ANL/ED/CP-99102

Speaker Recognition Through NLP and CWT Modeling

by

S. Alenka Brown-VanHoozer
Argonne National Laboratory - West
Engineering Division
P. O. Box 2528
Idaho Falls, ID 83403-2528

15th Annual NDIA Security Technology
Symposium & Exhibition

Norfolk, VA
June 14-17, 1999

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

SPEAKER RECOGNITION THROUGH NLP AND CWT MODELING

Stephen W. Kercel and Raymond W. Tucker
Oak Ridge National Laboratory
PO Box 2008, MS 6011
Oak Ridge, Tennessee 37831-6011
e-mail: rt4@ornl.gov
phone : (423) 574-5278 and 576-0947

Alenka Brown-VanHoozer
Argonne National Laboratory-West
PO Box 2528, MS 6000
Idaho Falls, Idaho 83403
e-mail: alenka@anl.gov
phone: 208-533-7926

ABSTRACT

The objective of this research is to develop a system capable of identifying speakers on wiretaps from a large database (>500 speakers) with a short search time duration (<30 seconds), and with better than 90% accuracy. Much previous research in speaker recognition has led to algorithms that produced encouraging preliminary results, but were overwhelmed when applied to populations of more than a dozen or so different speakers. The authors are investigating a solution to the "large population" problem by seeking two completely different kinds of characterizing features. These features are extracted using the techniques of Neuro-Linguistic Programming (NLP) and the continuous wavelet transform (CWT).

NLP extracts precise neurological, verbal and non-verbal information, and assimilates the information into useful patterns. These patterns are based on specific cues demonstrated by each individual, and provide ways of determining congruency between verbal and non-verbal cues. The primary NLP modalities are characterized through word spotting (or verbal predicates cues, e.g., see, sound, feel, etc.) while the secondary modalities would be characterized through the speech transcription used by the individual. This has the practical effect of reducing the size of the search space, and greatly speeding up the process of identifying an unknown speaker.

The wavelet-based line of investigation concentrates on using vowel phonemes and non-verbal cues, such as tempo. The rationale for concentrating on vowels is there are a limited number of vowels phonemes, and at least one of them usually appears in even the shortest of speech segments. Using the fast, CWT algorithm, the details of both the formant frequency and the glottal excitation characteristics can be easily extracted from voice waveforms. The differences in the glottal excitation waveforms as well as the formant frequency are evident in the CWT output. More significantly, the CWT reveals significant detail of the glottal excitation waveform.

1. INTRODUCTION

Wiretaps are an invaluable tool in conducting investigations into criminal operations at all levels of law enforcement, but especially in countering drug smuggling operations. It is only their irreplaceable investigatory value that justifies the heavy burden, both in terms of manpower and record keeping, that wiretaps impose on law enforcement. Even a small investigation may involve dozens of lines and scores of individuals.

The objective of this research is to develop a system capable of identifying speakers on wiretaps from a large data base (500+ individuals) with a short search time duration, and with greater than 90% probability of correct identification and less than 10% probability of misidentification. This is a problem that has been declared "solved" many times in the past, but only for distinguishing a few voices in a database of a few dozen. However, the real problem is to devise a method that reliable recognizes a speaker in a database of 500 and more.

Conventional automatic speaker-recognition systems often rely heavily on standard Fourier analysis to extract the frequency-domain characteristics (spectrum) of the speaker's voice. This presents two difficulties. First, it assumes that the signal is mathematically stationary, when in reality it is not.[1] Using a model whose behavior is fundamentally different from that of the underlying physical process, guarantees the introduction of predictive error. Second, it ignores other identifying cues that might be present in the signal.

Other efforts have focused on developing sophisticated Gaussian-based functions (called "Gaussian mixture models") for statistically modeling the spectral characteristics of the speaker's voice as revealed using one-dimensional Fourier analysis. In this approach, maximum-likelihood parameter estimation techniques are used to form a speaker model using a weighted sum of Gaussian functions to capture the variance within a population of speakers. This method has met with reasonable success for relatively large speaker populations. However, it lacks the flexibility to be consistently reliable over a wide range of environments and speakers.[2]

There are three other strategies that have met with varying degrees of success. These are, cepstral methods, autocorrelation methods, and wavelet-based methods. Cepstral methods are based on the proposition that channel distortion is multiplicative in some linear transform space, and that taking the cepstrum of a channel-distorted signal reduces the channel distortion to an additive constant that can be conveniently subtracted from distorted signal leaving the distinguishing features in the residual.[2] Autocorrelation methods attempt to look directly for self-similarities in the signal. Unfortunately, both assume stationarity, and both are vulnerable to noise. Autocorrelation works reasonably well for high-pitched speakers. Cepstral analysis works reasonable well for low-pitched speakers. However, neither method is suitable for a wide range of pitches. Even worse, both methods are critically dependent on the choice of segment length, with no objective means being provided for guessing in advance what the segment length should be.[1] (In practical terms, these methods only work if the user already knows the answer.)

Wavelet methods overcome many of these limitations.[1] They are usable over a wide range of pitch. They are robust to noise. And they are not critically dependent on segment length. Early attempts with wavelet methods used feature selection criteria that were somewhat arbitrary, making their use in practice something of an art. Subsequent work has sought to use adaptive wavelets to make this process more objective and automatic.[3] A major limitation is that the octave-resolution of the dyadic wavelet lacks the flexibility to provide a truly powerful and robust identification system.

The research currently under way at ORNL and ANL seeks to solve the problem of dealing with a large speaker population by combining two strategies. The first is to use a more flexible and powerful adaptive wavelet technique than those attempted previously by using the fast continuous wavelet transform (CWT).[4] This overcomes the octave-resolution limitation of the dyadic wavelet, and reduces computational cost by allowing a rapid "zoom-in" to the useful ranges of time, scale, and basis function. The second is to explore a previously ignored set of features in speech data, the cues to the primary representational system operating in the brain of the speaker.

This investigation is being performed on the TIMIT and NTIMIT databases, which consist of 630 speakers uttering 10 phrases each..[5,6,7] The NTIMIT database was formed by transmitting and recording the TIMIT database over a set of commercial telephone lines to introduce the effects of a typical communication channel to the voice signals. Using these standard data sets will also allow a direct comparison of the achieved recognition rates with the published results by researchers using conventional methods.

2. FEATURES IN CWT SPACE

A novel, but proven, method of computing the fast, continuous-wavelet transform will be used as the basis of the processing engine for this speaker recognition system. The advantage of using wavelet analysis (as opposed to conventional methods) is that the wavelet transform is well suited for detecting transient events, such as the start or end of a particular segment of pitch, and non-stationary events, such as a "pitch" event that really turns out to be a chirp.

The fast, continuous-wavelet algorithm is a result of several years of experience in applying wavelet methods to real-world problems. In contrast to discrete and dyadic wavelet transforms, whose scaling properties are less flexible, it offers the advantage of much higher resolution, where needed, in both time and scale. The continuous wavelet algorithm also offers the opportunity to tailor the choice of the wavelet function used to match those characteristics that distinguish individual speakers. Figure 1 shows a comparison of the continuous wavelet- and the discrete wavelet-transform representation of the same voice segment. As is evident in Figure 1, the continuous wavelet representation (top) reveals significantly more detail than the dyadic wavelet representation (bottom.)

[redacted]

Figure 1. Continuous vs. discrete wavelet representation of a voice segment.

The rationale for concentrating on vowels is that they are formed with the entire vocal tract, and thus should maximize the speaker-dependent features of the voice. Based on preliminary analysis of voice signals using the fast, continuous wavelet transform (CWT) algorithm, the details of both the formant frequency and the glottal excitation characteristics can be easily extracted from voice waveforms. The differences in the glottal excitation waveforms as well as the formant frequency are evident in the CWT output. More significantly, the CWT reveals significant detail of the glottal excitation waveform in the time-frequency display.

3. FEATURES IN NLP SPACE

"Every individual channels information differently based on our preference to the sensory modality of representational system (visual auditory or kinesthetic) we tend to favor most (...our primary representational system (PRS)). Therefore some of us access and store our information primarily visually first, some auditorily and others kinesthetically (through feel and touch), which in turn establishes our information processing patterns and strategies and external to internal (and subsequently vice versa) experiential language representation."[8] Identifying these primary states is the basis for using neuro-linguistic programming (NLP) techniques.

It is possible to use the techniques of NLP to extract specific patterns based on integral cues demonstrated by each individual. The mechanisms provided by NLP allow for extracting neurological, verbal and non-verbal information from an individual forming the basis in which to categorize individuals into one of the three primary modalities or systems. The collected information can then be assimilated into useful patterns; thus, allowing further categorization of the primary modalities into *visual detail, visual-general, auditory-tonal, auditory-digital, kinesthetic-tactile and kinesthetic-emotive* (olfactory and gustatory fall under the kinesthetic modality). For example, kinesthetically oriented individuals respond with a much slower voice tempo that may contain long pauses between words or sentences and often have a low, deep, and breathy tonality to their voice.[9]

The verbal and non-verbal cues of an individual's speech are distinguishing features that can be used for identification. For example, an individual whose predicates consist of such words as, *see, show, look, view*, etc., and whose tone is nasal and high pitched is visual oriented. These modalities are difficult to disguise since they are generated at the unconscious level. Word spotting (verbal predicates, e.g., see, sound, feel, etc.), speech transcription (non-verbal cues, e.g., breathing, spacing between words, pitch, etc.) and neurological strategies (eye accessing movements) provide the sensory based acuity required for this type of categorization.

NLP is a methodology that enables the collection of precise neurological, verbal and non-verbal information, abstraction of the information into useful patterns, and the use of the patterns for explicit outcomes, such as identification. These patterns are based on specific cues demonstrated by each individual, thus providing a great deal of information involving word spotting, speech transcription, congruency in verbal and non-verbal cues, accessing, processing and storing of information, etc.

NLP allows the investigators to first categorize individuals into three primary modalities or systems, visual, auditory and kinesthetic; then further categorize the primary modalities to visual-detail, visual-general, auditory-tonal, auditory-digital, feel and touch, olfactory and gustatory. The primary modalities would be accomplished through word spotting (or verbal predicates cues, e.g., see, sound, feel, etc.) while the latter would be characterized through the speech transcription used by the individual. This allows for the possibility of classifying the speakers in a database hierarchically. At the higher level they would be classified by NLP modality. At the lower level, they would be identified by a feature vector unique to each speaker. This would have the practical effect of reducing the size of the search space, and overcoming one of the greatest vulnerabilities of conventional speaker identification schemes..

The secondary NLP modalities can then be used to correlate the individual's non-verbal cues, e.g., breathing, tempo and tonality with that of the verbal cues extracted by the wavelet analysis to generate the feature vector. For example, breathing changes are different for each of the primary systems. Individuals that are auditory, would have an even breathing with a somewhat prolonged exhale in their responses, whereas, the kinesthetics would have deep, full breaths, and visual would breathe more quickly and shallow.

3.1 Verbal Cues

When communicating with others, people use specific words known as predicates to organize and make sense of their experience. These predicates can define a representation system by the words or phrases used by an individual.

a) *Look how high, see, observe, point of view, size, shapes, colors, distance*, etc., are characteristic of the words or predicates used for visual processing.

b) *Sounds rather loud, tone, click, hum-m-m, bang, tap of a pencil*, etc. are characteristic of the predicates used for auditory processing.

c) *Feels soft to the touch, laugh, grasp, handle, smooth, sour, smelly*, etc. are characteristic of the predicates used for kinesthetic processing.

Therefore, predicates paired with either of the other two modalities, (neurological or physiological cues) provides a means by which to identify the PRS of an individual.

3.2 Physiological (Non-Verbal) Cues

Breathing Changes.

"Breathing is one of the most profound and direct ways we have of changing or tuning our chemical and biological state to affect our neurology"[9] Associated with each of the modalities are the following characteristics.[9,10]

a) Shallow, quick breathing indicates visual processing.

b) Even or level breathing, including a sustained exhale indicates auditory processing.

c) Deep, full breathing indicates kinesthetic processing.

Tonal and Tempo Changes.

"Changes in voice tempo and tonality follow changes in breathing patterns. The amount of air, and the rapidity with which it pushed over one's vocal chords, will cause noticeable changes in voice quality."[9] Associated with each of the modalities are the following characteristics.[9,10]

a) Quick and choppy bursts of words in a high pitched, nasal and/or strained tonality with a typically fast tempo of speech indicates visual processing.

b) A clear, midrange tonality of words in an even, rhythmic tempo indicates auditory processing. Typically well-enunciated words will accompany the activity.

c) A slow voice tempo with long pauses and low, deep and often breathy tonality indicates kinesthetic processing.

Spacing

"Changes in voice tempo and tonality follow changes in breathing patterns. The amount of air, and the rapidity with which it pushed over one's vocal chords, will cause noticeable changes in voice quality."[9] Associated with each of the modalities are the following characteristics:

a) Short spacing between words indicates visual processing.
b) More even spacing between words indicates auditory processing.
c) Large spacing between words (versus visuals and kinesthetics) indicates kinesthetic processing.

These physiological non-verbal cues had been determined qualitatively by Bandler *et al*, over a course of approximately six years (1974-1982), and have been studied and applied over the past 26 years with consistency and accuracy, e.g., visuals have a faster tempo than kinesthetics and breathe higher in the chest. The cues provide the means to determine the PRS from a core sample set from the TIMIT database, and are instrumental in quantifying the data that are the subject of this investigation.

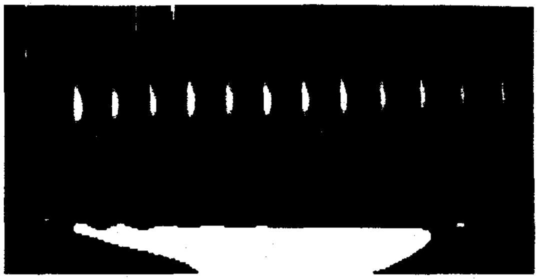
4. WORK IN PROGRESS

4.1 CWT Results

Pitch determination

A key speaker-dependent feature of interest is the fundamental formant frequency, or "pitch," of the speaker's voice. The pitch of the speaker's voice corresponds to the rate at which the glottis opens and closes as air is forced through the larynx during voiced speech. Previous work has shown that the pitch of a speaker's voice can be reliably estimated using the discrete wavelet transform with dyadic scale changes (D_yWT).[3] In the D_yWT, the wavelet scales take the form $a = 2^i$ for some practical range of i. Using this method, the discrete wavelet transform of the voice signal is computed for three scales that are chosen to cover the expected range of pitches for human speech. The pitch period is estimated by locating periodic peaks that appear across the three wavelet scales.

The CWT algorithm produces an approximation to a continuous change in wavelet scale as opposed to the power of two changes usually employed in the D_yWT. The practical effect of this difference is that with the proper choice of the mother wavelet, the pitch period can be directly observed and thus extracted from the resulting time-frequency CWT representation. Our preliminary results indicate that a Gauss-Hermite wavelet with an order between 2 and 5 works best for revealing the pitch information for vowel sounds. Figure 2 shows the amplitude of the time-frequency representation of a male and female speaker uttering the vowel sound /aa/. The results have been amplitude scaled and displayed in image format. The pitch period is readily evident as the distinct "blobs" that represent the release of acoustic energy as the glottis is forced open and then shuts. To extract the pitch period from the time-frequency representation, it is only necessary to locate the centroid of each blob and then determine the corresponding frequency.



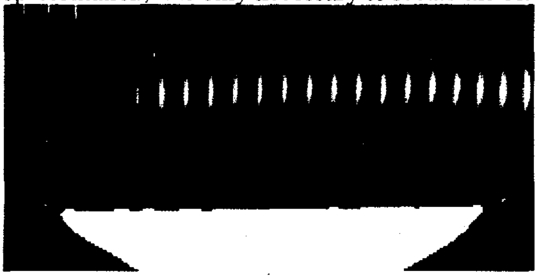


Figure 2. CWT of Female (Left) and Male (Right) Voices Uttering /aa/ (TIMIT Database)

Features in addition to pitch have been extracted from the time-frequency representation produced by the CWT. Although shown as images in Figure rwt1, the CWT algorithm produces a two-dimensional (2-D) surface as shown in Figure 3. As is evident in Figure 3, each of the blobs shown in Figure 2 is actually a peak in the 2-D time-frequency plane. More advanced, time-dependent features, such as the glottal opening and closing rate, the trajectory of the blob centroid in the time-frequency plane, and the evolution of the blob shape can be extracted from the time-frequency representation produced by the CWT.

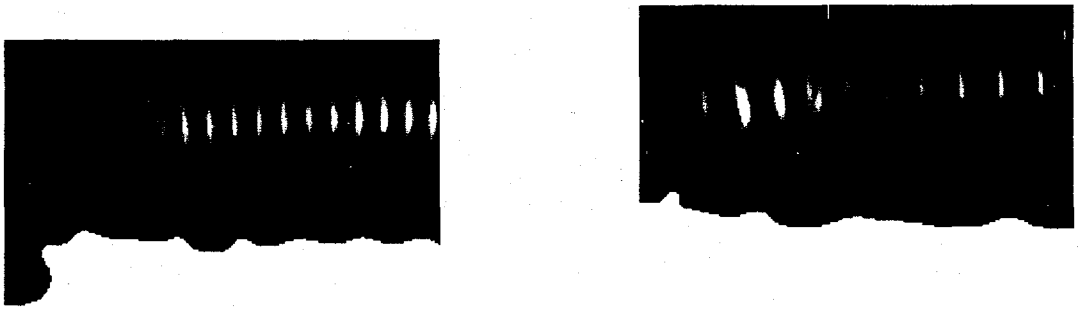


Figure 3. CWT of Female (Left) and Male (Right) Voices Uttering /aa/ (NTIMIT Database)

Impact of Telephone/Wiretap Communications

An important question for this research is whether speaker-dependent features can still be reliably extracted from voice signals obtained via wiretaps. In particular, the pitch of a speaker's voice typically lies in the range of frequencies that are attenuated or eliminated totally by telephone transmission, and thus may be difficult to determine from a wiretap signal. To determine if pitch can be used as a reliable speaker-dependent feature under these conditions, the analysis described above has been repeated with the corresponding signals from the NTIMIT database.[7] The NTIMIT speech database was generated by transmitting the TIMIT utterances over a variety of actual telephone lines. Since it contains the same speakers and phrases as the TIMIT database, a direct observation of the effect of the telephone transmission line can be made. Figure 3 shows the time-frequency representation of the voice segments from the NTIMIT database that approximately correspond to the segments shown in Figure 2. The pitch period can still be observed in the spacing of the distinct "blobs" despite the slight deformation of the blobs due to transmission distortion.

4.2 NLP Results

Study

The sentence, "*She had your dark suit in greasy wash water all year,*" was used to compare the speech patterns from three regions dr1 (New England); dr7 (New York City) and dr8 (Army Brat), and establish the PRS of the speakers. The software application, *Sound Sculpture II,* was used to display the visual and sound characteristics associated with each speech pattern. The physiological cues were easily identified, but needed to be compared against another variable, which was unavailable. Therefore, it was necessary to generate a control data set of speech patterns (where the PRS had been predetermined).

Control Data Set

A small sample size of 18 speech patterns (males and females) - seven visuals, five auditories, and six kinesthetics was used to establish qualitative parameters associated with each modality . Geographical site (dialect) was not a consideration, but mental states were, e.g., depression, anxiety, etc. The individuals read the selected sentence, *"She had your dark suit in greasy wash water all year;"* while their speech pattern was recorded using the software application, *Big Sound. Sound Sculpture II* was used to display the visual and sound characteristics associated with each speech pattern. Using qualitative measurements, the control data set was then compared against speech patterns selected from three regions of the TIMIT data test set.

Examples of the control data set are shown in Figures 4-6. The characteristics of a voice pattern of an individual whose preference is the visual representation system are shown in Figure 4. The overall length is shorter than the other two states, more uneven, words are less rounded, small spacing between words or phrases. The characteristic description of a voice pattern of an individual whose preference is the auditory representation system is shown in Figure 5. The overall length is a little longer in length than the visual state; more even and rounded words, with larger spacing between words or phrases. The characteristics of a voice pattern of an individual whose preference is the kinesthetic representation system are shown in Figure 6. Features are longer in length than the other two states; words are rounded more, more spacing between words or phrases, words more drawn out, etc.

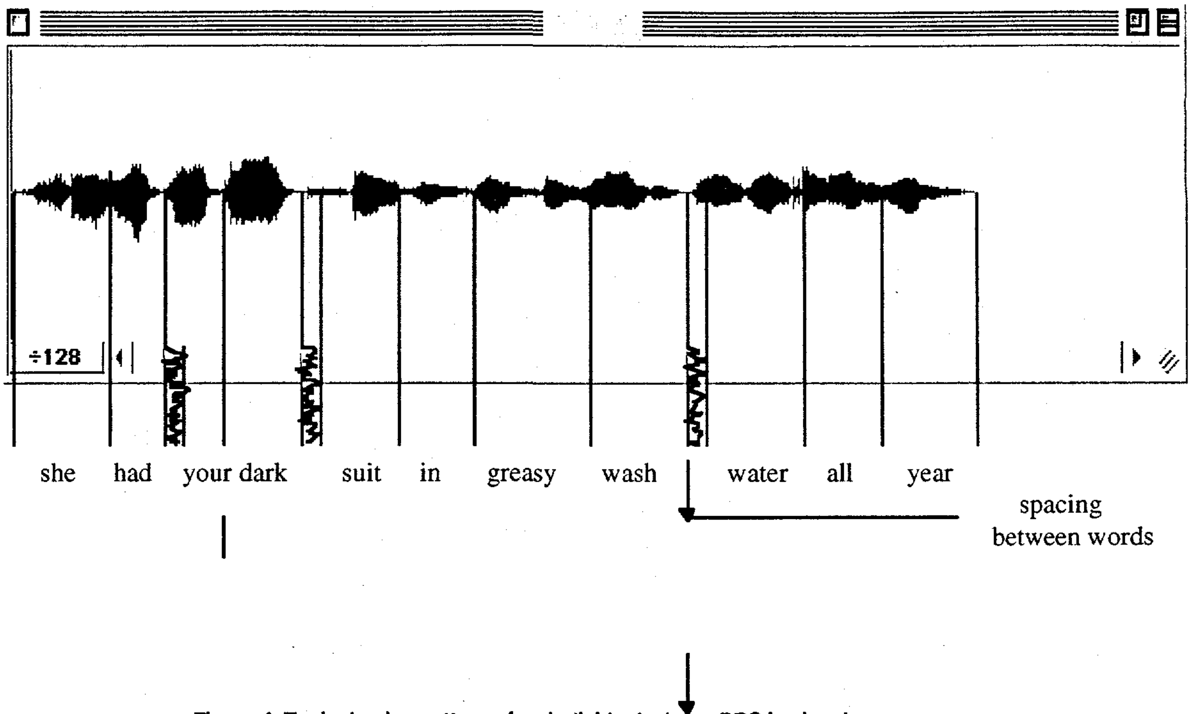


Figure 4. Typical voice pattern of an individual whose PRS is visual.
= word or phrase, quick, choppy bursts of words, etc.

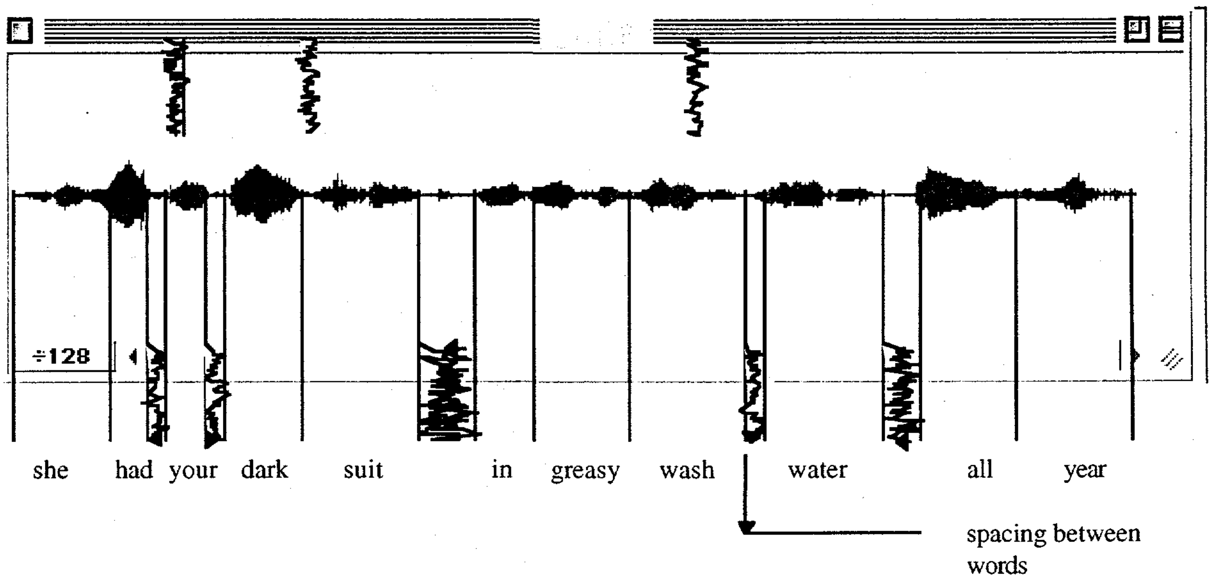


Figure 5: Typical voice pattern of an individual whose PRS is auditory

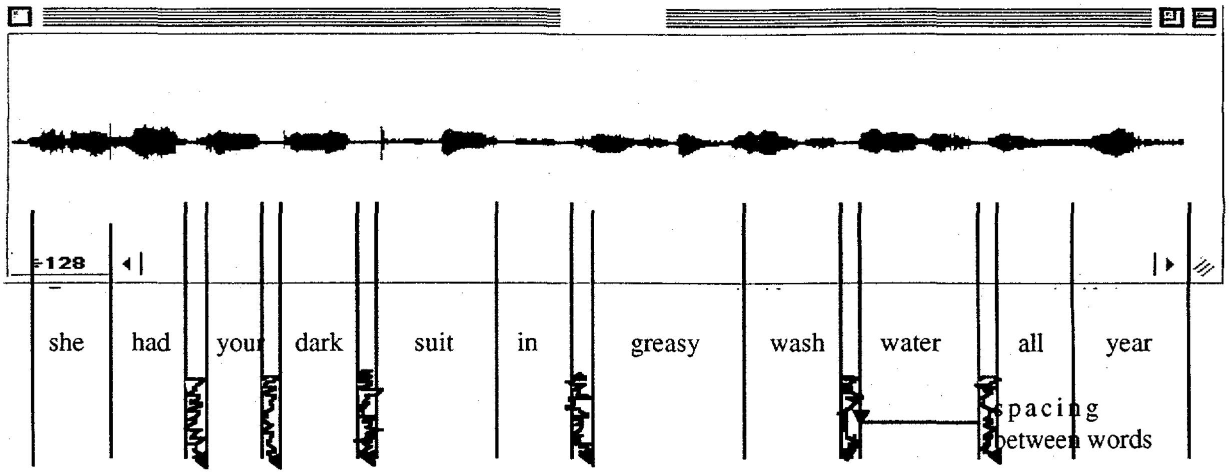


Figure 6: Typical voice pattern of an individual whose PRS is kinesthetic.

Results

Identifying the primary states from the speech patterns can be accomplished using the physiological cues of the NLP techniques. However, determining accuracy can not be established since only one set of cues is being implemented. Therefore, the control data sample is currently being analyzed using a classification algorithm. The program is comparing the characteristics of each speech pattern associated with a specific PRS. The intent is to establish quantitative measurements (or parameters) for each PRS.

Also, matching of the primary states or PRS to each speech pattern shows only the representational state used by the speaker when reading a prewritten script. The representational state may be the same as during random speech or the PRS may change during a reading task. A data test set would need to be generated consisting of speakers reading a prewritten script and random speech.

5. CONCLUSIONS

This is a preliminary report of the first few months work on a project with a scheduled duration of two years. The objective is to combine CWT and NLP techniques to devise a reliable text-independent speaker identifier for large databases. The CWT work to date has focused on extracting features such as pitch from voiced sounds in speech signals. Future investigation will be directed at extracting features from unvoiced sounds by adjusting the CWT order and wavelet scale specifications. Additional emphasis will be placed on extracting and representing time-dependent features that are unique to individual speakers.

There are two potential strategies for combining NLP and CWT techniques. One is to have NLP modality as the high level in a hierarchical database of the samples, with the CWT-derived features being a lower level descriptor of the sample. (i.e. We use NLP to provide a means to limit the search space. Our prior experience is that a large search space is a back breaker in speaker identification.) The other strategy for using NLP is to add features (extra dimensions) directly to the CWT-derived feature vector.

For the second strategy to add any value, it would be necessary to have high correlation between the speaker and the NLP features, high correlation between the speaker and the CWT features, and low (preferably no) correlation between NLP and CWT features. That would mean that the NLP features and CWT features would be orthogonal (or nearly so). Only if the added features are relatively independent, including the NLP features and CWT features in a single high-dimensional feature vector produce a classifier superior to either feature set alone.

A larger sample size of control data of speech patterns may need to be developed to verify the variables governing an established PRS of specific voice patterns. This data test set would need to consist of random speech and oral reading voice patterns. The patterns would then need to be analyzed for any characteristic similarities and differences in defining the PRS. Therefore, the data test set generated will more closely relate to that of real world application.

ACKNOWLEDGMENTS

This research is supported through the Department of Energy Special Technologies Program under STP Requirement Number STP-029-99.

REFERENCES

1. Kadambe, S., Boudreaux-Bartels, G. F., "Application of the Wavelet Transform for Pitch Detection of Speech Signal," *IEEE Transactions on Information Theory*, Vol. 38, No. 2, March 1992, pp.917-924.

2. Mammone, R.J., Zhang, X. and Ramachandran, R.P, "Robust Speaker Recognition A Feature-based Approach," *IEEE Signal Processing Magazine*, September 1996, pp. 58-71.

3. Kadambe, S. "Text Independent Speaker Identification System Based on Adaptive Wavelets," in *Wavelet Applications*, Harold H. Szu, Editor, Proc. SPIE 2242, pp. 669-677 (1994).

4. Dress, W. B., "Applications of a Fast, Continuous Wavelet Transform," in *Wavelet Applications IV*, Harold H. Szu, Editor, Proc. SPIE 3078, pp. 570-580 (12997).

5. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S, and Dahlgren, N.L, *Acoustic-Phonetic Continuous Speech Corpus CD-ROM*, NISTIR 4930, CD-ROM Released October 1990, Documentation Published February 1993.

6. Defense Advanced Research Projects Agency (DARPA) - Information Science and Technology Office - TIMIT Acoustic -Phonetic Continuous Speech Corpus, Training and Test Data, NIST Speech Disc CD1-1.1, Readme.doc., 10-12-1990.

7. Jankowski, C., "The NTIMIT Speech Database," printed documentation which accompanies the NTIMIT CD-ROM, January 1991.

8. Brown-VanHoozer, S.A. and VanHoozer, W.R. (1998). "Process vs. Content in Academic Learning." (unpublished work). E-mail: alenka@anl.gov.)

9. Bandler, R., Dilts, R., DeLozier, J., and Grinder, J. (1980). "*Neuro-Linguistic Programming: The Study of the Structure of Subjective Experience*." Vol. I., Real People Press, Moab, Utah.

10. Lewis, B.A. and Pucelik, F.R., *Magic Demystified: An Introduction to NLP*," Metamorphous Press, Lake Oswego, Oregon, (1982).