Submission Cover Sheet for 1999 IEEE SoutheastCon

Submission Type (Please Check One):

☐ *Full Length Paper - (Due 11/1/98)* ☑ *Concise Paper - (Due 12/1/98)*

Title of manuscript:
IMPLEMENTATION AND ANALYSIS OF SPEECH RECOGNITION FRONT-ENDS

Authors and affiliations (List authors in the order to appear on the manuscript. List affiliations (companies, universities, etc.) and addresses exactly as they should appear. If there are more than 4 authors, please attach a second sheet.

1. Name: **Vishwanath Mantha**
Affiliation: **Mississippi State University**
Address: **PO Box 9571**
Mississippi State, MS 39762

3. Name: **Yufeng Wu**
Affiliation: **Mississippi State University**
Address: **PO Box 9571**
Mississippi State, MS 39762

2. Name: **Richard Duncan**
Affiliation: **Mississippi State University**
Address: **PO Box 9571**
Mississippi State, MS 39762

4. Name: **Jie Zhao**
Affiliation: **Mississippi State University**
Address: **PO Box 9571**
Mississippi State, MS 39762

Corresponding Author:
Name: **Vishwanath Mantha**
Address: **PO Box 9571**
Mississippi State, MS 39762

Telephone: **601-325-8335**
FAX: **601-325-3149**
E-mail: **mantha@isip.msstate.edu**

Key Words: **digital signal processing, speech recognition, front-end**

For Technical Committee Use Only -- Do Not Write Below

Date Received: _____ Manuscript # Assigned _____

Author Notification Date: _____ Date Camera Ready Received: _____

Decision: _____ Session: _____

Notes: _____

Submission Cover Sheet for 1999 IEEE SoutheastCon

Submission Type (Please Check One):

☐ *Full Length Paper - (Due 11/1/98)* ☑ *Concise Paper - (Due 12/1/98)*

Title of manuscript:
IMPLEMENTATION AND ANALYSIS OF SPEECH RECOGNITION FRONT-ENDS

Authors and affiliations (List authors in the order to appear on the manuscript. List affiliations (companies, universities, etc.) and addresses exactly as they should appear. If there are more than 4 authors, please attach a second sheet.

5. Name **Aravind Ganapathiraju**
Affiliation **Mississippi State University**
Address **PO Box 9571**
Mississippi State, MS 39762

6. Name **Dr. Joseph Picone**
Affiliation **Mississippi State University**
Address **PO Box 9571**
Mississippi State, MS 39762

Corresponding Author:
Name: **Vishwanath Mantha**
Address: **PO Box 9571**
Mississippi State, MS 39762

Telephone: **601-325-8335**
FAX: **601-325-3149**
E-mail: **mantha@isip.msstate.edu**

Key Words: **digital signal processing, speech recognition, front-end**

For Technical Committee Use Only -- Do Not Write Below

Date Received: _____ Manuscript # Assigned _____

Author Notification Date: _____ Date Camera Ready Received: _____

Decision: _____ Session: _____

Notes: _____

IMPLEMENTATION AND ANALYSIS OF SPEECH RECOGNITION FRONT-ENDS

Vishwanath Mantha, Richard Duncan, Yufeng Wu, Jie Zhao, Aravind Ganapathiraju, Joseph Picone
Institute for Signal and Information Processing
Mississippi State University
Mississippi State, Mississippi 39762, USA
e-mail: {mantha, duncan, wu, zhao, ganapath, picone}@isip.msstate.edu
Ph (601) 325-3149 - Fax (601) 325-3149

Abstract **- We have developed a comprehensive front-end module integrating several signal modeling algorithms common to state-of-the-art speech recognition systems. The algorithms presented in this work include mel-frequency cepstra, perceptual linear prediction, filter bank amplitudes, and delta features. The framework for the front-end system was carefully designed to ensure simple integration into speech processing software. The modular design of the software along with an intuitive GUI provide a powerful tutorial by allowing a wide selection of algorithms. The software is written in a tutorial fashion, with a direct correlation between algorithmic lines of code and equations in the technical paper.**

INTRODUCTION

Before a computer can recognize human speech, the speech must first be processed into observation vectors. These vectors represent events in the feature space where there is better discrimination between the different types of sounds. This conversion process, known as signal modeling, is the function of the front-end module. Using these acoustic observation vectors and some language constraints, a network search algorithm (performed by a decoder) finds the most probable sequence of events to hypothesize the textual content of the audio signal [1].

This paper describes the development and analysis of a comprehensive front-end module for a speech recognition system. Several standard front-end algorithms have been implemented, including mel-frequency cepstral coefficients, linear prediction, filter bank amplitudes, and delta features. The framework for this system was carefully designed to ensure simple integration with the public domain speech recognition system [2] under development at the Institute for Signal and Information Processing.

SYSTEM STRUCTURE

The modular design of the front-end is shown in Figure 1. After pre-processing (windowing and pre-emphasis are not shown on the diagram), three basic operations can be performed on the speech signal. These general algorithms are filter bank amplitudes (FBA), the Fourier transform (FFT), and linear prediction (LP) [3]. From the digital filter bank a power estimation may be directly computed. Perceptual linear prediction (PLP) is a post-processing step for LP coefficients, acting as a cascaded filter.

The FFT, LP, and PLP algorithms compute the spectrum of the signal, which is then processed in one of two ways. The first method is filter bank amplitudes (similar to the general FBA algorithm which operated on the original signal), which computes a reduced number of averaged sample values from the spectrum. The cepstrum is an alternate method of processing this spectrum. Either cepstral or FBA coefficients may be used as observation vectors.

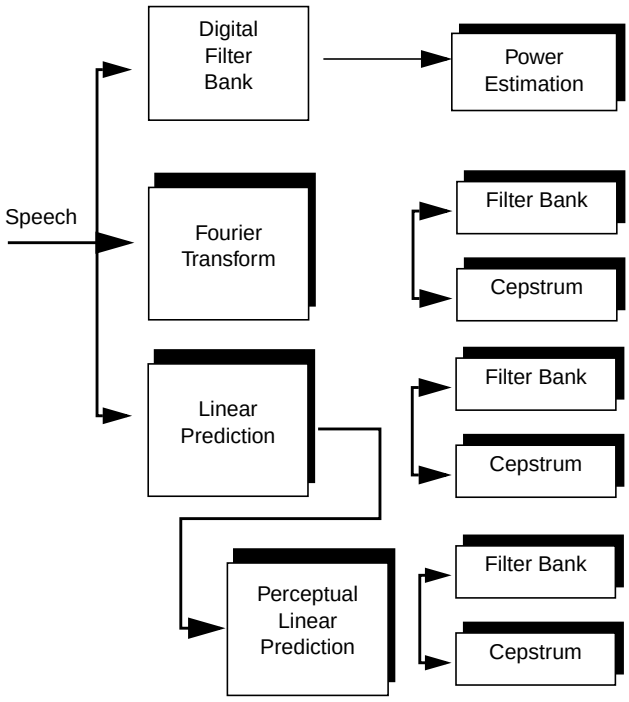

Figure 1. System block diagram

ALGORITHM IMPLEMENTATION

All algorithms are implemented to operate on a single window of speech data. The main control loop of the driver program windows and pre-emphasizes chunks of speech data, simplifying the implementation of each algorithm. This section serves as a brief overview of the different algorithms.

Filter Bank Amplitudes

The digital filter bank is one of the most fundamental concepts in speech processing [3]. A filter bank can be regarded as a crude model of the initial stages of transduction in the human auditory system. Each filter in this set is implemented as a linear phase filter. The filter equations for a linear phase filter can be summarized as:

$$s_i(n) = \sum_{j=\frac{(-(N_{FB_i}-1))}{2}}^{\frac{(N_{FB_i}-1)}{2}} a_{FB_i}(j)s(n+j), \quad (1)$$

where $a_{FB_i}(j)$ denotes the j^{th} coefficient for the i^{th} critical band filter.The number of filter banks normally is an odd number when implementing linear phase filters. The basic merit of this algorithm is that certain filter outputs can be directly correlated with certain classes of speech sounds.

Mel Frequency Spectral Coefficients

A mel is a psychoacoustic unit of measure for the perceived pitch of a tone, rather than the physical frequency. The correlation of the mel to physical frequency is nonlinear, since the human auditory system is nonlinear. A mapping between the mel scale and real frequencies was empirically determined by Stevens and Volkman [4] in 1940.

A homomorphic system is useful for speech processing because it offers a methodology for separating the excitation signal from the vocal tract shape. One feature space which offers this property is the cepstrum, computed as the inverse discrete Fourier transform (IDFT) of the log energy [3]. This signal is by definition minimum phase, another useful property. Cepstral coefficients are computed as:

$$c(n) = \frac{1}{N_s}\sum_{k=0}^{N_s} \log|S_{avg}(k)|e^{j2\pi k\frac{n}{N_s}} \qquad 0 \le n \le N_s - 1, \quad (2)$$

where $S_{avg}(k)$ is the average signal value in the k^{th} filter channel. In practice, the discrete cosine transform is used in lieu of the IDFT for computational efficiency. A liftering procedure is also used to weight the cepstrum and control the non information bearing variabilities [4].

Perceptual Linear Prediction (PLP) Analysis

The block diagram of PLP analysis is shown in Figure 2. The low order all-pole model of such an auditory spectrum has been found to be consistent with several phenomenon observed in speech perception [5]. This is an improvement over the LP analysis of speech, which estimates the smoothed spectral envelope of the speech power spectrum equally well at all frequencies. The latter is not consistent with human hearing. Also, the PLP-derived spectrum is more robust to noise compared to the LP-derived spectrum [5], as illustrated in Figure 3. It may be observed that the PLP-derived spectrum is able to model the second formant in the presence of noise whereas the LP-derived spectrum fails.

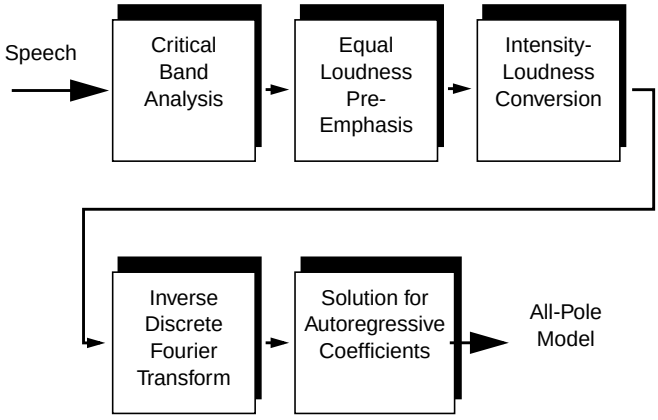

Figure 2. PLP Analysis

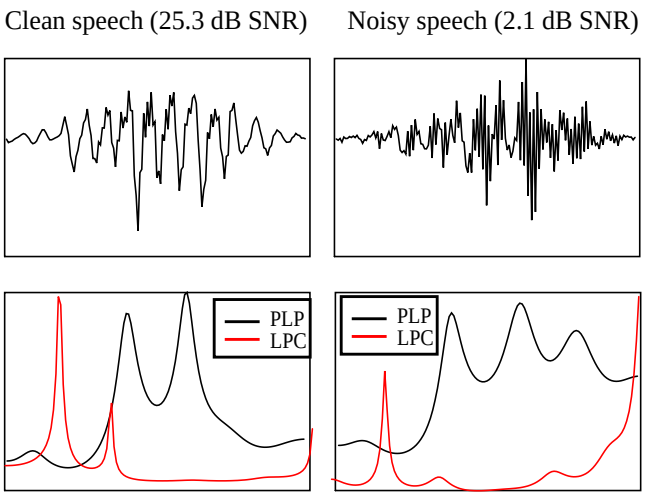

Figure 3. A comparison of PLP and LP-derived spectra

Delta Features

The performance of a speech recognition system is enhanced greatly by adding time derivatives to the static parameters. The first-order derivatives are referred to as delta features. Regression analysis is used to compute delta features [4]. The first formulation is:

$$d_n = \frac{\sum_{w=1}^{dw} w(c_{n+w} - c_{n-w})}{2\sum_{w=1}^{dw} w^2}, \quad (3)$$

where d_n is a delta coefficient at frame n, c_{n-w} and c_{n-w} are static parameters before and next to the current frame coefficient c_n, and dw is the delta window size. Since the regression technique requires past and future speech parameter values, suitable modifications are performed on the beginning and the end of the data stream.

Most state-of-the-art speech recognition systems use a front-end comprising of 12 Fourier transform-derived mel-frequency cepstral coefficients and mean energy as a first order model of the signal plus the first and second derivatives.

EXPERIMENTS AND RESULTS

We used the OGI Alphadigits [6] task to test our system. The corpus has a 40 word vocabulary which restricts the number of context-dependent models to a manageable number. Since the word error rates (WER) on this specific database are relatively low (about 10%), this task is useful for calibrating system performance.

Twelve cepstral features along with energy and their delta and acceleration coefficients comprised the feature set that we have employed to evaluate our front-end. We used the ISIP decoder [2] to perform our tests. Standard signal processing such as pre-emphasis and cepstral mean subtraction were employed. The training set consisted of 15,000 utterances averaging six alphanumeric elements in length, which we grouped into cross-word triphone models. As a point of comparison, a system of comparable complexity was built using HTK [4]. Table 1 summarizes the results obtained using both the systems. The main external difference in the two systems was the training algorithm used by the decoder. The HTK system used Baum-Welch training, we employed Viterbi training with the ISIP system.

GRAPHICAL USER INTERFACE

While the front-end is capable of producing output models consistent with other state of the art systems, it can also be used to study the differences between the different algorithms on real data. A Tcl-Tk based graphical user interface (GUI) is available to facilitate this user interaction. This utility inherits the signal display routine from the SWITCHBOARD Segmenter [7]. A snapshot of the GUI is shown in Figure 4.

The user can vary different parameters for each algorithm and study its effect on the output feature vectors. Since multiple algorithms may be run and displayed simultaneously, the user can directly compare the performance of different algorithms. The user may also play the audio as a secondary point of reference. All parameters (window type, LP order, etc.) can be varied via the configuration window.

CONCLUSIONS

Standard algorithms such as mean energy, digital filter banks, the Fourier transform, linear prediction, the cepstrum, and difference equations were incorporated in our front-end. Physiological knowledge of the human auditory and vocal articulatory systems is applied (the mel and Bark scales, perceptual linear prediction) to the standard signal processing techniques to better model speech and increase recognition performance.

All software for this project was developed in C++ using the public-domain GNU compiler. Our software comprehensively allows the user complete control over all aspects of the signal modeling process. This includes algorithm selection, frame and window duration, and internal parameters. The Tcl-Tk based graphical user interface (GUI) further facilitates user interaction with the numerous parameters. The performance of this module is competitive with other state-of-the-art speech recognition systems.

Table 1. Comparison of systems

	ISIP	HTK
WER	15.6%	14.5%
Substitutions	13.8%	12.1%
Deletions	1.0%	0.3%
Insertions	0.8%	2.0%

The front-end module described in this paper interfaces directly with the ISIP speech recognition system. All software and documentation for this project is available at the Institute for Signal and Information Processing's [8] web site.

REFERENCES

[1] J. Picone, "Signal Modeling Techniques in Speech Recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215-1247, September 1993.

[2] N. Deshmukh, A. Ganapathiraju, J. Hamaker and J. Picone, "An Efficient Public Domain LVCSR Decoder," *Proceedings of the Hub-5 Conversational Speech Recognition (LVCSR) Workshop*, National Institute for Standards and Technology, Linthicum Heights, Maryland, September 1998.

[3] L.R. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1993.

[4] S. Young, *The HTK Book: for HTK Version 2.0*, Cambridge University Press, Cambridge, England, 1995.

[5] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech." *Journal of the Acoustical Society of America*, vol. 4, pp. 1738-1752, 1990.

[6] R.Cole, et. al., "Alphadigit Corpus," http://www.cse.ogi.edu/CSLU/corpora/alphadigit, Center for Spoken Language Understanding, Oregon Graduate Institute, 1997.

[7] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker and J. Picone, "Resegmentation of Switchboard," *Proceedings of the International Conference on Spoken Language Processing*, vol. 4, pp. 1543-1546, Sydney, Australia, November 1998.

[8] R. J. Duncan, V. Mantha, Y. Wu and J. Zhao, "Implementation and Analysis of Speech Recognition Front-Ends," http://www.isip.msstate.edu/resources/ece_4773/projects/1998/group_signal, Institute for Signal and Information Processing, Mississippi State University, USA, November 1998.

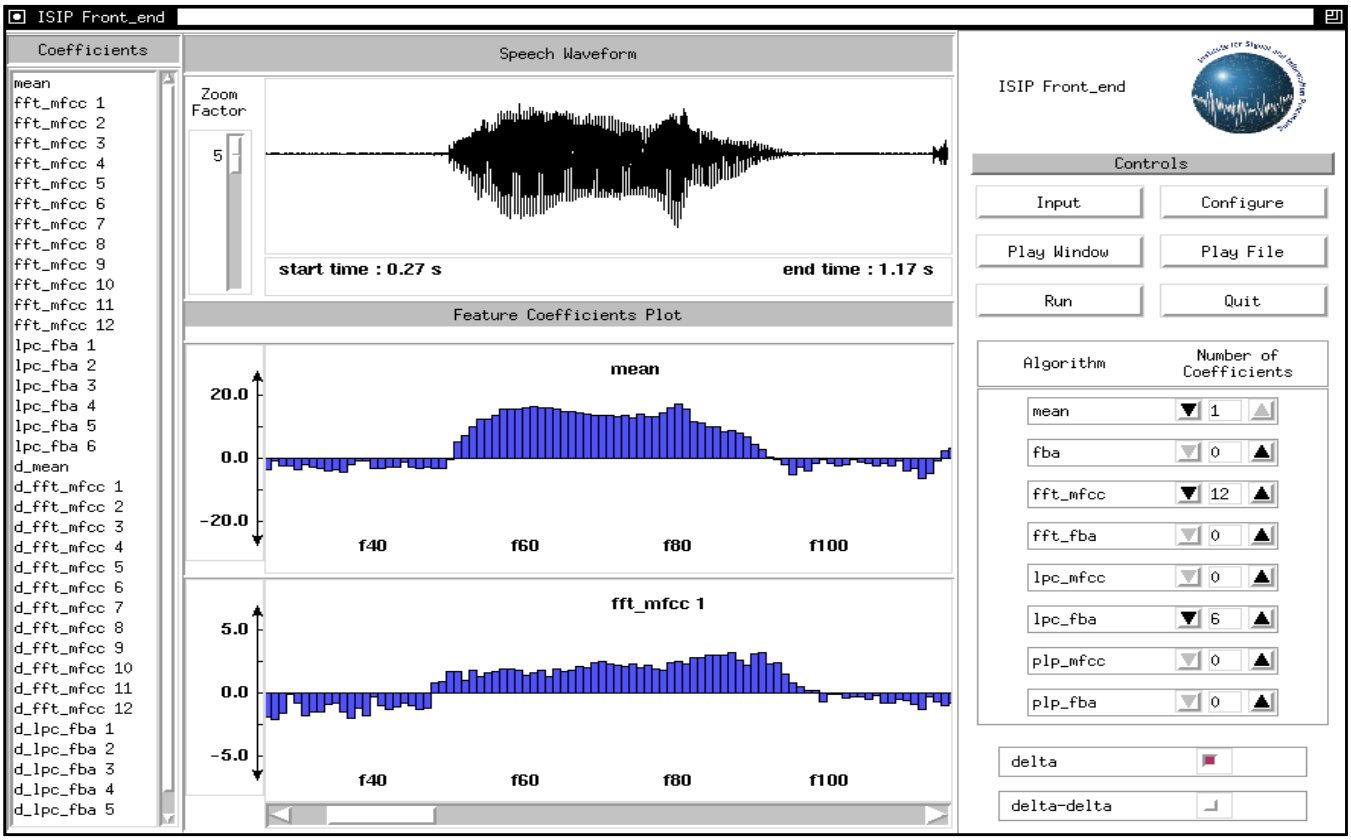

Figure 4. Snapshot of the graphical user interface