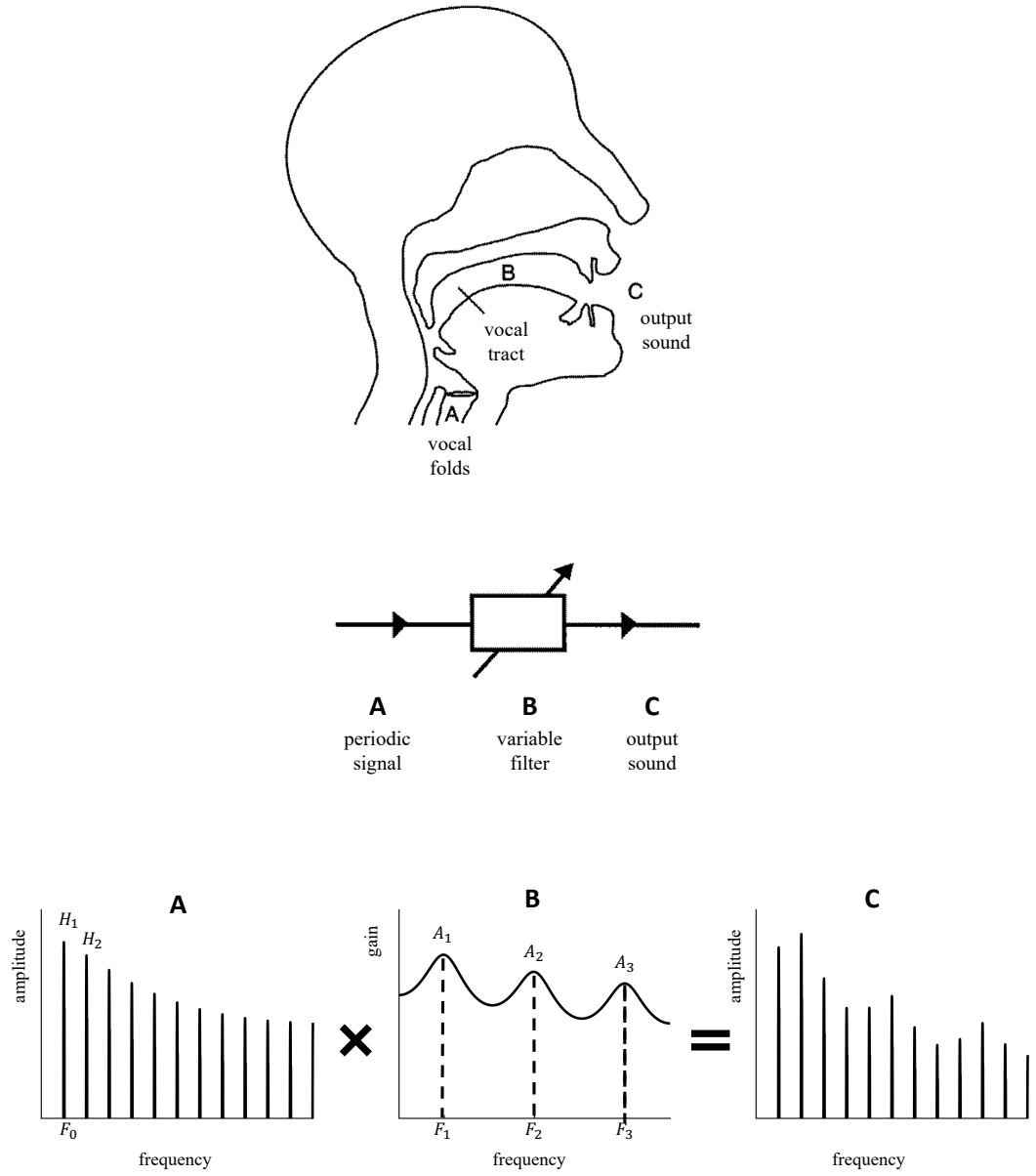


Figure 1.1: A schematic of the speech production system for a voiced sound segment. The schematic view of human speech production system (top panel) and its corresponding model are shown (middle panel). Example spectra that represent typical voice source, vocal tract transfer function, and the output sound (denoted as A, B, and C, respectively) are also provided (bottom panel). Adapted from [ESW97].

invariant system. Hence, the output speech spectrum can be considered as a product of the source spectrum and vocal tract transfer function as shown in Figure 1.1. The above mentioned acoustic features, the fundamental frequency, harmonic amplitude, formant frequencies and formant amplitudes, are highly dependent to the talker's physical traits. For example, the length and elasticity of vocal folds determine the fundamental frequency, and the vocal tract length influences formant frequencies. However, a talker can manipulate, to some extent, her/his vocal apparatus to produce a desired sound. For example, by changing the level of constriction of laryngeal muscles, the fundamental frequency can be higher or lower, and by changing the vocal tract constriction location, formant frequencies can be changed. Thus, these acoustic features can be related to the talker identity by providing information about the talker's physical traits and habitual style of speaking.

1.2.2 Acoustic Correlates to Perceived Identity

One way of analyzing the perceptually relevant acoustic aspects of voice identity is to correlate talker similarity judgments with a range of measured acoustic features from the stimuli. Based on these similarity judgments, multidimensional scaling (MDS, [KW78]) is often used to produce a solution based on a number of dimensions. The resulting MDS space can be thought of as a perceptual "voice space" where the stimuli are close if they are perceived as similar. The MDS axes can be interpreted by examining correlations between the coordinates of the stimuli and acoustic (or other) measures of those stimuli: a high correlation suggests the measure might be an important cue for distinguishing talkers.

For example, Baumann and Belin [BB10] had 16 female and 16 male speakers produce three vowels. In their experiments, F_0 was found to be the main parameter accounting for similarity judgments between both female and male voices. Nolan et al. [NMH11] had listeners judge similarity between two excerpts of telephone conversation speech with 15 young male talkers. They tried to find acoustic correlates to the MDS dimensions by using tokens of 6 different vowels in /hVd/ contexts in read speech from those 15 talkers. In that study, F_0 was found to correlate most strongly with similarity judgments, followed by mean

F_3, mean F_2, and mean F_1 across the different vowels.

1.2.3 Acoustic Features for ASV

The most widely-used acoustic feature set in automatic speech processing systems is the mel-frequency cepstral coefficients (MFCCs). MFCCs represent the smoothed spectral envelope of the speech signal, by applying an approach of deconvolution to speech [OS68]. To extract these coefficients from an audio recording, the audio samples are first divided into short (e.g., 25 msec) overlapping segments, or frames. The signal in this frame is often multiplied by a window function. Then, the Fourier power spectrum is calculated from the (windowed) speech excerpt. To the power spectrum, a mel-frequency filter bank is then applied, which has higher resolution in the lower frequency bands to reflect human speech recognition [DM80]. At each mel frequency, a logarithm is computed for the filtered spectrum. Finally, the discrete cosine transform is performed on the sequence of mel log powers. In the resulting sequence, the components near the origin reflect the overall envelope of the mel log power sequence, and the components far from the origin reflect the fine structure of that sequence. Thus, the information for overall spectral envelope can be represented by taking the coefficients near the origin. Recalling that the spectral envelope reflects the vocal tract shape of the talker, this representation may contain physical information of the talker identity. In addition, because the spectral envelope changes by each phone, a specific pattern of a talker's pronunciation can be reflected in this representation. For ASV applications, 20 or 24-order MFCCs are typically used, along with their first and second time derivatives.

In text-independent ASV tasks, however, the sensitivity of MFCCs to phonetic content might be a major cause of performance degradation [DP18]. Hence, various features that are thought to be less sensitive to such variability were proposed to improve system performance. For example, [DP16] used features derived from the linear prediction residual signal to represent voice source characteristics. These features improved the system performance by providing additional or complementary information to conventional cepstral features on text-independent tasks when the talkers were modeled with 2.5-min-long utterances and tested

with short utterances (2–10 sec). Other studies have shown that the phase components of the speech signal are important for talker identity [VRS16], and such information could be used for text-independent long-utterance ASV. Approaches to capturing talker-specific prosody have also been proposed [RAC03, SFK05, DDK07].

1.3 Voice Discrimination by Humans and Machines

1.3.1 Voice Discrimination by Humans

For humans, discriminating unfamiliar voices is a separate decision-making process from recognizing familiar voices [VK87]. While familiar talker recognition can be thought of as a gestalt-matching task, unfamiliar voice discrimination additionally involves acoustic feature comparisons. Several studies have shown that the perception of an unfamiliar voice requires both a generic talker pattern that acts as a mental reference and a talker-specific pattern that deviates from that reference [KS11, Chap. 5.3.4,]. Such a standard pattern, acquired over a lifetime, includes both how human voices generally sound and what aspects of speech are related to the talker's identity.

Internal response: the decision space

In human voice discrimination experiments, binary decisions (same versus different talker) of human listeners are collected. Resulting human responses can be evaluated and analyzed with signal detection theory (e.g., [MC05]). Signal detection theory assumes that a listener in the voice discrimination experiment is judging a single stimulus attribute; talker similarity between two utterances. A distribution of similarity values can be obtained by repeated presentations. A listener's internal response can be represented with such distribution, although it is not yet clear which and how acoustic features are used to construct a mental talker model, and how the similarity between voices is assessed.

The top panel of Figure 1.2 presents the probability density of similarity values between utterances in same-talker stimulus pairs (target trials, or S_T). On average, target trials are

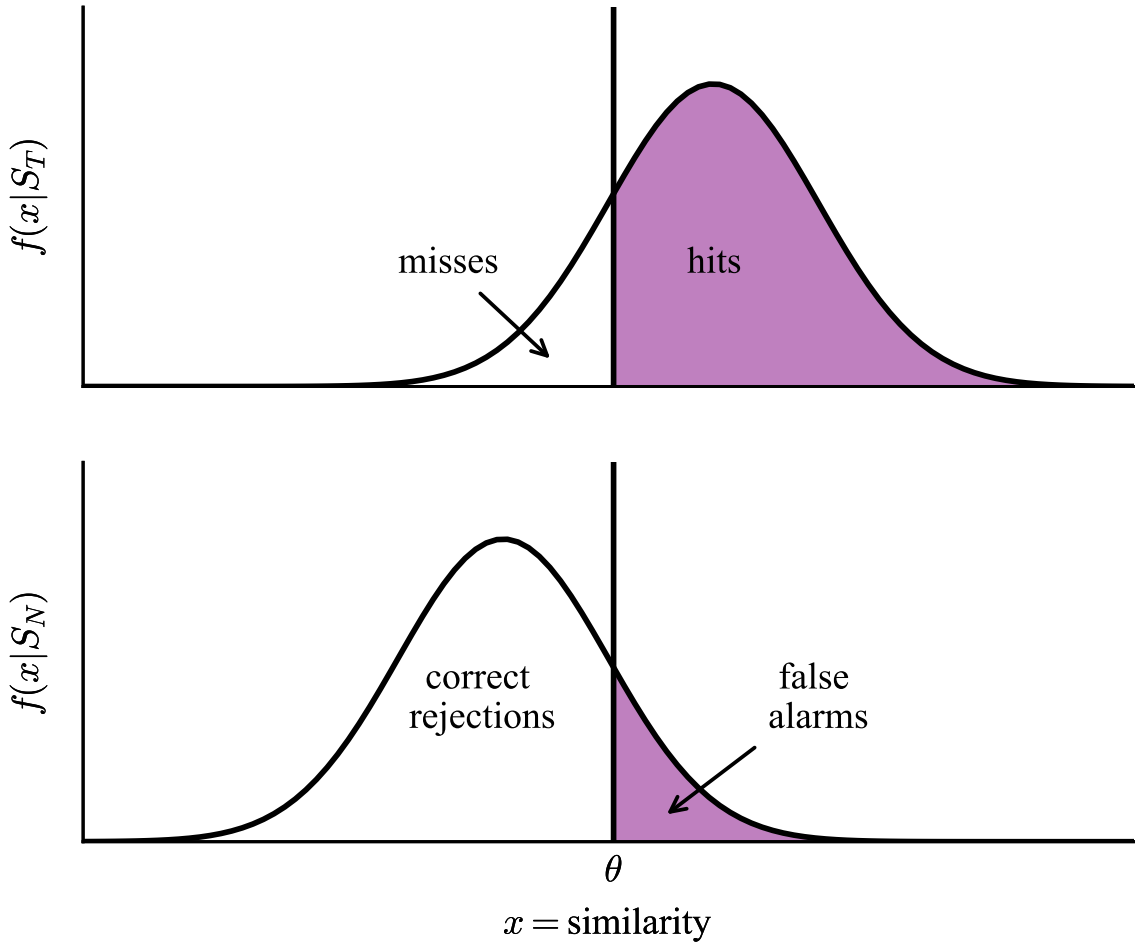


Figure 1.2: Underlying distributions of similarity of target and non-target trials. Top panel shows distribution due to target trials (S_T); values above decision criterion θ lead to hits (shaded area), those below to misses. Bottom panel shows distribution due to non-target trials (S_N); values above the criterion lead to false alarms (shaded), and those below to correct rejections. Adapted from [MC05].

Table 1.1: Stimulus-response matrix. A correct "same talker" response is termed a hit, while an incorrect such response is a false alarm. A correct rejection and a miss are defined similarly for a "different talker" response.

	Response	
Stimulus class	"Same talker"	"Different talker"
Same-talker pair (target trial; S_T)	hits	misses
Different-talker pair (non-target trial; S_N)	false alarms	correct rejections

more similar within a pair than non-target trials. Thus, the distribution of similarity for different-talker stimulus pairs (non-target trials, or S_N) is displaced to the left of target trials. Unless a listener show perfect performance, there must be some similarity values that the listener finds ambiguous, that could have arisen either from an target or a non-target trial. The two distributions together compose the *decision space* – the listener's internal or underlying response. The listener can assess similarity between the utterances in a pair, but does not know which distribution led to that value.

A listener can make a same versus different talker decision by applying a decision threshold θ on the assessed similarity value. A similarity value greater than θ will result in a "same talker" response; otherwise a "different talker" response. The probability that a similarity value above θ will occur is the proportion of the area under the curve above θ, which is indicated as shaded area in Figure 1.2. Correctly identifying a same-talker pair is termed a hit, and failing to do so is a miss. Incorrect "same talker" decision is a false alarm, and correct "different talker" decision is a correct rejection (see Table 1.1).

Performance evaluation metric: sensitivity

Listener performance can be analyzed in terms of the hit and false alarm rates. The hit rate is the proportion of "same talker" responses to target trials (S_T), and the false-alarm rate is the proportion of incorrect "same talker" responses to non-target trials (S_N). That

is, the hit and false alarm rates can be written as probabilities of "same talker" responses conditional on the possible stimulus event:

$$\begin{aligned} \text{hit rate} &= P(\text{"same talker" response}|S_T) \\ \text{false alarm rate} &= P(\text{"same talker" response}|S_N) \end{aligned} \tag{1.1}$$

In signal detection theory, the listener's ability to discriminate is assessed in terms of sensitivity. The sensitivity measure, d' (d-prime), is defined in terms of Z, the inverse of the normal distribution function:

$$d' = Z(\text{hit rate}) - Z(\text{false alarm rate}) \tag{1.2}$$

Listeners are assumed, by detection theory, to have fixed sensitivity when asked to discriminate a stimulus pair. One aspect of responding, however, is their *response bias*, which is their willingness to say "same talker" rather than "different talkers". The locus of possible hit and false alarm rates pairs that yield a constant sensitivity is called *receiver operating characteristics* or *relative operating characteristics* (ROC, [Swets,1973]). Figure 1.3 shows ROCs associated with d'. The horizontal axis corresponds to the false alarm rate, and the vertical axis to the hit rate. If a listener is highly sensitive, (e.g., $d' = 3$), the curve approaches the upper left corner. On the other hand, a wild guess (e.g., $d' = 0$) is represented by a straight line between the origin and the upper right corner. The area under the ROC curve (AUC) can be also used as a performance metric: the larger this value is, the more sensitive (or accurate) the listener is.

Duration and phonetic content effects on human performance

Even though results vary widely depending on the experimental protocol used, humans are reasonably accurate at distinguishing unfamiliar talkers even with short utterances. For example, [KP91] found that humans had a hit rate of 88.6% and a false alarm rate of 19.7% ($d' = 2.02$) in a voice discrimination task with single-sentence (≈ 2 sec) pairs. Human performance generally improves as the utterance length increases until it plateaus with utterances longer than 60 seconds [BP66, LGP84]. Authors disagree on why longer stimuli produce better results. [RW93] found evidence supporting the hypothesis that the advantage of longer

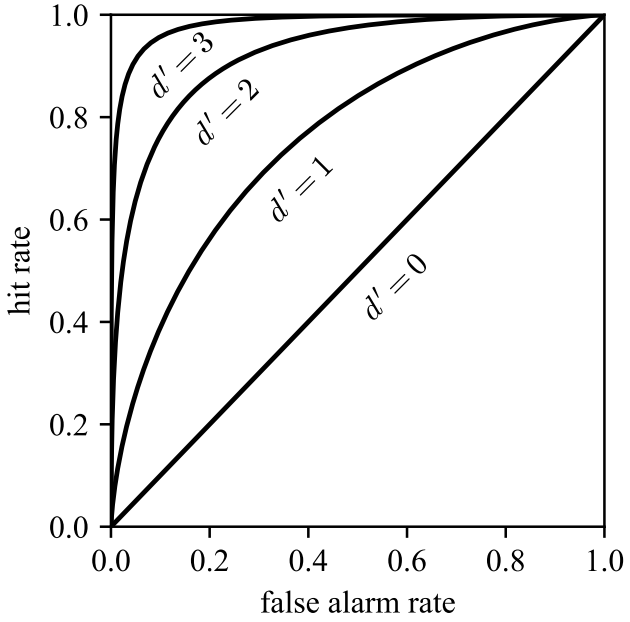


Figure 1.3: Receiver operating characteristics (ROC) curves. Curves connect locations with constant d'. Adapted from [MC05].

stimuli is in broader coverage of phonetic content. However, [CW97] argued that the critical factor was not the number of different sounds heard, but rather the duration of the utterances due to talker-specific prosody, speaking rate, and other non-phonemic aspects of the speech signal that are more pronounced in longer utterances [KS11, Chap. 7.3.1,].

Speaking style and emotion effects on human performance

The effect of speaking style variability on human voice discrimination has not been studied extensively. Studies in forensic talker identification note that speaking style mismatch between a criminal's voice heard at a crime scene and speech samples collected in a voice lineup (e.g., shouting versus reading) might confuse earwitnesses (see [Jes08]). In a voice discrimination context, we expect such speaking style variability to cause a significant performance degradation based on results from a few studies dealing with emotion variability (e.g., [SY80]). In that study, when a target voice changed tone (related to emotion or affect), mean 'hit-miss' and 'false alarm-correct rejection' scores decreased significantly.

1.3.2 Voice Discrimination by Machines

For machines, automatic speaker verification (ASV) is a task of "enrolling" a talker with one or more utterances from that talker, and "testing" whether a new utterance is from the same talker as the enrolled one or not. Automatic voice discrimination can be thought of as a special case of ASV where the talkers are enrolled with one utterance. In this dissertation, the term ASV is used to refer to automatic voice discrimination unless otherwise specified.

Talker representation: i-vector

Standard ASV systems are typically pre-trained with large amounts of data from a large number of talkers. Hundreds of hours of recordings are used to train a statistical model for human speech, called a *universal background model* (UBM, [RQD00]). A new utterance can be thought of as a deviation from the UBM. The nature and extent of the deviation, however, will be influenced by both talker-specific and utterance-specific information. Thus, these systems need to minimize within-talker variability while maximizing between-talker variability. Hundreds of additional hours of recordings are used to train a subspace onto which the deviation is projected. The projected low-dimensional vector, referred to as an *i-vector* [DKD11], is thought to represent talker identity. When the system receives a pair of speech samples as inputs, an i-vector is found for each utterance. Then, the likelihood that the i-vectors represent the same talker is calculated based on the pre-trained model and subspaces. Probabilistic linear discriminant analysis (PLDA, [KSO13]) is often used to calculate this likelihood. The system then applies a threshold to the likelihood to make a same versus different talker decision.

Automatic voice discrimination can be viewed as analogous to perceptual voice discrimination, although the latter is much more complicated than statistical pattern recognition based on frame-level features. That is, the pre-trained UBM and subspaces are analogous to a human's pre-existing idea of the average talker model and the manner in which a new voice differs from it. Such a model represents the life-long experience of the listener with voices and internal structuring that is not yet understood. Despite this analogy, however,

differences presumably exist between the talker-distinguishing strategies used by humans and machines as evidenced from their different performances for various challenging tasks. Challenging conditions include very short utterances ($\approx$ 2 sec), text-independent tasks, and speech spoken in different styles.

Evaluation metrics

There are mainly two types or errors in ASV. They are false acceptance (granting access to an imposter talker) and false rejection (denying access to a legitimate talker). These are equivalent to false alarm and miss, respectively, in the context of human performance evaluation discussed earlier. ASV systems generally yield a scalar output (e.g., PLDA score). This scalar variable represents the similarity between the enrollment and test utterances. To make a binary decision, the system applies a threshold (θ) to the similarity score.

Similar to the perception case, an ROC curve can be derived from the computed similarity score. That is, the pairs of false alarm and hit rates for different threshold values can be calculated, and in an ROC curve that represents the system's sensitivity or accuracy. A widely-used measure for ASV systems is the equal error rate (EER). The EER is defined as the false acceptance or false rejection rate value resulting from the threshold which makes the two rates equal.

Duration and phonetic content effects:

On very short utterances, machine performance degrades substantially, but humans are more robust than machines. For example, a state-of-the-art ASV system had an EER of 22.31% with 2-sec-long pairs, while its EER was 3.38% for 20-sec-log pairs [DJP16]. Human listeners, for single sentence ($\approx$ 2 sec) pairs, showed 11.4% miss and 19.7% false alarm rates [KP91]. As mentioned earlier, one reason for the degradation with shorter utterances could be that there is insufficient phonetic coverage for the machines to infer appropriate statistics. Text dependency also affects machine performance. For example, when utterances are short ($<$ 10 sec), matching phonetic content by using same-text pairs yields error rates that are

approximately half those of the text-independent pairs [DP16, PYK17]. One exception occurs when short digit sequences (< 2 sec) are used. In that limited-vocabulary case, performance can reach 95% accuracy or higher [LLM14].

Speaking style and emotion effects:

Although the effect of speaking style on ASV mismatch has not yet been studied extensively, some studies on emotion variability are available. For example, [PZH17] reported that an emotion mismatch between utterances degraded ASV system performance, which worsened as the utterance length decreased from 11 to 2.75 seconds for naturalistic (not acted) expressive voices. However, because that study did not compare matched emotion conditions, the amount of degradation that can be attributed to emotion variability is not clear. In [NB01], it was noted that the performance of a talker identification system degrades when trained with spontaneous speech and tested on read speech, compared to when spontaneous speech was used for both training and testing, even though the utterances were long (29 sec). The system was a closed-set talker identification task, which is not directly comparable to talker verification tasks, but it is expected that ASV performance might also decrease due to speaking style differences.

1.3.3 Comparison between Humans and Machines

In order to precisely understand which strategies are shared by and/or differentiate voice discrimination by humans and machines, a direct comparison between them using the same stimuli pair is needed. Due to the fact that evaluating human listeners with a large number of utterances demands much time and cost, making a direct comparison between humans and machines in a statistically significant manner has been challenging. Nonetheless, efforts have been made to make such comparisons in the past.

In 2010, the National Institute of Standards and Technology (NIST) presented the human assisted speaker recognition (HASR) task for evaluating systems that combined humans and machines [GMB10]. The task was designed in a way such that the most difficult test samples

are selected for the evaluation (channel mismatch, noise, same/different talker that sound highly dissmimilar/similar, etc.). However, the total number of trials in these experiments was low (15 trials for HASR1 and 150 trials for HASR2) compared to evaluations designed for automatic systems. The HASR study was repeated during the 2012 NIST SRE where both noisy and channel degraded speech data were encountered.

In the NIST HASR tasks, machines were able to perform better than human-assisted approaches [HKN10, SCS11a, RFG11, KAR11, GMD11]. In [GHR13], it was shown that human and machine decisions were complementary, meaning that in some cases the humans correctly identified a talker where the automatic system failed, and vice versa. However, the HASR tasks were exceptionally difficult for human listeners because of the severe channel mismatch, unfamiliarity with the talkers, noise, and other factors. Another important factor to note is that the trials consisted of long utterances (2.5 min). Because humans do not need such long utterances, and they even could become tired while listening to the stimuli, some studies presented human listeners shorter stimuli than those given to machines. In [KAR11], for example, human listeners were given 6-sec long excerpts while machines were given the full utterances. Considering that machine performance considerably degrades with short utterances, as noted above, the results could be very different if short utterances were used for the comparison.

In the forensic voice comparison literature, attempts to integrate human and machine responses exist. In [HHF17], authors analyzed falsely accepted (1 pair) and falsely rejected pairs (13 pairs), and found that forensic experts were able to resolve the classified pairs.

However, to our knowledge, a direct and detailed comparison between human and machine voice discrimination under various conditions of intrinsic within-talker variability has not yet been made, in part because a proper database was not available to undertake such studies. In this dissertation, the recently developed UCLA database enabled detailed analyses of the effects of intrinsic within-talker variability.

1.4 Psychoacoustic Model of Voice Quality

1.4.1 Previously Proposed Acoustic Features for Voice Quality

Voice quality (or timbre) is often understood, in a broad sense, as perceptual responses to a voice. Note that, although it is often related only to laryngeal activities, both laryngeal and supralaryngeal features contribute to quality. Voice quality has been frequently associated with identity, as "physical, psychological, and social characteristics [Lav80]" of a voice. For example, in an international survey on forensic speaker comparison practices, voice quality was reported most often (33%) as the most useful feature within linguistic, phonetic, and acoustic domains [GF11].

Despite its potential effectiveness, applying voice quality features to automatic speech processing applications has been difficult. One of the difficulties is that voice quality is often described with impressionistic terms, such as tense, harsh, and breathy. These terms can be interpreted differently based on the the researcher's understanding. In addition, it is difficult to automatically extract voice quality attributes directly from the speech signal. Numerous studies to automatically represent voice quality with acoustic feature vectors have been proposed (see [MRD09] for review). Yet, the reliability of automatic algorithms to extract voice quality attributes directly from the speech signal is still limited. One approach involves inverse-filtering, to identify the voice source characteristics, which is based on the assumption that voice quality can be measured by estimating the glottal source signal. Other techniques have been developed to estimate parameters that represent voice quality directly from temporal fluctuations of the signal's periodicity (e.g., jitter and shimmer [LTJ07]), but the relationship between such parameters and perceptual responses is questionable [KG05]. Other researchers proposed spectral features, such as the amplitude difference in dB between the first and second harmonic (H_1^*-H_2^*), between the first harmonic and first formant amplitude (H_1^*-A_1^*), and between the first harmonic and third formant (H_1^*-A_3*), where the asterisks (*) indicate the effects of formants were corrected [Han97, HC99]. In addition, cepstral peak prominence (CPP, [HCE94]) and harmonic-to-noise ratio (HNR, [Kro93]) that measure signal periodicity in the cepstral domain have also been proposed, and are often

related to breathiness.

The above mentioned acoustic features to quantify voice quality could be utilized in some applications for speech signal analysis and for automatic speech processing tasks. However, it is not clear if these features can represent all aspects of voice quality.

1.4.2 Development of the Psychoacoustic Model of Voice Quality

Although a number of acoustic features has been proposed to represent voice quality, studies of voice quality had limited explanatory power in terms of the relationship to voice production and perception. Kreiman et al. proposed a psychoacoustic model of voice quality [KGG14, GSG16]. They aimed at constructing an acoustic model that bridges the gap between voice production and perception. They noted that listeners perceive voice quality as an integral pattern, rather than a bundle of separate features [SHS97]. For example, the perceptual importance of a given feature depends on the values of other attributes of the pattern, and not solely on the value of the feature itself [VKE85, VKW85].

In order to quantify the entire voice pattern, that study applied analysis-by-synthesis to completely recreate the perceived voice pattern. Through an extensive series of perception experiments [KGA07, KG10, KG12, GKE13, GSK13], the authors showed that listeners are perceptually sensitive at six parameters, and that, as a set, the parameters are sufficient to quantify source contributions to normal and most pathological voice quality. They were harmonic source spectrum (4 parameters), inharmonic (noise) source, and temporal source frequency (F_0). For the harmonic source spectrum, four acoustic parameters of the voice source spectral model are proposed (see Figure 1.4): the differences in dB between the amplitudes of the first two harmonics (H_1-H_2), the second and fourth harmonics (H_2-H_4), the fourth harmonic and the harmonic nearest to 2 kHz, and the harmonic nearest 2 kHz and that nearest 5 kHz (H_{2k}-H_{5k}). The inharmonic source can be measured in terms of harmonic-to-noise ratio (HNR), which measures harmonic energy normalized by the shaped noise spectrum level [Kro93].

This approach provided a perceptually valid acoustic model using a rather small set of

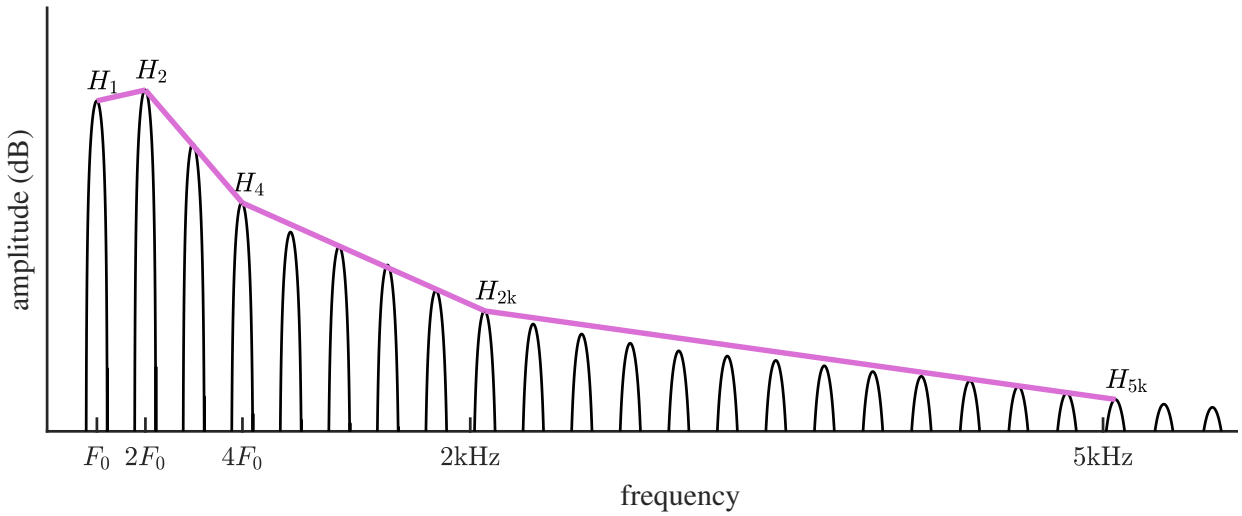


Figure 1.4: A schematic for the source spectral model for the voice quality.

parameters. The model can be represented with quantitative measures that can be directly extracted from speech signals, which in turn facilitates applications in automatic signal processing.

1.5 Dissertation Outline

The rest of this dissertation is organized as follows:

In Chapter 2, the databases used in the experiments reported in this dissertation are presented. A new database developed at UCLA to systematically study between- and within-talker variability is described in this chapter.

In Chapter 3, results from an initial set of experiments are reported. Those experiments aim at exploring the role of voice quality features, based on the psychoacoustic model, in predicting human voice discrimination responses and in improving automatic voice discrimination performance.

In Chapter 4, a modified voice quality feature set is introduced that was determined using continuous speech samples and with different speaking styles. Then, ASV experiments with a larger number of talkers are presented to analyze how much each type of variability

impact system performance.

In Chapter 5, our approaches to finding relevant acoustic information of talker identity and the way it is used by humans for continuous speech are discussed.

Additionally, inspired by the observations suggesting that voice quality features vary according to the talkers' emotional status, the features were applied to an emotion recognition task (Chapter 6).

In Chapter 7, key results of all studies are summarized, and directions for future work are suggested.

CHAPTER 2

Databases

Two databases were used in this dissertation. The first is the UCLA Speaker Variability Database, which was recently developed to analyze between- and within-speaker variability. The recordings in this database were used for perceptual voice discrimination experiments and automatic speaker verification (ASV) experiments. The second database was from the National Institute of Standards and Technology (NIST), and was used to pre-train ASV systems.

2.1 The UCLA Speaker Variability Database

In order to systematically study both between- and within-talker variability, a multi-talker speech database including multiple speech tasks per talker is needed. To our knowledge, none of the existing multi-talker speech databases offers the desired combination of a large number of talkers (both female and male), multiple recording sessions per talker, multiple speech tasks per talker, and very high quality audio (controlled recording conditions, good quality microphone, high sampling rate, etc). Thus, we developed the UCLA Speaker Variability Database [KPK15, KKA18] including multiple recordings of talkers recorded in a variety of speaking tasks and on multiple occasions.

Parts of this chapter were published in [KPK15].

2.1.1 Equipment

Audio recordings were made in a sound-attenuated booth using a 1/2" Brüel & Kjær microphone with a sampling rate of 22 kHz. The microphone was suspended from a baseball cap worn by the talker for a fixed mouth-to-microphone distance. All speech was elicited via on-screen displays.

2.1.2 Subjects

More than 100 female and 100 male UCLA undergraduate students were recorded across different recording sessions. None of the participants had diagnosed speech or hearing problems.

2.1.3 Materials

Talkers were recorded in three separate sessions on different days and in different speech styles. Note that the goal of recording speech from different conditions was to sample normal, daily-life voice variation. We did not try to elicit voice disguises, impersonations, acted emotions, or other dramatic acting. Instead, we focused on normal variability in real-life situations, to the extent that these could be elicited in a sound booth. The point of the different conditions was not to study them as such, but simply to enhance the likelihood of sampling realistic amounts of within-talker variability in voice quality.

At the beginning of each session, talkers repeated the sustained vowel /a/ (as in the word "spa") three times and read ten sentences. These tasks allow cross-section comparison. The vowel /a/ was chosen because its high F1 reduces errors when estimating voice source parameters. Read sentences consisted of 2 repetitions of 5 Harvard sentences [IEE69], read in all three recording sessions for a total of 6 repetitions of each sentence and 30 sentences overall. The sentences were "The boy was there when the sun rose.", "Kick the ball straight and follow through.", "Help the woman get back to her feet.", "A pot of tea helps to pass the evening.", and "The soft cushion broke the man's fall." These sentences were used to

study phonetic content variability.

Each session then included two further speech tasks, different in each session, for a total of six one-time-only speech tasks. In the first session, participants were instructed to talk to the research assistant (RA) who was outside the booth, giving her either directions on how to go somewhere, or instructions on how to do something. Some suggestions about what to talk about were provided on printed sheets (see Appendix A for examples). Talkers were told to speak for at least 30 seconds, and an on-screen display counted out 30 seconds. This task provided a sample of clear but unscripted speech. Next, participants were instructed to repeat to the RA a conversation they had recently that wasn't important – not exciting, nor upsetting, just normal. Again, some possible topics were provided, and again, the on-screen display prompted for 30 seconds of speech. This task provided a sample of unscripted low-affect speech.

In the second session, participants were instructed to repeat to the RA a conversation they had recently about something exciting that made them really happy. As before, some possible topics were provided and the on-screen display prompted for 30 seconds of speech. This task provided a sample of positive-affect speech. Next, participants used their cell phones to call a friend or relative and talked for at least two minutes. Only the participant's side of the conversation was recorded. This task provided a sample of unscripted conversational speech.

In the third session, participants were instructed to repeat to the RA a conversation they had recently about something that really annoyed them. As before, some possible topics were provided and the on-screen display prompted for 30 seconds of speech. This task provided a sample of negative-affect speech. Finally, a sample of pet-directed speech was collected. Talkers were instructed to talk to pets displayed in a video. They could choose between a kitten video (2 min 36 sec) and a puppy video (1 min 51 sec). Resulting utterances were often (but not always) characterized by exaggerated prosody, similar to infant-directed speech [BKV02]. The speech tasks in each session and the resulting amount of speech are summarized in Table 2.1.

Table 2.1: Summary of the UCLA Speaker Variability Database. The last column shows the total amount of speech for all 3 sessions for each speech task per speaker.

	Session A	Session B	Session C	Total amount
Sustained vowels	3 tokens (3 sec)	3 tokens (3 sec)	3 tokens (3 sec)	$\approx$ 10 sec
Read sentences	10 sentences ($\approx$ 25 sec)	10 sentences ($\approx$ 25 sec)	10 sentences ($\approx$ 25 sec)	$\approx$ 75 sec
Other speech task (unscripted)	instructions (25–30 sec)	phonecall (60–120 sec)	talk to pet video (60–120 sec)	$\approx$ 145–270 sec
Reported conversations (unscripted)	neutral (25–30 sec)	happy (25–30 sec)	annoyed (25–30 sec)	$\approx$ 75–90 sec
Total	$\approx$ 75–90 sec per speaker	$\approx$ 110–180 sec per speaker	$\approx$ 110–180 sec per speaker	$\approx$ 300–450 sec per speaker

2.2 NIST SRE Database

The Speaker Recognition Evaluation (SRE) databases developed by NIST are often used to train a universal background model (UBM) and speaker variability subspaces. We used the NIST SRE04, 05, 06, and 08 databases [PM04, PML06, MG09] for this purpose. These databases provide more than 3,000 hours of speech samples from 2,692 female and 1,115 male talkers, over a variety of channels including telephone speech, microphone, and "interview" speech.

Note that although the SRE databases offer many recordings from a large number of talkers with multiple speech tasks, they do not provide multiple speech tasks per talker under controlled recording environments. Thus, the UCLA Speaker Variability Database is more suitable for detailed performance analyses in terms of within- and between-speaker variability.

CHAPTER 3

Initial Experiments Based on the Psychoacoustic Model of Voice Quality

The first set of experiments evaluated voice discrimination abilities for humans and machines using the UCLA Speaker Variability Database. The main focus was on exploring the role of psychoacoustically-valid acoustic features of voice quality in predicting human responses and in improving automatic speaker verification (ASV) performance.

3.1 Voice Quality Feature Selection Using Sustained Vowels

3.1.1 Stimuli

The voices of five female talkers were selected at random from the UCLA Speaker Variability Database. In this section, only the 9 tokens of the sustained vowel /a/ were studied, given the explanatory nature of this study.

3.1.2 Selection of Measures and Data Reduction

Based on a series of psychoacoustic studies of voice quality [KG12, GSG16] described in Chapter 1, six parameters for the spectral model of the voice source were found to be most effective. They are F_0, H_1^*-H_2^*, H_2^*-H_4^*, H_4^*-$H_{2\mathrm{k}}^*$, $H_{2\mathrm{k}}^*$-$H_{5\mathrm{k}}$, and the harmonic-to-noise ratio (HNR). Here, H_N^* denotes the N-th source spectral harmonic magnitude, and the asterisk

Parts of this chapter were published in [KPK15] and [PSK16].

(*) indicates a correction for the influence of vocal tract resonances using the formula given in [ISA07]. Other measures proposed in the literature on voice quality were included for the sake of completeness. These included measures of HNR in 4 discrete frequency ranges, the root-mean-squared energy, and cepstral peak prominence (CPP), H_1^*-A_1^*, H_1^*-A_2^*, and H_1^*-A_3^*, where A_N denotes the amplitude of the harmonic closest in frequency to the N-th formant. Finally, measures of F_1, F_2, and F_3 were included because vowel quality differed substantially across (and occasionally within) talkers.

Measures for all parameters were made using a 25-msec Hamming window with a 1-msec frame interval across the entire duration of each vowel token. The parameters were moving-averaged over a 100-msec span and then were sampled every 100 msec. Source measures were extracted with the VoiceSauce toolkit [SKV11], and formant frequencies were measured using the Snack algorithm [Sjo04] which was included in the toolkit. Measures were screened for outliers (which were treated as missing values), and source spectral measures were validated using analysis-by-synthesis [KAG10].

Correlation and canonical correlation were used to examine patterns of association among the various acoustic variables along with F_0. The 4 HNR measures were significantly and substantially intercorrelated with each other and with CPP (mean $r = 0.95, p < .001$), so only CPP was retained for subsequent analyses. Similarly, canonical correlation indicated that a set comprised of H_1^*-H_2^*, H_2^*-H_4^*, H_4^*-$H_{2\mathrm{k}}^*$, and $H_{2\mathrm{k}}^*$-$H_{5\mathrm{k}}$ was highly correlated with another set of H_1^*-A_1^*, H_1^*-A_2^*, H_1^*-A_3^*, and energy ($R^2 = 0.88$), so only the first set of variables were retained. Finally, F_0 and the first three formant frequencies were retained in the final set of measures. Note that the final set of 9 acoustic measures is equivalent to the psychoacoustic model; the observed correlations between model parameters and other variables suggest that adding parameters to the model would not increase its explanatory power.

All nine measures were normalized to a 0–1 scale using a range for each variable for females' vowel /a/ sounds as observed in previous studies [KG12, BO00, HGC95]. The measures and their normalization ranges are shown in Table 3.1. This initial feature set was denoted as VQual1.

Table 3.1: Normalization ranges for the parameters for females' vowel /a/ sounds from [KG12, BO00, HGC95].

	F_0	CPP	H_1^*-H_2^*	H_2^*-H_4^*	H_4^*-$H_{2\mathrm{k}}^*$	$H_{2\mathrm{k}}^*$-$H_{5\mathrm{k}}$	F_1	F_2	F_3
Min	93	15	-2.64	-0.39	0	0	522	963	2138
Max	275	32	21.6	29.2	37.7	46.87	1163	2701	3490

3.2 Perceptual Voice Discrimination Experiments

3.2.1 Stimuli

Speech samples from the five female talkers selected in the previous section were used for analysis. The sustained /a/ sounds from the 5 talkers and read sentences from 3 talkers, among the 5 talkers, were studied. All vowel tokens from the 3 sessions were used, for a total of 9 vowels per talker. Read sentences were chosen from different sessions and with different content. Two sessions per talker were selected, with 2 repetitions of 2 different sentences per session, for a total of 8 sentences per talker. The selected sentences were "A pot of tea helps to pass the evening," and "The soft cushion broke the man's fall". Note that all samples are at most 3-sec long.

3.2.2 Method

Two sets of experiments, one with sustained vowels and the other with read sentences, were conducted. For the first set of experiments using the sustained vowel /a/ sounds, the full unedited vowel samples were used to ensure that idiosyncratic vocal features like final creak or pitch declination were represented. Given nine recordings from five talkers, the stimulus set included a total of 180 "same talker" pairs and 450 "different talker" pairs, for a total of 630 possible comparisons among stimuli. Two different randomizations of this set were created, each of which was divided into thirds to create 6 subsets of 210 listening trials.

Listeners were normal-hearing UCLA students and staff members. They received pay-

ment or class credit for their participation. Ten listeners were assigned at random to each subset, for a total of 60 listeners in 6 groups; but across groups, each pair of stimuli was judged by 20 listeners. Listeners heard the pairs of stimuli over Etymotic insert earphones (model ER-1) at a comfortable constant listening level (interstimulus interval = 250 msec). Each pair could be played only once in each presentation order (AB/BA). Listeners were not told how many talkers were represented in the trials. For each pair of stimuli, listeners judged whether the voices represented one talker or two different talkers, and reported their confidence in their response on a scale from 1 (positive) to 5 (wild guess). The experiment was self-paced and listeners were encouraged to take as many breaks as needed. Testing lasted, on average, about 45 minutes per listener.

The second set of experiments was conducted with read sentences from 3 female talkers. From a total of 24 tokens, 30 same-talker pairs and 48 different-talker pairs were created. For the sentence listening experiments, 15 normal-hearing UCLA students and staff members participated. As in the vowel listening experiments, the experiments were self-paced and listeners were encouraged to take breaks as needed.

3.2.3 Human Listener Performance

Results of the perceptual experiments, in terms of accuracy (hits & correct rejections), are shown in Figure 3.1. For sustained vowels, listeners averaged 69.0% accuracy (sd: 10.43%, range: 38.6% – 84.8%). In comparison, using sentence stimuli resulted in higher accuracy ranging from 65.4% to 97.4%, with a mean of 89.0% (sd: 8.21%).

The dissimilarity score δ from an individual listener was calculated from the listener's same/different talker response and the uncertainty u which was reported on a 1 (positive) to 5 (wild guess) scale in the following way:

$$\delta = \begin{cases} u & \text{, for ``same talker'' response} \\ 11 - u & \text{, for ``different talker'' response.} \end{cases} \tag{3.1}$$

The resulting dissimilarity δ ranged from 1 to 10. It can be inferred that the same versus

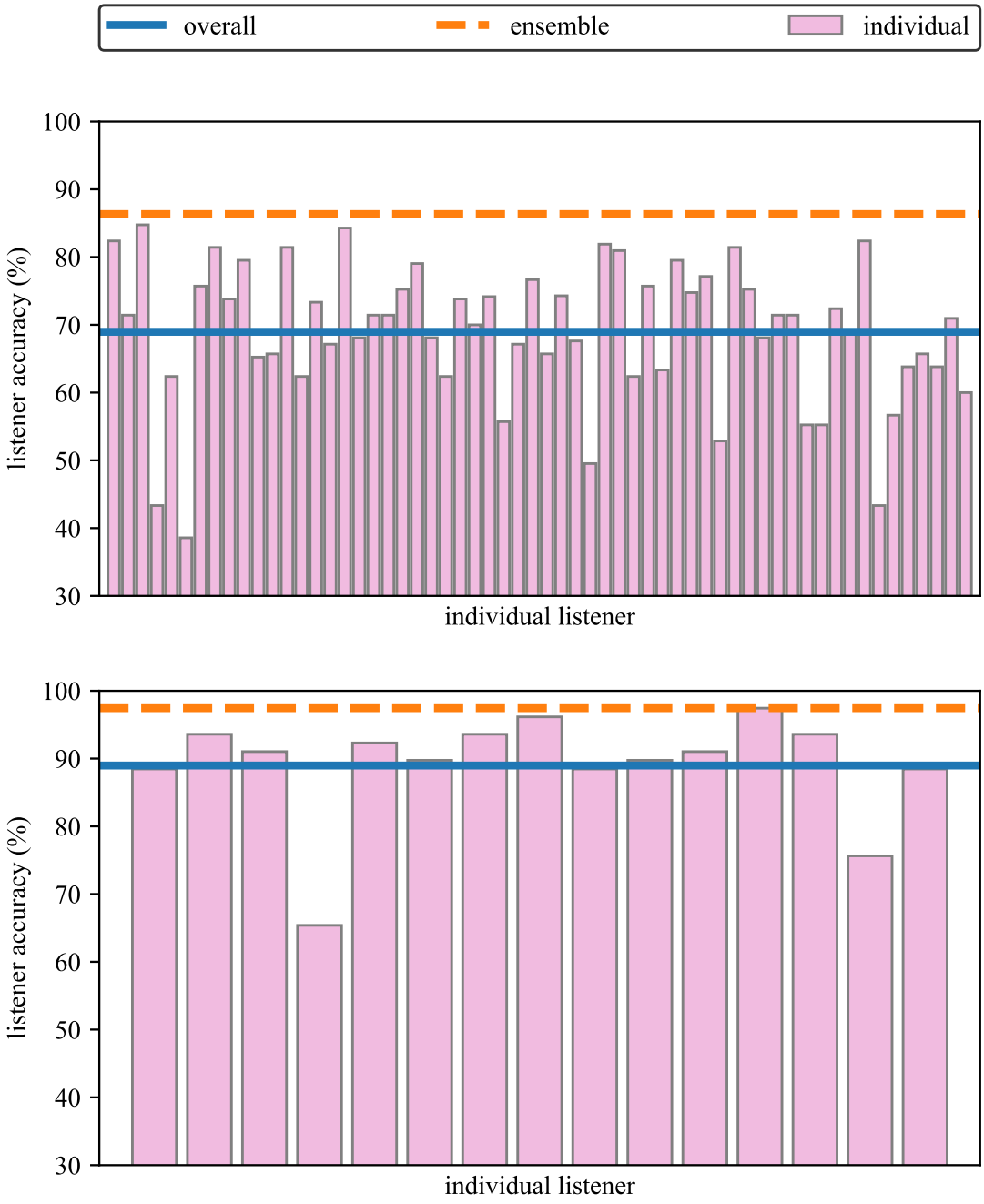


Figure 3.1: Listener performance (accuracy) in identifying same/different talker pairs with sustained vowel /a/ sounds and 60 listeners (top), and read sentences and 15 listeners (bottom). The mean value of accuracy of individual listeners is indicated with a solid line, and the recalculated accuracy from averaged voice dissimilarity score across listeners is indicated with a dashed line.

different speaker decision threshold was at 5.5. For example, if a listener responded that she/he was positive that the two tokens were from the same talker ($u = 1$), then δ is low (< 5.5). On the other hand, if the response was "positive these are different talkers", then δ is high (> 5.5). The δ scores for each stimuli pair were then averaged across listeners, and the averaged dissimilarity was denoted as $\bar{\delta}$. For identical token pairs, which were not included in the perception experiment, $\bar{\delta} = 0$ was assigned.

It was observed that when the same/different talker decision was re-calculated based on the ensemble score $\bar{\delta}$, accuracy increased substantially for both vowels and sentences, as shown in Figure 3.1. The accuracy gain was more obvious in the vowel case, for which ensemble accuracy reached 86.4% (compared to 69.0% for averaged individual listener accuracy), which was higher than the score of the best individual listener. For sentences, the ensemble accuracy increased to 97.4% versus 89.0% for average data. These results are consistent with the finding that although individual naïve listeners vary in their performance, an aggregation of their responses was more accurate than the best performing listener [SCS11b]. Therefore, it should be noted that the accuracy of combined listener responses does not represent the accuracy of an average listener, and the former can be substantially higher than the latter.

3.3 Modeling Human Responses

3.3.1 Method

Multi-dimensional scaling (MDS, [KW78]) was used to compute the distance between tokens in a perceptual space. Here, the objective of using MDS was to obtain a perceptual distance, and not to reduce dimensionality. Therefore, a higher dimensional MDS than usual was used to represent the dissimilarity between stimuli [JAA07, MNH15]. The averaged dissimilarity score $\bar{\delta}$ was normalized to have a value between 0 and 1, and the normalized score was analyzed using a 6-dimensional non-metric MDS (vowels: stress= 0.058, $R^2 = 0.8813$; sentences: stress= 0.0004, $R^2 = 0.9806$). The Euclidean distances between token pairs of all possible combinations were calculated in the MDS space. The resulting token distance had a 0 to 1

range.

Two sets of features were used separately and in combination to predict human responses. The first set consisted of 20-order MFCCs along with their first- and second-order derivatives (60-dim in total). This set is a standard feature set for ASV systems and was used as a baseline in this study. The second set was the 9-dim voice quality feature set (VQual1, described in Section 3.1.2) with their first- and second-order derivatives (27-dim in total). The mean and standard deviation of each feature, including derivatives, within a token were calculated, resulting in 120-dim for MFCCs and 54-dim for VQual1. The absolute differences in the feature means and standard deviations were then found between tokens. Perceptual dissimilarities were predicted with a linear regression framework. Here, the variable being predicted was the Euclidean distance between two tokens in the MDS perceptual space, and the predictors were the differences in means and standard deviations between the two tokens. MFCCs and VQual1 were used individually or combined by concatenating them together before linear regression.

3.3.2 Results and Discussion

The human response prediction results in terms of root-mean-squared error (RMSE) between the predicted value and the token distance in the MDS space, either with only the mean or with the mean and standard deviation of every feature, are summarized in Table 3.2.

Both the mean and standard deviation were important in modeling human responses. For instance, the perceptual distance best predicted by MFCC feature vectors consisted of the mean and standard deviation of each feature and its derivatives. The resulting RMSE were 0.140 for sentences and 0.121 for vowels, compared to 0.143 and 0.123 when only means were used.

VQual1 improved the performance for all conditions by providing complementary information to MFCCs, although the gain was rather small. For vowels, when VQual1 features were combined with MFCCs, the RMSE performance was improved by 4.07% and 3.14% for the mean only case and for the mean and standard deviation case, respectively. For

Table 3.2: Perceptual dissimilarity prediction performance, in terms of root-mean-squared error, using MFCCs, VQual1, and the combination of the two feature sets. Relative improvements by combining MFCCs and VQual1 features compared to the performance with only MFCCs are shown in parentheses.

	Vowels		Sentences	
	Mean	Mean&sd	Mean	Mean&sd
MFCC	0.123	0.121	0.143	0.140
VQual1	0.128	0.128	0.156	0.140
MFCC+VQual1	0.118 (4.07%)	0.117 (3.14%)	0.140 (2.24%)	0.123 (11.80%)

sentences, using only the means improved by 2.24%. Interestingly, when both mean and standard deviation were used, the performance gain was notable (11.80%). This is possibly because sentences vary more within a token than vowels do, and the deviation contains information related to perceived speaker identity.

Note that human listeners had higher accuracy on read sentences than on sustained vowels, but the acoustic features did less well in predicting human performance for the sentences. This might be because there are many other sources of information in connected speech that are not represented well by the current feature set.

Score-level fusion was tried as well, but no improvement was found over concatenating the features.

3.4 Automatic Voice Discrimination Experiments

In the previous analyses, the voice quality feature set provided valuable information to model perceived speaker identity. This led to a hypothesis that the feature set might be useful for automatically identifying talkers. Specifically, we were interested in investigating the effect

of several types of within-speaker variability on machine performance, and analyzing how the voice quality feature set contributes to system performance. In this section, standard ASV system performance is analyzed under recording session, speaking-style and affect variabilities. Because we were interested in evaluating the voice quality feature set on standard ASV system performance, an i-vector/PLDA framework was used instead of the method used for human response modeling in the previous section.

3.4.1 Stimuli

From the UCLA Speaker Variability Database, 25 female and 25 male talkers were randomly selected, and their read sentences in all 3 sessions and unscripted speech with different affect (*affective speech*) at each session were used. In each session per talker, there are ten read sentences, each 2–3 sec long, and an affective speech recording lasting 30–60 sec. Because ASV systems are sensitive to the utterance length of the enrollment and test data, it is important to balance the amount of data for a fair comparison. In order to balance the amount of data between read sentences and affective speech, we clipped a 30-sec segment in the middle of the affective speech and divided the segment into ten 3-sec segments. The resulting amount of data to enroll each talker was approximately 60 sec for session and affect variability experiments, and approximately 90 sec for style variability experiments. All test utterances were shorter than 3 sec.

3.4.2 Method

The effect of within-speaker variability was observed by comparing results from two different conditions. One was to enroll the talkers with data containing variability and test with known variability (*matched condition*), and the other was to test with unseen variability (*mismatched condition*). For example, in the session-matched condition, each talker was enrolled with randomly selected samples from all three sessions and tested on the remaining tokens. In the session-mismatched condition, the talkers were enrolled using only data from two sessions and tested on the third. Affect- and style-matched and mismatched conditions

were defined in a similar way.

Performance of a standard ASV system was evaluated under the conditions described above. The 20-order MFCCs along with their first- and second-order derivatives were used as baseline features. An i-vector [DKD11]/PLDA [KSO13] speaker verification system was implemented with the Kaldi toolkit [PGB11]. The variability matrices for the i-vector and PLDA were trained using the NIST SRE databases. The system was developed gender-dependently.

In order to examine the effect of using voice quality features on ASV performance, another system with the same back-end but with voice quality features was implemented. The measure $H^*_{2\mathrm{k}}$-$H_{5\mathrm{k}}$ could not be used, because it requires access to harmonic components close to 5 kHz, and the development data (NIST SRE) were band-limited to 4 kHz. Thus, the feature set used in this task excluded $H^*_{2\mathrm{k}}$-$H_{5\mathrm{k}}$ from the VQual1 set, and was referred to as VQual1*. The resulting feature vector dimension was 60 for MFCCs and 24 for VQual1*. The two systems were fused at the score level to obtain final results. The score-level fusion is analogous to averaging human listeners' dissimilarity scores (i.e., $\bar{\delta}$) and making a new decision based on the average score.

3.4.3 Results and Discussion

The baseline ASV system performance is reported in terms of equal error rates (EER) in Table 3.3. Session variability did not effect system performance much, but affect variability and speaking-style variability caused a notable degradation in system performance, both for female and male talkers. Affect variability among neutral, happy, and annoyed speech by female talkers caused the most degradation, more than doubling the error rate. Note that the affective speech recordings in the database also had session variability because they were recorded in different sessions. However, since the effect of the session variability was negligible, as mentioned earlier, the affect variability is most likely the main reason for the performance degradation.

It is possible that these results were influenced by phonetic content. Read sentences were

Table 3.3: Equal error rate (EER) for the ASV system, using only MFCC features, for the different conditions. The relative error increase in the mismatched compared to the matched conditions is shown in parentheses.

	Female	Male
Session-matched	2.81%	1.78%
Session-mismatched	2.73% (−2.85%)	2.19% (23.03%)
Affect-matched	3.64%	2.67%
Affect-mismatched	7.49% (105.68%)	4.00% (50.00%)
Style-matched	5.07%	2.37%
Style-mismatched	6.87% (35.64%)	3.64% (53.91%)

Table 3.4: Equal error rate (EER) for the fusion of MFCC and VQual1* systems. The relative error increase in the mismatched compared to the matched conditions is shown in parentheses, and the relative improvements over using only MFCCs are shown in separate columns.

	Female		Male	
	EER	% improvement	EER	% improvement
Affect-matched	3.22%	11.60%	2.16%	19.11%
Affect-mismatched	6.70%	10.49%	3.60%	9.89%
	(108.27%)	-	(67.10%)	-
Style-matched	4.65%	8.26%	2.16%	8.78%
Style-mismatched	6.90%	−0.37%	3.38%	7.25%
	(48.40%)	-	(56.50%)	-

distributed evenly into enrollment and test sets so that the system was enrolled with all 5 different sentences. For affective speech, however, phonetic imbalance could occur between enrollment and test data because phonetic content was not controlled for.

The performance of the ASV system with fused MFCC and VQual1* systems is summarized in Table 3.4. Results showed that voice quality features provided complementary information to MFCCs. In the affect-matched condition, fusing voice quality features notably improved the system performance for both genders. The improvement was 11.60% for female voices and 19.11% for males. A similar trend was observed for the affect-mismatched cases and style-matched conditions. However, the style-mismatched condition degraded slightly for female talkers when VQual1* was fused. It might be the case that the female speakers had larger differences in voice quality between read sentences and unscripted speech than male speakers did, resulting in larger within-speaker variability in VQual1*.

Even though system performance improved by adding the voice quality features, differences in EERs between matched and mismatched conditions widened. This suggests that

voice quality may be varying significantly according to the emotional status and speaking-style of the talker. Further analysis is needed with orthographic transcriptions and acoustic measures to check whether the degradation was due to VQual1* changes or to phonetic mismatch. Nevertheless, it was apparent that voice quality features provided talker-specific information which might not be sufficiently represented by MFCCs.

3.5 General Discussion

In this chapter, the first set of experiments was conducted to study human and machine abilities in discriminating different talkers. Our interest was to assess voice quality features in modeling human responses and in improving ASV performance. The voice quality feature set (VQual1) was determined using sustained vowel /a/ sounds, and the resulting set was equivalent to the psychoacoustic model of voice quality.

Human listeners were reasonably accurate in voice discrimination tasks. Not surprisingly, listeners were less accurate for vowels than for sentences (69.0% versus 89.0% accuracy). This was expected because sustained vowel sounds have much less information about talkers' identity than connected speech.

In modeling perceived talker dissimilarities, the voice quality feature set provided complementary information to MFCCs. The root-mean-squared error decreased as much as 3.14% for vowels and 11.80% for read sentences by combining the means and standard deviations of the voice quality features with MFCCs. Interestingly, human responses for vowels were better modeled than those for sentences. This can be partly explained by that the acoustic features used in this study might insufficiently represent the information human listeners are using in connected speech.

For machines, a standard ASV system (with MFCCs) was evaluated when there were session, affect, and speaking-style variability for both female and male voices. It was shown that the system performed worse when there was affect or speaking-style variability for the short test utterances used ($\leq$ 3 sec). However, it was not clear how much each variability contributes to performance degradation. For example, affect variability degrades the sys-

tem performance significantly, but the content mismatch between the enrollment and test utterances might also be an important factor.

The voice quality features (without H_{2k}^{*}-H_{5k}) were then applied to the ASV system, and they improved ASV system performance in 7 out of 8 conditions, by providing complementary information to MFCCs. For example, the relative error rate decrease was as much as 19.11% for affect-matched condition for male voices. There was one condition where fusing voice quality information with MFCCs slightly degraded system performance. That was when the speaking style was mismatched between the enrollment and test utterance for female voices. It suggests the voice quality features vary significantly between different speaking styles, especially for female voices.

In conclusion, the first set of experiments showed that psychoacoustically-valid voice quality features are promising both for modeling human responses and for improving ASV performance. Further studies were conducted based on the results of these initial experiments and reported in the following chapters.

CHAPTER 4

Improving ASV Performance Using Voice Quality Features

As shown in Chapter 3, voice quality features were effective for human response modeling and automatic speaker verification (ASV). The study presented in this chapter extends the analysis using a larger number of talkers. Specifically, the extent to which features can automatically separate different talkers is investigated for continuous speech when content and speaking-style vary considerably. A method to modify the voice quality feature set so that it can better separate different talkers is presented. The performance of ASV systems using the modified feature set was evaluated on the UCLA database under content and style variability, and on the NIST SRE evaluation data to verify its effectiveness on general ASV tasks.

4.1 Development of the VQual2 Feature Set

In Chapter 3, acoustic features inspired by a psychoacoustic model of voice quality [GSG16] were introduced. The feature set, VQual1*, consisted of F_0, F_1, F_2, F_3, H_1^*-H_2^*, H_2^*-H_4^*, H_4^*-$H_{2\mathrm{k}}^*$ and CPP. This feature set was effective for modeling human responses, and it improved automatic speaker verification performance, except for one condition where speaking style was mismatched between enrollment and test utterances for female talkers. This section describes the method used to modify the feature set to better represent talker identity under

Parts of this chapter were published in [PYK17].

contrasting speaking styles.

4.1.1 Stimuli

Speech samples from 100 female and 100 male talkers in the UCLA Speaker Variability Database were used in this study. Among the various speaking styles included in the database, sustained vowel /a/ sounds, read sentences and pet-directed speech were used. Vowel sounds were used for a preliminary analysis. Read sentences are text-constrained clear speech, while pet-directed speech is spontaneous and includes exaggerated prosody. These two speaking styles differ the most in the database.

The set of read sentences contains 6 repetitions of 5 different sentences per speaker and has been transcribed, so that the effect of content variability can be assessed. Since the read sentences and pet-directed speech represent contrasting speaking styles, using them together was suitable for examining the effect of speaking style variability.

All utterances were downsampled to 8 kHz for consistency with the development databases (NIST SRE databases).

4.1.2 Method

4.1.2.1 Candidate features

The main purpose of analyzing talker separability of features under within-talker variability was to improve the voice quality feature set to better differentiate between talkers. Candidate features were selected based on two considerations. The first consideration was whether formant correction should be performed when estimating harmonic amplitude differences. It was noted that formant correction without formant bandwidth information can be more erratic than no correction at all [ISA07]. Because it is difficult to accurately estimate bandwidths directly from the signal [HC99], bandwidths calculated with formant frequency information were used as in [IA04]. However, pilot experiments showed that over-correction may occur when formants are close to a harmonic. For example, the first formant of the vowel /i/ (mean

$F_1 = 437$ Hz [HGC95]) can be close to the second harmonic for a female talker. In that case, the effect of the formant on H_2 is overly corrected, resulting in unreliable H_1^*-H_2^* and H_2^*-H_4^* values. Such inaccuracies may weaken the ability of the features to separate talkers.

The second consideration was which form of the formant amplitudes should be added. Formant amplitudes are frequently used to represent voice quality, and they may have important talker-specific information. Frequently-used features include H_1^*-A_1^*, H_1^*-A_2^*, and H_1^*-A_3^* [Han97, VS14] where A_1, A_2, and A_3 are the amplitudes for the first, second, and third formants. Although these features were excluded because they were highly correlated to other measures for the vowel /a/ sounds (Section 3.1), they might be important for continuous speech. All features mentioned above, as well as A_1, A_2, and A_3, were chosen as candidate features.

The features were extracted pitch-synchronously every 10 msec using Voice Sauce software [SKV11], with Praat [BW17] chosen as the method for extracting pitch and formant frequencies. All features were automatically extracted, and no manual refinements were made.

4.1.2.2 Feature separability measures

The ability of each candidate feature to separate talkers was examined using the f-ratio [NMC97, LD08] separability measure. This criterion is widely used to measure how well an individual feature separates classes of stimuli. It implies that if the spread of class means increases, or if the clusters themselves become narrower, then the separability will increase. In this sense, the f-ratio identifies features which have large between-class variance and small within-class variance:

$$f = \frac{\text{between class variance}}{\text{within class variance}} = \frac{\sum_{i=1}^{M} P_i(\mu_i - \mu)^2}{\sum_{i=1}^{M} P_i \sigma_i^2}, \tag{4.1}$$

where M is the number of classes, μ_i is the within-class mean of the i-th class, μ is the global mean, P_i is the *a priori* probability of the class, and σ_i^2 is the within-class variance of a single feature.

Note that the f-ratio only measures the class separability of individual features. To

evaluate an entire feature set, an extension to the f-ratio for a multi-dimensional feature set is needed. The between-class variance and the within-class variances can be extended to the between-class scatter (S_b) and within-class scatter (S_w), respectively [TK09, pp. 280–283].

$$S_b = \sum_{i=1}^{M} P_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{4.2}$$

$$S_w = \sum_{i=1}^{M} P_i \Sigma_i \tag{4.3}$$

where Σ_i is the covariance matrix for the i-th class. Note that the traces of S_b and S_w, $\mathrm{Tr}\{S_b\}$ and $\mathrm{Tr}\{S_w\}$, are measures of the average distance, over all classes, of the mean of each individual class from the global mean, and the average variance of the features, respectively.

The mixed scatter matrix S_m is defined to be the covariance matrix of the feature vector x with respect to the global mean:

$$S_m = E\big[(x - \mu)(x - \mu)^T\big] = S_w + S_b. \tag{4.4}$$

Its trace, $\mathrm{Tr}\{S_m\}$, is the sum of the average, over the classes, variance of the features around their respective global mean. Using these extensions of between- and within-class variances, the class separability of a multi-dimensional feature set can be defined as:

$$J = \mathrm{Tr}\{S_w^{-1} S_m\}. \tag{4.5}$$

This measure is called the J-measure, and is often used for evaluating class separability of a feature set (e.g., [NMC97]). Note that the J-measure is large when samples in the feature space are well clustered around their mean, within each class, and the clusters of the different classes are well separated. Thus, it can be used as a separability measure for multi-dimensional feature sets.

4.1.2.3 Preliminary analysis with vowel sounds

Vowel /a/ sounds were used to preliminarily analyze the effects of formant correction on talker separability of voice quality features. This vowel has a high first formant frequency

Table 4.1: f-ratio values for individual harmonic amplitude difference features and J-measure values for sets of the three features with and without formant correction. Vowel /a/ sounds were used to calculate f-ratios and J-measures.

	With correction				Without correction			
	H_1^*-H_2^*	H_2^*-H_4^*	H_4^*-$H_{2\mathrm{k}}^*$	J	H_1-H_2	H_2-H_4	H_4-$H_{2\mathrm{k}}$	J
Female	0.27	0.21	0.24	3.86	0.32	0.50	0.33	4.15
Male	0.93	0.57	0.44	5.00	0.90	1.31	0.39	5.67

(mean $F_1 = 936$ Hz for female talkers [HGC95]), and thus it is expected that there is little effects of formants on harmonics, compared to other sounds. In this context, if formant correction yields worse f-ratios for vowel /a/ sounds than no correction does, it is likely that formant correction decreases f-ratios for other sounds.

The f-ratio values for individual harmonic amplitude difference features with or without formant correction, and J-measure values for a set of the features with or without formant correction are shown in Table 4.1. In most cases, not using formant correction resulted in a higher f-ratio values, and J-measures were higher without correction for both genders than with correction. These results indicate that automatic formant correction rather decreases talker separability of the harmonic amplitude difference features.

Unfortunately, it is difficult to find why correction resulted in decreased f-ratio and J-measure values. Over-correction and other inaccuracies in automatic correction might be a reason. However, correction accuracy could not be measured without the "ground truth" voice source spectrum from given data. The effects of formant correction needs to be further analyzed.

Even though formant correction was less effective for vowel sounds than no correction, it might still be effective for high phonetic content variability, by compensating for the phonetic effects. Therefore, experiments were continued using continuous speech samples: read sentences and pet-directed speech.

4.1.2.4 Experimental conditions

Phonetic content and speaking-style variability were considered in designing experimental conditions. For content variability, a large content difference condition and a small content difference condition were created for comparison. The f-ratios for the two conditions can be used to analyze the effect of content variability on talker separability. First, all repetitions of the 5 sentences (from all 100 talkers for each gender) were randomly divided into 5 subsets, and the mean and standard deviation of the f-ratio were found across those subsets. Each subset in this condition represented a high phonetic content variability case. This condition is denoted as the *multiple text* condition. Second, 5 subsets were made so that each subset included the same sentence. The f-ratios were calculated again for these 5 subsets and the mean and standard deviation were computed across the subsets. In this case, each subset represented a dataset with low phonetic content variability. This condition is denoted as the *single text* condition.

Other conditions were created analogously for speaking style variability. In the *multiple style* condition, read sentences and pet-directed speech were distributed randomly into 2 subsets, and the mean and standard deviation of the f-ratio were calculated across the subsets. These subsets represented high speaking style variability. In the single style condition, speech samples from the two speaking styles formed separate subsets (*read* and *pet*). These represented low style variability within the data subsets.

4.1.3 Results and Discussion

The f-ratios calculated for each experimental condition for females and males are shown in Figure 4.1 and Figure 4.2, respectively. The top panels represent the mean and standard deviation of f-ratio across the 5 subsets of the multiple text condition, and those of the single text condition. The bottom panels represent the mean of f-ratio across the 2 subsets of the multiple style condition, and the f-ratios of read and pet-directed style conditions.

For phonetic content variability conditions for female talkers, the standard deviations for the single text condition were larger than those for the multiple text condition. For

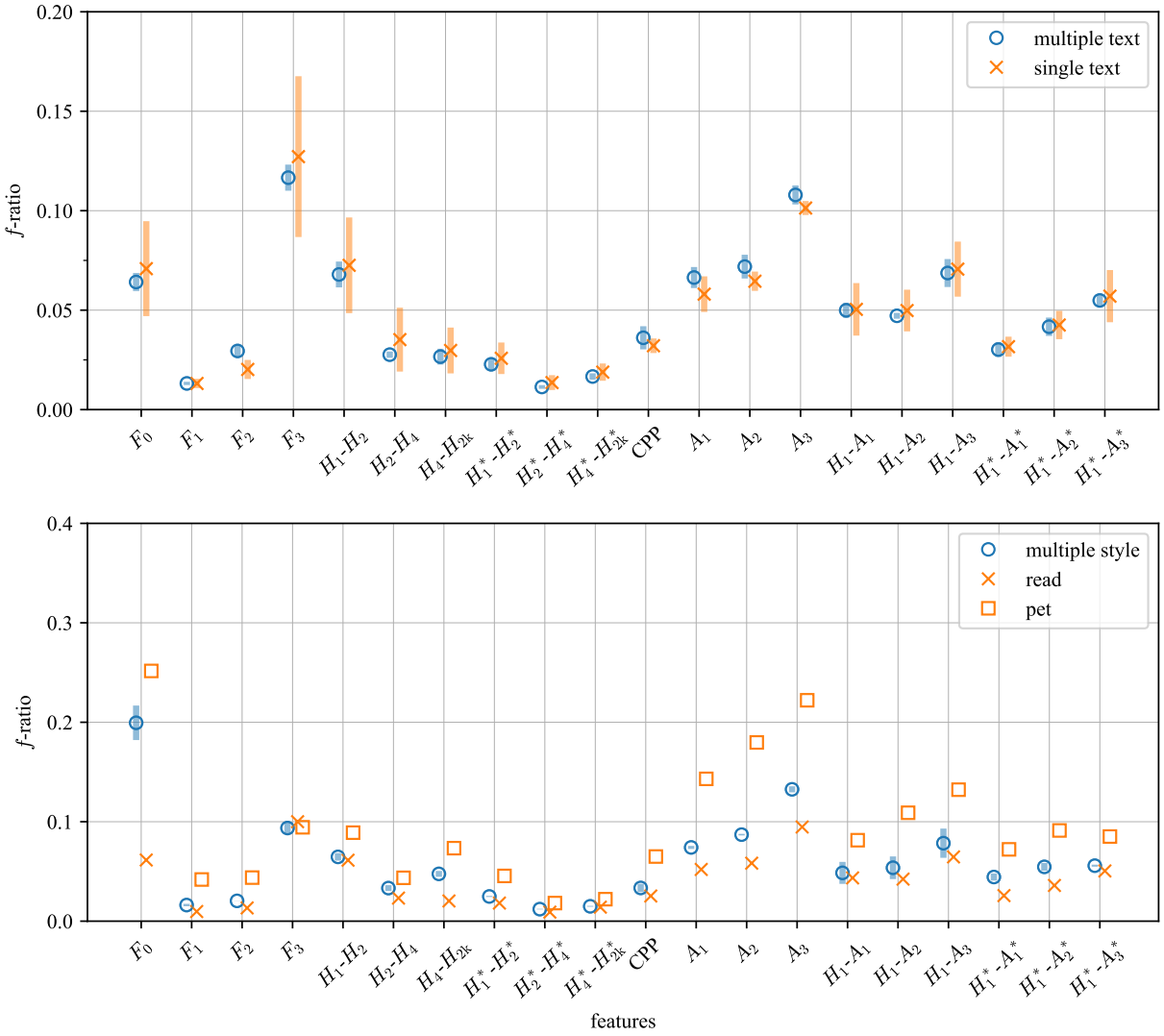


Figure 4.1: Computed f-ratios of various voice quality features using the female voices in the UCLA Speaker Variability Database for phonetic content (top) and speaking-style (bottom) variability. "Single text" in the top panel indicates that the f-ratios were computed for data subsets containing the same sentence, and "multiple text" indicates the subsets contained 5 different sentences. "Read" and "pet" in the bottom panel indicate the subsets only contained read sentences or pet-directed speech, and "multiple style" indicates that the subsets contained both speaking styles.

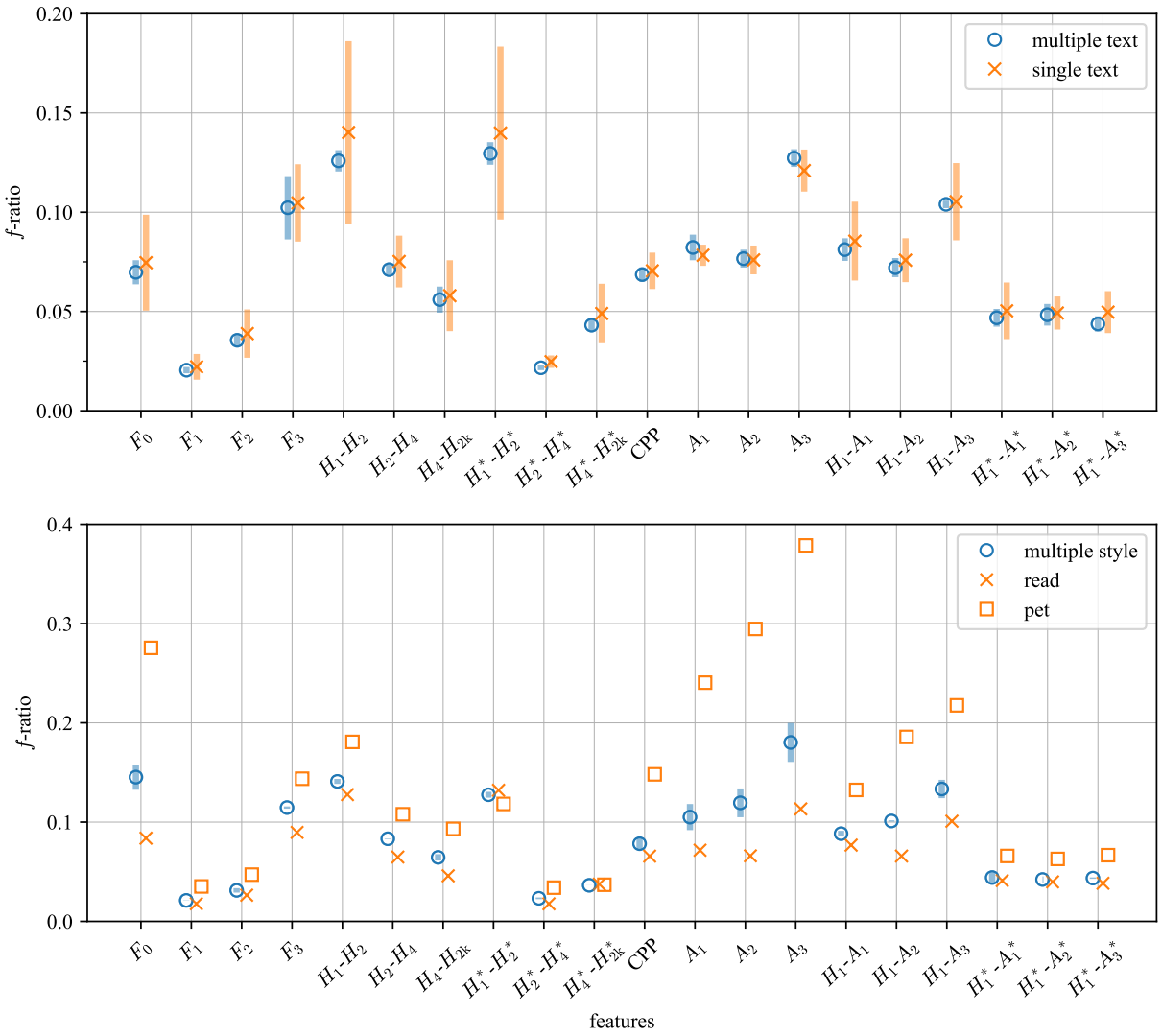


Figure 4.2: Computed f-ratios of various voice quality features using the male voices in the UCLA Speaker Variability Database for phonetic content (top) and speaking-style (bottom) variability. "Single text" in the top panel indicates that the f-ratios were computed for data subsets containing the same sentence, and "multiple text" indicates the subsets contained 5 different sentences. "Read" and "pet" in the bottom panel indicate the subsets only contained read sentences or pet-directed speech, and "multiple style" indicates that the subsets contained both speaking styles.

example, the standard deviations of F_0, F_3, and H_1-H_2 for the single texts were notably larger than those for multiple texts. Considering that the single text condition consisted of different sentences across subsets, while the multiple text condition contained similar text distributions, a higher standard deviation across subsets for the single text condition indicates that the features' talker separability depends on the phonetic content. In addition, single text subsets had higher overall mean f-ratios than multiple text subsets, although the differences were not overwhelming. This implies that the acoustic features vary according to phonetic content, and such phonetic content variability increases within-talker variability more than between-talker variability for most features. However, CPP, A_1, A_2, and A_3 had higher mean f-ratios for the multiple text condition. Those four features could be capturing idiosyncratic information related to phonetic content, and/or talker separability was not influenced much by phonetic content. Similar trends were observed for male talkers.

The features with the highest f-ratio were similar across the genders and content variability conditions, although they had different ranks. For the multiple text condition, the features with the highest f-ratio were F_3, A_3, A_2, H_1-A_3 and H_1-H_2 for females, and those for males were H_1^*-H_2^*, A_3, H_1-H_2, H_1-A_3, and F_3. For the single text condition, they were F_3, A_3, H_1-H_2, F_0, and H_1-A_3 for female talkers, and H_1-H_2, H_1^*-H_2^*, A_3, H_1-A_3, and F_3 for males. Note that F_3, A_3, H_1-A_3, and H_1-H_2 always appear in the top 5 highly-ranked features.

Interestingly, the f-ratios in the pet-directed subset were higher then for the read sentences subset. This suggests that each talker had her or his own unique style of talking to pets, making talkers more distinct, increasing between-talker variability. This uniqueness in speaking style between talkers will likely result in decent performance in same-style ASV experiments which will be discussed in Section 4.2.

The most effective features were also similar for each gender and for each style variability condition. For female talkers, the features with the highest f-ratios for multiple style condition were F_0, A_3, F_3, A_2 and H_1-A_3, those for read sentences were F_3, A_3, H_1-A_3, F_0, and H_1-H_2, and those for pet-directed speech were F_0, A_3, A_2, A_1, and H_1-A_3. For males, they were A_3, F_0, H_1-H_2, H_1-A_3, and H_1^*-H_2^* for multiple style condition, H_1^*-H_2^*, H_1-H_2,

A_3, H_1-A_3, F_3 for read sentences, and A_3, A_2, F_0, A_1, H_1-A_3 for for pet-directed speech. Here, F_0, A_3, and H_1-A_3 appear in the top 5 features, except for read sentences from male voices, where F_0 had the 6th highest f-ratio.

To select features, two sets of harmonic amplitude difference features were compared: with formant correction (H_x^*-H_y^*) and without formant correction (H_x-H_y). Formant amplitude features were compared among three sets: raw amplitudes A_1, A_2, and A_3 (A_z), differences between H_1 and formant amplitudes H_1-A_1, H_1-A_2, and H_1-A_3 (H_1-A_z), and differences between H_1 and formant amplitudes with formant correction H_1^*-A_1^*, H_1^*-A_2^*, and H_1^*-A_3^* (H_1^*-A_z^*). For all conditions, the f-ratios of harmonic amplitude differences were higher for H_x-H_y than H_x^*-H_y^*, and A_z had highest f-ratios among formant amplitude features. Thus, H_x-H_y and A_z features can be thought to have higher talker separability than their variants.

However, even if individual features have high f-ratios, their talker separability as a set might not be high. For example, if they are highly intercorrelated with each other, the talker separability might be lower than another set of independent features with lower f-ratio values. Thus, the J-measures were calculated using the read sentences in the UCLA database to analyze talker separability of a feature set. All six possible combinations between two harmonic amplitude difference features (H_x-H_y and H_x^*-H_y^*) and three formant amplitude features (A_z, H_1-A_z, and H_1^*-A_z^*) were made.

J-measures were calculated on the combined sets and shown in Table 4.2. Here, the highest J-measure value was obtained by combining harmonic amplitude differences without formant correction, and raw formant amplitudes for each gender. For instance, if H_x-H_y features were combined with A_z, the J-measure value was 6.326 for female talkers, while combining with H_1-A_z result in a J-measure of 6.245. Including formant-corrected features resulted in lower J-measures than including their formant-uncorrected counterparts. These results are consistent with the f-ratio results.

The final feature set included F_0, F_1, F_2, F_3, H_1-H_2, H_2-H_4, H_4-$H_{2\mathrm{k}}$, CPP, A_1, A_2, and A_3. These features will be called the VQaul2 features. We then hypothesized that VQual2

Table 4.2: J-measures for sets of features that combining the features in the top row and the features in the second row. H_x-H_y indicates H_1-H_2, H_2-H_4, and H_4-H_{2k}; A_z indicates A_1, A_2, and A_3. The highest J-measure values were boldfaced for each gender.

	H_x-H_y			H_x^*-H_y^*		
	A_z	H_1^*-A_z^*	H_1-A_z	A_z	H_1^*-A_z^*	H_1-A_z
Female	**6.326**	6.231	6.245	6.230	6.160	6.177
Male	**6.532**	6.396	6.453	6.438	6.327	6.381

features would better separate talkers than VQual1* features, and would further improve ASV system performance. Recall that the original VQual1* feature set was generated from a psychoacoustic model of voice quality, and evaluated on sustained vowel sounds. The modified VQual2 set, on the other hand, was chosen to automatically separate different speakers, and was evaluated on continuous speech signals.

The J-measure value were computed for MFCCs and subsets of VQual2 features, using the read sentences, and are shown in Figure 4.3. As the voice quality features are added to MFCCs, the J-measure increases. Thus, we expect that VQual2 features to provide complementary information to MFCCs in ASV tasks. This will be evaluated in the following sections.

4.2 ASV Performance Analysis Under Content and Speaking-Style Variability

Recall that the compounded effect of content and style variability made ASV system performance degrade in the first set of experiments of Chapter 3. In this section, the impact of content and style variability were further investigated, and the VQual2 feature set was applied to verify its effectiveness.

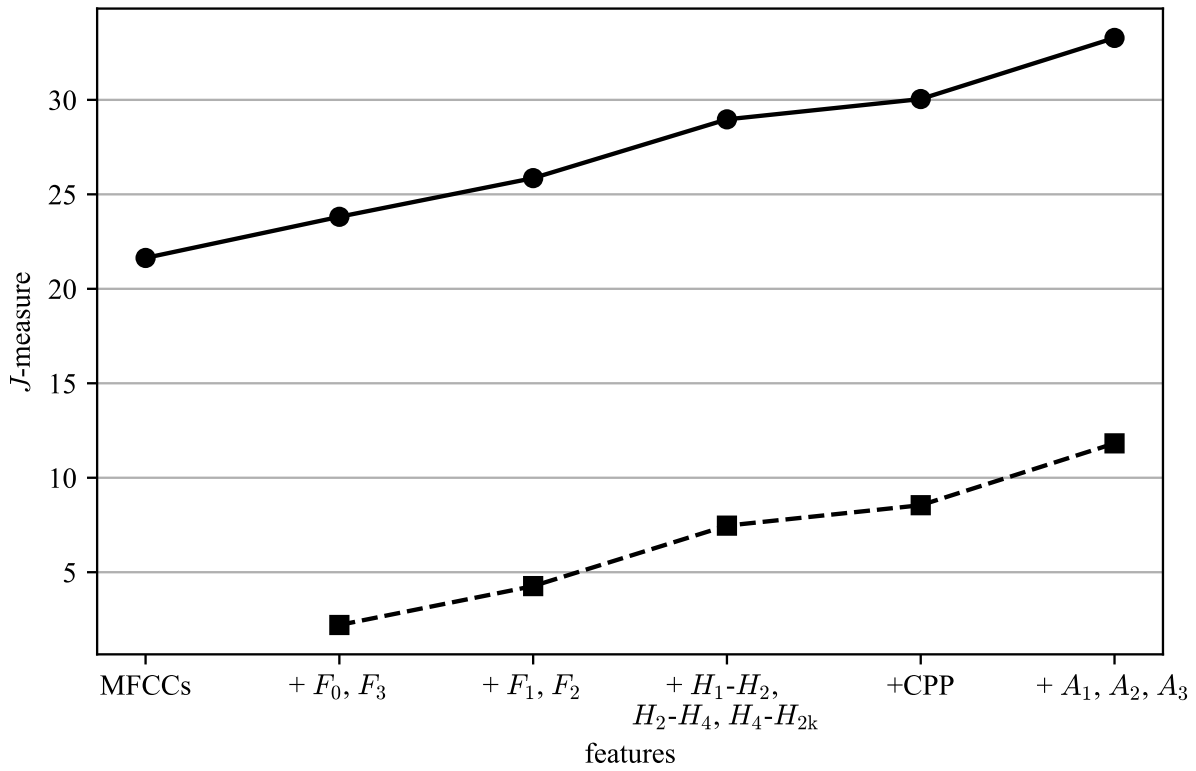


Figure 4.3: Computed J-measures of VQual2 and MFCC features using read sentences (solid line). Each point on the chart is the J-measure of all features below and to the left of that point. The dashed line shows VQual2 features without MFCCs for comparison. The J-measure was computed for both female and male talkers.

4.2.1 Method

4.2.1.1 Experimental conditions and data

ASV systems were evaluated using recordings from 100 female and 100 male talkers from the UCLA database. For content variability, the effect of varying phonetic content was analyzed with five sentences. Two different conditions were compared: same-text trials and different-text trials. In both conditions, one sentence was chosen to enroll talkers, and one sentence was chosen to test the system. In the *same-text* trials, the same sentence was used for enrollment and testing, but from different speakers or different repetitions from the same speaker. In the *different-text* trials, the sentences used for enrollment and testing were different. Each talker had 6 repetitions per sentence, resulting in a total of $100 \times 5 \times 6 = 3000$ sentences for each gender. All possible combinations of enrollment–test utterance pairs were used except for the case when the enrollment and test utterances were identical (in terms of speaker and repetition number).

The effect of varying speaking style was examined with both read sentences and pet-directed utterances. The read sentences were randomly concatenated together for each talker until 5 seconds of speech were collected. This was done twice to create enrollment and testing sets for every talker. No utterance was used for both enrollment and testing. Note that this task can be regarded as a text-constrained task. Pet-directed speech samples were cut into two non-overlapping segments containing 5 seconds of speech for enrollment and testing. One female and six male talkers were removed due to poor recording quality or low amounts of speech in the pet-directed recordings. The total number of speech samples was 396 and 376 for females and males, respectively. Again, all possible combinations of enrollment-test utterance pairs were used except for the pairs of identical samples.

Four types of trials were designed, denoted as "enrollment data–testing data": read–read, read–pet, pet–read, and pet–pet. In the read–read and pet–pet trials, the same speaking style was chosen for both enrollment and testing. In the read–pet trial, read sentences were used for enrollment and pet-directed speech was used for testing, and in the pet–read trial, the reverse was true.

4.2.1.2 ASV system setup

Performance of ASV systems depends, in part, on the use of appropriate features to distinguish speakers. MFCCs of dimension 20, along with first derivatives, were used as baseline features. The second derivatives were not used because they did not provide significant performance gain. VQual1* and VQual2 features, along with first and second derivatives, were used as alternative feature sets.

A standard i-vector [DKD11]/PLDA [KSO13] ASV system was used for the experiments.[1] Three systems using MFCCs, VQual1*, or VQual2 were developed, and were pre-trained on the NIST SRE04, 05, 06, and 08 databases. The i-vector dimension was 600 and it was reduced to 200 after PLDA. The UBM (modeled with 2048 Gaussians) and subspaces were trained using the NIST SRE training databases.

After obtaining the PLDA scores from each system, score fusion was used for further improvements [RFR02]. Since VQual2 features performed better than VQual1* features in most cases, only fusion performance with MFCCs and VQual2 features is reported. The fusion system outputs were linearly combined using the following equation:

$$s = \alpha s_v + (1 - \alpha) s_m \tag{4.6}$$

where s_m is the PLDA score using MFCCs, s_v is the PLDA score using VQual2 features, and α is the coefficient of s_v chosen from the range of 0 to 1. PLDA scores using both MFCCs and VQual2 features were first scaled to have zero-mean and unit-variance before fusion was performed.

4.2.2 Results and Discussion

The overall performance in terms of equal error rate (EER) is shown in Table 4.3. Relative improvement by a score-level fusion between MFCC- and VQaul2-based system over using only MFCCs is also shown in the table. The p values of the improvement were computed using

[1]The i-vector/PLDA system and computational resources were provided by the Johns Hopkins University Human Language Technology Center of Excellence.

Table 4.3: ASV system performance in terms of EER (%) with content and style variability using the UCLA database. The fusion performance is obtained by fusing the PLDA scores from the MFCC- and VQual2-based systems. Relative improvement by fusion compared to MFCCs is denoted along with its p value in parentheses. Human voice discrimination results reported in Chapter 3, are added in the last column for comparison.

		Utterance length	MFCC	VQual1*	VQual2	Fusion	% imp.	Human
Female	Same text	2–3 sec	7.71	15.66	12.67	6.22	19.33 ($p < 0.01$)	10.00
	Different text	2–3 sec	28.14	31.48	28.23	24.41	13.97 ($p < 0.01$)	12.22
Male	Same text	2–3 sec	5.97	15.15	13.63	4.93	17.42 ($p < 0.01$)	-
	Different text	2–3 sec	28.33	28.33	27.73	23.07	16.80 ($p < 0.01$)	-
Female	Read–read	5 sec	3.65	10.10	6.40	3.03	16.99 ($p < 0.01$)	-
	Pet–pet	5 sec	19.19	18.18	18.18	12.79	33.35 ($p < 0.01$)	-
	Read–pet	5 sec	30.30	36.77	35.06	29.29	3.33 ($p < 0.01$)	-
Male	Read–read	5 sec	2.13	5.48	3.19	1.06	50.23 ($p < 0.01$)	-
	Pet–pet	5 sec	6.38	12.72	11.70	4.25	33.39 ($p < 0.01$)	-
	Read–pet	5 sec	19.30	30.85	28.72	19.15	0.78 ($p = 0.59$)	-

McNemar's test [McN47]. For comparison, the table includes results from human perceptual voice discrimination experiments presented in Chapter 3, where the listeners were asked to determine if a given pair of read sentences was spoken by the same talker or two different talkers. The results from pet–read trials are omitted as they were almost identical to those from read–pet trials.

4.2.2.1 Phonetic content variability

ASV performance degraded severely when the enrollment and test utterances had different texts compared to when they had the same text. Using MFCCs, error rates increased dramatically when comparing same and different-text trials. For male talkers, for example, the EER was 5.97% in the same-text condition, and it increased to 28.33% in the different-text condition. A similar pattern was observed for female talkers. Note that human listeners were not affected by the content difference as much as the ASV systems were. These results suggest that by understanding human listeners' strategies to discriminate different talkers, ASV robustness to content variability might be improved.

Although VQual2 features did not perform as well as MFCCs in the same-text trials, in the different-text trials, they performed almost as well as MFCCs for females and even better than MFCCs for males. This indicates the effectiveness of VQual2 for text-independent ASV tasks.

In all trials designed for phonetic content variability, fusion with the VQual2 scores provided relative improvements of at least 13.97% ($p < 0.01$), suggesting that VQual2 features contain complementary information to MFCCs.

4.2.2.2 Speaking style variability

The EER from using MFCCs increased dramatically both for female and male talkers in the style-mismatched condition (read–pet) compared to the read–read and pet–pet trials. For example, the EER for female talkers increased from 3.65% in the read–read case to 30.30% in the read–pet case. This might be due to signal processing errors as an effect of exaggerated

prosody. Another issue for these pairs might have been the limited phonetic content of the pet-directed speech excerpts. While the read sentences were phonetically rich, pet-directed speech was largely limited to phrases such as "Awww, cute," with stereotyped intonation contours that lacked the idiosyncrasies of the read–read pairs. However, the amount of performance degradation is greater than the text variability conditions, suggesting that style variability was another important factor that significantly degraded ASV performance.

Score fusion with the VQual2 features showed decent improvements for the style-matched conditions, notably 33% ($p < 0.01$) improvement for the pet–pet trials, and both for female and male talkers. However, there was little improvement for the style-mismatched condition (3.33%, $p < 0.01$ for female and 0.78%, $p = 0.593$ for male talkers). This is expected since VQual2 features themselves partially reflect the speaking style of a talker and are susceptible to style changes.

4.3 ASV Performance on Short Utterances

For long utterances, phonetic overlap between enrollment and test utterance can be large, but for short utterances, it is more likely that the phonetic content does not match well between enrollment and test utterances. If the voice quality feature set is helpful when phonetic content mismatch occurs between enrollment and test utterances, as shown in the previous section, the feature set might be especially effective for text-independent short utterance ASV. As an attempt to test this possibility, the voice quality feature set was evaluated on standard ASV evaluation data with different utterance lengths.

4.3.1 Method

The ASV system described in Section 4.2.1 was used. Here, the NIST SRE10 database condition 5 extended task [MG10], which consisted of telephone-channel speech, was employed to evaluate the ASV system.

In addition to the full utterance (approximately 5 min) evaluation, new enrollment and

testing datasets were created using the SRE10 data by cutting the utterances to 10, 5, and 2 seconds of speech, in order to investigate the impact of various utterance lengths on system performance.

4.3.2 Results and Discussion

System performance in terms of EER is shown in Table 4.4. The fusion system is the linear score fusion between the MFCC and VQual2 systems as described before. The optimal choice of α is also shown in the table, where α is the weight of the VQual2 scores in the linear score fusion as in Eqn. (4.6).

As predicted, ASV systems performed worse as utterances became shorter. This is consistent with literature that measured performance of systems with MFCCs. For example, in a study using the SRE 2003 database, the EER degraded from 2.48% for full-length utterances to 22.31% for 2-sec utterances [DJP16]. Similarly, VQual1* and VQual2 system performance degraded with short utterances.

VQual2 features showed 1.05–2.37% absolute improvement compared to the VQual1* features in all conditions. When only VQual2 was used, the performance was worse than MFCCs, but the VQual2 features were able to improve the performance of the system through score fusion by providing complementary information to MFCCs. The weight of the VQual2 scores in the fusion (α) increased as utterances became shorter, suggesting that the system relied more on the VQual2 scores as the amount of speech decreased.

4.4 General Discussion

This study analyzed ASV system performance when content and style variability were large, especially when utterances were short. The VQual2 feature set was chosen for its effectiveness in separating talkers in these large variability conditions using the f-ratio and J-measure.

System performance according to content and speaking-style variability with the UCLA Speaker Variability Database was analyzed. ASV experiments using an i-vector/PLDA sys-

Table 4.4: ASV system performance in terms of EER (%) with the NIST SRE10 database. The relative improvement by fusion (MFCC+VQual2) is noted in parentheses. The coefficient α is the optimal weight of the VQual2 scores in the fusion. The utterance length in seconds for enrollment and test utterances is denoted in an enrollment–test form.

Utterance length	MFCC	VQual1*	VQual2	Fusion (% imp.)	α
Full	2.89	8.96	7.91	2.80 (3.11 %)	.10
10–10	10.88	19.60	17.23	9.53 (12.41 %)	.29
5–5	16.90	25.18	22.82	14.91 (11.78 %)	.35
2–2	28.47	32.95	30.92	25.95 (8.85 %)	.46

tem with MFCCs and VQual2 features were conducted, along with linear score fusion. It was found that when content and/or style conditions were mismatched, the system error rate increased significantly. Thus, approaches that minimize the effect of content variability are worth exploring.

VQual2 features might be useful for high-content variability cases. In the mismatched content conditions, score fusion with VQual2 features improved performance. But unlike humans, there was still a large difference in performance between matched and mismatched content conditions. This suggests that understanding human listeners' strategies to discriminate speakers might provide insights to improve machine performance.

For style variability experiments, VQual2 showed notable performance gain in the pet–pet trials. This suggests that voice quality features might be able to capture talkers' idiosyncratic way of exaggerating prosody. However, in the mismatched style conditions, score fusion did not improve performance much.

To examine the effects of using voice quality features with a larger set of utterances and talkers, ASV experiments were conducted using the NIST SRE10 telephone speech database. Results suggest that short utterances ($<$ 10 sec) benefit from using VQual2 features more than long utterances (about 5 min).

In conclusion, the voice quality feature set was modified to better separate different talkers. This feature set was especially effective for text-mismatched trials, and style-matched trials. It also successfully improved general telephone channel ASV performance on a standard NIST SRE evaluation database. The application of this feature set will be presented in later chapters, including human response analysis and emotion recognition.

CHAPTER 5

Comparing Human and Machine Abilities in Voice Discrimination for Short Utterances

In Chapter 3, the voice quality feature set VQual1, which was derived from a psychoacoustic model, was promising in modeling human responses. The feature set also improved automatic speaker verification (ASV) performance for short test utterances when read sentences and affective speech were used. The feature set was modified in Chapter 4, denoted as VQual2, to further improve ASV performance. It was noted that speaking style mismatch (read sentences versus pet-directed speech) degrades ASV performance significantly for 5-second speech segments, both with MFCCs and VQual2. However, ASV performance analysis under phonetic content variability showed that VQual2 could improve short-utterance text-independent ASV performance. VQual2 was also effective at standard telephone conversation ASV tasks, especially when utterances were short. Experimental setups and key results are summarized in Table 5.1.

The results suggested that humans are more accurate than machines for short-utterance, text-independent voice discrimination tasks. Hence, comparing responses from humans and machines might provide insights to further improve machine performance. However, a direct comparison between humans and machines cannot be made with the experiments presented in Chapters 3 and 4 because the experimental designs and resulting stimuli pairs for humans were different from those for machines. In addition, humans' ability to distinguish talkers under style variability was not tested in those previous experiments. Thus, in this chapter,

Parts of this chapter were published in [PYV18] and [PAK19].

a new set of perception experiments were conducted to investigate comparative effects of within-talker variability in phonetic content and speaking style on human and machine performance. Speech samples from a larger number of talkers were used, and high speaking-style variability between read sentences and pet-directed speech was employed.

5.1 Perception Experiments

Human listeners' ability to discriminate among talkers across the two speaking styles (read sentences and pet-directed speech) was investigated. The most contrasting speaking styles were chosen in the UCLA database to probe the limits of human perception with high within-talker variability.

5.1.1 Stimuli

Fifty female self-reported native speakers of English were randomly selected from the UCLA database. Female talkers were chosen because they used more prosodic exaggeration when talking to pets than did male talkers, leading to larger differences between the read sentences and the pet-directed speech. *Post hoc* listening by two linguists indicated that utterances from nine talkers were perceptually "marked" by a non-American dialect, overly-precise articulation and/or unusual dysfluencies in reading. The remaining 41 talkers lacked such personal idiosyncrasies and will be referred as "unmarked".

For each talker, three read sentences were selected from each of the three recording sessions. Each speech sample lasted less than 2 seconds. Two excerpts were taken from the pet-directed speech, matched in length to the average duration of the sentences. Stimuli were downsampled to 8 kHz to match the bandwidth of the NIST SRE databases.

5.1.2 Method

Stimuli were assembled into 100 pairs of voices in which both voice samples came from the same person (50 pairs of read sentences and 50 pairs where a read sentence was paired with

Table 5.1: Summary of experiments (Ex.) reported in Chapters 3 and 4. The VQual1* feature set contains the features in VQual1, except H^*_{2k}-H_{5k}. Fusion indicates a weighted combination of the scalar responses from individual systems. Speech samples were drawn from the UCLA database unless otherwise specified. Utterance lengths for ASV experiments are denoted as the length of enrollment and test utterances in seconds.

		Chapter 3	Chapter 4
Perception	Speaking style	Ex.1: vowel /a/ (1–3 sec) Ex.2: read ($\approx$ 2 sec)	N/A
	No. talkers	Ex.1: 5 females Ex.2: 3 females	N/A
	No. listeners	Ex.1: 60 Ex.2: 15	N/A
Feature selection	Stimuli	vowels	read, pet-directed
	No. talkers	5 females	100 females, 100 males
	Resulting feature set	VQual1 (F_0, F_1, F_2, F_3, CPP, H^*_1-H^*_2, H^*_2-H^*_4, H^*_4-H^*_{2k}, H^*_{2k}-H_{5k})	VQual2 (F_0, F_1, F_2, F_3, CPP, H_1-H_2, H_2-H_4, H_4-H_{2k}, A_1, A_2, A_3)
ASV	Speaking style	Ex.1: read Ex.2: affective Ex.3: read, affective	Ex.1: read Ex.2: read, pet-directed Ex.3: phonecall (SRE database)
	Utterance length	Ex.1: 60sec–3sec Ex.2: 60sec–3sec Ex.3: 90sec–3sec	Ex.1: 2sec–2sec Ex.2: 5sec–5sec Ex.3: various
	No. talkers	25 females, 25 males	Ex.1: 100 females, 100 males Ex.2: 100 females, 100 males Ex.3: $\gg$ 200 (SRE database)
	Systems	MFCCs, VQual1*, fusion	MFCCs, VQual1*, VQual2, fusion
Predicting human responses	Features	MFCCs, VQual1	N/A
	Method	linear regression	N/A
Key results		VQual1 was promising for predicting human responses (vowels, read sentences) and ASV (read sentences and affective speech).	VQual2 was developed, and it was effective for text-independent short utterance ASV.

pet-directed speech), and 2,450 pairs where the two talkers were different (half including two read sentences and half including one read sentence and one pet-directed speech sample), for a total of 2,550 pairs of stimuli. Stimuli were always drawn from different recording sessions, and each pair included two different read sentences. Thus, this task was always text- and recording session-mismatched.

To minimize listener fatigue, stimuli were divided at random into 12 subsets of 200 pairs of voices and 1 subset of 150 pairs. Thirteen groups of five normal-hearing UCLA students and staff members (aged 18–28; mean age 19.91; standard deviation 2.28; 65 listeners total) were recruited, of whom 30 considered themselves L1 English speakers. Participants listened to the pairs of stimuli over Etymotic insert earphones (model ER-1) at a comfortable constant listening level. Each pair could be played only once in each presentation order (AB/BA). Listeners were asked whether the two speech samples were produced by the same talker or by two different talkers. They also reported their confidence in their responses on a 1–5 scale (1 = positive, 5 = wild guess). They were not told how many talkers were represented in the trials. The experiment was self-paced, and listeners were encouraged to take breaks as needed. Total testing time was less than one hour per listener.

5.1.3 Evaluation Metric

Hit rates and false alarm rates were calculated by defining a hit as a correct "same talker" response and a false alarm as an incorrect "same talker" response. Additionally, listeners' same versus different responses were combined with their confidence ratings to create a scale ranging from "positive, same talker" (= 1) to "positive, different talkers" (= 10) as in Eqn. 3.1. These scalar responses were used to derive receiver operating characteristic (ROC) curves using SYSTAT software [Sys]. d' (d-prime, e.g., [MC05]) and the area under the receiver operating characteristic curve (AUC) were calculated for each ROC curve. Note that d' values calculated from ROC curves can differ from values directly calculated from hit and false alarm rates. The metric used for ASV in Chapters 3 and 4, the EER, was also computed from ROC curves.

Note that the AUC and EER measures are not always correlated because the two measures reflect different properties of the curve. The AUC is calculated from the entire ROC curve, and it reflects overall accuracy regardless of a specific decision threshold. On the other hand, the EER only focuses on the point where the false rejection rate and the false acceptance rate are equal, and it summarizes the system performance in terms of the error rate. These measures can differ especially when the ROC curves are skewed. Skewed ROC curves can result when the variance of the distribution in the decision space of same-talker pairs is different from that of the different-talker pairs [MC05].

5.1.4 Results

Hit rates, false alarm rates, d' (from the ROC curve), AUC, and EER are shown in Table 5.2. Because listener performance could be influenced by the talkers' perceptual markedness, results when the stimuli were pairs from the 41 unmarked talkers, pairs from the 9 marked talkers, pairs consisting of one marked and one unmarked talker, and pairs from all 50 talkers are shown separately in the table. The pairs of read sentences are denoted as *read–read* and the pairs of one read sentence and one pet-directed speech excerpt are denoted as *read–pet*.

Human listeners were reasonably accurate in distinguishing unmarked talkers when stimuli were pairs of read sentences ($d' = 1.81$). As expected, accuracy decreased when listeners heard read speech paired with pet-directed speech ($d' = 0.50$). The decrease in the hit rate (33.2%) was greater than the increase in the false alarm rate (9.4%), suggesting that the responses for read–pet pairs were biased to "different talker" responses. Because there were many more unmarked talkers than marked talkers, the "all talker pairs" results are very similar to those for the unmarked talkers for the read–read pairs.

Although the marked talkers had idiosyncrasies in their speech, they were in fact harder to discriminate. d' equaled 1.48 for read–read pairs (compared to 1.81 for unmarked talkers), and 0.16 for read–pet pairs (compared to 0.50). In addition to a decrease in sensitivity, the performance degradation reflected a large decrease in the hit rate and a smaller decrease in the false alarm rate, suggesting a stricter response criterion. For trials including one marked

Table 5.2: Composite human voice discrimination performance for the 41 perceptually-unmarked talkers, 9 perceptually-marked talkers, pairs consisting of one marked and one unmarked talker, and all 50 talkers in terms of hit rates (HR, %), false alarm (FA) rates (%), d' (calculated from ROC curves), AUC, and EER (%). Read–read and read–pet indicate that the token pair presented to the listener was composed of two different read sentences or one read sentence and one pet-directed speech segment, respectively. All tokens were approximately 2-sec long. Note that there were no "same talker" pairs when listeners compared a marked talker to an unmarked talker, so that the hit rate could not be calculated. Boldfaced numbers indicate the best performing condition in terms of each metric.

	No. same talker pairs	No. different talker pairs	HR	FA	d'	AUC	EER
Read–read, unmarked talker pairs	41	820	**87.3**	25.8	**1.81**	0.885	**19.02**
Read–pet, unmarked talker pairs	41	820	54.1	35.2	0.50	0.644	39.23
Read–read, marked talker pairs	9	36	68.9	21.7	1.48	0.844	24.86
Read–pet, marked talker pairs	9	36	37.8	34.4	0.16	0.538	46.23
Read–read, marked/unmarked pairs	N/A	369	N/A	**20.7**	N/A	N/A	N/A
Read–pet, marked/unmarked pairs	N/A	369	N/A	32.1	N/A	N/A	N/A
Read–read, all talker pairs	50	1,225	84.0	24.2	1.73	**0.876**	20.19
Read–pet, all talker pairs	50	1,225	51.2	34.2	0.46	0.628	40.34
All pairs	100	2,450	67.6	29.2	1.11	0.766	30.58

talker and one unmarked talker, only false alarm rates could be calculated because stimuli always came from different talkers. Those marked/unmarked pairs had the lowest observed false alarm rate: 20.7% for read–read pairs and 32.1% for read–pet pairs.

5.1.5 Discussion

Humans were reasonably accurate in distinguishing talkers from read–read pairs, consistent with results from other studies (e.g., [KP91]). In contrast, human voice discrimination accuracy decreased considerably for read–pet pairs, with d' less than 1.0 for all such comparisons. One reason for the low accuracy for read–pet pairs might be related to limited phonetic content in pet-directed speech, as noted in Chapter 4. A perceptual speaker identification study supports this view, by showing a greater variety in vowel sound in a speech sample improves identification accuracy [RW93]. Another study claimed that speech sample duration was more important than phonetic variety to accurately remember a voice ([CW97]). This suggests that speaker-specific prosodic characteristics (e.g., pitch, loudness, and speaking rate), pauses between words, and other non-phonemic aspects of voice signal might play a crucial role to assess voice similarity. These cues might sound different between read sentences and pet-directed speech, leading to a bias toward "different talker" responses. For example, there is a significant difference in F_0 between the read sentences and pet-directed speech. The mean F_0 for the read sentences was 221.23 Hz, while that of pet-directed speech was 313.02 Hz [$F(1, 548) = 575.2$, $p < 0.01$]. The extraordinarily high F_0 of the pet-directed speech might have confused listeners, who typically rely heavily on F_0 when assessing talker identity [NMH11, BB10].

Differences in perception when listening to marked versus unmarked talkers emphasize the importance listeners place on specific cues, such as an unfamiliar accent or disfluency, even when stimuli are short ($\approx$ 2 sec). Note that talkers' word choice was not a cue in this experiment, because the sentences were given and the pet-directed speech did not include much lexical variety. Out of the 9 marked talkers, 5 were perceived to have a non-American dialect. In this context, decreases in performance when talkers were perceptually marked, in

part, is consistent with previous findings that accented talkers are more difficult to identify than unaccented talkers, especially when the utterances are short ($<$ 1 sec) [GKB81], and that listeners are better at discriminating among talkers when they are familiar with the phonetic inventory used by particular talkers [KS11, Chap. 7.2.3].

Responses to the speech of the marked talkers were not only less accurate, but may also have been biased to "different talker" decisions, suggesting a stricter criterion was applied by listeners. This is possibly related to that listeners are worse in remembering accented talkers than unaccented talkers as reported by Goldstein et al. [GKB81].

5.2 ASV Experiments

This section describes application of an i-vector/PLDA ASV system to the stimuli described in Section 5.1.1. The same tasks presented to the human listeners were given to the ASV system, permitting a fair comparison between humans and machines.

5.2.1 Method

An i-vector/PLDA ASV system described in Section 4.2.1.2 was used to analyze machine performance. Here, only the recordings from female talkers were used to train the UBM and subspaces because the evaluation utterances were all from female talkers.

Two feature sets, MFCCs and VQual2, were used in the experiments. Second derivatives were not used because they did not provide notable performance gain in preliminary experiments.

We tried to match the utterance duration between the training data and evaluation data for i-vector and PLDA training by truncating the 2.5-minute long original recordings to 2-second segments. However, it did not show any notable performance differences, possibly due to decreased phonetic coverage. Thus, the original recordings were retained. After obtaining PLDA scores from each system, score fusion was performed as described in Section 4.2.1.2.

The AUC and the EER were calculated to measure system performance. AUC values

were estimated, using SYSTAT software, to facilitate comparisons with human performance.

5.2.2 Results and Discussion

Machine and human results are shown in Table 5.3. Similar to results in other studies that improved ASV performance by fusing complementary features (e.g., [DP18]), score fusion generally improved machine performance in the present study. For read–read pairs using all talkers, for example, the AUC for the MFCC feature set, VQual2 feature set, and for the fusion of the two were 0.776, 0.683, and 0.791, respectively. Thus, while the performance of VQual2 alone does not exceed the performance of MFCCs, fusing the two systems seemingly provided complementary information that improved performance. This pattern was observed in most of the other comparisons.

The decrease in performance of the VQual2 features due to style mismatches was smaller than that observed for MFCCs, although overall performance was generally worse for VQual2 features. For unmarked talkers, the EER for VQual2 increased from 36.08% for read–read pairs to 44.09% for read–pet pairs (for a 22.20% relative decrease in performance), while the EER for MFCCs increased from 30.31% to 44.17% (for a 45.73% relative decrease in performance). For marked talkers, the VQual2 EER increased from 41.58% to 44.91% (a 8.01% relative decrease in performance), while the MFCC EER increased from 32.03% to 39.31% (a 22.73% relative decrease in performance).

Robustness to style variability suggests that voice quality features might be useful for conditions that are challenging to conventional cepstral features. Note, however, that experimental results reported in Chapter 4 showed that performance degradation of the VQual2 features due to style mismatches was similar to or worse than that of MFCCs. Unfortunately, a direct comparison with that study is not appropriate because the speech samples used in that study were 5-sec long while the speech samples in this study were about 2-sec long. Since longer utterances benefit both MFCC and VQual2 systems, especially if the phonetic content is richer, it might be the case that the advantage of having more phonetic content, especially in read sentences, outweighed the within-talker variability in speaking style. Thus,

Table 5.3: ASV performance evaluated using the same stimuli as in the perception experiments. The AUC was measured, and the EER (%) was calculated from the ROC curve. Human perception results in terms of AUC and EER are repeated from Table 5.2 in the last column for comparison. The best performance for each condition is boldfaced.

	MFCC		VQual2		Fusion		Human	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Read–read, unmarked talker pairs	0.765	30.31	0.679	36.08	0.780	29.21	**0.885**	**19.02**
Read–pet, unmarked talker pairs	0.587	44.17	0.581	44.09	0.601	47.54	**0.644**	**39.23**
Read–read, marked talker pairs	0.687	32.03	0.657	41.58	0.683	31.78	**0.844**	**24.86**
Read–pet, marked talker pairs	0.593	39.31	0.531	44.91	**0.601**	**37.35**	0.538	46.23
Read–read, all talker pairs	0.776	29.17	0.683	36.18	0.791	28.71	**0.876**	**20.19**
Read–pet, all talker pairs	0.594	43.44	0.587	43.55	0.615	42.79	**0.628**	**40.34**
All pairs	0.716	35.97	0.627	43.18	0.714	36.52	**0.766**	**30.58**

the compounded effect of utterance length, phonetic content and style variability requires further analysis both for humans and machines. In addition, the experiments in Chapter 4 did not consider talker markedness, which might also have been a factor impacting system performance.

Unexpectedly, MFCC performance for read–pet, marked talker pairs (EER = 39.31%) was better than that for unmarked talkers (EER = 44.17%). For VQual2, the performance degraded, but the difference was small (from 44.09% to 44.91%). These results suggest that machines might be somewhat more robust to talkers' markedness than humans are. In that case, machines might play an important role to supplement human decisions when human listeners are confused by markedness. Thus, the effect of markedness on human and machine performance is worth further analysis with a larger number of talkers.

5.3 A Comparison between Human and Machine Performance

This section compares the human and machine voice discrimination results in the face of within-talker variability as presented in Section 5.1.4 and Section 5.2.2. The purpose of the

comparison is to investigate performance differences between humans and machines when large within-talker variability makes the task difficult for both, and to analyze the factors that affect performance. Recall that all tasks are text- and recording-session-mismatched.

For humans, although the EER could be computed from the ROC curve derived from listeners' confidence ratings, this metric might be misleading because humans cannot precisely adjust their decision threshold to make the miss and false alarm rates equal. Additionally, as noted earlier, because EER only focuses on a specific decision threshold, it might not represent overall accuracy. Thus, the AUC is used to compare human and machine performance, and the EER is used only to compare machine performance in different conditions in the rest of the paper.

As shown in Table 5.3, humans performed better than machines in most conditions. For instance, with unmarked talkers, the AUC for ASV score fusion was 0.780 for read–read pairs, compared to AUC = 0.885 for humans. Performance differences between humans and machines could be due to many factors. First, humans can utilize multiple levels of information from the audio signal, but machines rely on frame-level features. For example, humans routinely attend to individual talkers' unique prosody, idiosyncrasies in voice onset time, and so on, but ASV systems consider the distribution of features extracted from 25-msec frames and at most their time derivatives. Second, it is likely that even when humans and machines use similar acoustic information, they process the information in different ways to make same versus different talker decisions.

For read–read pairs, machines were less robust to markedness than humans were. Fusion performance on read–read pairs from unmarked talkers resulted in an AUC of 0.780, while the AUC for marked talkers equaled 0.683 (12.44% relative decrease in performance). Human performance resulted in AUCs of 0.885 and 0.844 (1.24% relative decrease in performance) for the unmarked and marked talkers, respectively. Because the UBM represents the overall smoothed distribution of the acoustic features from a large number of talkers, idiosyncrasies due to talker markedness might not be well-represented with this model. In addition, if similar idiosyncrasies are not well-represented in the pre-training data, the machine will fail to model the between-talker variability from these idiosyncratic differences, leading in turn

to performance degradation.

On the other hand, machines performed better than humans for read–pet pairs from marked talkers. Fusion AUCs for read–pet pairs from unmarked talkers and from marked talkers were both 0.601. However, the AUC for human listeners decreased from 0.644 (unmarked talkers) to 0.538 (marked talkers), a 16.46% relative decrease in performance. These results imply that machines are less sensitive to talker markedness than humans are when the acoustic characteristics of the speech change due to prosody exaggeration. However, it might be difficult to generalize because there were only 9 marked talkers, and the effects of talker markedness on pet-directed speech is not clear. The compounded effect of talker markedness and speaking style on human and machine performance can be explored in the future by including recordings from L2 English speakers.

It was consistently observed that the performance gap between humans and machines was smaller for mismatched speaking styles. For instance, with read–pet, unmarked pairs, the AUC for fusion was 0.601 and the AUC for humans was 0.644, while the AUCs for the read–read, unmarked pairs was 0.780 for fusion and 0.885 for humans. The interesting small performance gap between humans and machines for the read–pet condition will be analyzed in detail in future studies.

5.4 Relationship between Human and Machine Decision Spaces

If humans are more accurate in discriminating different talkers, understanding how humans make decisions might provide insights to improve machine performance. In this section, human and machine performance were investigated with their decision spaces inferred using multi-dimensional scaling (MDS). The relationship between human and machine decision spaces and the relationship between the decision spaces and VQual2 features are analyzed.

5.4.1 Performance Analysis for Subsets with a Smaller Number of Talkers

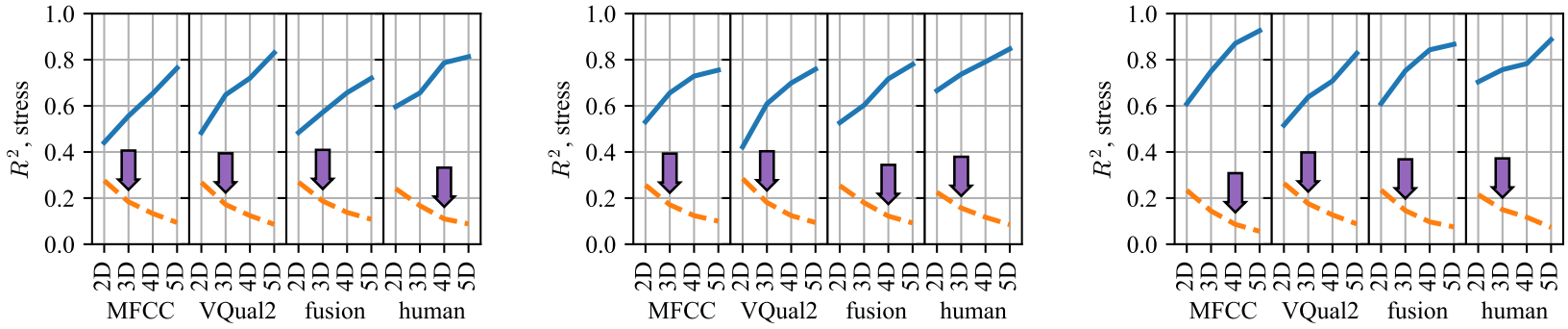
Previous studies [KG96] have shown that listener performance in discrimination tasks is characterized by flexible, idiosyncratic perceptual strategies, such that a feature may be important for distinguishing some pairs of talkers but not others. Given this situation, combining too many talkers in a single analysis obscures the strategies used by listeners, because relations in the "perceptual talker space" become too complicated to summarize even with a large number of parameters. For this reason, we conducted analyses using small ($n = 15$) subsets of the original set of 41 unmarked talkers in this section. The analyses were restricted to read–read sentence pairs, because the main purpose is to investigate the difference in decision strategies between humans and machines, and performance between humans and machines differed most for these pairs. With 15 talkers, each subset had 15 same-talker pairs and 105 different-talker pairs. Ten sets of 15 talkers were randomly selected from the 41 unmarked talkers. Three of the ten subsets (RAND1, RAND2, and RAND3) were chosen for multi-dimensional scaling (MDS) analysis so that each unmarked talker was included in at least one of the subsets. Discrimination data for the read sentence pairs used in the perception experiment were extracted, and the performances of humans and machines were calculated for each subset.

As shown in Table 5.4, the AUC for humans varied between 0.851 and 0.909, and the EER varied between 16.10% and 21.53%. MFCC performance was worse than humans' and more variable across subsets: its AUC varied between 0.679 and 0.772, and the EER varied between 26.61% and 40.57%. The AUC for score fusion varied between 0.713 and 0.792, and the EER varied between 24.18% and 34.02%. VQual2 performance was most consistent (although not the best) among the three ASV systems; its AUC ranged from 0.678 to 0.684 and its EER ranged from 34.44% to 36.75%.

The three subsets showed different rankings of the three ASV systems. In RAND1, the MFCC system had a much better EER (26.61%) than VQual2 (36.75%), and fusion showed the best performance (24.18%). In RAND2, MFCC performance (30.50%) was better than that of VQual2 (34.44%), and was similar to fusion (30.31%). In RAND3, VQual2

Table 5.4: Human and machine performance in terms of EER (%) and AUC. Performance is measured for ten subsets of 15 randomly selected talkers reading sentences. The mean and standard deviation (std) across the ten subsets are shown in the first two rows. Performance on three of the ten subsets (RAND 1, RAND2, and RAND3) used for MDS analysis is shown in the bottom three rows. There were 15 same-talker pairs and 105 different-talker pairs in each subset. Fusion indicates that a linear score fusion is used between the MFCC and VQual2 systems. Performance of the best performing ASV system, is boldfaced for each subset.

	MFCC		VQual2		Fusion		Human	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Mean	0.726	32.99	0.680	35.87	0.744	29.82	0.890	18.49
Std	0.056	6.21	0.040	2.65	0.039	3.80	0.021	2.42
RAND1	0.772	26.61	0.680	36.75	**0.792**	**24.18**	0.851	21.53
RAND2	0.703	30.50	0.678	34.44	**0.722**	**30.31**	0.898	17.61
RAND3	0.679	40.57	0.684	36.35	**0.713**	**34.02**	0.909	16.10



(a) RAND1 (b) RAND2 (c) RAND3

Figure 5.1: Calculated R^2 values (solid line) and stress (dashed line) for the MDS solutions for human data and for the ASV systems using MFCC, VQual2 features, and their score fusion. Arrows point to the elbow in each curve.

performance (36.35%) exceeded MFCC (40.57%), and was improved by fusion (34.02%).

5.4.2 Method

Nonmetric MDS [KW78] was applied to provide insight into the differences in the information utilized by humans and machines. MDS is often used in forensic studies to objectively measure perceived talker similarity to construct fair voice lineups [McD13]. The MDS space can be thought of as a "(perceptual) talker space" where the stimuli are close if they are perceived as similar. The MDS axes can be interpreted by examining correlations between the coordinates of the stimuli and acoustic or other measures of those stimuli: a high correlation suggests the measure might be an important cue for distinguishing talkers.

5.4.2.1 MDS space determination

For each 15-talker subset, confidence ratings from human listening data were combined with same versus different judgements, such that a value of 1 (positive, same talker) was assumed to mean the voices were very similar, and a value of 10 (positive, different talkers) meant they were maximally dissimilar. These scores were averaged across listeners and assembled into lower-half dissimilarity matrices. For the three ASV systems (MFCC, VQual2, and fusion), the dissimilarity between a pair was calculated as the negated PLDA score. Nonmetric MDS was then performed on the human data and on ASV systems for each talker subset. MDS solutions were calculated in 2–5 dimensions for each subset of the data, and solutions were chosen by reference to plots of the number of dimensions extracted versus R^2 and stress ([KW78, pp. 48–60]). R^2 measures the variance in dissimilarities explained by the MDS solution, and stress measures the overall fit of the scaling model to the data. Solutions were chosen based on elbows in plots of stress and R^2 versus the number of dimensions (Fig. 5.1). A four-dimensional solution best fits human data for RAND1, while the solutions are three-dimensional for RAND2 and RAND3.

Table 5.5: R^2 scores of the CCA between the MDS space from the three ASV systems (MFCC, VQual2, and fusion) and human MDS space in each talker subset (RAND 1–3).

	MFCC	VQual2	Fusion
RAND1	0.295	0.300	0.563
RAND2	0.151	-0.099	0.220
RAND3	0.503	0.125	0.284

5.4.2.2 Relationship between human and machine decision spaces

Canonical correlation analysis (CCA, e.g., [TF13]) was used to evaluate the extent to which human and machine talker spaces were related. Here, one set of the variables consists of the MDS coordinates from each of the three ASV systems (MFCC, VQual2, and fusion) for a talker subset, and the other is the MDS coordinates from human responses for the same subset. The R^2 values were calculated in a sense that one set of the variables were predicted by a CCA model from another set of variables. If the first set of variables can be predicted perfectly by the model, the R^2 value is 1. A constant model, which always predicts same values regardless of the input, would get an R^2 value of 0. If a model is worse than the constant model, then the variance of residuals can be larger than the total variance of the data, resulting in a negative R^2.

5.4.2.3 Acoustic correlates of MDS axes

We analyzed how the 11 VQual2 acoustic measures were correlated with MDS spaces for both sets of similarity data. The mean value of each of the 11 acoustic measures was calculated for each utterance from all 50 talkers, and a factor analysis of dimension 5 was undertaken to reduce the number of predictor variables. A similar procedure was applied to the standard deviations of the acoustic measures. The absolute factor loadings, which reflect the correlations between the acoustic measures and factors, are shown in Fig. 5.2.

For the acoustic means, factor 1 was mostly related to the formant amplitudes (A_1, A_2,

and A_3), CPP, and F_2, and factor 2 was related to F_2, and A_3. Factor 3 was highly correlated with H_2-H_4 and F_1, and factor 4 showed a strong relationship with F_0. Factor 5 was related to F_3. For the standard deviations, factor 1 was highly correlated with formant amplitudes (especially A_2) followed by H_4-H_{2k} and CPP, and factor 2 was related to F_0, and to the first and the second formant frequencies. No feature was correlated with factor loading magnitude greater than 0.5 for factors 3, 4, and 5.

Next, factor scores calculated at the utterance level were averaged within each talker, after which we constructed 5-dimensional acoustic talker spaces for each subset. Finally, the relationship between the acoustic space and the MDS spaces was analyzed using multiple regression.

5.4.3 Results

5.4.3.1 Relationship between human and machine decision spaces

The resulting R^2 score using 3-component CCA is shown in Table 5.5. Dimensions of the machine MDS spaces were insufficiently interpretable in terms of the dimensions of the human perceptual space, suggesting that machines and humans used different strategies to discriminate talkers. For RAND1, at most 56.3% of the variance in the ASV talker space was explainable using the dimensions from the speaker space derived from perceptual data. For RAND2/VQual2, the negative R^2 value indicates that the estimated model was worse than the constant model. The overall low R^2 values suggested that there was little relationship between human and machine talker spaces, at least when a linear model was used.

If we compare the CCA results in Table 5.5 with the EER performance in Table 5.4, we notice that the relationship between the model fit and system performance was weak. For example, the best performing ASV system in Table 5.4 was fusion for all subsets. In Table 5.5, however, fusion showed the highest R^2 value for RAND1 and RAND2, but not for RAND3. In addition, even though the R^2 value of RAND3 MFCC was the second highest ($R^2 = 0.503$), its performance was the worst (EER = 40.57%) among all three subsets. The weak relation between human and machine talker spaces suggests that acoustic information

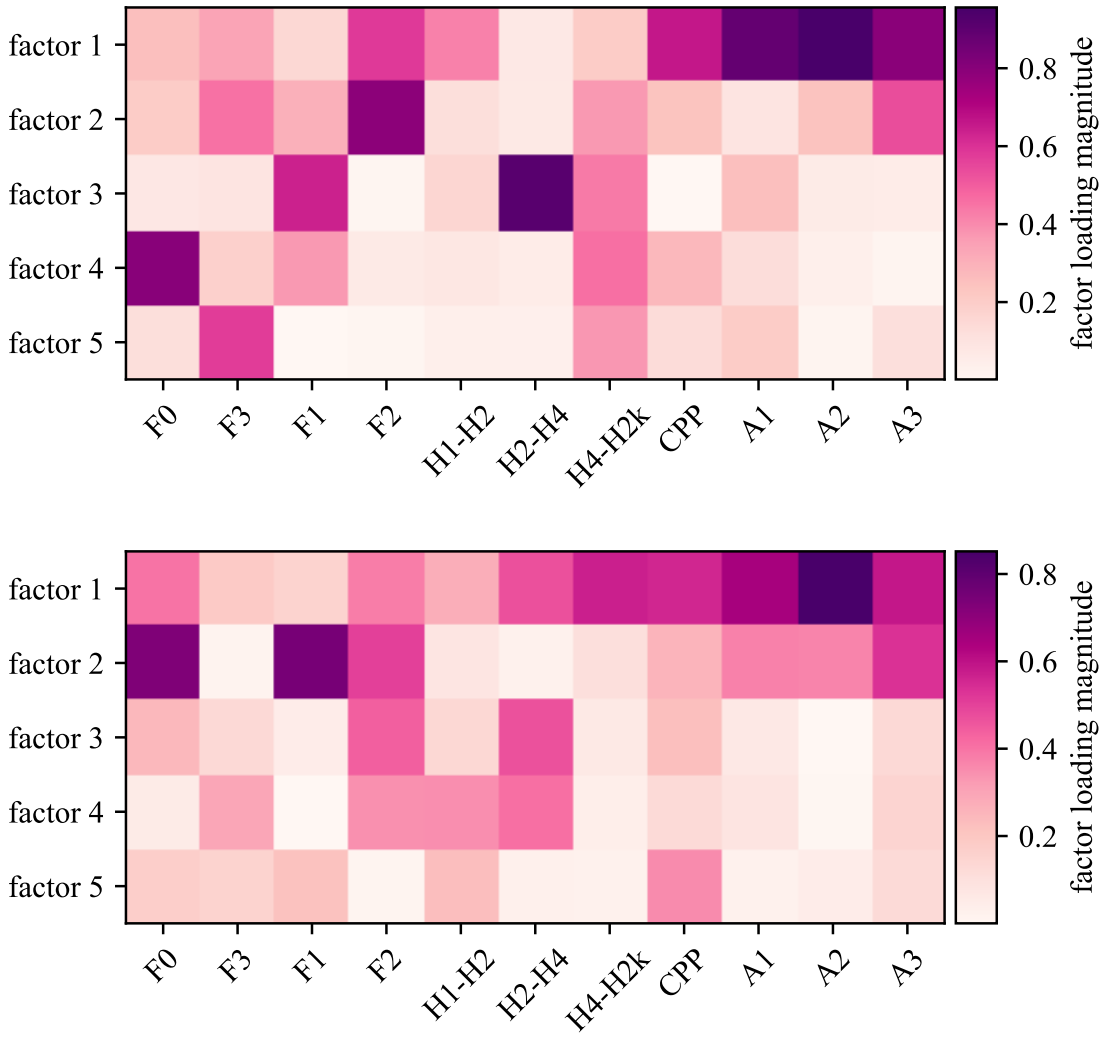


Figure 5.2: Absolute values of the factor loadings for acoustic measures. Darker color indicates greater factor loadings. A 5-dimensional factor analysis was performed using the means (top) and standard deviations (bottom) of the acoustic measures for each utterance for dimensionality reduction.

might be used differently by humans versus machines.

5.4.3.2 Acoustic correlates of MDS axes

Multiple regression results between the MDS spaces and acoustic factors are shown in Table 5.6. Interestingly, factors estimated from the means of the acoustic measures were related to the most important dimension ($D1$) of the perceptual talker spaces for all subsets, and the factors from the standard deviations, which can be related to the within-utterance variability, were related to $D1$ of the MFCC talker spaces for all subsets. For humans, factors 4 and 5, derived from mean acoustic measures, were statistically significant ($p < 0.05$) for the multiple regression model in RAND1 and RAND2, and RAND1 and RAND3, respectively. Recall that factor 4 was highly related to F_0 and factor 5 was related to F_3. These results are consistent with previous studies that reported F_0 and F_3 being the most important acoustic predictors of human judgements (e.g., [BB10, NMH11]). In RAND2, factor 2 from the mean data, which was related to F_2, F_3 and A_3, was significantly related to human $D1$. For the standard deviations, RAND3, factor 1 was related to formant amplitudes, and was significantly related to human $D1$. These results suggest that formant amplitudes might also provide important information for human decision making.

For MFCCs, factor 5 from the standard deviation data was significantly related to $D1$ for subsets RAND1 and RAND3. For VQual2, which was derived from a psychoacoustic model of voice quality, none of the MDS dimensions was significantly associated with any factor(s), even though the factors were estimated using voice quality features. R^2 values in linear regression only reflect the linear part of the decision-making process, but there are other parts that are not linear. Thus, even though the VQual2 system makes decisions based on the VQual2 feature set, its decision space might not be fully interpretable as a linear combination of those features.

Table 5.6: Multiple regression results on human and machine MDS coordinates (dependent variables) with acoustic talker spaces (independent variables). The first three columns show R^2, F-statistics, and p-values of the multiple regression models. Only the MDS dimensions which can be modeled with $p < 0.05$ are shown in the table. SE, T and p, which indicate the standard error, t-statistics, and p-values of the independent variables, are shown for each of the factors. The independent variables with $p < 0.05$ are boldfaced.

	Model			Factor 1			Factor 2			Factor 3			Factor 4			Factor 5		
	R^2	F(5,9)	p	SE	T	p	SE	T	p	SE	T	p	SE	T	p	SE	T	p
Means																		
RAND1 human D1	0.76	5.82	0.01	0.25	-0.48	0.64	0.31	-0.96	0.36	0.32	-2.05	0.07	**0.22**	**-2.74**	**0.02**	**0.15**	**-3.77**	**0.00**
RAND2 MFCC D2	0.74	5.13	0.02	0.41	2.17	0.06	**0.35**	**3.21**	**0.01**	0.28	-2.23	0.05	0.19	0.54	0.60	0.26	0.50	0.63
RAND2 fusion D1	0.70	4.18	0.03	0.42	2.22	0.05	0.36	1.49	0.17	0.29	1.10	0.30	**0.19**	**2.66**	**0.03**	0.27	-0.39	0.71
RAND2 human D1	0.86	11.44	0.00	0.35	-1.44	0.19	**0.30**	**-3.25**	**0.01**	0.24	-0.17	0.87	**0.16**	**-4.30**	**0.00**	0.22	-2.01	0.08
RAND3 human D1	0.90	15.66	0.00	0.19	-0.13	0.90	0.22	-1.93	0.09	0.15	0.88	0.40	0.15	-0.29	0.78	**0.13**	**-5.18**	**0.00**
Standard deviations																		
RAND1 MFCC D1	0.71	4.36	0.03	0.31	0.38	0.71	0.27	1.72	0.12	0.28	0.40	0.70	0.34	1.71	0.12	**0.35**	**-3.23**	**0.01**
RAND2 MFCC D1	0.78	6.44	0.01	0.30	-0.20	0.85	0.25	-2.25	0.05	**0.28**	**-3.03**	**0.01**	**0.26**	**2.34**	**0.04**	0.26	-0.70	0.50
RAND3 MFCC D1	0.67	3.64	0.04	0.33	-0.59	0.57	0.32	-0.49	0.64	0.36	1.18	0.27	0.28	-1.37	0.20	**0.38**	**-3.27**	**0.01**
RAND3 MFCC D2	0.92	19.81	0.00	**0.15**	**-5.32**	**0.00**	0.15	0.49	0.64	**0.17**	**-3.76**	**0.00**	0.13	-2.19	0.06	0.18	-1.74	0.12
RAND3 fusion D1	0.78	6.43	0.01	**0.30**	**2.39**	**0.04**	0.29	-0.73	0.48	0.33	1.37	0.20	0.26	1.57	0.15	0.34	-1.49	0.17
RAND3 human D1	0.66	3.51	0.05	**0.40**	**2.30**	**0.05**	0.39	0.20	0.84	0.44	0.63	0.54	0.34	0.72	0.49	0.46	-1.64	0.14

5.4.4 Discussion

Across all three subsets of the read–read, unmarked stimuli, humans were more accurate and consistent at voice discrimination than were machines. However, subsets differed in how difficult they were for humans versus machines, and human and machine talker spaces were not strongly related in terms of the features that explained stimulus confusability. Differences between humans and machines could have occurred because humans utilized information that was not explicitly given to machines, such as spectro-temporal information and linguistic knowledge, and/or they used similar acoustic features but processed them differently.

The present results do not allow us to evaluate these possibilities. To evaluate the first possibility, an automatic system that can process supra-segmental information is needed. The widely used, frame-level feature based ASV system used in the current study is not explicitly given such information. Other systems that utilize prosodic information to model talkers (e.g.,[DDK07]) need to be developed. Evaluating the second possibility would require a more complex model of how features are processed and used in decision making. For example, even though the VQual2 system made decisions based on voice quality features, the decisions did not appear to depend on the linear combination of the means and standard deviations of the features that explained human and MFCC system performance.

Instead, the results highlight differences in human and machine decision making. For example, the most important dimensions underlying human responses were highly related to the means of voice quality features, while MFCC responses were more closely related to standard deviations of the same features. This might indicate that humans perform best with the talkers whose voices are separated apart in mean values, while MFCC-based systems work best when the within-utterance variance is large so that the acoustic information coverage in an utterance is sufficient to model the talker.

5.5 Comparison between Human and Machine Responses and Reliability

In the previous section, talker spaces inferred from the human and machine responses were compared using MDS. Here, we analyze a different aspect of human and machine responses, focusing on the direct relationship between them and the responses' reliability. Neurological data showed that speaker recognition and discrimination are separate abilities [VK87]. Considering that, perceptual strategies to identify similarity might be different from those to detect dissimilarity between talkers. For example, a talker clustering experiment showed that human listeners' performance for 'telling people together' and 'telling people apart' differed significantly [LBG18]. In this context, we relate responses by humans and the ASV systems for target (same-talker pairs) and non-target trials (different-talker pairs), separately.

5.5.1 Method

The human and machine voice discrimination results presented in Section 5.1.4 and Section 5.2.2 were used. In this study, read sentence pairs from all 50 female speakers were investigated. Read sentence pairs were selected because the main focus was to analyze the difference in responses and reliability between human and machine, and performance between them differed more for these pairs than read–pet pairs.

5.5.1.1 Evaluation metric

Although EER and AUC are widely-accepted metrics for evaluation, they are not suitable for analyzing target and non-target trials separately. Thus, in this study, the detection cost function (C_{det}), commonly known as DCF, and the log-likelihood-ratio cost function (C_{llr}) were used for performance evaluation [LB07]. The dissimilarity scores δ for human responses were calculated as defined in Eqn. 3.1, and these scores were averaged across listeners. For the ASV systems, the PLDA score, which represents the ratio of the likelihood that the given pair of utterances are from the same talker to the likelihood that the pair is from two

different talkers, was used. After obtaining the dissimilarity scores from human listeners and PLDA scores from each of the automatic systems, the scores were calibrated using standard logistic regression. The resulting calibrated log-likelihood-ratio (LLR or L) represents the scalar responses by humans and the two automatic systems.

C_{det} is defined as the expected cost of detection errors. It is a measure of discrimination suitable for evaluating application-dependent performance. For our application, C_{det} was obtained with cost of misses set at 25 and cost of false alarms set at 1 (25 was the ratio between non-target and target trials).

On the other hand, C_{llr} is defined as an integral over a spectrum of operating points of C_{det}. Thus, C_{llr} is an application-independent measure for evaluating scalar responses. It can be interpreted as a measure of loss of information, thus the lower the C_{llr}, the more the average information per trial (in bits) increases by applying the system. C_{llr} has an analytic solution as shown in [LB07]:

$$C_{\text{llr}} = \frac{1}{2}\left(\sum_{t\in\text{tar}} \frac{\log_2(1+e^{-L_t})}{N_{\text{tar}}} + \sum_{t\in\text{non}} \frac{\log_2(1+e^{L_t})}{N_{\text{non}}}\right) \tag{5.1}$$

where L_t is the log-likelihood-ratio for trial t; 'tar' is a set of N_{tar} target trials and 'non' is a set of N_{non} non-target trials. The two normalized summation terms represent expectations of 'log costs' for target trials (first term) and for non-target trials (second term), respectively.

For example, consider a trial t_1. If it is a target trial, the cost is $C_{\text{llr}}^{\text{tar}} = \log_2(1+e^{-L_{t_1}})$. If a system correctly gives a high degree of support for the target hypothesis, i.e. $L_t \gg 1$, then the cost is close to zero ($C_{\text{llr}}^{\text{tar}} \approx 0$). On the other hand, if the system incorrectly gives a high degree of support for the non-target hypothesis ($L_t \ll -1$), then the cost becomes high ($C_{\text{llr}}^{\text{tar}} \approx |L_t|$). The cost for non-target trials, $C_{\text{llr}}^{\text{non}} = \log_2(1+e^{L_t})$, can be understood in a similar way. For a neutral log-likelihood ratio ($L_t = 0$), $C_{\text{llr}}^{\text{tar}} = C_{\text{llr}}^{\text{non}} = 1$. That is, the reference system, which does not process speech and simply outputs $L_t = 0$ for every trial, will have a cost of $C_{\text{llr}} = 1$. A poor system might result in $C_{\text{llr}} > 1$, indicating a worse performance than the reference system.

The Bosaris toolkit [BD11] was used to calibrate the raw scores and for calculating C_{det} and C_{llr}. As limited amount of data were analyzed, and as the main purpose of the study

was to analyze calibration-independent performance, the calibration was trained and used on the same dataset.

5.5.1.2 System fusion

Systems were fused based on the logistic regression method [BBC07] using the Bosaris toolkit [BD11]. The fusion trains combination weights to fuse multiple systems providing a calibrated set of log likelihood ratios. Note that human and machine responses were all converted to L scores. This consistency in metrics enabled fusion between human and machine responses.

5.5.1.3 Talker-level analysis

The L and C_{llr} values were analyzed at the talker level. In this way, we investigated how responses by humans and machines as well as their reliability differ talker-by-talker. L_t is a measure of how much the system considers the pair in the trial t to be from a single talker. That is, this score has a larger value if the voices in the pair sound "similar" to the system.

For each of the 50 talkers, the L_t values for the trials including that talker were computed. Then, mean values of L_t for target and non-target trials were calculated separately, denoted as L^{tar} and L^{non}, respectively. If L^{tar} is large for a talker, this indicates that the talker has small within-talker variability. Similarly, if L^{non} is large for a talker, it indicates that the talker has small between-talker variability, and it is difficult for the system to distinguish her from others.

$C_{\text{llr}}^{\text{tar}}$ and $C_{\text{llr}}^{\text{non}}$, at the talker level, were calculated in a similar manner. C_{llr} can be representative of the reliability of the L score. The lower the C_{llr}, the more reliable the system responses are for the talker.

5.5.2 Results and Discussion

5.5.2.1 Human and machine performance

Human and machine performances are summarized in Table 5.7. Consistent with the previous results using EER and AUC, humans performed better than machines. For example, for humans C_{det} was as low as 0.273, while values for the MFCC-based system and VQual-based system were 0.500 and 0.682, respectively. Humans performed even better than fusion of the two automatic systems, which had $C_{\text{det}} = 0.513$. In addition, fusing human responses with any automatic system improved the performance, consistent with [GHR13]. This trend was preserved with different false alarm cost values, and for the C_{llr} values.

Fusion between systems improved the performance, suggesting complementarity among systems. VQual2 contributed to decreasing different types of cost when fused with another system. For example, when MFCC was fused with VQual, the $C_{\text{llr}}^{\text{non}}$ decreased from 0.739 to 0.721, without changing $C_{\text{llr}}^{\text{tar}}$. On the other hand, when humans' scores were fused with VQual2, the $C_{\text{llr}}^{\text{non}}$ was not affected while the $C_{\text{llr}}^{\text{tar}}$ decreased from 0.417 to 0.405. MFCCs provided more complementary information to human responses than VQual2 features did; they reduced $C_{\text{llr}}^{\text{tar}}$ and $C_{\text{llr}}^{\text{non}}$ from 0.417 to 0.342 and from 0.434 to 0.368, respectively.

5.5.2.2 Log-likelihood-ratio analysis

Talker-level L^{tar} and L^{non} are shown in Fig. 5.3. For humans, the target trial distribution had a smaller variance compared to that of the ASV systems. Additionally, the L^{tar} and L^{non} distribution for humans were well-separated. This explains higher human accuracy compared to machines.

Interestingly, the distributions of human responses were non-Gaussian and skewed towards correct responses. This tendency was more evident for the L^{tar} than L^{non}. That is, humans were quite sure when they made decisions, and they were more positive when they made "same talker" responses than "different talker" responses. Listeners were more conservative in making "different talker" responses, and the responses had more Gaussian-like

Table 5.7: ASV performance for all 50 talkers in terms of detection cost functions (C_{det}), log-likelihood-ratio cost (C_{llr}), log-likelihood-ratio cost for target trials ($C_{\text{llr}}^{\text{tar}}$), and log-likelihood-ratio cost for non-target trials ($C_{\text{llr}}^{\text{non}}$). The plus ('+') symbol indicates a fusion between the systems. Best performance among individual systems and among fused systems are boldfaced.

	C_{det}	C_{llr}	$C_{\text{llr}}^{\text{tar}}$	$C_{\text{llr}}^{\text{non}}$
MFCC (M)	0.500	0.737	0.736	0.739
VQual2 (V)	0.682	0.884	0.897	0.872
Human (H)	**0.273**	**0.425**	**0.417**	**0.434**
M+V	0.513	0.728	0.736	0.721
H+M	**0.216**	0.355	0.342	0.368
H+V	0.273	0.419	0.405	0.434
H+M+V	0.231	**0.353**	**0.341**	**0.365**

distribution than that of "same talker" responses.

Next, the correlations between the L^{non} for humans and the two ASV systems were analyzed to understand which acoustic information was related to human responses (see Table 5.8). Compared to MFCCs, VQual2 had a higher correlation with the human responses for L^{non} ($r = 0.610$). This might be related to the finding that human experts' decisions based on voice quality information could resolve false acceptance by an MFCC-based system [HHF17]. Interestingly, for the 9 marked talkers, the correlation was even

Table 5.8: Correlation coefficients of L^{tar} and L^{non} per speaker between each of the two ASV systems (MFCC and VQual) and humans.

	MFCC	VQual2
L^{tar}	0.127	0.216
L^{non}	0.273	0.610

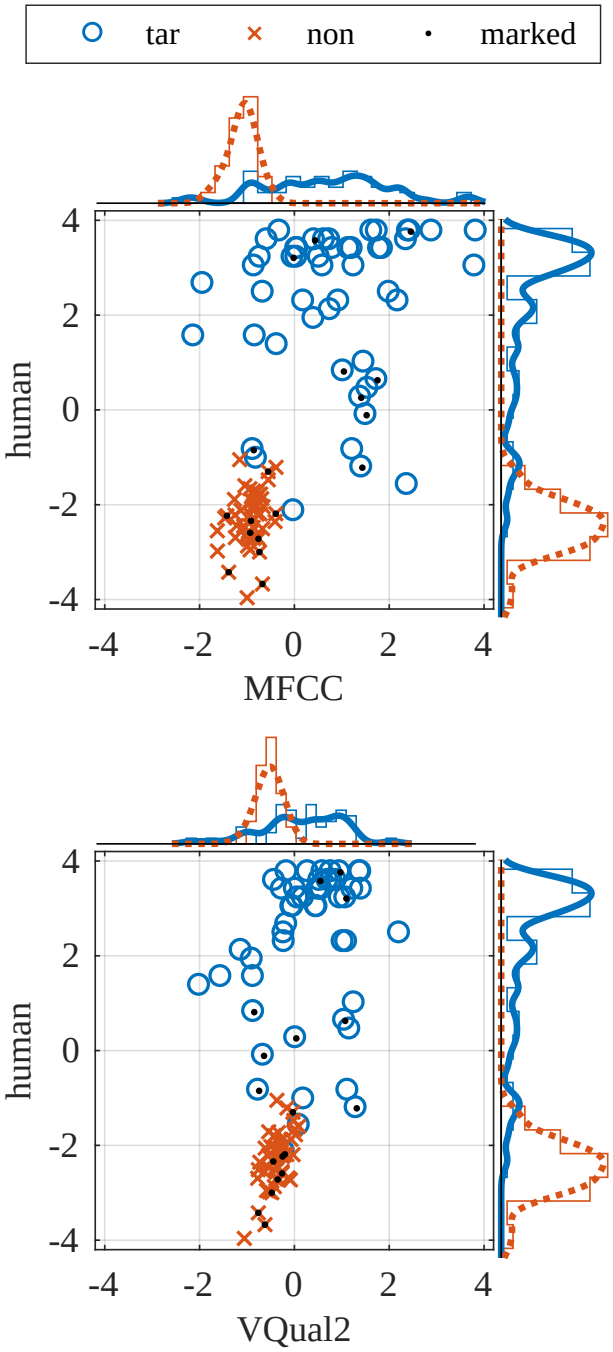


Figure 5.3: Scatterplots of L^{tar} and L^{non} per talker comparing MFCC vs humans (top) and VQual2 vs humans (bottom). L^{tar}s are denoted with discs ('$\circ$'), and L^{non}s are denoted with crosses ('$\times$'). Dots ('$\cdot$') indicate the talkers are perceptually marked.

higher ($r = 0.912$). This might be related to findings that when linguistic cues are limited in the stimuli, human listeners assess talker similarity of non-target pairs by relying on voice quality [SFH16]. These results suggest that the VQual2 features are related to human responses for non-target trials, especially when the talker has perceptual markedness. This hypothesis will be tested in a follow-up study including more marked talkers.

This tendency was not apparent for L^{tar}. It is possible that humans' "same talker" decisions involve more information than "different talker" decisions. It can be hypothesized that humans make a "same talker" decision only when all (or most) aspects of the voices match, while they make a "different talker" decision when any aspect does not match. Unfortunately, because only one target trial per talker was made in this experiment, as opposed to 25 for the non-target trials, it is difficult to analyze what acoustic information was correlated with human responses for target trials.

5.5.2.3 Log-likelihood-cost analysis

In Section 5.1.4, it was noted that human performance degraded when talkers were perceptually marked. To analyze the relationship between talker markedness and system reliability in terms of the information loss, C_{llr} values were analyzed. Table 5.9 shows C_{llr}, $C_{\text{llr}}^{\text{tar}}$, and $C_{\text{llr}}^{\text{non}}$ values for the 9 perceptually marked talkers, 4 marked talkers who were monolingual English speakers, and 5 marked talkers who had non-American accents. The 4 monolingual marked talkers did not have non-American accents, but they had unusual dysfluencies in reading.

For humans, the mean $C_{\text{llr}}^{\text{tar}}$ among the marked talkers was 0.784 compared to 0.417 for all talkers. For MFCCs, the mean $C_{\text{llr}}^{\text{tar}}$ among the marked talkers was 0.574 compared to 0.736 for all talkers. VQual2 showed no significant difference between the marked talkers and all talkers (0.909 and 0.897). That is, the MFCC-based system could take advantage of acoustic information for "same talker" decisions from talker markedness, while humans and VQual2-based system could not.

Moreover, when selecting the 4 monolingual English speakers among those marked talk-

Table 5.9: C_{llr}, $C_{\text{llr}}^{\text{tar}}$, and $C_{\text{llr}}^{\text{non}}$ values for perceptually marked talkers ($n = 9$), monolingual marked talkers ($n = 4$), and marked talkers with non-American accents ($n = 5$). Values for all 50 talkers are reported in Table 5.7.

	Marked			Monolingual			Accent		
	C_{llr}	$C_{\text{llr}}^{\text{tar}}$	$C_{\text{llr}}^{\text{non}}$	C_{llr}	$C_{\text{llr}}^{\text{tar}}$	$C_{\text{llr}}^{\text{non}}$	C_{llr}	$C_{\text{llr}}^{\text{tar}}$	$C_{\text{llr}}^{\text{non}}$
MFCC	0.671	0.574	0.767	0.746	0.664	0.828	0.611	0.503	0.718
VQual2	0.883	0.909	0.856	1.015	1.121	0.919	0.777	0.740	0.814
Human	0.576	0.784	0.368	0.996	1.445	0.548	0.240	0.256	0.224

ers, human $C_{\text{llr}}^{\text{tar}}$ increased to 1.445. The other 5 talkers' C_{llr} was much lower (0.256). This suggests that humans were not able to detect a consistent pattern in dysfluencies, but they could detect patterns for making "same talker" decisions for the 5 talkers with non-American dialects. This hypothesis will be tested in future studies by including more target trials and marked talkers.

5.6 General Discussion

Experiments reported in this chapter and key results are summarized in Table 5.10. Human and machine voice discrimination performance on short-utterance, text-independent stimulus pairs were investigated in this study. Read sentences, about 2-sec long, were used to evaluate performance with clear speech, and excerpts from pet-directed speech of similar duration were used to investigate the effect of exaggerated prosody. Analyses compared performance when pairs matched (read–read) or mismatched (read–pet) for speaking style.

Results showed that human listeners were reasonably accurate at discriminating voices based on read–read pairs, but performance degraded significantly with style-mismatched pairs. Contrary to expectations, humans performed worse when discriminating between two marked talkers than when discriminating between two unmarked speakers both for read–read pairs and read–pet pairs. The effect of talker markedness on voice discrimination is worth

exploring in detail in the future. The UCLA Speaker Variability Database includes many non-native speakers of English whose speech could be useful for this purpose.

The machines tested here were less accurate than humans for read–read pairs, which is consistent with previous studies that reported poor ASV performance with short-utterance text-independent tasks. Performance degraded even more with pet-directed speech for unmarked talkers, especially for MFCC- and VQual2-based systems, either because prosody exaggeration distorted acoustic features or because the databases used for the pre-training did not have a similar speaking style. Score-level fusion of the two systems improved performance, suggesting that VQual2 provides information that is complementary to MFCCs. This feature set may be especially valuable when within-talker variability is large. Interestingly, with style-mismatched pairs, talker markedness had little effect on VQual2 features, and MFCC and fusion performance even improved for these pairs, to such an extent that machines outperformed human listeners. Unfortunately, the number of marked talkers in this study was not large enough to ensure that this result is robust. A follow-up study will analyze what advantage machines have when human performance is critically affected, and how to utilize that advantage in speaker verification tasks.

Human and machine performance on read–read pairs of unmarked talkers was further investigated with MDS on smaller subsets of talkers. CCA results between human and machine talker spaces showed a weak relationship between the human and machine spaces. Further, better machine performance did not lead to an increase in the strength of this association. These results suggest that humans and machines use different strategies to distinguish talkers. Multiple regression between acoustic feature factors and MDS spaces for humans and machines found that human MDS axes were reasonably well-modeled as linear combinations of means of voice quality features. On the other hand, neither MFCC nor VQual2 MDS spaces could be well-modeled using mean values. These findings suggest that investigating how voice quality feature means are related to human responses might provide valuable insights into perceived talker identity. The knowledge could also prove useful for improving machine performance, by exploring how to process acoustic features effectively.

Different aspects of voice discrimination decisions by humans and machines were inves-

tigated, focusing on analyzing their responses and reliability. System responses were re-evalutated in terms of the log-likelihood-ratio, and the reliability was calculated in terms of the log-likelihood-ratio cost function. Target and non-target trials were analyzed separately.

Consistent with previous results, human listeners were considerably more accurate than machines. Higher confidence for correct target decisions compared to correct non-target decision was observed in human response distribution. For non-target trials, system responses per talker were highly correlated between humans and VQual2, especially when the talkers were perceptually marked. In addition, when humans are distinguishing between talkers, they seem to use an approach similar to the VQual2 system.

For target trials, humans response reliability decreased for marked speakers compared to when all the speakers were considered. However, MFCC response reliability was higher for marked talkers than all talkers. This suggests that MFCCs could extract information from talker markedness for target trials, while VQual2 response reliability was not affected by talker markedness. Results suggest that machines might be able to supplement human listeners in such conditions.

In future studies, perception experiments will include more target trials, as well as more marked talkers. In addition, machine performance can be examined by varying the training data conditions, such as talkers' language background, gender, and/or recording conditions. Modeling prosodic features might also be a promising research direction for ASV, as is examining how effectively human and machine decisions can be combined.

Table 5.10: Summary of experiments (Ex.) reported in Chapter 5. Fusion indicates a weighted combination of the scalar responses from individual systems.

Perception	Speaking style	read sentences, pet-directed (both ≈ 2 sec)
	No. talkers	50 females
	No. listeners	65
ASV	Speaking style	read sentences, pet-directed (both ≈ 2 sec)
	No. talkers	50 females
	Systems	MFCCs, VQual2, fusion
Human vs. machine comparison	Speaking style	read sentences
	Method	Ex.1: decision spaces inferred with multidimensional scaling
		Ex.2: log-likelihood-ratio cost function (C_{llr})
Key results		Human and machine decision spaces were weakly correlated. VQual2 system responses were most related to human responses for non-target pairs.

CHAPTER 6

Applications of Voice Quality Features in Affect Recognition

In Chapters 3 through 5, the effectiveness of voice quality features for improving automatic speaker discrimination performance and for predicting human speaker discrimination responses was shown. In this chapter, the application of the voice quality feature set (VQual2) is extended to affect recognition. Voice quality has been frequently associated with affect [Sch86, MA93], and a strong relationship between voice quality and perceived affect was found through experimental studies with human listeners [GN03]. Thus, we expect the VQual2 feature set to be correlated with perceived affect, and hence may improve automatic classification of it. The study presented here aims to automatically classify perceived affect from speech samples spoken by mentally, neurologically, and/or physically disabled individuals, as a participation in the Interspeech 2018 Atypical Affect subchallenge [SSB18].

6.1 Data

The Atypical Affect challenge provided speech samples from the EmotAsS (EMOTional Sensitivity ASsistance System for people with disabilities) database [HSC17]. The database consists of spontaneous speech samples from 8 female and 7 male German talkers with mental, neurological, and/or physical disorders. Among the 15 talkers, 12 had mental, 2 had neurological, and 1 had multiple disabilities. No further details on the disabilities was

Parts of this chapter were published in [PAC18].

Table 6.1: Number of utterances per class in training/development/testing subsets for the Atypical Affect challenge [SSB18].

	Training	Development	Test
Angry	125	50	272
Happy	743	2,287	650
Neutral	2,287	2,842	2,024
Sad	187	329	153
Total	3,342	4,186	3,099

provided, due to strict privacy restrictions. The talkers were recorded in a familiar room at their workplace while speaking about their personal and health issues. Recordings were made with a ZoomH6 and a Jabra Speak 510 microphone, both at a sampling rate of 44.1 kHz. A total of 10,627 segments of speech (9.2 hours) were collected, and they were annotated by, on average, 12 volunteering human listeners through a gamified crowdsourcing platform iHEARuPLAY [HSC17]. The annotators were asked to choose among 6 emotions (anger, disgust, fear, happiness, sadness, and surprise) and neutral. Because only a few samples were annotated as disgust, fear, and surprise, these samples were discarded, resulting in 4 classes: anger, happiness, sadness, and neutral. The data split into training/development/test subsets is shown in Table 6.1. This dataset is denoted as the Atypical Affect dataset.

In order to cope with the limited amount of data in the Atypical Affect dataset, supplemental data for training were collected from the dataset for the Self-Assessed Affect subchallenge [SSB18]. This corpus was chosen because it contained German spontaneous speech, consistent with the Atypical Affect dataset. The Self-Assessed Affect dataset included recordings from 100 talkers (85 females, 15 males), who told two negative and two positive stories, each with a duration of about 5 minutes. The sampling rate was 44.1 kHz.

6.2 Method

6.2.1 Acoustic Features

The computational paralinguistics challenge provided a baseline feature set [SSB16, WES13] that can be extracted using the OpenSMILE toolkit [EWS10]. The set consists of F0, energy, spectral, cepstral coefficients and voicing-related frame-level features which are referred to as low-level descriptors. They also include the zero-crossing rate, jitter, shimmer, harmonic-to-noise ratio, spectral harmonicity and psychoacoustic spectral sharpness. The complete feature set had a dimension of 65. This feature set is denoted as the ComParE16 feature set.

Mel-frequency cepstral coefficients (MFCCs) of dimension 20, including the zeroth coefficient, were extracted using a 25 msec window and a 10 msec frame shift. VQual2 features of dimension 11 were extracted as described in Chapter 4. The first and second derivatives were also computed for MFCCs and VQual2. The complete feature set had dimensions of 60 and 33 for MFCCs and VQual2, respectively.

6.2.2 Utterance Representation

The baseline system provided by the challenge calculated various statistics within an utterance, such as the mean and standard deviation, of the ComPareE16 feature set representing each utterance. We propose to use the supervector framework [CSR06] to model the distribution of acoustic features within an utterance. The utterance-level representations were then used for affect classification.

Compensating for the effect of a particular disorder on speech signals was not explicitly attempted in this study because the available metadata did not include information to perform such analysis. Instead, we focused on predicting perceived affect regardless of the kind of disorder.

6.2.2.1 Supervector construction

In the supervector framework [KMD03], each utterance is represented with a single vector that is constructed by concatenating the mean vectors of a Gaussian mixture model (GMM) representing the feature distribution within an utterance. The mixture model is often adapted from the universal background model (UBM), which is a statistical model for average speech sounds, usually trained with a large amount of recordings from a large number of speakers. The supervector summarizes the feature distribution within an utterance, and thus can be used as an utterance representation. A further process can be made using the i-vector framework. In that framework, the supervector is projected to a low-dimensional subspace in order to efficiently represent the utterance-specific characteristics [HH15]. The i-vector approach is most effective when a large database including a wide range of affect and speaker variability is available. Considering that the available amount of data was limited to train both the UBM and the i-vector subspaces, we decided to directly use supervectors for this task.

A UBM was constructed using the 'neutral' class of the Atypical Affect dataset, and the data provided for the Self-Assesed Affect subchallenge. After the UBM was trained, the feature distribution of each utterance was modeled with a GMM adapted from the UBM using the maximum a posteriori criterion [CSR06]. The supervector, which is the concatenated mean vectors of the GMM, was used to represent each utterance.

6.2.2.2 Utterance clustering based on F0

F0 distributions for the training and development datasets were bimodal, as shown in the first column of Figure 6.1. The bimodal distribution suggests the possible effect of gender differences (females having, on average, a higher F0 than males). It was also observed that the F0 distribution for the training and development datasets were not similar. For example, there were two peaks in the F0 distribution of the 'happy' class in the training dataset, while the F0 distribution of the same class in the development dataset did not show a clear second peak. The mismatch in the F0 distribution between the datasets suggests

that the gender distribution in the datasets might differ significantly. Considering that affect representation can be different across gender [VA06], gender distribution mismatch could degrade classification performance. For example, if the majority of the training data were from males while the development data had more females than males, the statistical model built on the features extracted from the training data might not be able to accurately predict the affect for the development data.

One possible approach to alleviate the mismatch problem is gender-dependent modeling, but gender labels were not available. Hence, we used F0 to group the utterances into two clusters. A Gaussian mixture model with two mixtures was used for clustering, based on the median value of F0 within each utterance. The resulting cluster size differed significantly across datasets. For example, in the training dataset, the number of utterances for the 'happy' class was 165 and 578 in low-F0 cluster and high-F0 cluster, respectively, while in the development dataset, the corresponding numbers were 632 and 333.

Example feature distributions (F0 and H4-H2k) without and with clustering are shown in Figure 6.1. Note that clustering was performed at the utterance level, while the feature distribution is computed at the frame level. Because utterances can have low-F0 frames while the utterance F0 median value is high, frame-level F0 distribution might not show a clear separation between clusters. Clustering resulted in a more matched distribution between the training and development datasets, especially for 'happy' and 'angry' F0 distributions, and for the H4-H2k distribution in the 'happy' class.

However, this clustering inevitably results in a reduced amount of data for training within each cluster, which might introduce limitations to classification performance.

6.2.3 Affect Classification

6.2.3.1 The Gaussian mixture model classifier

Because the number of utterances in each affect class is limited, training the Gaussian mixture models (GMMs) directly from the samples within each class is prone to overfitting. In order to mitigate the overfitting problem, a GMM was trained using the 'neutral' class,

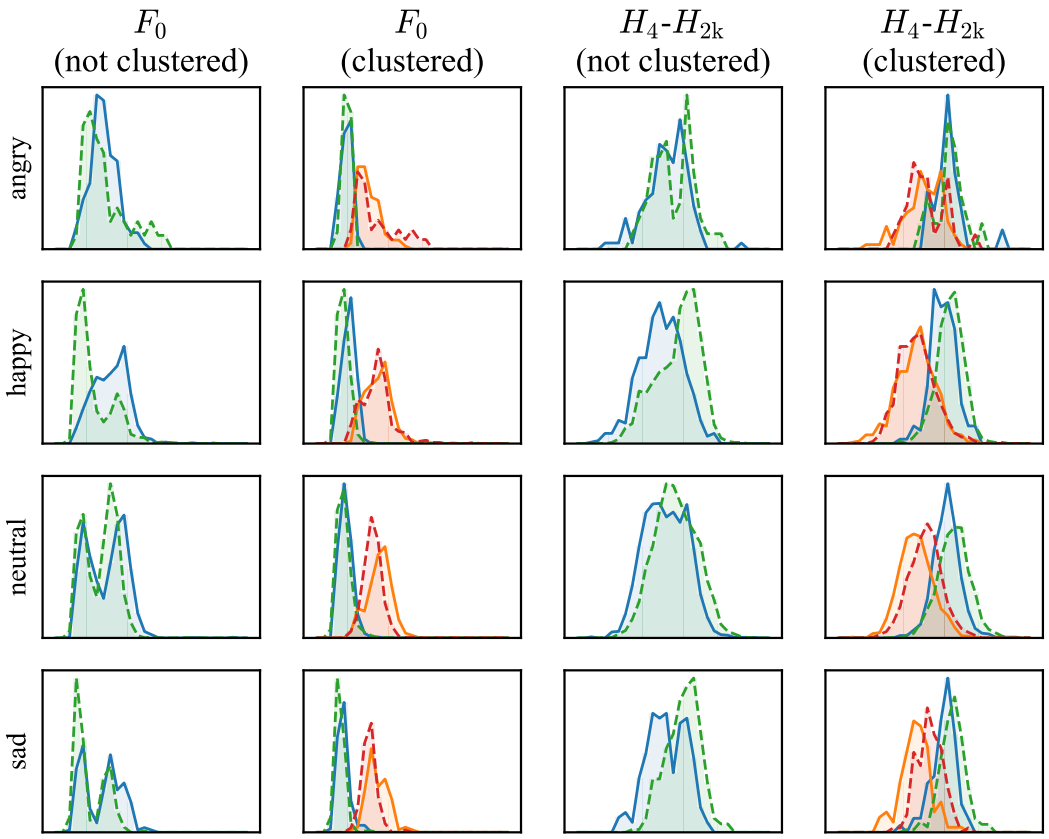


Figure 6.1: Example feature distributions without and with unsupervised clustering. The distribution for the training dataset (solid line) and the development dataset (dashed line) are shown separately. Blue and green curves show the distribution of features from utterances in the first cluster, while read and orange curves show those in the second cluster. The distributions within clustered subsets match better between the training and development datasets compared to the non-clustered distributions.

not only because it is the class with the highest number of samples, but also because emotional speech could be regarded as a variation of neutral speech [YBL04]. The models for the remaining three classes ('sad', 'angry', and 'happy') were adapted from the 'neutral' model. The classification decision, based on the log-likelihood criteria, dictated whether a test supervector was drawn from each class.

6.2.3.2 The support vector machine classifier

A standard support vector machine (SVM) using a linear kernel implemented in the Weka toolkit [HFH09] was used. The supervector configuration that performed the best for each feature set was used for the SVM classification. The complexity parameter C was chosen between the values $10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$, and 10^{-6}, so that it maximizes the system performance on the development dataset for each feature set. Data upsampling was carried out for the under-represented classes to address the data imbalance problem.

6.2.3.3 System fusion

The best performing configuration for each feature set/classifier combination on the development dataset was selected as the representative system for that combination. The system fusion was performed on the n-best performing representative systems.

In order to fuse the results from the GMM and SVM classifiers, the log-likelihood output from the GMM classifier was converted into the confidence score so that it was consistent with the baseline SVM classifier results. The SVM classifier's confidence score was the probability that the test utterance belonged to a class, and it was calculated so that scores for the classes added up to one. On the other hand, the GMM likelihood was calculated for each class independently, hence, there was no guarantee that the likelihoods for the classes added up to one. Thus, the confidence for the i-th class, c_i, was calculated as follows:

$$c_i = \frac{\exp\left(l_i - \mu\right)}{\sum_{i=1}^{M} \exp\left(l_i - \mu\right)} \tag{6.1}$$

where l_i is the log-likelihood for the i-th class, M is the number of classes, and $\mu =$

$1/M \sum_{i=1}^{M} l_i$ is the mean of the log-likelihoods across the classes. The confidence scores from the classifiers were averaged and used for the combined class decision.

6.2.4 Evaluation Metric

The evaluation measure for the challenge was unweighted average recall (UAR). Unweighted average recall was used instead of a weighted one, since it is commonly used where there are highly unbalanced distributions of speech samples among classes. Using this metric, no matter how imbalanced the data is, the chance-level performance is $\frac{100}{M}\%$, where M is the number of classes. Therefore, making a random guess among the 4 emotion classes in this task will result in a UAR of 25%.

Note that the metrics used in the previous chapters (e.g., EER) are not appropriate for this task, because those metrics evaluate binary decisions while this task is a multi-class classification one.

6.3 Results and Discussion

6.3.1 Individual Supervector-Based System Performance

The performances of individual systems in terms of UAR are summarized in Table 6.2. VQual2 performed better than both MFCCs and the baseline ComParE16 feature set in all conditions.

Contrary to expectation, clustering did not provide a performance gain. For the MFCC feature set, the performance degraded from 41.37% UAR to 36.78%. The degradation might be due to overfitting because of insufficient amount of data to train within each cluster. Because MFCCs had 60-dim features while VQual2 had 33-dim features, the shortage of data points might have affected the MFCC-based system more critically.

It is interesting to note that the ComParE16 feature set performance was improved to 40.7% UAR by using the supervector, compared to the OpenSMILE baseline system with a UAR of 37.8%; recall that the baseline system uses a statistics vector for utterance

Table 6.2: Individual system performance in terms of unweighted average recall (UAR, %). The performance was measured on the development dataset. The system configurations chosen for fusion are denoted with asterisks (*), and the ranking among them is shown in the last column. The SVM parameter $C = 10^{-6}$, 10^{-5}, and 10^{-3} was used for the VQual2, MFCCs and ComParE16 features, respectively.

Feature set	Clustering	Classifier	UAR	Ranking
VQual2	Yes	GMM	39.40	-
VQual2	No	GMM	*41.37	2
VQual2	No	SVM	*41.92	1
MFCC	Yes	GMM	36.78	-
MFCC	No	GMM	*41.21	3
MFCC	No	SVM	*40.95	4
ComParE16	Yes	GMM	36.19	-
ComParE16	No	GMM	*40.21	6
ComParE16	No	SVM	*40.71	5

representation. Because both systems used the same acoustic feature set, these results can be used to compare the effect of different methods in modeling the utterances.

6.3.2 Fused System Performance

The configurations selected for each feature set/classifier combination are denoted with asterisks (*), and their performance ranking is shown in Table 6.2. The two best systems were both VQual2-based systems, one with an SVM and the other with a GMM classifier. The n-best system fusion performance is shown in the second column of Table 6.3. Note that a fusion of the two best systems did not improve performance (from 41.92% to 41.71%). This might be due to fact that both systems used the same acoustic information. Adding

Table 6.3: Fused system performance on the development dataset, in terms of unweighted average recall (UAR, %). The best performing fusion is boldfaced.

		+ baseline
2-best (VQual2/SVM, VQual2/GMM)	41.71	42.24
3-best (2-best, MFCC/GMM)	42.60	**43.92**
4-best (3-best, MFCC/SVM)	43.89	43.78
5-best (4-best, ComParE16/SVM)	**44.42**	42.96
6-best (5-best, ComParE16/GMM)	42.69	41.04

the third and the forth best systems, which were based on MFCCs, the UAR improved by 2.18%, providing complementary information to VQual2-based systems. The 5-best system combination, by adding the ComParE16/SVM system, performed the best (UAR= 44.42%).

The OpenSMILE baseline system, with a UAR of 37.8%, used different utterance representation from the supervector framework. Even though the performance was lower than the systems introduced in this study, the baseline system might be complementary. Thus, the fusion of the baseline system in addition to the n-best systems was investigated. The performance with the baseline system is shown in the second column of Table 6.3. Fusing the baseline system with the 2 and 3 best systems improved the performance, suggesting a complementary effect. However, fusing it with the 4, 5 and 6 best systems degraded the system performance.

6.3.3 System Performance Evaluation on the Test Dataset

The complete system block diagram is shown in Figure 6.2 and its performance on the development and test dataset is reported in Table 6.4. As the number of evaluation trials reached the limit, the best performing system on the development dataset could not be evaluated on the test dataset. However, the evaluation result on the test dataset is available for the second best system, which was the fusion of OpenSMILE baseline and the 3 best

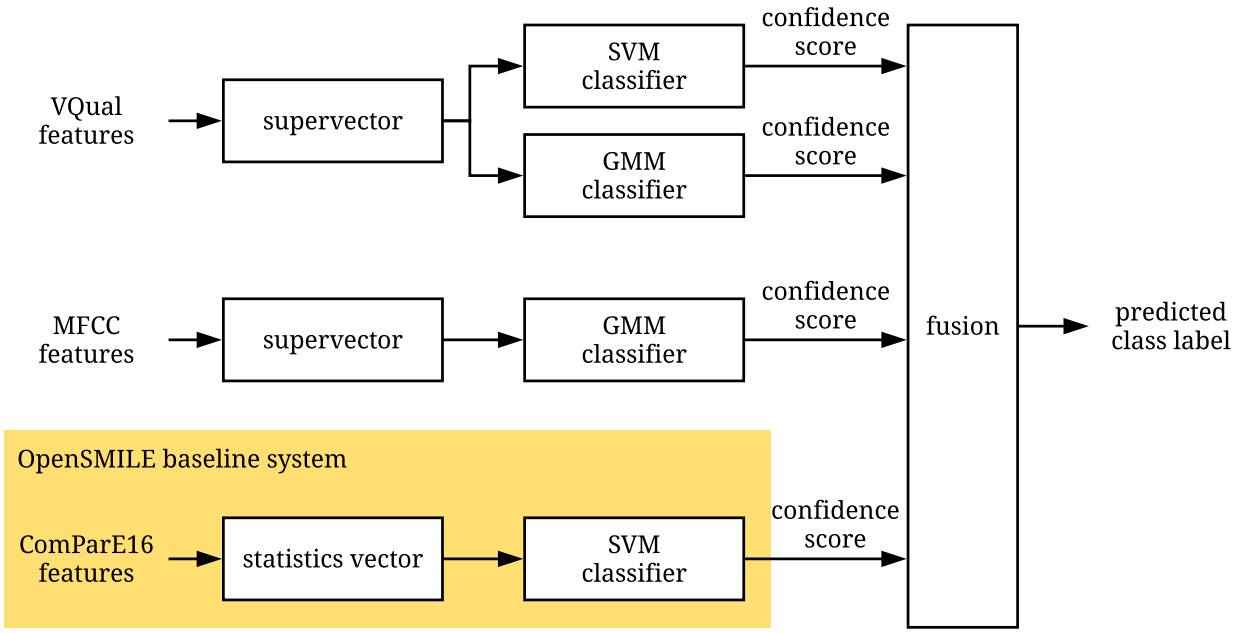


Figure 6.2: The complete system block diagram. The 3-best systems and the OpenSMILE baseline system were fused.

systems.

The proposed system notably outperformed OpenSMILE baseline system on the development dataset: the system performance improved from 37.8% to 43.9%. On the test dataset, the proposed system did not show similar trends: its UAR was 41.0%, while the baseline was at 43.1%.

Confusion matrices of the proposed system for the development and test datasets are

Table 6.4: System performance in terms of unweighted average recall (UAR, %) on the development and test datasets.

	Development	Test
OpenSMILE baseline	37.8	43.1
OpenSMILE + 3-best	43.9	41.0

shown in Figure 6.3. For the development dataset, the recall improvement was evident for the 'angry' and 'sad' classes compared to the OpenSMILE baseline. The 'angry' and 'sad' recalls improved from 30.00% to 46.00%, and from 43.16% to 61.40%, respectively. The precisions for those classes showed small difference between the baseline and the proposed systems compared to the recall improvements: the 'angry' class precision slightly increased from 4.35% to 4.91%, whereas the 'sad' class precision decreased from 17.77% to 16.81%.

The performance pattern, unfortunately, was not consistent in the test dataset. For the test dataset, the 'angry' recall and precision increased to 77.94% and 21.74%, respectively. However, the 'sad' recall and precision decreased to 16.34% and 5.13%, respectively. Because those two had the least amount of data in the training dataset, overfitting might have yielded these results. For example, there were only 125 'angry' voices and 187 'sad' voices while there were 2,287 'neutral' voices in the training dataset. Thus, it is likely that the two classes did not have sufficient data to construct reliable models.

For both datasets, the 'happy' class was the least recalled class. One possible explanation for this confusion is the mismatch across the training and development datasets, which was observed in Section 6.2.2.2.

6.4 Conclusion

The VQual2 feature set was applied to affect classification on the recordings collected from individuals with mental, neurological, and/or physical disorders. No explicit compensation for the effects of a particular disorder on speech signal was attempted in this study. Instead, perceived affects were classified regardless of the kind of disorder.

Compared to the baseline ComParE16 feature set, VQual2 showed better affect classification performance in all experimental configurations. The VQual2 feature set also performed equivalently well, or better in some configurations, compared to MFCCs. It is noteworthy that the proposed VQual2 feature set had lower dimension than the MFCCs and the baseline ComParE16 feature set, but it could outperform those feature sets. As noted earlier, perceived affects are known to strongly related with voice quality. The efficiency of VQual2

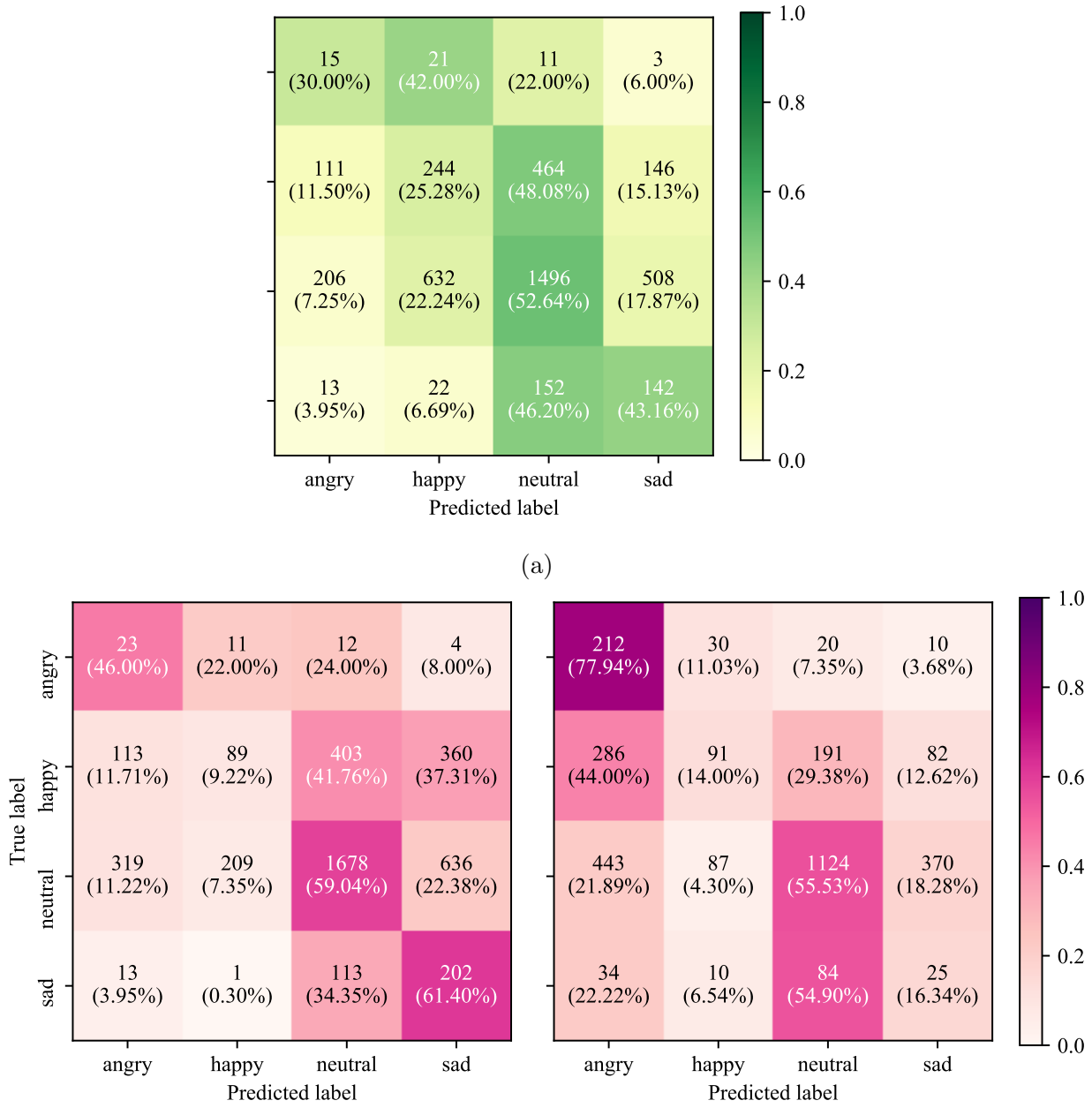


Figure 6.3: Confusion matrices for the results from the (a) OpenSMILE baseline system on the development dataset and the proposed system on (b) the development dataset and (c) the test dataset. Numbers in each cell represents the number of speech samples and corresponding recall values (%).

might be in providing information that is directly related to perception.

The supervector approach used in this study showed its effectiveness in representing the utterances. The utterance-level distribution of VQual2 features and MFCCs was effectively modeled with this approach, resulting in a system which outperformed the OpenSMILE baseline system on the development dataset. Additionally, using a supervector derived from ComParE16 feature set resulted in a better performance than the baseline system which used the statistics vector for the same feature set. These results suggest that in the cases when the amount of data is insufficient to apply the i-vector framework, the supervector approach can be a viable alternative to represent the local feature distribution within an utterance.

The confidence score that an utterance was drawn from a class was used for system fusion. When the systems using different features were fused, the performance improved, suggesting complementary effect between feature sets. The system fusion configuration was finalized based on the single system performance, and the complete system performance was analyzed based on confusion matrices. The performance gain was obtained by improving the recall for 'sad' and 'angry' classes. However, the proposed system was less effective on the test dataset, suggesting overfitting due to insufficient amounts of data especially in the 'sad' and 'angry' classes.

In conclusion, VQual2 was effective in representing perceived affect, in addition to representing acoustic and perceived talker identity. Analysis of the system performance suggests that further improvements could be made by better modeling the classes with limited training data. Addressing acoustic mismatch across datasets would be another important direction for future studies.

CHAPTER 7

Summary and Future Work

Voice discrimination abilities of humans and machines under text, affect, and speaking-style variabilities were discussed in this dissertation. A new speech database including a large number of talkers and multiple speech tasks per talker was developed at UCLA to study the effects of within- and between-talker variability systematically. Specifically, the role of a feature set, which was based on a psychoacoustic model of voice quality [KGG14, GSG16], in representing perceptual and acoustic talker identity was explored.

7.1 Summary

Experimental setups and key results presented in Chapters 3 through 5 are summarized in Table 7.1.

Ch. 3: Voice quality features were promising for representing talker identity

Preliminary experiments were conducted using sustained vowel /a/ sounds from 5 female talkers and read sentences from 3 female talkers. The initial voice quality feature set (VQual1) was determined based on correlation and canonical component analysis, using the selected vowel sounds. The resulting set was equivalent to the psychoacoustic model of voice quality, and it contained F_0, F_1, F_2, F_3, harmonic amplitude differences with formant correction (H_1^*-H_2^*, H_2^*-H_4^*, H_4^*-$H_{2\mathrm{k}}^*$, and $H_{2\mathrm{k}}^*$-$H_{5\mathrm{k}}$), and cepstral peak prominence (CPP). Human voice discrimination experimental results showed that humans were reasonably accurate in distinguishing talkers.

In predicting human responses to voice discrimination tasks, VQual1 features provided

complementary information to MFCCs. The root-mean-squared error decreased as much as 3.14% for vowels and 11.80% for read sentences. Human responses for vowels were better predicted than those for sentences. It may be that the acoustic features used in this study, or the way they were used, might insufficiently represent the information human listeners are using in connected speech. For example, because the prediction was based on the mean and standard deviation of the acoustic features over an utterance, spectro-temporal information and/or prosodic information might be insufficiently represented. A more sophisticated model for prediction including such information is worth exploring.

The voice quality features (without H_{2k}^{*}-H_{5k}; VQual1*) were also applied to a standard i-vector/PLDA ASV system. First, an ASV system was evaluated using MFCCs under session, affect, and speaking-style variabilities for both female and male voices. As expected, the system performed worse with affect or speaking-style mismatch between enrollment and test utterances with the short test utterances ($\leq$ 3 sec). The VQual1* improved ASV system performance in 7 out of 8 conditions, by providing complementary information to MFCCs. One condition when fusing VQual1* slightly degraded system performance was when the speaking style was mismatched between the enrollment and test utterances for female voices. This might indicate that voice quality features vary significantly between different speaking styles, especially for female voices.

Ch. 4: The modified voice quality feature set improved ASV performance for short utterance text-independent tasks

To better represent talker identity for ASV, VQual1 was modified to another set (VQual2). Speech samples with two different speaking styles (read sentences and pet-directed speech) from 100 female and 100 male talkers were taken from the UCLA database.

Based on the f-ratio criterion, harmonic amplitudes without formant corrections were more effective in differentiating talkers than with correction. This might be due to over-correction which can occur when harmonics are close to a formant frequency. However, this possibility was not fully explained with the present study, and further analysis is needed.

Formant amplitudes (A_1, A_2, A_3) were included because they showed high f-ratio across gender and speaking style. The resulting modified feature set included F_0, F_1, F_2, F_3, harmonic amplitude differences without formant correction (H_1-H_2, H_2-H_4, H_4-H_{2k}), A_1, A_2, A_3, and CPP.

In preliminary experiments presented in Chapter 3, ASV performance degradation was observed under within-talker variability, but it was not clear how much each kind of variability contributes to performance degradation. For example, the degradation for affect mismatch might as well be influenced by text mismatch between enrollment and test utterances. Thus, the effect of text mismatch was analyzed by comparing same-text pairs and different-text pairs. The effect of style variability was also analyzed by including pet-directed speech as well as read sentences. Results showed that when content variability was large and utterances were short ($\approx$ 2 sec), MFCC-based system performance degrades substantially.

The VQual2-based system outperformed the VQual1*-based system in all conditions, and it outperformed MFCC-based system in some conditions. The system performance improved 14–19% by fusing the two systems in all conditions, compared to using only MFCCs. These results indicate that VQual2 can be effective for ASV tasks under high text variability. However, unlike humans, there was a still a large performance gap between same- and different-text conditions for the 2-sec utterance pairs. This suggests that understanding human listeners' strategies to discriminate voices might provide insights to improve ASV performance.

Not surprisingly, the improvement by fusion was not significant with high style variability for 5-sec utterance pairs. This again suggests that voice quality features vary significantly across speaking styles. However, if a consistent pattern of voice quality for each speaking style can be found, voice quality features could be utilized to detect style differences and/or to normalize these effects for better ASV or automatic speech recognition performance. This can be a direction for future studies.

The effectiveness of VQual2 were evaluated on a standard ASV database (NIST SRE10), with various utterance lengths. Results suggests that utterances shorter than 10 sec benefit

from using VQual2 features more than long utterances do.

Ch. 5: Human and machine talker spaces were weakly correlated, but human responses for different-talker pairs were strongly correlated with voice quality-based system responses

Voice discrimination abilities of humans and machines for very short utterances ($\approx$ 2 sec) under text and style variability were analyzed. Read sentences and pet-directed speech from 50 female talkers were used for perceptual and automatic voice discrimination experiments. As expected, humans were more accurate than machines for read sentence pairs, but the performance difference became small for style-mismatched pairs and for perceptually marked talkers.

Although not every marked talker were accented, human performance degradation with marked talkers might be at least in part related to the finding that accented talkers are more difficult to identify than unaccented talkers [GKB81]. For machines, the effects of a non-native accent or other markedness on voice discrimination have not yet been extensively studied, but results in this study suggests that machine performance might degrade with talker markedness. Therefore, the effects of markedness on voice discrimination are worth exploring in detail in the future. The UCLA database includes many non-native speakers of English whose speech could be useful for this purpose.

Using read sentence pairs, the talker spaces were inferred from human and machine responses. Talker spaces were weakly correlated, indicating a weak or non-linear relationship between talker representations by humans and machines. Moreover, a high correlation between them was not related to better machine performance, suggesting disparity between human and machine strategies used in discriminating between talkers. Interestingly, human talker spaces were reasonably well-modeled as linear combinations of means of voice quality features, while machine talker spaces were not. These findings suggest that investigating how voice quality feature means are related to human responses might provide valuable insights into perceived talker identity.

Results were further analyzed based on the log-likelihood ratio and response reliability. Same-talker and different-talker pairs were analyzed separately based on an assumption that perceptual strategies to identify similarity might be different from those to detect dissimilarity between talkers. For different-talker pairs, the VQual2-based system responses were highly correlated with human responses. Results also suggested that machines could supplement human decisions for perceptually marked talkers.

Ch. 6: Voice quality features were effective in affect recognition

VQual2 was effective in representing perceived affect recognition, in addition to representing acoustic and perceived talker identity. VQual2 was the most effective feature set among three sets of features, and fusion provided further performance improvements. This suggests another application where voice quality features can contribute to predict human decisions. Another successful application in a similar direction was depression detection [AGP18].

7.2 Future Work

When the voice quality feature set was modified, formant correction was omitted. Although the modification was effective for improving ASV performance, where the advantage comes from is not clear. In fact, formant correction is expected to benefit text-independent tasks by attenuating the effect of phonetic variaiblity. This unexpected result might be related to inaccuracies in measurement due to over-correction as noted before. However, this possibility could not be evaluated because the accuracy could not be measured without the "ground truth" voice source spectrum, which is unattainable from natural speech signal. Detailed analysis on advantages and disadvantages of formant correction using synthetic speech is worth considering for future studies.

Comparison between human and machine responses showed that human responses for different-talker pairs were highly correlated with VQual2-based system responses. However, acoustic correlates to human responses for same-talker pairs were insufficiently analyzed because only a few such pairs were included in the experiments. Another interesting finding

was that human performance was highly influenced by perceptual markedness of talkers. However, it was difficult to generalize these results because the number of talkers was small. A new set of experiments including more same-talker pairs and/or marked talkers might provide better understanding about human responses.

It was also noted that one dimension of information that is used by humans but not by machines might be spectro-temporal information. Considering that standard ASV systems used in this dissertation make decisions based on the distribution of static features and at most their time derivatives, temporal information might be insufficiently utilized by machines. Thus, a spectro-temporal talker representation for machine might bring about further improvements in performance.

Another possible approach for further improvement is using voice quality features to normalize style and/or affect variability. It was noted that the voice quality features varied across speaking styles, and the features were effective to represent different affect. If machines can learn how acoustic features change across style and/or affect, they can be more robust to within-talker variability.

Table 7.1: Summary of experiments (Ex.) reported in Chapters 3 through 5. The VQual1* feature set contains the features in VQual1, except H^*_{2k}-H_{5k}. Fusion indicates a weighted combination of the scalar responses from individual systems. Speech samples were drawn from the UCLA database unless otherwise specified. Utterance lengths for ASV experiments are denoted as the length of enrollment and test utterances in seconds.

		Chapter 3	Chapter 4	Chapter 5
Perception	Speaking style	Ex.1: vowel /a/ (1–3 sec) Ex.2: read ($\approx$ 2 sec)	N/A	read, pet-directed ($\approx$ 2 sec)
	No. talkers	Ex.1: 5 females Ex.2: 3 females	N/A	50 females
	No. listeners	Ex.1: 60 Ex.2: 15	N/A	65
Feature selection	Stimuli	vowels	read, pet-directed	N/A
	No. talkers	5 females	100 females, 100 males	N/A
	Resulting feature set	VQual1 (F_0, F_1, F_2, F_3, CPP, H^*_1-H^*_2, H^*_2-H^*_4, H^*_4-H^*_{2k}, H^*_{2k}-H_{5k})	VQual2 (F_0, F_1, F_2, F_3, CPP, H_1-H_2, H_2-H_4, H_4-H_{2k}, A_1, A_2, A_3)	VQual2
ASV	Speaking style	Ex.1: read Ex.2: affective Ex.3: read, affective	Ex.1: read Ex.2: read, pet-directed Ex.3: phonecall (SRE database)	read, pet-directed
	Utterance length	Ex.1: 60sec–3sec Ex.2: 60sec–3sec Ex.3: 90sec–3sec	Ex.1: 2sec–2sec Ex.2: 5sec–5sec Ex.3: various	$\approx$ 2 sec
	No. talkers	25 females, 25 males	Ex.1: 100 females, 100 males Ex.2: 100 females, 100 males Ex.3: $\gg$ 200 (SRE database)	50 females
	Systems	MFCCs, VQual1*, fusion	MFCCs, VQual1*, VQual2, fusion	MFCCs, VQual2, fusion
Predicting human responses	Features	MFCCs, VQual1	N/A	N/A
	Method	linear regression	N/A	N/A
Human vs. machine comparison	Speaking style	N/A	N/A	read
	Method	N/A	N/A	Ex.1: decision spaces inferred with MDS Ex.2: log-likelihood-ratio cost function

APPENDIX A

Examples of Topics Used to Elicit Speech (Chapter 2)

A.1 Session A

Imagine you are talking to someone you don't know, like the RA. Give her either directions on how to go somewhere, or instructions on how to do something (your choice – anything you like). For example:

- A tourist on campus stops you near here, and asks you how to get to the Bruin Bear
- You're in a classroom building (pick one), and someone asks you how to get to one at the other end of campus (pick one)
- Someone asks you about options for printing out an assignment in the campus computer labs
- Tell someone how a Bruin Card works

Think of a conversation you've had recently about something that wasn't important – not exciting, not upsetting, just normal. Repeat that conversation to the RA as best you can, in a "First he said..., then I said ..." style. Some possible topics:

- Ordering food, where the server had to ask you about your choices (like what you wanted for protein, and sides, and condiments; or for build-your-own pizza)
- Describing your day to someone at home – nothing really interesting happened, but they want to hear all about it

- Deciding with a friend about where to go out – which restaurant, or which movie

A.2 Session B

Think of a CONVERSATION you've had about something exciting, that made you really happy. Repeat that conversation to the RA as best you can, in a "FIRST SHE SAID...THEN I SAID" style. Some possible topics:

- You interviewed really well for a great internship or job or summer program
- A close friend or relative got engaged and talked to you about wedding plans
- You were planning a special vacation trip with someone

A.3 Session C

Think of a CONVERSATION you've had about something that really annoyed you. Repeat that conversation to the RA as best you can, in a "FIRST HE SAID..., THEN I SAID ..." style. Some possible topics:

- Talking to a friend about something another friend insisted on doing despite your objections
- Talking to a roommate about a housekeeping disagreement (e.g. everyone is supposed to do their own dishes)
- A time-wasting conversation with someone working in a store or business, e.g. a billing dispute, or cable service

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 *Proc. Interspeech*, pp. 1676–1680, Hyderabad, India, sep 2018. ISCA.

[BB10] Oliver Baumann and Pascal Belin. "Perceptual Scaling of Voice Identity: Common Dimensions for Different Vowels and Speakers." *Psychological Research*, **74**(1):110–120, 2010.

[BBC07] Niko Brümmer, Lukáš Burget, Jan Honza Černocký, Ondej Glembek, František Grézl, Martin Karafiát, David A. Van Leeuwen, Pavel Matějka, Petr Schwarz, and Albert Strasheim. "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006." *IEEE Transactions on Audio, Speech and Language Processing*, **15**(7):2072–2084, 2007.

[BD11] Niko Brümmer and Edward De Villiers. "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing." Technical report, AGNITIO Research, South Africa, 2011.

[BFB04] Pascal Belin, Shirley Fecteau, and Catherine Bédard. "Thinking the Voice: Neural Correlates of Voice Perception." *Trends in Cognitive Sciences*, **8**(3):129–135, 2004.

[BKV02] Denis Burnham, Christine Kitamura, and Uté Vollemer-Conna. "What's New, Pussycat? On Talking to Babies and Animals." *Science*, **296**(5572):1435–1435, may 2002.

[BO00] Ronald J Baken and Robert F Orlikoff. *Clinical Measurement of Voice and Speech*. Cengage Learning, 2000.

[BP66] Peter D. Bricker and Sandra Pruzansky. "Effects of Stimulus Content and Duration on Talker Identification." *The Journal of the Acoustical Society of America*, **40**(6):1441–1449, dec 1966.

[BW17] Paul Boersma and David Weenink. "Praat: Doing Phonetics by Computer.", 2017.

[CSR06] William M. Campbell, D. E. Sturim, and Douglas A. Reynolds. "Support Vector Machines Using GMM Supervectors for Speaker Verification." *IEEE Signal Processing Letters*, **13**(5):308–311, 2006.

[CW97] Susan Cook and John Wilding. "Earwitness Testimony: Never Mind the Variety, Hear the Length." *Applied Cognitive Psychology*, **11**(2):95–111, apr 1997.

[DDK07] Najim Dehak, Pierre Dumouchel, and Patrick Kenny. "Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification." *IEEE Transactions on Audio, Speech and Language Processing*, **15**(7):2095–2103, sep 2007.

[DJP16] Rohan Kumar Das, Sarfaraz Jelil, and S. R. Mahadeva Prasanna. "Significance of Constraining Text in Limited Data Text-Independent Speaker Verification." In *Proc. International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5. IEEE, jun 2016.

[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.

[DM80] S. Davis and P. 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, aug 1980.

[DP16] Rohan Kumar Das and S. R. Mahadeva Prasanna. "Exploring different attributes of source information for speaker verification with limited test data." *The Journal of the Acoustical Society of America*, **140**(1):184–190, jul 2016.

[DP18] Rohan Kumar Das and S. R. Mahadeva Prasanna. "Speaker Verification from Short Utterance Perspective: A Review." *IETE Technical Review*, **35**(6):599–617, nov 2018.

[ESW97] J. Epps, J. R. Smith, and J. Wolfe. "A Novel Instrument to Measure Acoustic Resonances of the Vocal Tract During Phonation." *Measurement Science and Technology*, **8**(10):1112–1121, oct 1997.

[EWS10] Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: the Munich Versatile and Fast Open-Source Audio Feature Extractor." In *Proc. ACM Multimedia*, pp. 1459–1462, Firenze, Italy, 2010.

[Fan60] G Fant. *Acoustic Theory of Speech Production*. Gravenhage, 1960.

[GF11] Erica Gold and Peter French. "An International Investigation of Forensic Speaker Comparison Practices." In *Proc. International Congress of Phonetic Sciences (ICPhS)*, pp. 751–754, Hong Kong, 2011.

[GHR13] Rosa González Hautamäki, Ville Hautamäki, Padmanabhan Rajan, and Tomi Kinnunen. "Merging Human and Automatic System Decisions to Improve Speaker Recognition Performance." In *Proc. Interspeech*, pp. 2519–2523, Lyon, France, 2013.

[GKB81] Alvin G. Goldstein, Paul Knight, Karen Bailis, and Jerry Conover. "Recognition Memory for Accented and Unaccented Voices." *Bulletin of the Psychonomic Society*, **17**(5):217–220, 1981.

[GKE13] Marc Garellek, Patricia Keating, Christina M. Esposito, and Jody Kreiman. "Voice Quality and Tone Identification in White Hmong." *The Journal of the Acoustical Society of America*, **133**(2):1078–1089, feb 2013.

[GMB10] Craig Greenberg, Alvin Martin, Linda Brandschain, Joseph Campbell, Christopher Cieri, George Doddington, and John Godfrey. "Human Assisted Speaker Recognition in NIST SRE10." In *Proc. Odyssey - The Speaker and Language Recognition Workshop*, pp. 180–185, Brno, Czech Republic, 2010.

[GMD11] Craig S. Greenberg, Alvin F. Martin, George R. Doddington, and John J. Godfrey. "Including Human Expertise in Speaker Recognition Systems: Report on a Pilot Evaluation." In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5896–5899. IEEE, may 2011.

[GN03] Christer Gobl and Ailbhe Ní Chasaide. "The Role of Voice Quality in Communicating Emotion, Mood and Attitude." *Speech Communication*, **40**:189–212, 2003.

[GSG16] Marc Garellek, Robin Samlan, Bruce R. Gerratt, and Jody Kreiman. "Modeling the Voice Source in Terms of Spectral Slopes." *The Journal of the Acoustical Society of America*, **139**(3):1404–1410, mar 2016.

[GSK13] Marc Garellek, Robin A. Samlan, Jody Kreiman, and Bruce R. Gerratt. "Perceptual Sensitivity to a Model of the Source Spectrum." In *Proc. Meetings on Acoustics*, volume 19, pp. 1–5, Montreal, Canada, 2013.

[Han97] Helen M. Hanson. "Glottal Characteristics of Female Speakers: Acoustic Correlates." *The Journal of the Acoustical Society of America*, **101**(1):466–481, jan 1997.

[HC99] Helen M. Hanson and Erika S. Chuang. "Glottal Characteristics of Male Speakers: Acoustic Correlates and Comparison with Female Data." *The Journal of the Acoustical Society of America*, **106**(2):1064–1077, aug 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, aug 1994.

[HFH09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA Data Mining Software: An Update." *SIGKDD Explorations Newsletter*, **11**(1):10–18, 2009.

[HGC95] James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler. "Acoustic characteristics of American English vowels." *The Journal of the Acoustical Society of America*, **97**(5):3099–3111, may 1995.

[HH15] John H. L. Hansen and Taufiq Hasan. "Speaker Recognition by Machines and Humans: A Tutorial Review." *IEEE Signal Processing Magazine*, **32**(6):74–99, 2015.

[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, aug 2017. ISCA.

[HKN10] Ville Hautamäki, Tomi Kinnunen, Mohaddeseh Nosratighods, Kong-Aik Lee, Bin Ma, and Haizhou Li. "Approaching Human Listener Accuracy with Modern Speaker Verification." In *Proc. Interspeech*, pp. 1473–1476, Makuhari, Chiba, Japan, 2010.

[HSC17] Simone Hantke, Hesam Sagha, Nicholas Cummins, and Björn Schuller. "Emotional Speech of Mentally and Physically Disabled Individuals: Introducing the EmotAsS Database and First Findings." In *Proc. Interspeech*, pp. 3137–3141, Stockholm, Sweden, 2017.

[IA04] Markus Iseli and Abeer Alwan. "An Improved Correction Formula for the Estimation of Harmonic Magnitudes and Its Application to Open Quotient Estimation." In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pp. I–669–672, 2004.

[IEE69] IEEE Subcommittee on Subjective Measurements. "IEEE Recommended Practice for Speech Quality Measurements." *IEEE Transactions on Audio and Electroacoustics*, **17**(3):225–246, sep 1969.

[ISA07] Markus Iseli, Yen-Liang Shue, and Abeer Alwan. "Age, Sex, and Vowel Dependencies of Acoustic Measures Related to the voice Source." *The Journal of the Acoustical Society of America*, **121**(4):2283–2295, apr 2007.

[JAA07] Jintao Jiang, Edward T. Auer, Abeer Alwan, Patricia Keating, and Lynne E. Bernstein. "Similarity Structure in Visual Speech Perception and Optical Phonetic Signals." *Perception & Psychophysics*, **69**(7):1070–1083, 2007.

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

[KAG10] Jody Kreiman, Norma Antoñanzas-Barroso, and Bruce R Gerratt. "Integrated software for analysis and synthesis of voice quality." *Behavior research methods*, **42**(4):1030–1041, 2010.

[KAR11] Juliette Kahn, Nicolas Audibert, Solange Rossato, and Jean-François Bonastre. "Speaker Verification by Inexperienced and Experienced Listeners vs. Speaker Verification System." In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5912–5915. IEEE, may 2011.

[KG96] Jody Kreiman and Bruce R. Gerratt. "The Perceptual Structure of Pathologic Voice Quality." *The Journal of the Acoustical Society of America*, **100**(3):1787–1795, 1996.

[KG05] Jody Kreiman and Bruce R. Gerratt. "Perception of Aperiodicity in Pathological Voice." *The Journal of the Acoustical Society of America*, **117**(4):2201–2211, apr 2005.

[KG10] Jody Kreiman and Bruce R. Gerratt. "Perceptual Sensitivity to First Harmonic Amplitude in the Voice Source." *The Journal of the Acoustical Society of America*, **128**(4):2085–2089, oct 2010.

[KG12] Jody Kreiman and Bruce R. Gerratt. "Perceptual Interaction of the Harmonic Source and Noise in Voice." *The Journal of the Acoustical Society of America*, **131**(1):492–500, jan 2012.

[KGA07] Jody Kreiman, Bruce R. Gerratt, and Norma Antoñanzas-Barroso. "Measures of the Glottal Source Spectrum." *Journal of Speech, Language, and Hearing Research*, **50**(3):595–610, jun 2007.

[KGG14] Jody Kreiman, Bruce R. Gerratt, Marc Garellek, Robin Samlan, and Zhaoyan Zhang. "Toward a Unified Theory of Voice Production and Perception." *Loquens*, **1**(1):e009, jun 2014.

[KKA18] Patricia A Keating, Jody Kreiman, and Abeer Alwan. "The UCLA Speaker Variability Database.", 2018.

[KMD03] Patrick Kenny, M Mihoubi, and Pierre Dumouchel. "New MAP Estimators for Speaker Recognition." In *Proc. Interspeech*, pp. 1–4, 2003.

[KP91] Jody Kreiman and George Papcun. "Comparing Discrimination and Recognition of Unfamiliar Voices." *Speech Communication*, **10**(3):265–275, aug 1991.

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

[Kro93] Guus de Krom. "A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals." *Journal of Speech, Language, and Hearing Research*, **36**(2):254–266, apr 1993.

[KS11] Jody Kreiman and Diana Sidtis. *Foundations of Voice Studies*. Wiley-Blackwell, Oxford, UK, 2011.

[KSO13] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Md. Jahangir Alam, and Pierre Dumouchel. "PLDA for Speaker Verification with Utterances of Arbitrary Duration." In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7649–7653, 2013.

[KW78] Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling*. Sage, Beverly Hills, 1978.

[Lav80] John Laver. *The Phonetic Description of Voice Quality*. Cambridge University Press, 1980.

[LB07] David A. van Leeuwen and Niko Brümmer. "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems." In Christian Müller, editor, *Speaker Classification I: Fundamentals, Features, and Methods*, pp. 330–353. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[LBG18] 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*, pp. 1–18, sep 2018.

[LD08] Xugang Lu and Jianwu Dang. "An Investigation of Dependencies Between Frequency Components and Speaker Characteristics for Text-Independent Speaker Identification." *Speech Communication*, **50**(4):312–322, 2008.

[LGP84] Gordon E Legge, Carla Grosmann, and Christina M Pieper. "Learning Unfamiliar Voices." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**(2):298–303, 1984.

[LLM14] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li. "Text-Dependent Speaker Verification: Classifiers, Databases and RSR2015." *Speech Communication*, **60**:56–77, may 2014.

[LTJ07] Xi Li, Jidong Tao, Michael T. Johnson, Joseph Soltis, Anne Savage, Kirsten M. Leong, and John D. Newman. "Stress and Emotion Classification using Jitter and Shimmer Features." In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. IV–1081–1084, Honolulu, Hawaii, USA, apr 2007. IEEE.

[MA93] Iain R. Murray and John L. Arnott. "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion." *The Journal of the Acoustical Society of America*, **93**(2):1097–1108, 1993.

[MC05] Neil A. Macmillan and C. Douglas Creelman. *Detection Theory: A User's Guide*. Erlbaum, Mahwah, NJ, 2 edition, 2005.

[McD13] Kirsty McDougall. "Assessing Perceived Voice Similarity Using Multidimensional Scaling for the Construction of Voice Parades." *International Journal of Speech, Language and the Law*, **20**(2):163–172, 2013.

[McN47] Quinn McNemar. "Note on the Sampling Error of the Difference between Correlated Proportions or Percentages." *Psychometrika*, **12**(2):153–157, 1947.

[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.

[MG10] Alvin Martin and Craig Greenberg. "The 2010 NIST Speaker Recognition Evaluation (SRE10).", 2010.

[MNH15] Kirsty McDougall, Francis Nolan, and Toby Hudson. "Telephone Transmission and Earwitnesses: Performance on Voice Parades Controlled for Voice Similarity." *Phonetica*, pp. 257–272, 2015.

[MRD09] Youri Maryn, Nelson Roy, Marc De Bodt, Paul Van Cauwenberge, and Paul Corthals. "Acoustic Measurement of Overall Voice Quality: A Meta-Analysis." *The Journal of the Acoustical Society of America*, **126**(5):2619–2634, nov 2009.

[NB01] Hirotaka Nakasone and Steven D. Beck. "Forensic Automatic Speaker Recognition." In *Proc. Odyssey - The Speaker and Language Recognition Workshop*, Crete, Greece, 2001.

[NMC97] Simon Nicholson, Ben Milner, and Stephen Cox. "Evaluating Feature Set performance using the F-ratio and J-measures." In *Proc. Eurospeech*, pp. 413–416, 1997.

[NMH11] Francis Nolan, Kirsty McDougall, and Toby Hudson. "Some Acoustic Correlates of Perceived (Dis) Similarity between Same-Accent Voices." In *Proc. International Congress of Phonetic Sciences (ICPhs)*, pp. 1506–1509, Hong Kong, 2011.

[OS68] A. Oppenheim and R. Schafer. "Homomorphic Analysis of Speech." *IEEE Transactions on Audio and Electroacoustics*, **16**(2):221–226, jun 1968.

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

[PAK19] Soo Jin Park, Amber Afshan, Jody Kreiman, Gary Yeung, and Abeer Alwan. "Target and Non-Target Speaker Discrimination by Humans and Machines." In *Proc. ICASSP (in press)*, 2019.

[PGB11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. "The Kaldi Speech Recognition Toolkit." In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

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

[PML06] Mark A. Przybocki, Alvin F. Martin, and Audrey N. Le. "NIST Speaker Recognition Evaluation Chronicles - Part 2." In *Proc. Odyssey - The Speaker and Language Recognition Workshop*, pp. 1–6. IEEE, jun 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 *Proc. Interspeech*, pp. 1044–1048, San Francisco, USA, sep 2016.

[PYK17] Soo Jin Park, Gary Yeung, Jody Kreiman, Patricia A. Keating, and Abeer Alwan. "Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems." In *Proc. Interspeech*, pp. 1522–1526, Stockholm, Sweden, aug 2017.

[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." *The Journal of the Acoustical Society of America*, **144**(1):375–386, jul 2018.

[PZH17] Srinivas Parthasarathy, Chunlei Zhang, John H. L. Hansen, and Carlos Busso. "A Study of Speaker Verification Performance with Expressive Speech." In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5540–5544, New Orleans, USA, mar 2017. IEEE.

[RAC03] D. Reynolds, Walter Andrews, Joseph Campbell, Jiri Navratil, Barbara Peskin, Andre Adami, Qin Jin, David Klusacek, Joy Abramson, Radu Mihaescu, Jack Godfrey, Doug Jones, and Bing Xiang. "The SuperSID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition." In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pp. IV–784–787. IEEE, 2003.

[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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5908–5911. IEEE, may 2011.

[RFR02] Ravi P. Ramachandran, Kevin R. Farrell, Roopashri Ramachandran, and Richard J. Mammone. "Speaker Recognition-General Classifier Approaches and Data Fusion Methods." *Pattern Recognition*, **35**(12):2801–2821, 2002.

[Ros02] Philip Rose. *Forensic Speaker Identification*. International Forensic Science and Investigation. Taylor & Francis, London, jul 2002.

[RQD00] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models." *Digital Signal Processing*, **10**(1-3):19–41, jan 2000.

[RW93] Rebecca Roebuck and John Wilding. "Effects of Vowel Variety and Sample Length on Identification of a Speaker in a Line-up." *Applied Cognitive Psychology*, **7**(6):475–481, nov 1993.

[Sch86] Klaus R Scherer. "Vocal Affect Expression: A review and a Model for Future Research." *Psychological Bulletin*, **99**(2):143–165, 1986.

[SCS11a] Reva Schwartz, Joseph P. Campbell, Wade Shen, Douglas E. Sturim, William M. Campbell, Fred S. Richardson, Robert B. Dunn, and Robert Granville. "USSS-MITLL 2010 Human Assisted speaker Recognition." In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5904–5907. IEEE, may 2011.

[SCS11b] Wade Shen, Joseph Campbell, Derek Straub, and Reva Schwartz. "Assessing the Speaker Recognition Performance of Naive Listeners Using Mechanical Turk." In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5916–5919. IEEE, 2011.

[SFH16] Eugenia San Segundo, Paul Foulkes, and Vincent Hughes. "Holistic Perception of Voice Quality Matters more than L1 when Judging Speaker Similarity in Short Stimuli." In *Proc. Australasian Conference on Speech Science and Technology (SST)*, pp. 309–312, Parramatta, Australia, 2016.

[SFK05] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. "Modeling Prosodic Feature Sequences for Speaker Recognition." *Speech Communication*, **46**(3-4):455–472, jul 2005.

[SHS97] Stefan R. Schweinberger, Anja Herholz, and Volker Stief. "Auditory Long term Memory: Repetition Priming of Voice Recognition." *The Quarterly Journal of Experimental Psychology*, **50**(3):498–517, aug 1997.

[Sjo04] Kåre Sjölander. "Snack sound toolkit.", 2004.

[SKV11] Yen-Liang Shue, Patricia A. Keating, Chad Vicenik, and Kristine Yu. "VoiceSauce: A Program for Voice Analysis." In *Proc. International Congress of Phonetic Sciences (ICPhs)*, pp. 1846–1849, Hong Kong, 2011.

[SSB16] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language." In *Proc. Interspeech*, pp. 2001–2005, 2016.

[SSB18] Björn Schuller, Stefan Steidl, Anton Batliner, Peter B. Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian B. Pokorny, Eva-maria Rathner, Katrin D. Bartl-Pokorny, Christa Einspieler, Dajie Zhang, Alice Baird, Shahin Amiriparian, Kun Qian, Zhao Ren, Maximilian Schmitt, Panagiotis Tzirakis, and Stefanos Zafeiriou. "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats." In *Proc. Interspeech*, pp. 122–126, Hyderabad, India, sep 2018. ISCA.

[SY80] Howard Saslove and A. Daniel Yarmey. "Long-Term Auditory Memory: Speaker Identification." *Journal of Applied Psychology*, **65**(1):111–116, 1980.

[Sys] Systat Software Inc. "Systat Version 12.".

[TF13] Barbara G. Tabachnick and Linda S. Fidell. *Using Multivariate Statitstics*. Pearson, 6 edition, 2013.

[TK09] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Elsevier, 4 edition, 2009.

[VA06] Thurid Vogt and Elisabeth André. "Improving Automatic Emotion Recognition from Speech via Gender Differentiation." In *Proc. Language Resources and Evaluation Conference*, pp. 1123–1126, Genoa, Italy, 2006.

[VK87] Diana Van Lancker and Jody Kreiman. "Voice Discrimination and Recognition are Separate Abilities." *Neuropsychologia*, **25**(5):829–834, 1987.

[VKE85] Diana Van Lancker, Jody Kreiman, and Karen Emmorey. "Familiar Voice Recognition: Patterns and Parameters. Part I: Recognition of Backward Voices." *Journal of Phonetics*, **13**(1):19–38, 1985.

[VKW85] Diana Van Lancker, Jody Kreiman, and Thomas D. Wickens. "Familiar voice recognition: Patterns and parameters. Part II: Recognition of Rate-Altered Voices." *Journal of Phonetics*, **13**(1):39–52, 1985.

[VRS16] Karthika Vijayan, Pappagari Raghavendra Reddy, and K. Sri Rama Murty. "Significance of Analytic Phase of Speech Signals in Speaker Verification." *Speech Communication*, **81**:54–71, jul 2016.

[VS14] Jitka Vaková and Radek Skarnitzl. "Within- and Between-Speaker Variability of Parameters Expressing Short-Term Voice Quality." In *Proc. Speech Prosody*, pp. 1081–1085, 2014.

[WES13] Felix Weninger, Florian Eyben, Björn W. Schuller, Marcello Mortillaro, and Klaus R. Scherer. "On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common." *Frontiers in Psychology*, **4**:1–12, 2013.

[Wol72] Jared J. Wolf. "Efficient Acoustic Parameters for Speaker Recognition." *The Journal of the Acoustical Society of America*, **51**(6B):2044–2056, jun 1972.

[YBL04] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. "An acoustic study of emotions expressed in speech." In *Proc. Interspeech*, pp. 2193–2196, Jeju, Jeju Island, Korea, 2004. ISCA.