SPEAKER RECOGNITION THROUGH NLP AND CWT MODELING

Stephen W. Kercel and Raymond W. Tucker
Oak Ridge National Laboratory*
P.O. Box 2008
Oak Ridge, Tennessee 37831-6011
Phone: (423) 574-5278
Fax: (423) 574-6663
E-mail: kzo@ornl.gov and rt4@ornl.gov

S. Alenka Brown-VanHoozer
Argonne National Laboratory-West
P.O. Box 2528
Idaho Falls, Idaho 83403
Phone: (208) 533-7926
Fax: (208) 533-7863
E-mail: **alenka@anl.gov**

To be presented at the
15th Annual Security Technology Symposium
Security-Related Research and Methodology Session
Norfolk, Virginia
June 16, 1999

* This research was performed at OAK RIDGE NATIONAL LABORATORY, managed by LOCKHEED MARTIN ENERGY RESEARCH CORP. for the U.S. DEPARTMENT OF ENERGY under contract DE-AC05-96OR22464.

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: kzo@ornl.gov and 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

The objective of this research is to develop a system capable of identifying speakers on wiretaps from a large database. 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 reliably recognizes a speaker in a database of 500 and more.

Fourier-based speaker-recognition systems typically encounter two difficulties. First, they wrongly assume that the signal is mathematically stationary.[1] Using a model whose behavior is fundamentally different from that of the underlying physical process, guarantees the introduction of predictive error. Second, they ignore many identifying cues present in the signal.

Other strategies that have met with varying degrees of success include cepstral methods, autocorrelation methods, Gaussian mixture, and wavelet-based methods. Gaussian mixture lacks the flexibility to be consistently reliable over a wide range of environments and speakers.[2] Cepstral methods are based on the proposition that in the cepstrum of a channel-distorted signal, the channel distortion is an additive constant.[2] Autocorrelation methods attempt to look directly for self-similarities in the signal. Autocorrelation works reasonably well for high-pitched speakers, and cepstral analysis works reasonably well for low-pitched speakers, but neither method is suitable for a wide range of pitches.

Wavelet methods are usable over a wide range of pitch. They are robust to noise, and are not critically dependent on segment length. Early attempts with wavelet methods used feature selection criteria that were somewhat arbitrary. Subsequent work has sought to use adaptive wavelets to make this process more objective and automatic.[1,3]

Our research seeks to solve the problem of dealing with a large speaker population by combining two strategies. One is to use the fast continuous wavelet transform (CWT),[4] a more flexible wavelet technique than those attempted previously. CWT overcomes the octave-resolution limitation of the dyadic wavelet. The second strategy is to explore previously ignored features in speech data, the cues to the primary representational system (PRS) operating in the brain of the speaker.

This investigation is being performed on the TIMIT and NTIMIT databases. These consist of 630 speakers uttering 10 phrases each.[5,6,7] The NTIMIT database was formed by transmitting and recording the TIMIT database over commercial telephone lines to introduce the effects of a typical communication channel to the voice signals.

2. FEATURES IN CWT SPACE

In contrast to discrete and dyadic wavelet transforms, whose scaling properties are less flexible, the CWT offers the advantage of much higher resolution, where needed, in both time and scale. The CWT 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 CWT and the discrete wavelet-transform representation of the same voice segment. As is evident in Figure 1, the CWT representation (top) reveals significantly more detail than the dyadic wavelet representation (bottom.)

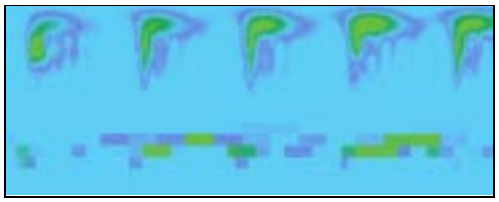

Figure 1. Continuous vs. discrete wavelet representation of a voice segment

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. The details of both the formant frequency and the glottal excitation characteristics can be easily extracted from voice waveforms with the CWT. More importantly, the CWT reveals significant detail of the glottal excitation waveform.

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. 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 the PRS is the basis for using neuro-linguistic programming (NLP) techniques.

It is possible for NLP to extract specific patterns based on integral cues demonstrated by each individual. NLP can extract neurological, verbal and non-verbal information from an individual forming the basis for identifying the PRS (and ancillary representation systems) of the individual speaker. 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*. 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]

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.

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

3.2 Physiological (Non-Verbal) Cues

"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 breathing 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.

"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 tempo 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.

Associated with each of the modalities are the following spacing 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 applied over the past 26 years with consistency and accuracy. The cues provide the means to determine the PRS from a core sample set from the TIMIT database.

4. WORK IN PROGRESS

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 can be reliably estimated using the discrete wavelet transform with dyadic scale changes (D_yWT).[3] The pitch period is estimated by locating periodic peaks that appear across three (D_yWT) 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 pitch period can be directly observed and thus extracted from the resulting CWT representation. Figure 2 shows the amplitude of the time-frequency representation of a male and female speaker uttering the vowel sound /aa/. 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 CWT representation, it is only necessary to locate the centroid of each blob and then determine the corresponding frequency.

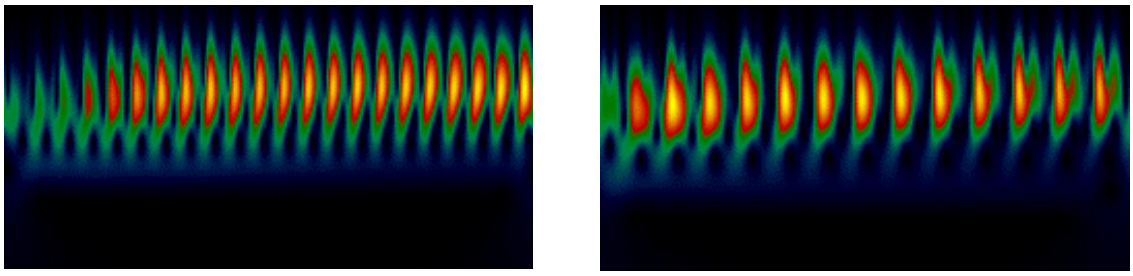

Figure 2. CWT of Female (Left) and Male (Right) Voices Uttering /aa/ (TIMIT Database)

An important question for this research is whether speaker-dependent features can 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 distorted by the channel. To answer this question, the analysis described above has been repeated with the

corresponding signals from the NTIMIT database.[7] 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 channel distortion.

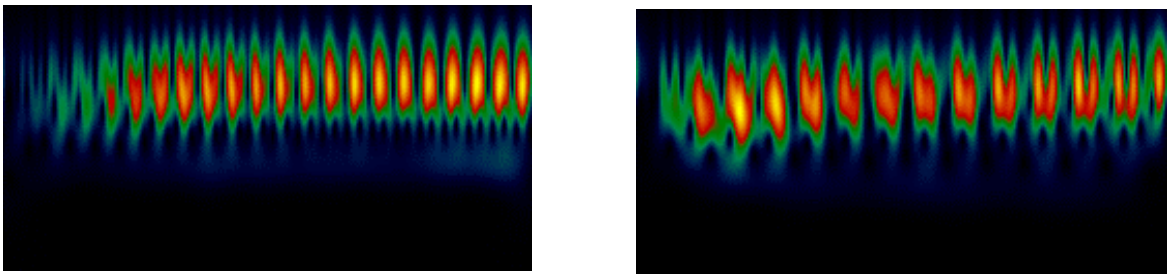

Figure 3. CWT of Female (Left) and Male (Right) Voices Uttering /aa/ (NTIMIT Database)

For NLP analysis, a control data set consisting of 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. The individuals read the sentence, *"She had your dark suit in greasy wash water all year*;" while their speech pattern was recorded and analyzed.

Examples of the control data set are shown in Figures 4-6. A typical voice pattern of a visual individual is shown in Figure 4. A typical voice pattern of an auditory individual is shown in Figure 5. A typical voice pattern of a kinesthetic individual is shown in Figure 6.

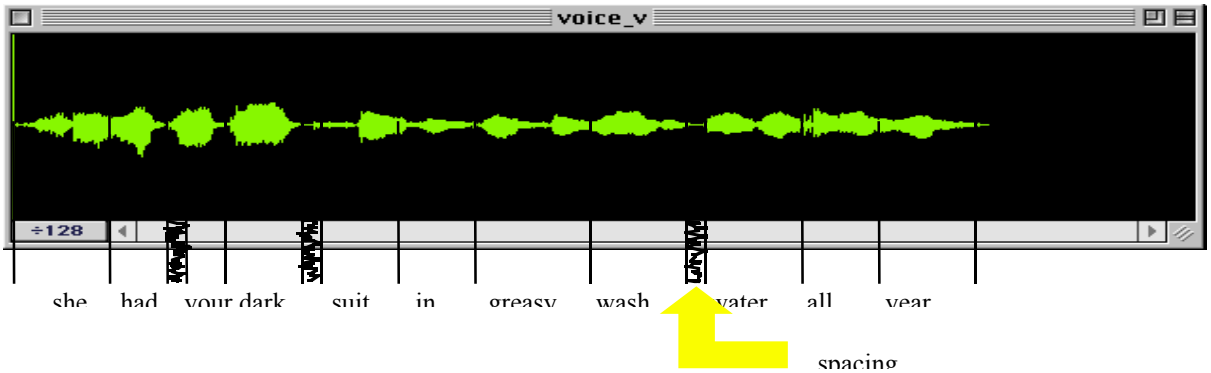

Figure 4. Typical voice pattern of an individual whose PRS is visual
= word or phrase, quick, choppy bursts of words, etc.

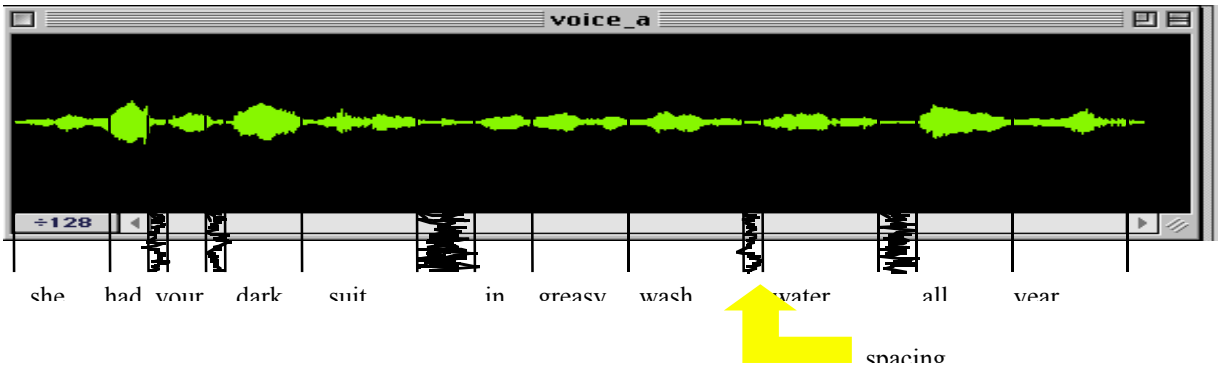

Figure 5: Typical voice pattern of an individual whose PRS is auditory

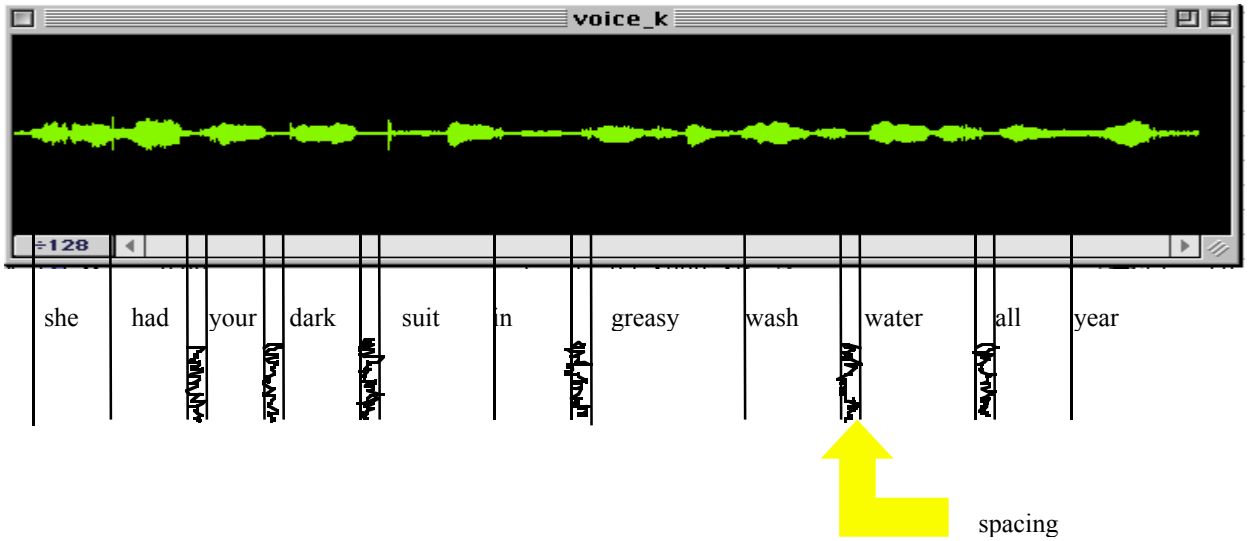

Figure 6: Typical voice pattern of an individual whose PRS is kinesthetic

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.

5. CONCLUSIONS

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.) 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 that the NLP features and CWT features 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.

ACKNOWLEDGMENTS

This research is supported by US-DOE 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).