ORIGINAL RESEARCH

Using a predefined passphrase to evaluate a speaker verification system

Jonathan Leet *, John Gibbons, Charles Tappert, Vinnie Monaco

School of Computer Science, Pace University, White Plains, New York, USA

Received: July 31, 2014 **Accepted:** November 6, 2014 **Online Published:** November 26, 2014
DOI: 10.5430/air.v4n1p1 **URL:** http://dx.doi.org/10.5430/air.v4n1p1

Abstract

This article presents a standardized and repeatable process used to evaluate the performance of a speaker verification system. Through the use of a common passphrase and a subset of extracted feature vectors that outperforms other combinations, the study limits the exposure to potential experimental flaws, while measuring true biometric performance more effectively than existing evaluation methodologies. After collecting a dataset of 33 participants, the researchers achieved a performance rate of 99.8% for the 22 users who contributed at least 20 text-dependent samples. The primary focus of the research, however, was to illustrate a variety of testing techniques that can be used to efficiently analyze the performance of a speaker verification system and advocate the use of a common passphrase in this process.

Key Words: Speaker verification system, Feature vectors

1 Introduction

This paper will analyze current methodologies used to evaluate the performance of a speaker verification system. There appears to be a significant lack of standardization in the processes used to compare these systems, and researchers are often encouraged to implement large datasets to provide a greater reliability in the experimental results.[1–5] Unfortunately, the use of multiple, text-dependent speech samples can have a negative effect on the ability to analyze the phonetic information contained in a particular phrase. An evaluation process that uses a single, common passphrase can be more effective at analyzing the performance of a speaker verification system, while limiting exposure to potential experimental flaws and permitting the measurement of true biometric performance.

This paper includes the design, implementation, and testing of a universal evaluation methodology for speaker verification systems. Through the use of a common passphrase and a subset of extracted feature vectors that outperforms other combinations, this study will demonstrate a standardized and repeatable process to evaluate a speaker verification system. After collecting a dataset of 33 participants, the verification system used for experimentation achieved an overall performance rate of 99.8% for the 22 users who contributed at least 20 text-dependent samples. The primary focus of the research, however, was to illustrate a variety of testing techniques that can be used to efficiently analyze the performance of a speaker verification system, in addition to advocating the use of a common passphrase during the evaluation process.

A main concern of the researchers performing these experiments was the limited dataset used for the experimental research. Other relevant academic research like that described and inspired by CMU's Roy Maxion and col-

[*] **Correspondence:** Jonathan Leet; Email: jonathan.leet@gmail.com; Address: School of Computer Science, Pace University, White Plains, New York, USA

leagues in keystroke and mouse movement studies, including the 2013 study by Monaco, et al. published in the IEEE 6th International Conference on Biometrics proceedings, also used a limited number of text samples. In the research by Monaco, et al., 300 samples (consisting of 10 samples collected from each of 30 authors) were retrieved for experimental use from Project Gutenberg (http://en.wikipedia.org/wiki/Project_Gutenberg) and were taken from books published between 1880 and 1930. The samples were not restricted geographically, and authors were included from Great Britain, Ireland, and the United States. Although the user populations available for experimentation was less extensive than the dataset used in these studies, the research still possessed enough academic relevance to be published by one of the leading organizations in the field of biometrics.[6, 7]

This study did not include further experiments to substantiate the performance results, as the expressed intention of the research was not to produce a more sophisticated speaker verification system. Although the performance figures were an important aspect of the research, they were not central to the arguments made in this paper. The main contributions of this study were to demonstrate that a subset of the features extracted from a speech sample may be more efficient at authenticating a particular individual, and that phonetic information can be measured more effectively using a single, text-dependent passphrase.

The pressure to increase the speech database size is also a fundamental motivation for evaluation standardization, as the implementation of different speech databases could easily lead to the inflation of experimental results. The resulting state of the existing research related to the performance of a speaker verification systems is an ecosystem of disparate architectures, which cannot be adequately compared or contrasted against one another. This is a byproduct of the various, non-standardized techniques used measure the performance of these systems and creates a marketplace open to the potential misrepresentation of true biometric performance.

The reviewers of this study deliberately avoided the use of large speech databases for the evaluation of speaker verification systems. This paper will argue that these collections will only serve to dilute the potential value of measurements made though "benchmark" testing. The granularity of this research is compromised through the use of multiple passphrases, especially during the enrollment phase of the authentication process. The ability to analyze phonetic information is made more difficult and complex, not only because this will increase the segmentation problem often encountered, but because it will also make the comparison between individual phonemes more cumbersome and potentially impossible. Without the ability to compare the performance of individual sound units, the simple choice of a particular passphrase could artificially "boost" performance results and, consequently, lead to an inherent security threat. This article advocates the use of a text-dependent phrase in analyzing the performance of a speaker verification system. Although text-dependent phrases are no longer used in commercial systems and could be considered outdated, the implementation of these concepts can address specific issues related to the assessment of pure biometric performance and can positively impact the validity and reliability of any study or evaluation of these particular systems. The main contribution of this study is meant to illustrate a variety of testing techniques that can be used to efficiently analyze the performance of a speaker verification system, while limiting exposure to potential experimental flaws and permitting the measurement of true voiceprint biometric performance. This is achieved through the use of a common phrase to allow for the testing and evaluation of existing speaker verification systems in a controlled and repeatable manner.[7]

Similar highly-controlled experiments providing insight on the pure performance achievable by biometric systems were conducted by Maxion in 2011, and this paper argues that it is novel and meaningful to introduce and explore these concepts within the field of speaker verification. While a specially-designed phrase, "My name is", is a text-dependent phrase used in this research, the purpose of the study is to provide a framework or specific methodology that can be used to evaluate an existing speaker verification system. The text-dependent phrase can also be used for efficient imposter testing, which is one main advantage that Maxion also realized in his keystroke biometrics studies that serve as a precedent to the research contained in this paper. This is one of the clear benefits of a methodology relying on text-dependent phrases, considering other enrolled users samples can be easily used as potential imposters for testing purposes.[7]

A variety of researchers continue to find applications for text-dependent phrases, regardless of a lack of interest in these concepts as related to current or future commercial offerings. Applications, like those described by Gu and Thomas, which evaluate or "benchmark" speaker verification systems, use text-dependent phrases regularly to accomplish their testing goals.[8] This study compared two different text-dependent phrases during imposter testing, while use of the exact phrase spoken by a different user was a unique approach explored only in the research presented later in this article. This paper advocates a more standardized approach to auditing these systems be outlined, as this will allow for the comparison of commercial systems in a manner that would be transparent and reliable.

At this time, the only reliable measurement of biometric performance related to speaker verification systems would be the third party evaluations conducted by independent researchers or the proprietary/real-world measurements, which may be insufficient at providing performance details that can be compared to other systems in the commercial arena.[3,9] These evaluation methods are similar to those included in this paper, but are presented in a clear and highly repeatable manner in the study included in this research.

Current evaluation methodologies lack standardization and could lead to experimental flaws that compromise the integrity of performance testing. Furthermore, the architecture and processes used to evaluated speaker verification systems are much more clearly defined then the testing methodology itself. Phonetic information can be extracted in a variety of methods using a collection of feature vectors, much like Reynolds illustrated in 1995 and Shriberg, et al. demonstrated in 2005.[10,11] Studies like Kato and Shimizu's research on balancing phonemes can be paired with existing evaluations techniques and system architectures to streamline the enrollment and verification process as described by Wagner, et al. by using text-dependent phrases to create a powerful and reusable testing methodology.[9,12]

Additionally, concepts like phoneme classes are explored by Hebert and Heck and could also be implemented in parallel to a text-dependent evaluation methodology.[7] This study focuses primarily on phonetic research, in addition to common algorithms which evaluate the fundamental frequency of a particular word or phrase spoken by a user. Considering a majority of the existing research in the field of speaker verification tends to explore the potential of phoneme related information, the paper intends on exploring the potential feature combinations to extract during a system evaluation, while providing a very specific evaluation process that ensures the pure biometric performance is accurately captured in a consistent manner.

This research could make an immediate impact throughout the field of speaker verification, as the commercial software provided by the industry's leading companies is already used in a variety of client applications.[14–20] For example, Nuance provides clients in financial services and other industries with voiceprint verification during live customer service calls. The American Safety Council's iAM BioValiation product is also used in customer service authentication but has a stronger adoption rate in government agencies including a number of state departments of motor vehicles. Smaller niche companies like Voicevault and Voice Biometrics Group offer extensible APIs for client integration/customization and a wide variety of verification options depending on client requirements, such as high-volume Interactive Voice Response (IVR) systems. The highly niche company Authentify markets its solutions primarily for "out of band", phone-based verification of web-initiated transactions.

The remaining sections of the paper are as follows. Section 2 presents an analysis of the passphrase types used in commercial voiceprint systems, section 3 advocates the use of a common phrase for analyzing the performance of a voiceprint system, section 4 describes the system developed for this study, section 5 presents the experiments performed and the results obtained, and the final section draws conclusions.

2 Commercial speaker verification (voiceprint) systems

2.1 Testing methods

Known testing methodologies remain mostly based on proprietary evaluations, which are not made public to potential consumers. A variety of academic studies also use non-standardized processes and procedures to evaluate speaker verification systems, which leads to an inability to compare and contrast performace between competitive offerings. By standardizing a testing framework, it would be possible to isolate factors that lead to the possibility of "boosting" performance metrics, through passphrase selection and system tuning to a particular user set. These techniques, while not always applied maliciously, can lead to enhanced results that could not be used to accurately measure the true performance of a speaker verification system. Therefore, the standardization of this process would help provide enhanced security throughout the consumer base for this technology, which we have already illustrated to be reasonable widespread.

A main source of performance information for the commercial systems is through third party studies like those conducted by Wagner, et al. in their 2006 study that included Nuance, Scansoft SpeechWorks Speaker Verification SDK pro 3.0, and the Persay VocalPassword Build 5.0.5.0. The samples used in this study were collected over several months in 3 different data collection sessions. Each participant provided multiple samples, all used to increase the potential value of the enrollment process, and not that of the authentication algorithms themselves.[9] The premise of this paper challenges these results, as the methodology presents a highly artificial testing scenario that could easily invalidate the results as described throughout this study.

Again, this paper identifies flaws in the current studies, as they incorporate multiple samples from each individual and use datasets containing potentially unrealistic numbers of users. Additional studies were also conducted using large, extensive databases of speech samples, like that of the RSR2015 collection, on a variety of commercial offerings.[1] Through the use of over 300 samples and over 30 phrases from user participant, this research is entirely susceptible to the same experimental flaws that were identified previously in this study, specifically that the advanced training of the system creates and artificial condition that allows for inflated performance results.

Further studies like those conducted by The University of Edinburgh in 2001 also used a large dataset of over 700 speakers with each providing multiple samples throughout the enrollment process.[5] Once again, this paper argues that this will lead to the same experimental flaws that will lead to highly artificial conditions that could lead to an inaccurate understanding of the actual performance of a system. It was a central concept of this study to avoid the use of extensive collections of training information, in an effort to identify the potential security flaws in systems, when proper processes are not implemented to avoid weak enrollment and training methodologies.

The use of an extensive database, with multiple samples from each subject, could lead to boosting and tuning of these systems, which is the exact rationale for use of a simple limited set of data. The use of a limited set of participants and a single text-dependent utterance is neither a weakness nor an oversight of this paper, as other academic research has used more realistic data sets to express the true biometric value of a particular system, specifically in the arena of keystroke biometrics.[14, 15]

A fundamental concern recognized during the research was that the evaluation methodologies currently being explored allow for the boosting of results by providing extensive and unrealistic amounts of data. It is possible that the volume of data collected or the number of users in these systems will not be in the hundreds, but far less in number. Furthermore, the collection of phrases used for training may be limited in comparison to the study referenced in this article, which leads to an even greater potential for inaccurate or inflated results. The simplicity of this study is its inherent strength, as it eliminates the potential tuning effect that these large studies can create.

2.2 Passphrase types and selection criteria

The careful selection of a passphrase can allow for the incorporation of "boosting" techniques that will increase overall system security by using passphrases, which tend to demonstrate higher level of lexical individuality. The research that is included in the study will demonstrate the different types of decomposition of the speaker sample to achieve different results. This paper also advocates the use of a single phrase with the intention is to limit experimental biased that could be the result of phrase selection. A true measurement of the biometric performance for a system would be achieved in more controlled and realistic manner using the methodology suggested in this paper. As noted earlier, the usability of a system cannot be impacted by elaborate enrollment or verification process.

There are several types of passphrases used in commercial voiceprint systems. Similar to a password, a passphrase is the phrase spoken by the user to gain access to an application. In the companies surveyed, the passphrase types, in vendor terminology, span two dimensions – active versus passive and open versus closed:

- Active-Open
- Active-Closed
- Passive

Active speech collection systems prompt the user for a specific passphrase, and vendors also refer to this method as "text dependent" or "text prompted". Passive speech collection systems allow the user to say anything, and vendors also refer to this method as "text independent" or "free-form". In an active-open speech collection system the user is asked to speak a passphrase defined by the system. In an active-closed system the passphrase is determined by the user and kept secret, although the system prompts the user to say their passphrase. Allowing the user to select a secret passphrase is similar to users selecting passwords to enter on a keyboard.

Passive speech collection systems allow client applications to take passive speech – such as a conversation between a caller and a customer service representative – and send as much speech as practical (usually several minutes) to the service platform to build a rich phonetic model. Other voice authentication systems allow the user to select the speech utterance to be input to the system, which is similar to users selecting passwords to enter on a keyboard. The primary speech utterance used in this study will be the same for all users being authenticated, which is implemented specifically for the reasons that will be expanded on throughout this section.

The most popular passphrase types are active-open and passive, with active-open currently dominating the speaker verification arena. The active-open approach prompts the user for a different passphrase from session to session for enhanced security. The passphrases typically consist of numbers, dates, or other common words extensively analyzed by the speech community over many years to improve automatic recognition over all users in speech dialogue applications.

As the technology has improved in recent years, however, an increasing number of companies are becoming interested in the potential usability of other passphrase types. Six companies that produce software in the space of speech recognition and speaker identification were surveyed: Nuance, Authentify, Persay VocalPassword Build 5.0.5.0, VoiceVault, iAM BioValidation, and VoiceBiometrics Group. Table 1 shows five company's referenced with specific passphrase type support.

In the next several sections we will take a closer look at the existing systems currently being used for speaker verification and authentication. Each section will present as much material on the evaluation process used to reach the offered biometric performance measurement. The ability of the researcher to gather this information was impeded by the inherent proprietary nature of these systems and primarily it could not be obtained. This strengthens the argument made in this paper, because this is the type of ambiguity in true performance metrics that could translate into the selection of an inferior system.

Table 1: Voice biometric company passphrase survey

Company	Passphrase Type
Nuance	Active Open Passive
Authentify	Active Open Passive
iAM BioValidation	Active Open
VoiceVault	Active Open Active Closed
Voice Biometrics Group	Active Open Passive

2.3 Nuance

Nuance had 23 million of the total 28 million voiceprints worldwide as of the end of 2012. Nuance's agent assisted authentication software listens to a live user-agent conversation and provides the agent with a user-identity confirmation. The company TD Waterhouse employs Nuance and uses an active-open passphrase: 10-digit phone number + month-day date.[18]

Wagner et al. conducted a study in 2006 that included Nuance in their research.[9] As previously stated, the methodology used in this study would be subject to the same experimental flaws that the processes included is this research helps eliminate in the evaluation process. This paper argues that the published results, including those related to the nuance offering, cannot be trusted based on the highly artificial testing environment and overly complex enrollment phases. The evaluation techniques presented in this literature would have an advantage over the previous research, as it take into account flaws that have been identified in other biometric disciplines.

2.4 Persay VocalPassword Build 5.0.5.0

Persay VocalPassword Build 5.0.5.0 is the commercial offering from the New York based Persay Company. The solution is currently used in applications such as mobile banking, social networks, payment services and membership clubs, and it can utilize a variety of login ID and passphrase combinations to tailor a unique authentication experience. A variety of platforms currently offer support of this software including mobile platforms like iOS devices. It was also a subject of the study conducted by Wagner, et al. in 2006, which this paper argues is subject to the experimental flaws described in this research.[9]

2.5 Authentify

Authentify claims they are the worldwide leader in voice authentication through a phone to verify the identity of users making web transactions.[15] With users from different cultures or having different accents or vocabulary comfort levels, numbers are usually the most accessible way to get consistent voice data. In this active-open model, verification is performed against a randomly-generated phrase to reduce chances of a fraudulent user able to match the generated phrase. They also have a text-independent (passive) model.

A study in 2006 conducted by Elliot, et al. provided performance metrics for the Authentify product using a non-standardized methodology for evaluation.[5] This results of this article would suggest, after extensive academic research consistent with the experimental results of this study, that the measurements they achieved cannot be effectively copared against other research into competing products. The inability to compare dissimilar evaluations nullifies the potential value of this research. The standardized methodology recommended in this paper will attempt to resolve these issues by advocating the use of a single, text-dependent utterance for the evaluation of speaker verification systems.[6]

2.6 iAM BioValidation

iAM BioValiation is a product provided by the American Safety Council (ASC), a market leader in engineering, authoring, and delivery of e-Learning training solutions. ASC currently implements voice or keystroke biometrics for The New York Department of Motor Vehicles, The New Jersey Motor Vehicle Commission, The University of California at San Diego, American Automobile Association, and The Florida Department of Highway Safety and Motor Vehicles.[14] iAM BioValidation employs a text-directed speech model (Active-Open). The system prompts for training and authorization using randomized sets of numbers, such as the sequence of two-digit numbers "57 96 95 93 97 77 54 45", spoken as "fifty seven, ninety six,$\cdots$, forty five".

Currently, there was no information related to the evaluation methodology used to determine the relevant performance measurements. This only further serves as rationale to utilize the methodology proposed in this paper, as the use of the process outlined in this article will not expose proprietary elements of the system. It will, however, assist in the evaluation of a particular system and help "benchmark" the individual performance in a comparable fashion with other commercial or proprietary offerings in the marketplace. This is powerful aspect of this research, as each individual system can follow the steps presented in this research to compare and contrast the potential performance, in respect to the competing products on the market.

2.7 VoiceVault

VoiceVault is a smaller but more agile voice biometrics vendor that allows developers to implement their voice biometric engines on cloud-based enterprise and mobile platform solutions.[19] VoiceVault specializes in text-dependent digit and secret passphrase voice biometric solutions for identity verification using small amounts of speech. VoiceVault provides "multi-factor identity authentication solutions that enhance the something you know (a PIN or password) with something you are (your unique voice)".[22]

The VoiceVault text dependent solution encompasses two types of user experience – active-open and active-closed passphrases. An example text-prompted passphrase is: "seven four eight three". At the time of this paper's publication, it was not evident that performance information for this offering was available. This only supports the articles central argument, specifically, that a standardized evaluation methodology for use with disparate speaker evaluation systems needs to be identified to make these studies worthwhile.

2.8 Voice Biometrics Group (VBG)

Voice Biometrics Group provides a custom solution for every client, featuring broad production support for both text-dependent and text-independent techniques, and using multiple languages in multiple countries.[20] There are no specific preferences and they don't favor one engine configuration over another. Their VMM-1 voice biometric decision engine has internal support for active and passive passphrases and is fully configurable to support whatever operating mode is best for their client applications. Their active, text-prompted use cases tend to be favored in high-volume applications where it is desirable to keep things quick and easy for end users, and keep IVR handle time low. Again, performance information related to this offering did not appear publically available at the time of this article's publication.

2.9 Analysis of passphrase types

The most popular passphrase types are active-open and passive, with active-open currently dominating. The active-open approach prompts the user for a different passphrase from session to session for enhanced security. The passphrases typically consist of numbers, dates, or other common words extensively analyzed by the speech community over many years to improve automatic recognition over all users in speech dialogue applications.

As the technology has improved in recent years, however, an increasing number of companies are becoming interested in passive speech collection. This approach allows client applications to take passive speech – such as a conversation between a caller and a customer service representative – and send as much speech as practical (usually several minutes) to the service platform to build a rich phonetic model.[23]

The common identical-for-all-user passphrase, an active-open type where the user is simply prompted to say the common passphrase, is rarely if ever used today because it can be easily compromised by an attacker. Nevertheless, it is ideal for experimental studies as explained in the next section.

3 Common passphrase testing approach

Choosing a passphrase type for an experimental study in the speaker verification arena is not easy. It might be realistic to use a variety of passphrases as in the commercial systems, either an active-open approach where a different phrase is used for each session or an active-closed approach where each user chooses a secret passphrase. However, using a common passphrase identical for all users is ideal for system testing purposes for the following reasons.

A common passphrase greatly simplifies data collection and biometric system evaluation. This approach facilitates testing for imposters since the speech samples obtained from non-authentic users can be employed as "zero-effort" imposter samples. In contrast, multiple passphrases would require imposter samples for each passphrase and corresponding multiple system evaluations to obtain, for example, ROC curves for each passphrase. It also simplifies segmenting the passphrase utterance into its words, syllables, and individual phonetic sound units for extraction of important authentication information-bearing features.

Segmenting one known utterance into its smaller linguistic units is much easier than segmenting many unknown utterances into their smaller units. For example, segmenting the utterance into its phonetic units provides feature measurements at the individual sound level which has been shown to be more effective than global feature measurements alone.[8] Additionally, the combination of using the same authentication utterance for all users, and one that consists of frequently used words that are easy to pronounce, avoids many of the experimental flaws in measuring the performance of biometric systems.[14] Finally, in contrast to the exaggerated performance figures touted by the vendors, the common passphrase approach also permits the measurement of true voice authentication biometric performance.

Finally, it allows for the careful selection of the common phrase to optimize the variety of phonetic units for their authentication value. The passphrase used in this study, "My name is", was designed to have a short duration of about one second for fast authentication and to contain a reasonable variety of different sound types to characterize the individual users. This phrase contains seven phonetic units: three nasal sounds, the two [m]'s and one [n]; three vowel sounds, [aI], [eI], and [I]; and one fricative [z]. The nasal and vowel sounds characterize the user's nasal and vocal tracts, respectively, and the fricative characterizes the user's teeth and front portion of the mouth.

4 Speaker verification system

The speaker verification system developed for this study consists of speech signal processing, feature extraction, and authentication classification.

4.1 Speech signal processing

This study employed the commonly used mel frequency cepstral coefficients (MFCC).[24] The analog speech signal was converted into digital form by sampling at 44,100 Hz and passing the signal through a pre-emphasis (high pass) filter to boost the energy in the high frequencies. The digital time signal was then converted into a sequence of spectral frames. This was performed using the fast Fourier transform (FFT) operating on sequences of 1,024 digitized amplitude measurements to obtain 23 msec windowed spectral frames with hamming windows shifted by 10 msec.

The FFT frequencies were then warped onto the mel scale to obtain 13 frequency bands modeling the frequency response of the human hearing system, and finally the cepstral components of the windowed frames were computed.

For each utterance the system isolated the sequence of spectral frames corresponding to the phrase "My name is" by identifying the endpoints – the starting point of the speech from background noise and the high energy fricative [z] in the word "is".

The spectral frames corresponding to the seven phonetic sounds in the phrase – [m], [aI], [n], [aI], [m], [I], [z] – were automatically determined by using the dynamic time warping (DTW) algorithm to align (warp) each phrase against manually segmented phrases as shown in Figure 1.

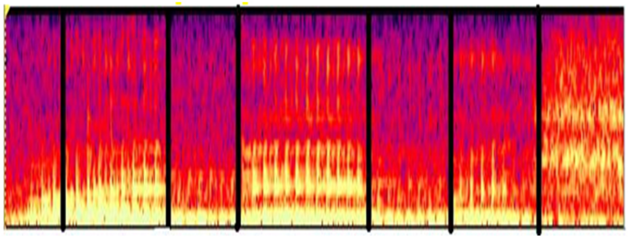

Figure 1: "My name is" segmented into its seven sounds

4.2 Feature extraction

Features were extracted from the sequence of spectral frames of the common authentication phrase, "My name is". There were a total of 227 feature measurements obtained from the phrase – 29 at the phrase level, 9 at the word level (also syllable level in this case since each of the three words consisted of a syllable), and 189 at the individual sound level. The 29 phrase-level features were:

- mean energy in the 13 frequency bands (13)
- energy variance in the 13 freq. bands (13)
- total energy in the phrase (1)
- length of the phrase in time samples (1)
- leverage phrase fundamental frequency (F0) from the three vowels (1)

The features obtained from the words (syllables) relate primarily to speech prosody. Prosody involves the intonation and stress of speech, and the corresponding measurable features are energy and length for stress (emphasis), and fundamental frequency for intonation. The 9 word-level features were:

- relative energy in the three words (3)
- relative length of the three words (3)
- relative F0 of vowels of the three words (3)

Note that the F0's in the three vowels were chosen to represent the F0's in the words because vowels dominate syllables and because obtaining accurate estimates of F0 in the vowels was found to be easier than in the voiced consonants.

A total of 189 individual phonetic sound features were obtained, 27 from each of the seven sounds as follows:

- relative length of sound re phrase length (1)
- mean energy in each of the 13 freq bands (13)
- variance of energy in each freq band (13)

4.3 Authentication classification

The classification procedure is based on a vector-difference authentication model which transforms a multi-class problem into a two-class problem. The resulting two classes are *within-person* ("you are authenticated") and *between-person* ("you are not authenticated"). As originally developed this dichotomy model is a strong inferential statistics method found to be effective in large open biometric systems. A closed-system variation of this model was recently developed and is used in this study to provide a detailed biometric performance analysis of the experimental results.[7]

In the simulated authentication process, a claimed user's speech sample requiring authentication is first converted into a feature vector. The differences between this feature vector and all the earlier-obtained enrollment feature vectors from this user are computed. The resulting query difference vectors are then classified as within-person (authentication) or between-person (non-authentication) by comparing them to the previously computed difference vectors for the claimed user.

A k-nearest-neighbor algorithm with Euclidean distance is used to classify the unknown difference vectors, with a reference set composed of the differences between all combinations of the claimed user's enrolled vectors (within-person) and the differences between the claimed user and every other user (between-person). Thus, *differences of difference vectors are being calculated*. A leave-one-out cross fold validation (LOOCV) is used to obtain system performance. The LOOCV procedure simulates many true users trying to get authenticated and many imposters trying to get authenticated as other users. For n users each supplying m samples, $m \times n$ positive (one for each sample) and $m \times n \times (n-1)$ negative (each sample versus the other users) tests can be performed.

4.4 Biometric system performance analysis

Receiver operating characteristic (ROC) curves characterize the performance of a biometric system and show the tradeoff between the False Accept Rate (FAR) and the False Reject Rate (FRR). In this study, the ROC curves were obtained using a linear-weighted decision procedure of the k nearest neighbors with k=21. Each neighbor is assigned a weight, from k to 1, with the closest neighbor weighted by k, the second by $k-1, \cdots$, and the farthest by 1. With k fixed, another parameter, l, is varied from 0 to $k(k+1)/2$, resulting in 232 points on the ROC curve. At each point, the query sample is accepted as within if the weighted sum is greater than or equal to l and between otherwise. The error rates are then calculated as $FAR = FP/(FP+CN)$ and $FRR = FN/(FN+CP)$, where FP = # false positives, FN = # false negatives, CP = # correct positives, and CN = # correct negatives. The equal error rate (EER), where $FAR = FRR$, is used as a single measure of performance.

Because the mean population performance does not give the complete picture of a biometric system, the varied performance over the population of users was analyzed and described using the animal designations of the biometric zoo: sheep (easy to verify), goats (difficult to verify), lambs (easy to imitate), and wolves (good at imitating).[25]

The Fisher scores of each of the feature measurements were also computed to provide an independent measure of the value of each feature. Because of the interdependencies of the various features, however, the Fisher scores provide only a rough indication of the feature values.

5 Experiments

5.1 Data collection

Using the built-in microphone on standard Dell laptop computers, voice samples were recorded in sets of five samples per person per day with a minimum of two-day intervals between recordings. Voice samples were obtained from 33 adult participants, 13 females and 20 males, ranging in age from 19 to 34 (average 29). From these participants, 22 provided 20 samples each, 5 provided 10 samples each, and 6 provided 5 samples each.

5.2 Experimental results

The primary experiments obtained performance as a function of the number of samples per user and the user population size (see Table 2). Because the participants of experiments A and B provided different numbers of samples (e.g., in experiment A some provided 5, some 10, and some 20), the EER for experiments A and B were obtained by averaging the results of three runs of randomly chosen five and ten samples per participant. As anticipated, performance increased (EER decreased) as the population decreased and the number of training (enrollment) samples increased.

Table 2: Primary experimental results

Exp	Number of Participants	Number of Samples Each	EER(%)	Perf. (%)
A	33	5	1.82	98.18
B	27	10	0.74	99.26
C	22	20	0.16	99.84

Figure 2 shows the ROC curves for experiments A, B, and C. Figures 3-5 show the histograms over the user population of FRR, FAR of attack receivers, and FAR of attackers for experiments A, B, and C, respectively. The histograms for A and B are the cumulative result of all 3 runs in each case. There were no significant goats (difficult to verify), lambs (easy to imitate), or wolves (good attackers) for the participants in experiment C. However, among the participants of experiment A and to a lesser extent experiment B, there are indications of several goats, lambs, and wolves.

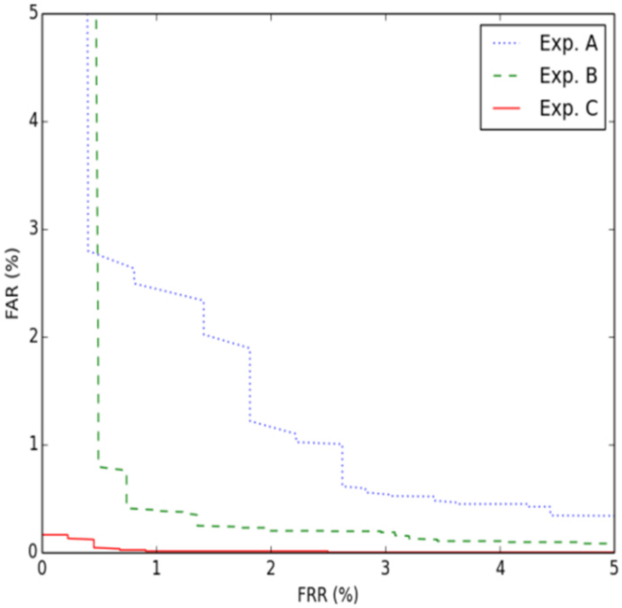

Figure 2: ROC curves for experiments A, B, and C

Each of the sub-experiments involved positive and negative authentication tests – the number of positive tests = *number-of-samples* and the number of negative tests = *number-of-*

samples times (n-1). For example, for Experiment C, the 440 speech samples allowed for the evaluation of 440 positive and 9,240 (440 × 21) negative tests. The negative tests were zero-effort imitations by other participants in the database.

Further experiments were conducted to analyze the performance contributions of the various feature subsets. These experiments were performed on the 22 participant data having the best performance and the most samples (Experiment C). Table 3 shows the performance of the three feature subset types at the phrase level, word level, and sound level. Drilling deeper into the sound-level features, Table 4 shows the performance of the three sound-level feature subsets.

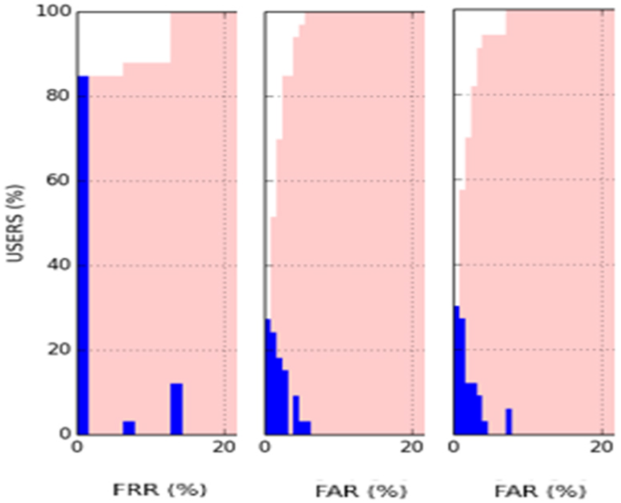

Figure 3: Experiment A: histograms of FRR (left), FAR of attack receivers (middle), and FAR of attackers (right)

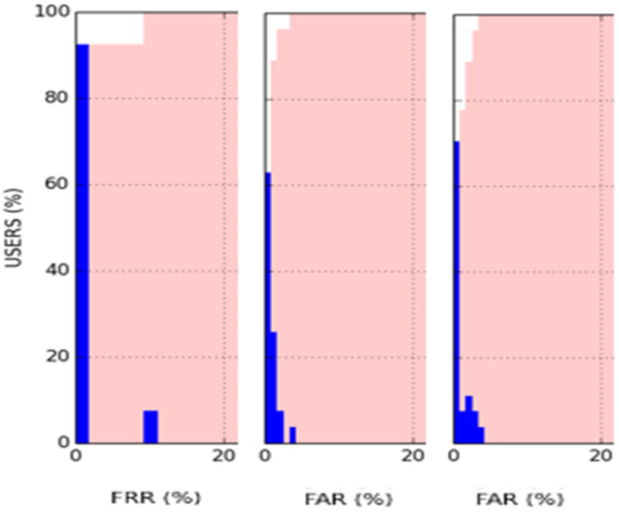

Figure 4: Experiment B: histograms of FRR (left), FAR of attack receivers (middle), and FAR of attackers (right)

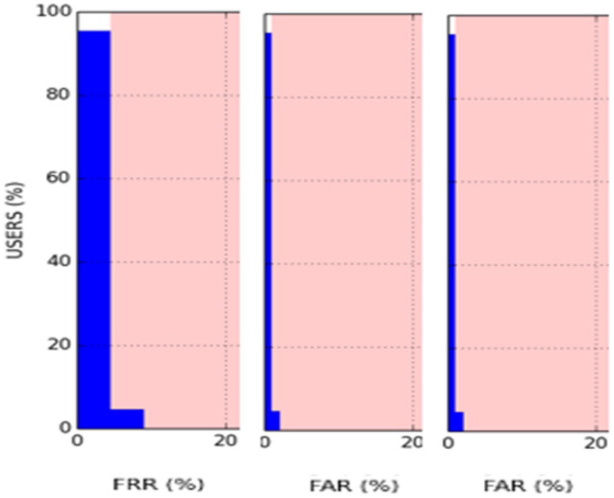

Figure 5: Experiment C: histograms of FRR (left), FAR of attack receivers (middle), and FAR of attackers (right)

Table 3: Performance by feature set

Feature Level Set	Number Features	EER(%)
Phrase	29	0.92
Word	9	22.10
Sound	189	0.21

Table 4: Performance by feature set

Sound Features	Number Features	EER(%)
Means	91	0.45
Variances	91	4.12
Lengths	7	22.18

Another way of examining the contribution of a feature set is by measuring the system performance of all the features minus that set – that is, a subtractive rather than an additive method. Table 5 shows the performance of the various phrase, word, and sound-level feature subsets relative to the all-feature baseline.

As anticipated, the sound-level energy means were the highest contributing feature subset since the change was the greatest when omitting them. Most interesting, however, was the negative contribution change for the energy, length, and F0 measurements at the phrase and word levels. This means that these features added no discriminative value, in fact a negative one, and were essentially contributing noise. On analysis of the data from which these features were derived we found a number of outliers that were not handled properly, but too late to correct the analysis for this paper. Nevertheless, it is interesting that this method of system performance analysis discovered these errors.

Omitting the poor features found above yielded an EER of 0.097% or 99.90% performance, which is a little better than the 99.84% performance shown in Table 2. However, since this increased performance was obtained by analyzing and adjusting the results, it would need to be verified on new data.

The remaining experiments investigate ways of analyzing the biometric value of individual features. The fisher scores of the individual feature measurements and system performance (1-EER) on single features were obtained to provide a rough estimate the value of single features, and these values on the top ten fisher scores are listed in Table 6, showing reasonable correlation between the Fisher score and system performance as anticipated.

Table 5: Feature set contributions relative to baseline. Negative contributions are highlighted.

Feature Subset via Subtraction from Baseline	Number Features	EER(%)	ΔEER(%)
All Features (Baseline)	227	0.16	0.00
All-Phrase energy means (13)	214	0.23	0.07
All-Phrase energy variances (13)	214	0.22	0.06
All-Phrase total energy (1)	226	0.15	-0.01
All-Phrase length (1)	226	0.15	-0.01
All-Phrase average F0 (1)	226	0.12	-0.04
All-Word relative energies (3)	224	0.16	0.00
All-Word relative lengths (3)	224	0.14	-0.02
All-Word relative F0(3)	224	0.10	-0.06
All-Sound relative lengths (7)	220	0.22	0.06
All-Sound energy means (91)	136	0.91	0.75
All-Sound energy variances (91)	136	0.23	0.07

Table 6: Top-ten Fisher-score features with corresponding individual feature performance by the system. The top three valued features by each of the two methods are highlighted.

Fisher Rank	Feature Measurement	Fisher Score	System Perf.
1	Phrase-Mean-Freq Band 5	13.4	64.7
2	Phrase-Mean-Freq Band 1	10.8	62.2
3	[n]-Mean-Freq Band 5	10.1	59.0
4	[i]-Mean-Freq Band 5	10.0	61.1
5	Phrase-Mean-Freq Band 9	8.9	58.3
6	[ai]-Mean-Freq Band 5	7.8	59.0
7	[z]-Mean-Freq Band 5	7.4	60.7
8	[ei]-Mean-Freq Band 5	7.4	59.3
9	[z]-Mean-Freq Band 1	7.1	61.4
10	[ai]-Mean-Freq Band 9	6.9	55.8

6 Conclusions

This study developed a speaker verification system using state-of-the-art speech signal processing, standard feature extraction methods, and a unique backend classification system to achieve a performance of 99.8% on 22 participants. The main contribution of the study, however, was to illustrate a variety of testing techniques that can be used to efficiently analyze the performance of a speaker verification system, while limiting exposure to potential experimental flaws and permitting the measurement of true biometric performance.

Of the major feature sets – phrase, word, and sound level – the sound level features appeared to provide the most discriminative value, then came the phrase-level and finally the word-level features, but this order also corresponds to the number of features in each of the sets.

The Fisher scores of the features indicated the following. Since all of the top ten Fisher-score features involved frequency band mean energies, this verifies the importance of the frequency band mean energies as shown above. Four of the ten Fisher-score features involved frequency band 1 (supported by two of the top three system performances among the ten), indicating that band 1 may contain more biometric value than any of the other 12 bands, which makes sense since the first cepstral band corresponds to overall energy. Three of the top ten (and the top two) involved phrase-level features. Of the seven individual sounds in the phrase (three vowels: [aI], [eI], and [I]; three nasals: two [m]'s and one [n]; and one fricative: [z]), the top ten Fisher-score features involved the fricative [z] twice, the vowels three times, and a nasal once.

In future work the following greedy algorithm might be explored using the subtraction method of section 5.2 to eliminate poor features and obtain a near optimal subset of features:

Greedy Weak Feature Elimination Algorithm

```
(1) Calculate baseline EER for full set of n features
(2) j = n
(3) Calculate j EER's leaving out each feature
(4) Select the subset of j-1 features that yields the lowest EER value (greedy choice) of those lower than the EER of the previous best set of features
(5) j = j - 1
(6) Go to step 2 and repeat until eliminating any single feature does not decrease the EER
```

This algorithm should eliminate most of the non-contributing features from the original feature set. With the number of features and data employed in this study, the running time of this algorithm was too long to be completed for this paper submission. A variation of this algorithm – evaluate all subsets having a lower EER in step 4 – could produce an even better final subset because more combinations could be explored but with longer running time. Neither of these algorithms is optimal because there could be a reduction in EER by eliminating a pair, triplet, etc. of features at one time rather than just one.

Acknowledgements

We thank the student teams in the Master's projects course that contributed to this effort over recent years.

References

[1] Larcher A., Lee, K. A., Ma, B., and Li H. Text-dependent speaker verification: Classifiers, databases and RSR. 2015. Speech Communication. 2014.

[2] Luan, J., Hao, J. Method and Apparatus for Estimating Discriminating Ability of a Speech Method and Apparatus for Enrollment and evaluation of speaker Authentication. United States Patent Application Publication. 2007; 0124145.

[3] Martin, A.F., Greenberg, C.S. NIST 2008 speaker recognition evaluation: performance across telephone and room microphone channels, Annual Conference of the International Speech Communication Association (Interspeech). 2009; 2579–2582.

[4] Meisel, W. Speech in the User Interface: Lessons from Experience, Trafford Publishing.

[5] The University of Edinburgh. Evaluation of Nuance v7.0.4 Speaker Verification Performance on the Dialogues Spotlight UK English Database. The Centre for Communication Interface Research. Dialogues Spotlight Consortium. 2001.

[6] Maxion, R.A. Making Experiments Dependable. The Next Wave/NSA Magazine, Vol. 19, No. 1, pp. 13-22. Reprinted from Dependable and Historic Computing, LNCS 6875, 2011; 344-357.

[7] Monaco, J.V., Bakelman, N., Cha, S.-H., and Tappert, C.C. Recent advances in the development of a long-text-input keystroke biometric authentication system for arbitrary text input. Proc. Euro, Intelligence and Security Informatics Conf. (EISIC), Sweden. 2013.

[8] Gu, Y., Thomas, T. An implementation and evaluation of an on-line speaker verification system for field trials. Annual Conference of the International Speech Communication Association (Interspeech). 1998; 125–128.

[9] Wagner, M., Summerfield, C., Dunstone, T., Summerfield, R., Moss, J. An evaluation of commercial off-the-shelf speaker verification systems. In: Odyssey Speaker and Language Recognition Workshop. 2006; 1–8.

[10] Reynolds, D., Automatic Speaker Recognition Using Gaussian Mixture Speaker Models, Volume B, Number 2, The Lincoln Laboratory Journals 5. 1995.

[11] Shriberg, E., Ferrera, L., Kajarekara, S., Venkataramana, A., Stolcke, A. Modeling prosodic feature sequences for speaker recognition. Speech Communication. 2005; 46(3-4): 455-472.

[12] Kato, T., Shimizu, T. Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP. 2003; 57–60.

[13] Hebert, M., Heck, L. P. Phonetic class-based speaker verification, EUROSPEECH-2003, 2003; 1665-1668.

[14] American Safety Council. BioValidation. Available from: `http://biovalidation.com/`. Accessed on September 11, 2013.

[15] Authentify. Available from: `http://www.authentify.com/`. Accessed on September 11, 2013.

[16] Ben-Asher, N., Kirschnick, N., Sieger, N., Meyer, J., Ben-Oved, A. and Möller, S. On the need for different security methods on mobile phones. Proc. 13th Int. Conf. Human Computer Interaction with Mobile Devices and Services. Mobile HCI '11. New York, NY. 2011; 465-473.

[17] Nuance Communications, Inc. Available from: `www.nuance.com`. Accessed on September 11, 2013.

[18] Opus Research. Nuance Communications Named the Global Voice Biometrics Leader. Available from: `http://www.nuance.com/company/news-room/press-releases/WEb_Nuance-Communications-Named-the-Global-Voice-Biometrics-Leader.docx`. Automatic Speaker Recognition Using Gaussian Mixture Speaker Models. Accessed on September, 11, 2013.

[19] The Editors of Speech Technology. Speech Technology. Speech Technology Media, a division of Information Today, Inc., July 2012. Available from: `http://www.speechtechmag.com/Articles/?ArticleID=83629&PageNum=2`. Accessed on September 11, 2013.

[20] Voice Biometrics Group. Available from: `http://www.voicebiogroup.com`. Accessed on September 11, 2013.

[21] Elliot, S. and Rolfe, A. Case Study Phone-based Voice Biometrics for Remote Authentication, Authentify Inc. 2010; ASEC-10.

[22] VoiceVault Inc. Online: http://www.voicevault.com. Accessed on September 11, 2013.

[23] Jun, L., He, Z. Spectral Subtraction Speech Enhancement Technology Based on Fast Noise Estimation, ICIEC. 2001.

[24] Jurafsky, D. and Martin, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd Edition, Prentice Hall. 2008.

[25] Di Crescenzo, G., Cochinwala, M., and Shim, H. S. Modeling cryptographic properties of voice and voice-based entity authentication, In Proceedings of the 2007 ACM workshop on Digital identity management. DIM '07. ACM. New York, NY. 2011; 53-61.