Defining an Enhanced Feature Subset to Assess the True Biometric Performance of Speaker Verification Systems

By

Jonathan M. Leet

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies in Computing

at

The Seidenberg School of Computer Science

Pace University

May 2015

Dissertation Signature (Approval) Page

We hereby certify that this dissertation, submitted by Jonathan M. Leet, satisfies the dissertation requirements for the degree of Doctor of Professional Studies in Computing and has been approved.

_____-_____
Charles Tappert Date
Advisor

_____-_____
Ronald Frank Date
Dissertation Committee Member

_____-_____
Juan Shan Date
Dissertation Committee Member

School of Computer Science and Information Systems
Pace University 2015

Abstract

Defining an Enhanced Feature Subset to Assess the True Biometric Performance of Speaker Verification Systems

By

Jonathan M. Leet

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies in Computing

May 2015

The speaker verification system industry requires increased standardization in testing processes and evaluation methodology, specifically those that select the best combination of feature vectors (or enhanced feature subset) to use as input for analysis in these systems. There are numerous methods currently used to "benchmark" these offerings, which translates to a marketplace of solutions that cannot be compared against each other on equal terms. Organizations like NIST have also failed to standardize these evaluation methodologies, which is necessary to create an environment that fosters reliable performance metrics. These inconsistencies indicate that the current research is flawed and unreliable. This dissertation will demonstrate a reusable methodology to identify the best subset of feature information to isolate in the measurement of the actual biometric performance of a speaker verification system with the expressed intent of standardizing the manner that testing is conducted.

Acknowledgements

I have profoundly changed in the past five years, in which I have been enrolled as a doctoral student at Pace University. Through the experiences and relationships made, I have become a focused professional with a better understanding of my core competencies in information technology. I am thankful for the incredibly knowledgeable professors and educational support staff that has made my time at Pace rewarding and fulfilling. I also thank my advisor Dr. Tappert, specifically, because without his encouragement, wisdom, and direction, I could not have accomplished this amazing feat. I appreciate the time and effort that he has invested in me more than he could ever imagine.

Additionally, I appreciate the access I had to other graduate students willing to help conduct research and hypothesize on potential adjustments that could increase the accuracy of our system. Although the point of this study was not to create a more sophisticated speaker verification system, the graduate group was fundamental during the discovery process, and their contribution was invaluable. For this I am eternally thankful.

Sincerely,

Jonathan

Jonathan M. Leet

Dated: May 11, 2015

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Current Speaker Verification Evaluation and Feature Extraction Methodologies

This section describes the existing state of speech feature vector selection in speaker verification system evaluation and illustrates the need for a standardized approach to identifying an enhanced subset of features vectors.

The main contribution of this dissertation is clear. The speaker verification system industry requires increased standardization in testing processes and evaluation methodology, specifically those that select the best combination of feature vectors (or enhanced feature subset) to use as input for analysis in these systems. There are numerous methods currently used to "benchmark" these offerings, which translates to a marketplace of solutions that cannot be compared against each other on equal terms. Organizations like NIST have also failed to standardize these evaluation methodologies, which is necessary to create an environment that fosters reliable performance metrics. These inconsistencies indicate that the current research is flawed and unreliable. This dissertation will demonstrate a reusable methodology to identify the best subset of feature information to isolate in the measurement of the actual biometric performance of a speaker verification system with the expressed intent of standardizing the manner that testing is conducted.

According to Saquib et al. in 2011, individual voices have measurable differences in characteristics like the length of the vocal chords, specific elements of the vocal tract, and variations in speaking habits. With the combined effect of numerous speech productions organs like the laryngeal pharynx, oral pharynx, oral cavity, and the nasal cavity, the complexity of human speech provides adequate components required for a biometric feature to remain applicable in an authentication study. Articulators like the size and shape of the mouth, throat, nose, and teeth, combined with the manner of vocalizing, contribute to a distinct and unique vocal pattern. These studies have even demonstrated that measurements like vocal tract shape can be accurately estimated from the spectral shape, represented as feature vectors, and used as input to a speaker verification system [50].

In 2000, Markowitz stressed the importance of understanding the fundamental differences between voice recognition and voice biometrics: "The most significant difference in the technologies is that voice biometrics technologies do not know what a person is saying, relying on speech recognition to do that". The researchers interested in speaker verification are concerned with the evaluation and comparison of pre-enrolled speaker profiles, when compared to another speech sample that may, or may not, be the individual attempting authentication. The semantic content is negligible and becomes considerably less important than the numerical feature information that can be extracted through automation, which is demonstrated in current peer-reviewed literature [34].

Moreover, the trend toward speaker independence that characterizes speech recognition cannot exist for voice biometrics. By definition, voice biometrics is linked to a particular speaker. As

a result, speaker verification technology requires some type of enrollment for each user. The need for enrollment is a specific attribute of voice biometrics, which is a fundamental difference from voice recognition. The user must have prior interaction with the system to build a speaker profile for comparison. Thus, one of the significant areas of research is in the enrollment process itself, as this becomes a feature of the system that will have an early and significant impact on performance, which studies have illustrated by comparing the EER measurements of different passphrases [34].

Empirical evidence indicates that in recent years the use of speech technology applications has grown rapidly, and this growth is expected to continue, allowing humans and computers to connect more efficiently [38]. The key, security-related speech technology of speaker verification is the primary focus of this study. Science fiction tends to suggest that this technology is fully developed and has been used in a variety of applications related to national security, but this is not exactly the case. Although there is widespread use of this technology in these arenas, the true biometric performance of speaker verification systems is poorly measured in a variety of non-standardized methodologies, because the relevant studies and experiments fail to use common datasets, enrollment processes, or feature extraction methods to identity an enhanced subset of feature vector information, as illustrated in Table 1. This negatively impacts to the value of performance measurements, because the results of these studies cannot be adequately compared and contrasted, as the methodologies implemented were entirely inconsistent with one another.

Table 1. Speaker Verification System Evaluation Studies:

Study	Enrollment Methodology	Feature Extraction Methodology
Guo & Wang, 2006 [19]	-24 male and 24 female subjects -3000 Total Samples -Used 8 phrases and normalization	-Word-based testing only (single lexical level) -Used average scores over utterances -Digits only
Han & Gau, 2010 [20]	-20 Male subjects -Score normalized values extracted from a 10 minutes speech sample -8 different microphones	-Values averaged from 8 mikes (unrealistic for everyday use) -Individual lexical-levels are not isolated during speech modeling
Kato & Shimizu, 2003 [28]	-81 male and 74 female subjects -3 six digit utterances used for enrollment -Score normalized values used	-Digits only (full phonetic spectrum not covered) -Individual lexical-levels are not isolated during speech modeling
Pandit & Kittkr, 1998 [46]	-Two different data sets (one consists of 33 French male and female speakers and the other consists of 40 Spanish speakers) -4 repetitions of 0-9 digit sequence recorded at 1 week intervals (unrealistic for everyday use)	-Digits only (full phonetic spectrum not covered) -Discusses the concept of an enhanced feature vector set, but does not discuss its implementation methodology
Reynolds, 1995 [49]	-438 males and 192 females subjects -Each test was conducted using 50 randomly selected speakers from the overall pool of 630 -8 utterances used during enrollment	-Score normalization -Lack of detail regarding feature information -Averages used, phonetic information and other lexical-levels of information combined to form speech model
Schriberg, 2005 [52]	-20 male and 20 female subjects -8 recorded training conversations for enrollment -2 minute samples	-Score normalization -Cohort background scoring -Use of spectral averages, no isolation of lexical-levels

The University of Edinburgh, 2000 [56]	-700 participants -Multiple phrase used during the enrollment	-Score normalized results -Phonetic information averages -Fail to isolate the sound-level information before processing
Wagner, et al., 2006 [61] – *Evaluation included Nuance and Persay*	-91 male and 231 female subjects (RSR2015 Speech Database [35]) -Samples recorded in three sessions over 9 months	-Score normalized -Averages calculated from entire feature classes -Individual lexical-levels are not isolated during speech modeling

According to Martin's 2008 presentation, even the evaluation methodologies used by the National Institute of Standards and Technology (NIST) have been inconsistent over time. These changes have resulted in an inability to compare these studies, as they are not conducted on equal terms. From 1991-2008, NIST Evaluations have varied significantly in the collection, processing, and feature extraction of speech samples. For example, original testing was completed using speech read from an individual and captured from a microphone, while the more recent speaker verification system analysis has incorporated speech from meetings with multiple individuals, conversational speech between subjects, broadcasted speech, speech over cellular networks, and speech over landline telephones [35]. The incongruent channels used to gather and enroll subjects will have an impact on the performance of the system, as this study will demonstrate quantitatively and qualitatively. Furthermore, it is impossible to contrast this research against other relevant academic studies. The NIST experiments are fundamentally inconsistent with those presented in Table 1 of this dissertation, which leads to an inability to compare existing results to cover the broadest range of offerings in the marketplace.

1.2 Articulatory Phonetics

This section describes the science of articulatory phonetics, which is a fundamental discipline used in this dissertation. References to the ARPABET are also essential to the understanding of this material.

Articulatory phonetics is used to understand the production of speech and design an optimal common speech utterance, while acoustic phonetics is used in the processing and analysis of the speech utterance waveforms. Biometric measurements are then used to design and evaluate an authentication system that operates on short speech utterances. The combined understanding of both linguistic and biometric disciplines proved essential to this dissertation [27]. Additionally, the automated decomposition of speech is central to all arguments made throughout this dissertation and requires a multi-disciplined approach to test design, which is an argument that also exists in other scholarly work including the 2009 work by Chakroborty and Saha [10].

A particular phrase spoken by an individual consists of distinct voice components called phonemes. A phoneme is the smallest comparable unit in the sound system of a language. Each phoneme has a pitch, cadence, and inflection, giving each person a unique voice. Based on the individual's vocal tract, features like tract length, ratio of larynx to sinuses cadence, pitch, tone, frequency, range, and duration of voice, the frequency values of a speech sample will vary considerably from user to user. Similarities in voice often come from cultural and regional influences in the form of accents. In different cultures and regions, the lexical

individuality of a speech sample may require the decomposition of a phrase into different sound-level feature [27].

As described by Jurasky, et al. in 2008, speech has specific points and manners of articulation, which can be used to create the combination of utterances that form human speech. The place of articulation can vary as follows [27].

Labial: Consonants formed by the two lips coming together are referred to as bilabial and include [p], [b], and [m]. Labiodental consonants like [v] and [f] are made by pressing the bottom lip against the upper row of teeth and letting the air flow through the space of the upper teeth.

Dental: These sounds are made by placing the tongue against the teeth and include the [th] and [dh] sounds.

Alveolar: The tip of the tongue is placed against the alveolar ridge, which creates sounds like the [s], [z], [t], and [d] sounds.

Palatal: The palato-alveolar sounds are created with the blade of the tongue against the rising back of the alveolar ridge and creates sound like [sh], [ch], [zh], and [jh].

Velar: The sounds [k], [g], and [ng] are made by pressing the back of the tongue against the velum.

Glottal: The glottal stop [q] is made by closing the glottis.

Furthermore, the manner of articulation also has an impact on the variation in speech from speaker to speaker [27]:

Stop: a constant in which airflow is completely blocked for a short time. The period of blockage is called closure, while the explosion is called release. Sounds like [b], [d], and [g] are voice stops, while sounds like [p], [t], and [k] are considered unvoiced stops.

Nasal: The nasal sounds like [n], [m], and [ng] are made by lowering the velum and allowing air to pass into the nasal cavity.

Fricatives: Airflow is constricted, but not cut off completely. The lower lip is pressed against the upper teeth to create sounds like [f] and [v].

Approximates: The two articulators are close, but not close enough to cause turbulent airflow. The tongue moves close to the roof of the mouth, but not close enough to create a fricative sound. Sounds like [y] and [w] are created using this manner of articulation.

Tap or Flap: This is a quick motion of the tongue against the alveolar ridge that creates sounds like [dx].

The phonetic alphabet is a representation of the sound units available for use in the creation of individual words. The International Phonetics Alphabet (IPA) provides an alphabet of phonetic information, primarily represented in Latin. Another alphabet used in this study is the Arpabet, which is listed as follows (as accessed on September 27, 2014 from Wikipedia.com) in Table 2 [27]:

Table 2. The Arpabet, IPA, and Word Examples of English Speech Sounds:

Table 2a. Monophthongs:

Arpabet	**IPA**	**Word examples**
AO	ɔ	off (AO1 F); fall (F AO1 L); frost (F R AO1 S T)
AA	ɑ	father (F AA1 DH ER), cot (K AA1 T)
IY	I	bee (B IY1); she (SH IY1)
UW	U	you (Y UW1); new (N UW1); food (F UW1 D)
EH	ɛ	red (R EH1 D); men (M EH1 N)
IH	ɪ	big (B IH1 G); win (W IH1 N)
UH	ʊ	should (SH UH1 D), could (K UH1 D)
AH	ʌ	but (B AH1 T), sun (S AH1 N)
	ə	sofa (S OW1 F AH0), alone (AH0 L OW1 N)
AX		discus (D IH1 S K AX0 S); note distinction from discuss (D IH0 S K AH1 S)
AE	Æ	at (AE1 T); fast (F AE1 S T)

Table 2b. Diphthongs:

Arpabet	IPA	Word Examples
EY	eɪ	say (S EY1); eight (EY1 T)
AY	aɪ	my (M AY1); why (W AY1); ride (R AY1 D)
OW	oʊ	show (SH OW1); coat (K OW1 T)
AW	aʊ	how (HH AW1); now (N AW1)
OY	ɔɪ	boy (B OY1); toy (T OY1)

Table 2c. R-colored vowels:

Arpabet	IPA	Word Examples
ER	ɝ	her (HH ER0); bird (B ER1 D); hurt (HH ER1 T), nurse (N ER1 S)
AXR	ɚ	father (F AA1 DH ER); coward (K AW1 ER D)
EH R	ɛr	air (EH1 R); where (W EH1 R); hair (HH EH1 R)
UH R	ʊr	cure (K Y UH1 R); bureau (B Y UH1 R OW0), detour (D IH0 T UH1 R)
AO R	ɔr	more (M AO1 R); bored (B AO1 R D); chord (K AO1 R D)
AA R	ɑr	large (L AA1 R JH); hard (HH AA1 R D)
IH R *or* IY R	ɪr	ear (IY1 R); near (N IH1 R)
AW R	aʊr	*This seems to be a rarely used r-controlled vowel. In some dialects* flower (F L AW1 R; *in other dialects* F L AW1 ER0)

Table 2d. Stops:

Arpabet	IPA	Word Examples

P	P	pay (P EY1)
B	B	buy (B AY1)
T	T	take (T EY1 K)
D	D	day (D EY1)
K	K	key (K IY1)
G	ɡ	go (G OW1)

Table 2e. Affricates:

Arpabet	**IPA**	**Word Examples**
CH	tʃ	chair (CH EH1 R)
JH	dʒ	just (JH AH1 S T); gym (JH IH1 M)

Table 2f. Fricatives:

Arpabet	**IPA**	**Word Examples**
F	F	for (F AO1 R)
V	V	very (V EH1 R IY0)
TH	Θ	thanks (TH AE1 NG K S); Thursday (TH ER1 Z D EY2)
DH	Ð	that (DH AE1 T); the (DH AH0); them (DH EH1 M)
S	S	say (S EY1)
Z	Z	zoo (Z UW1)
SH	ʃ	show (SH OW1)
ZH	ʒ	measure (M EH1 ZH ER0); pleasure (P L EH1 ZH ER)

HH	H	house (HH AW1 S)

Table 2g. Nasals:

Arpabet	IPA	Word Examples
M	M	man (M AE1 N)
EM	m̩	keep 'em (K IY1 P EM)
N	N	no (N OW1)
EN	n̩	button (B AH1 T EN)
NG	Ŋ	sing (S IH1 NG)
ENG	ŋ̍	Washington (W AO1 SH ENG T EN)

Table 2h. Liquids:

Arpabet	IPA	Word Examples
L	l	late (L EY1 T)
EL	l̩	bottle (B AO1 DX EL)
R	ɹ *or* ɻ	run (R AH1 N)
DX	ɾ	wetter (W EH1 DX AXR)
NX	ɾ̃	wintergreen (W IY2 NX AXR G R IY1 N)

Table 2i. Semivowels:

Arpabet	IPA	Word Examples
Y	J	yes (Y EH1 S)
W	W	way (W EY1)

Q	ʔ	glottal stop (uh-oh - ʔʌʔoʊ)
(missing)	hw *or* ʍ	"when" etc. in some dialects

The frequency response generated from the speech of a particular individual is called a spectrogram. The spectrogram consists of frequency and amplitude plotted over time. Frequency is the number of times per second a wave repeats in itself (or cycles), while amplitude measures the amount of air pressure variation. Once the individual phonemes or words extracted from the common phrase are identified (segmented from the entire sample), the frequency and amplitude values for each such segment are distinct per individual. Other research indicates that spectrographic analysis simplifies the identification of individual phonemes in a particular phrase, as well as increasing the likelihood of the successful identification of a speaker based on the underlying feature vectors [27].

Various sources have suggested that the following variables could impact the level of accuracy achievable through the use of a particular sample. This poses an inherent obstacle in the utilization of such technology for the following reasons [27]:

• The individual (sex, age, origin, etc.) that produces it will obviously have an influence over the measured results

• Our voices tend to be different at different times of the day and will affect voice print produced from sample to sample

• The noise condition under which the sample produced will affect the ability to extract meaningful feature information

• Longer samples may produce less reliable measurements, although it is suggested that this in counter-intuitive to conventional thought [13]

There are a variety of methods used to combat the interference of noise, which were developed for the purpose of speech recognition, as it's understood in today's marketplace. These techniques will prove useful, as they are unrelated to the functional concepts of speaker recognition. According to existing research, the most common approaches to attenuating noise are voiceprint adaptation, cohort modeling, and world models [27]. The research in this dissertation did not include a pre-processing methodology to enhance the sample, which may have contributed to a decrease in overall performance. It was also noted by the researchers that a handful of samples in this study contained measurements classified as "extreme outliers". Although not substantiated with automated analysis, it was hypothesized that values falling outside the reasonable norms could have been affected by observable defects in the speech samples.

In the enrollment phase, an outlier removal method can be used to reject the enrollment of samples that contain feature measurements that fall outside of the normal, observed range for a particular speaker. Since only a handful (potentially 1 in 100 depending on the identification criteria) of the samples would contain an "extreme outlier," this process would not be overly cumbersome, nor add much additional overhead to the enrollment process. This is not a main component of the testing methodology recommended in this dissertation, but it could be a useful means to standardize enrollment for future feature vector selection methodology

standardization, or simply as a systemic means to increase the efficiency of the enrollment phase itself.

The second phase of the authentication process is the verification of the utterance, when compared against the enrolled speech. "At the verification stage, the user's biometric data is compared with the template stored in the system and decision is made according to the result" [34]. As this study will clearly expand upon, the verification of a speech sample can be implemented in a variety of manners, all of which could yield different levels of success or failure; again, as Table 1 demonstrates. As both existing academic research and the NIST Standards illustrate, there is no exact standard on how to properly authenticate a speech sample against an enrolled profile, so there is a tremendous benefit that can be achieved by identifying a common methodology for future studies.

Research suggests that a balance between the enrollment process and verification stage will need to be achieved to reach optimal system performance [27]. Voiceprint adaptation involves the modification of an enrolled voiceprint to incorporate additional data into the original model. Adaptation makes it possible to include additional acoustic information, from the complete range of devices a person may use, in the voiceprint. Adaptation can also update the voiceprint with regard to variations in the person's voice. For example, a person may speak quite differently when under stress or when tired. This technique can allow for plotting of information within a specific user range and could provide a larger spectrum in which a user's sample may be considered authentic, as illustrated in existing academic research [6].