AN AUTOMATIC TEXT-INDEPENDENT SPEAKER RECOGNITION SYSTEM

Joseph A. Karakowski, Nilu P. Patel

US Army CECOM Center for EW/RSTA Fort Monmouth, NJ

ABSTRACT

A Speaker Recognition System was developed which recognizes different voices in a context-free speaking environment. It is composed of an ART II Neural Network coupled to a pair of Fuzzy Expert Systems which classify voices hierarchically from signal features and clusters. It has been laboratory tested using digitized voices with performance of over 60% correct for about 4 seconds of voiced speech.

1. INTRODUCTION

An in-house investigation performed by the EW/RSTA Systems Engineering Division resulted in the development of a Speaker Recognition System (SRS) which recognizes different speakers in a context-free speaking environment. This paper represents the results of the first year of this R&D effort.

Speaker recognition of text independent information has had limited success to date using short time samples. Markel and Davis[7] obtained text independent speaker recognition results of 98%; this however required an average of 39 seconds of speech. Additionally, these experiments were performed using large input bandwidths and low noise backgrounds. The described system has more stringent requirements; recognition must be made with an input bandwidth of 4.0 KHz, and short bursts of speech. The Speaker Recognition System classifies and sorts like speakers according to the characteristics of their voices. An introduction to the problems can be found in [1]. The voice characterization is performed in two stages, Signal Processing and Glob Processing.

A) SIGNAL PROCESSING. Voiced / unvoiced speech detection, CEPSTRUM, LPC, PARCOR transforms, and pitch are calculated from individual chunks of 32 ms of voiced audio. See Figure 1-1. Selected transform features of each chunk are averaged and formed into globs of 31 chunks, or about 1 second of voiced speech.

B) GLOB PROCESSING. The globs are processed by a neural net / expert system pair. The functional processing which occurs from bursts of speaker audio to voice classification is shown in Figure 1-2, where globs are formed for individual speakers, and transformed into classes of voices.

B.1 NEURAL-NET PROCESSING. An ART-II model clusters globs through a set of selected transform features into preliminary speaker classes.

B.2 EXPERT SYSTEM PROCESSING. The preliminary speaker classes are analyzed, rated, and merged as nodes in a fuzzy relational network. They are processed using a pair of fuzzy expert systems which produces the final voice classification.
2. SYSTEM DESIGN

The overall block diagram of the SRS is shown in Figure 2-1. It is composed of the functions of Digitization, Signal Processing, Neural Network, Expert Systems, Additional Processing and the Processing Environment which will now be described in more detail. The File and Program Control are background programs which track the system state and are not described further.

A. DIGITIZATION

Speech was recorded from four different speakers: three male and one female. Two types of speech was recorded for each; the first was text-dependent and consisted of repetitions of the phrase, "May we all learn a yellow lion roar." This phrase was used by Atal [1] in his speaker recognition experiments. The second consisted of excerpts from "Alice In Wonderland" and represents text-independent speech.

B. SIGNAL PROCESSING

A set of signal features were chosen which include Pitch[8], Average Log Energy, Linear Predictive Coding(LPC), Partial Correlation(PARCOR) and Cepstral coefficients[6].

The chunk processing proceeds as follows. The Pitch and Average Log Energy are calculated to determine if the current chunk represents voiced or unvoiced speech. If voiced, it is passed through a Hamming Window and further processed to determine the LPC, PARCOR and Cepstral coefficients. If unvoiced, then it is discarded.

The globs are created as follows: A running average of selected transform coefficients is done over 31 chunks. An averaged coefficient set, representing about 1000ms of voiced speech, is passed to the Neural Network as a glob.

C. NEURAL NETWORK

The Neural Network utilizes an enhanced Adaptive Resonance Theory(ART) architecture to segregate each input glob of signal features into preliminary speaker categories. The functional details can be found in [3,4]. The opinion(s) of the neural net are sent on to the expert system. See Figure 2-2. Note the feedback between the neural network and expert systems.

D. EXPERT SYSTEMS

The Expert Systems processing checks and merges preliminary speaker clusters to create a final speaker list. It also returns an
Encode/NoEncode command to the neural network for inclusion into a given category. It contains two fuzzy expert systems which perform the analysis and decision making functions in the network. Details of each expert system can be found in [2,5] and has the general structure of Figure 2-3.

Expert Systems has four functions as shown in Figure 2-4: Create Data Base, Merge Expert, Decision Expert, and Data Base Limiter.

For each classification cluster from the neural network a running average of four selected signal features is kept. This is the basis of a Speaker Data Base. The Merge Expert uses the Speaker Data Base to detect and merge identical classification categories. The time averages of the selected signal features in classification clusters are used by the expert system to develop fuzzy linkages between the clusters over time. The Decision Expert utilizes the results of the Merge Expert in the Speaker Data Base and forms a new set of clusters as the Merged Data Base using time as the major parameter. The Limiter checks on the Merged Data Base to limit false class creation and provides the results in the form of the Final Merged Data Base.

E. ADDITIONAL PROCESSING A Results Analysis module develops system performance measurements for each glob processed. A Graphics Display System presents data from a Results Analysis program in two windows; one shows percent correct for each individual speaker and a second shows total percent correct classification for all speakers.

F. PROCESSING ENVIRONMENT. The signal processing was performed on a Spectrum TMS320C30 signal processor board hosted on an IBM PC/AT running SPOX / C. The neural network and fuzzy expert systems were programmed in C and execute on a SUN 4/370. The IBM and SUN were netted using PC/NFS.

3. TESTING

A series of system tests were performed using four speakers with both text dependent and text independent speech.

A. TEST CASES Various test cases were run with different data sets from the same four speakers to determine the robustness of the system by varying several basic parameters. The input voice data were interleaved for maximum variability so there were never two speech samples (globs) from the same speaker occurring sequentially. The effect on overall classification performance was measured, as well as extraneous class creation. The test data varied were: A)Text-independent vs text dependent speech. B)Neural net vigilance parameters. C)“Chunk” and "Glob" times.

From the test data variations, the following results were determined. A)Cumulative percent correct speaker classification. B)Length of speech necessary for a certain percent correct classification. C)Ratio of correct / incorrect classifications per speaker class created, which checks the system's average performance per created class.

B. TEST RESULTS The following are representative of some of the system test
results of A) and B). Figure 4-1 displays cumulative percent correct speaker classification, with a 1 second glob and 32 ms chunk of text independent speech. Additional series of curves varying the vigilance parameter are also shown on this Figure. The length of voiced speech for 60% correct is about 4 seconds minimum with a vigilance of 0.85. Note that the effects of varying vigilance were also tested and were shown to be minimal over the range 0.85 to 0.99. This of course depends on the signal energy, which should be normalized.

A series of bar charts summarizing the overall effects of specific parameters were also generated. In a representative bar chart below, text-independent(black) vs text-dependent(speckled) were compared showing cumulative percent correct. The results again indicate a minimum of about 60% correct for varied glob times and fixed vigilance factor / chunk time.

The results of C) were measured by segmenting test results into two sets: One for correct speaker class and the second to “other” class. For an ideal system, the “other” class would always have no members. This test was run also while varying system parameters. Results were similarly obtained indicating the correct speaker set dominated the “other” set by 50% or more.

![Cumulative Percent Correct Speaker Classification vs Time](image)
4. RESULTS/DISCUSSION

The test results show that one can distinguish each of four voices with over a 60% degree of confidence using about 4 second sample times of text-independent voiced speech. It appears that from these limited results it is feasible to use signal features to characterize some of the invariant features of individual's voices and their production to successfully classify voices in this context. However, since the test data were limited, additional cycles of testing and modification need to be done to determine the robustness of the system performance in the cases with more than four speakers.

Within the SRS, the feature set is an area to be investigated further. Not only current features but also others not yet utilized can provide even better voice discrimination. This can be done by analysis and optimization which provides the maximum distance between speaker classes. This also applies to short and long time averages of signal features, since the neural net is concerned with short time data (across one glob) and the expert systems long time data (across many globs).

BIBLIOGRAPHY

[1] Atal, Bishnu S., "Automatic Recognition of Speakers from Their Voices," Proc. of the IEEE, Vol. 64, No. 4, April 1976, pp. 460-475.
[2] Buckley, J.J., Siler, W., Tucker, D., "A Fuzzy Expert System," Fuzzy Sets and Systems, Vol. 20 (1986), pp 1-16 North Holland.
[3] Carpenter, Gail A. and Stephen Grossberg, "ART 2: Self-organization of Stable Category Recognition Codes for Analog Input Patterns," Applied Optics, Vol. 26, No. 23, December 1987, pp. 4919-4930.
[4] Eck, J. Thomas, "An Enhanced ART 2 Architecture and an Application To Automatic Speaker Recognition," submitted to Department of Computer Information Science, New Jersey Institute of Technology, August 1991.
[5] Karakowski, J., "A Fuzzy Expert System for Real-Time Sidelobe Suppression," 24th Annual Asilomar Conference on Signals, Systems, and Computers, Nov. 1990.
[6] Markel, J.D. and A.H. Gray, Linear Prediction of Speech, Springer-Verlag, New York, 1976.
[7] Markel, John D., Davis, S.B., "Text Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Database," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-27 No.1, Feb. 1979.
[8] Ross, Myron J. et. al., "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-22, No. 5, October 1974, pp. 353-362.