proposal for

Implementation and Analysis of Speech Recognition Front-Ends

submitted to fulfill the semester project requirement for

EE 4773/6773: Digital Signal Processing

September 8, 1998

submitted to:

Dr. Joseph Picone

Department of Electrical and Computer Engineering
413 Simrall, Hardy Rd.
Mississippi State University
Box 9571
MS State, MS 39762

submitted by:

Richard Duncan, Vishwanath Mantha, Yufeng Wu, Jie Zhao

Signal Processing Group
The Institute for Signal and Information Processing

Mississippi State University
Box 9671
Mississippi State, Mississippi 39762
Tel: 601-325-8335
Fax: 601-325-8192
email: {duncan, mantha, wu, zhao}@isip.msstate.edu

I. ABSTRACT

The aim of this project is to develop a standard comprehensive front-end module in a speech recognition system. Several standard front-ends will be implemented, including mel cepstra, perceptual linear prediction, filter bank amplitudes, and delta features. The framework for this system will be carefully designed to ensure simple integration with the speech recognition system. The modular design of the software along with an intuitive GUI will allow a student of speech processing to easily interchange algorithms and vary every aspect of each feature vector, providing a powerful tutorial. The code itself will be written in tutorial fashion, with a direct correlation between algorithmic lines of code and equations in the technical paper. The effectiveness of the different front-end algorithms will be evaluated on a common set of speech data.

II. INTRODUCTION

There has been a considerable amount of interest shown in the field of computer speech recognition over the past two and a half decades. The front-end is required to extract important features from the speech waveform that are relatively insensitive to talker and channel variability unrelated to speech message content. This stage also attempts to decrease redundancy inherent in the speech waveform.

The most useful parameters in speech processing are those derived from a frequency domain representation of the signal. The vocal tract produces signals that are more consistently analyzed in the frequency domain rather than the time domain.The common mode of speech production involving a noisy or periodic waveform exciting a vocal tract filter corresponds well to separate spectral models for the extraction and the vocal tract. Hence all the standard front-ends use various frequency domain parameters for processing.

Linear Predictive (LP) analysis gives us an estimate of the autoregressive all pole model $A(w)$ of the short term power spectrum of speech $P(w)$. We can also view LP analysis as a means for obtaining the smoothed spectral envelope of $P(w)$. The major disadvantage of the LP all pole model in speech analysis is that $A(w)$ approximates $P(w)$ equally well at all frequencies of the analysis band. This property is inconsistent with human hearing, which tends to be more logarithmic above 800 Hz. Consequently, the spectral details of $P(w)$ are not always preserved or discarded by LP analysis according to their auditory prominence. The algorithms described in this section have been shown to improve on the basic LP model.

Mel Filter Bank Cepstra

The LPC cepstral coefficients can be derived directly from the LPC coefficient set to provide a more robust, reliable feature set. The resultant feature set is less dependent on the audio properties of the channel and emphasizes the acoustic differences in the different audible phonemes. Simply stated, the cepstral coefficients are the coefficients of the Fourier transform representation of the

log magnitude spectrum. The conversion is shown below:

$$\begin{aligned} c_0 &= \ln\sigma^2 \\ c_m &= a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad 1 \le m \le p \\ c_m &= \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k} \qquad m > p \end{aligned} \tag{1}$$

Where a_m are the LPC coefficients, σ^2 is the gain term in the LPC model, and p is the order of the LPC analysis.

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 the physical frequency is not linear, as the human auditory system is a non linear system. A mapping between the mel scale and real frequencies was empirically determined by Stevens and Volkman in 1940. The scale is roughly linear below 1000 Hz, then decays logarithmically. This nonlinear scale is invaluable to speech coding in that it reduces the sample space with minimal perceptual loss.

An alternate (and more common) approach to calculate the MFCCs is to start with the FFT. Taking the log of each coefficient will force the signal to be minimum phase. A set of triangular filters evenly spaced along the mel-scale smooths and averages the signal into a smaller number of coefficients. The discrete cosine transform can then be used to derive the mel frequency cepstral coefficients.

Delta Features

In addition to the MFCCs, another popular feature that can be used is the delta cepstrum, which is found by using the difference equation. If $c_s(n;m)$ denotes the mel cepstrum for the frames of the signal s_n ending at time m, then the delta cepstrum at frame m is defined as

$$\Delta c_s(n;m) = c_s(n;m + \delta Q) - c_s(n;m-\delta Q) \tag{2}$$

for all n. Here Q represents the number of samples by which the window is shifted for each frame. The parameter δ is chosen to smooth the estimate and typically takes a value of 1 or 2 (look forward and backward one or two frames). The delta (1st derivative) and delta delta (second derivative) provide valuable inter-frame temporal information of the speech signal. Typically 8-14 cepstral coefficients and their "derivatives" are used for speech recognition systems.

Perceptual Linear Predictive Analysis

The PLP analysis is a relatively new method for the analysis of speech signals. It is an improvement over the widely used LP (Linear Predictive) analysis. In PLP analysis, the all pole

modeling is applied to an auditory spectrum derived by (a) convolving $P(w)$ with a critical band masking pattern, followed by (b) resampling the critical band spectrum at approximately l Bark intervals, (c) pre-emphasis by a simulated fixed equal loudness curve, and (d) compression of the resampled and pre-emphasized spectrum through the cubic root non-linearity, simulating the intensity - loudness power law. The low order all pole model of such an auditory spectrum has been found to be consistent with several phenomena observed in speech perception. The block diagram of PLP Analysis is given in Figure 1.

The speech segment is weighted by the Hamming window

$$W(n) = 0.54 + 0.46\cos[2\pi n/(N-1)] \tag{3}$$

where N is the length of the window. The typical length of the window is about 20 ms. The FFT of this is used to smoothly transform the speech segment into the frequency domain.

The real and imaginary components of the short term speech spectrum are squared and added to get the power spectrum

$$P(w) = Re[S(w)]^2 + Im[S(w)]^2 \tag{4}$$

The spectrum $P(w)$ is warped along its frequency axis into the Bark frequency Ω by

$$\Omega(w) = 6\ln\{(w/(1200\pi)) + [w/(1200\pi)^2 + 1]^{0.5}\} \tag{5}$$

where w is the angular frequency in rad/s. The resulting warped power spectrum is then convolved with the power spectrum of the simulated critical band masking curve $\psi(\Omega)$.

$$\psi(\Omega) = \begin{cases} 0 & for(\Omega < -1.3) \\ 10^{2.5(\Omega + 0.5)} & for(1.3 \le \Omega \le -0.5) \\ 1 & for(-0.5 \le \Omega \le 0.5) \\ 10^{-1.0(\Omega - 0.5)} & for(0.5 \le \Omega \le 2.5) \\ 0 & for(\Omega > 2.5) \end{cases} \tag{6}$$

The discrete convolution of $\psi(\Omega)$ with (the even symmetric and periodic function) $P(w)$ yields samples of the critical band power spectrum

$$\theta(\Omega_i) = \sum_{i=-1.3}^{2.5} P(\Omega - \Omega_i) \cdot \psi(\Omega) \tag{7}$$

This convolution significantly reduces the spectral resolution of $\theta(\Omega)$ in comparison with the original $P(w)$. This also allows for down sampling.

The sampled $\theta(\Omega(w))$ is pre-emphasized by a simulated equal loudness curve, which is an

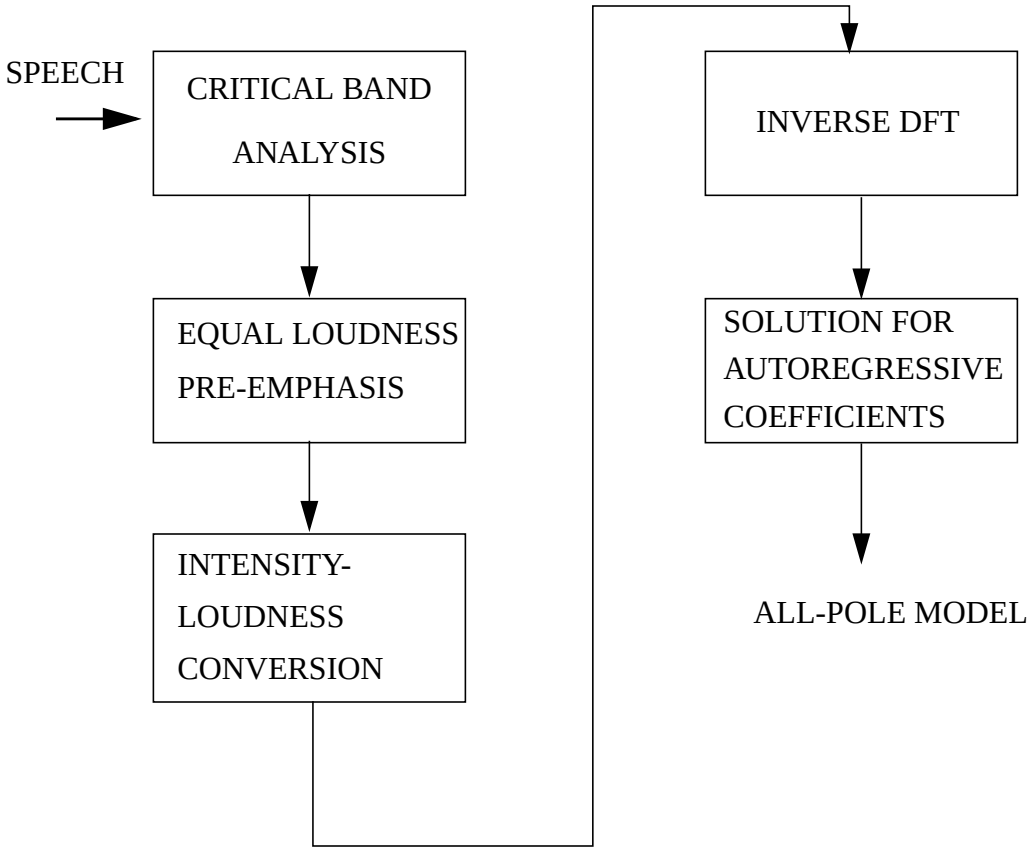

Figure 1. Block Diagram for PLP Analysis

approximation of the non equal sensitivity of human hearing at different frequencies. This pre-emphasized function is then amplitude compressed using cubic root amplitude compression.

The final operation of PLP analysis is the approximation of $\theta(\Omega)$ by the spectrum of an all pole model using the autocorrelation method of all pole spectral modeling. The principle is to apply the inverse DFT to $\theta(\Omega)$ in order to yield its autocorrelation function dual.

Filter Bank Spectral Analysis

A filter bank(FB) is just a bank of filters with a downsampler, upsampler, and sometimes delay. A typically filter bank is shown in Figure 2.. This structure is used nowadays in most subband coding systems. Here, $x(n)$ is the input signal, $y(n)$ is the output of reconstructed signal, the Hs are called the analysis filters, with system functions $H_k(z)$ arranged in a parallel bank. Fs are called the synthesis filters, with system functions $F_k(z)$. The outputs of the filters are summed to form the synthesized signal $y(n)$.The $V_k(n)$s are called the subband signals.The processing block indicates the subband signals which can be coded for storage or transmission.

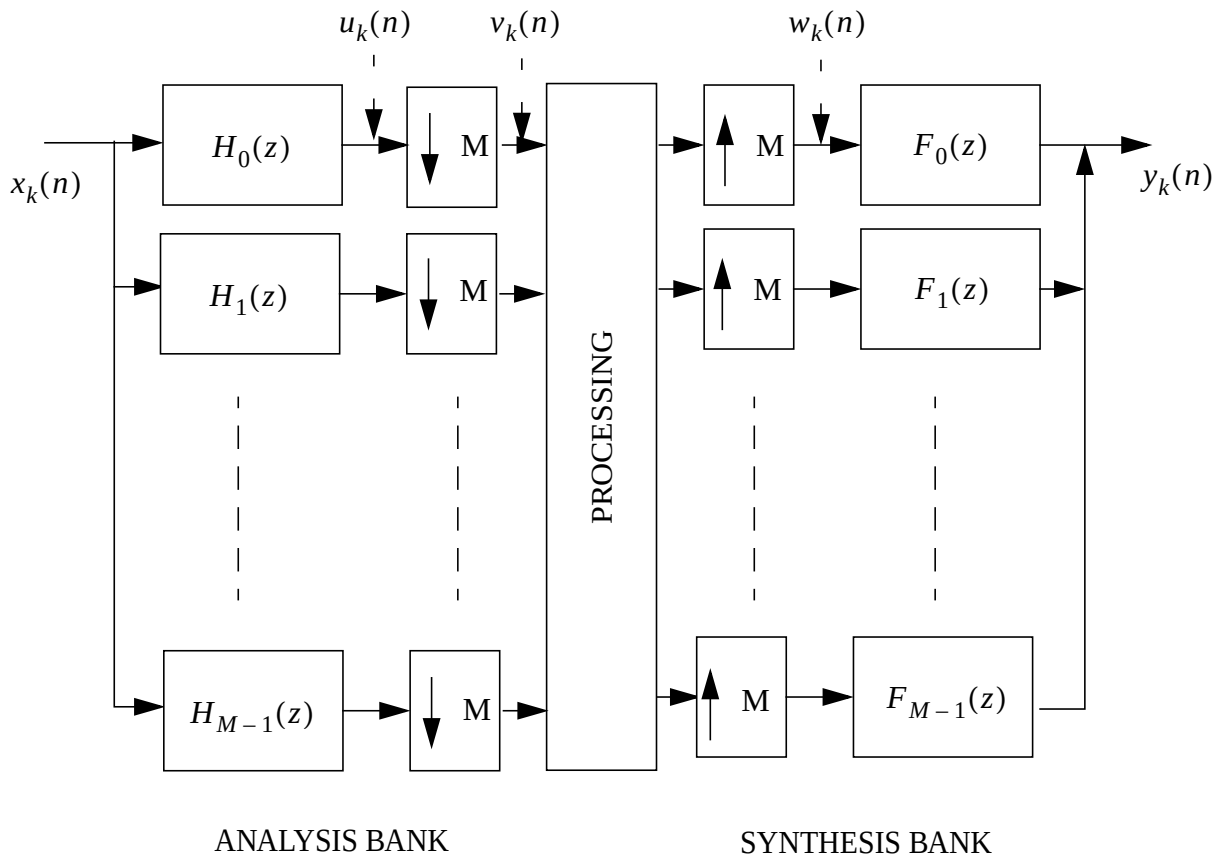

Figure 2. Structure of a typical filter bank

Probably the most important parametric representation of speech for the purpose of recognition is the short time spectral envelope. Spectral analysis methods are therefore generally considered as the core of the signal-processing front-end in a speech recognition system. The filter-bank spectral analysis model is one of the dominant methods of spectral analysis.

The overall structure of the filter-bank model for speech processing is shown in Figure 3.

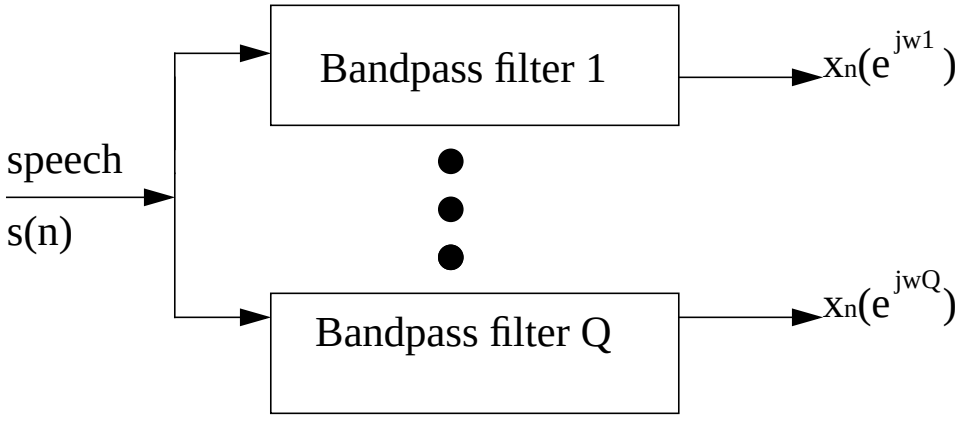

Figure 3. Filter-banks spectral analysis model

The speech signal, $s(n)$ is passed through a bank of Q bandpass filters whose coverage spans the frequency range of interest in the signal (100Hz-3kHz for telephone-quality signals, 100Hz-8kHz for broadband signals). The individual filters can and generally do overlap in frequency. The output of the ith bandpass filter, $X_n(j\omega_i)$ (where $\omega_i = 2\pi f_i / F_s$ is the normalized frequency, and F_s is the sampling frequency) is the short term spectral representation of the signal $s(n)$, at time n, as seen through the ith bandpass filter with center frequency ω_i. It can readily be seen that in the filter banks model each bandpass filter processes the speech signal independently to produce the spectral representation X_n.

III. PROJECT SUMMARY

The purpose of this project is to develop a modular framework for speech recognition front-ends in the public domain. The code will be developed under stringent quality control to facilitate integration directly into a real speech recognition system currently under development at the Institute for Signal and Information Processing. The main tasks of this project are set forth below:

- Implement several standard speech recognition front-ends including mel cepstra, perceptual linear prediction, filter bank amplitudes, delta and delta-delta features. The code for these algorithms will be written in C++ and compiled using the gnu compiler. It is necessary for this code to be in perfect form and well documented to facilitate ISIP's charter as a educational center of excellence in speech to text processing.

- Interface the front-end speech signal processor into a real large vocabulary continues speech recognition system under development at the Institute for Signal and Information Processing. The ISIP Trace Projection Decoder will be used to evaluate the effectiveness of the different front-end algorithms on real speech data.

- Development of a Graphical User Interface to be used as an educational tool in a fundamentals of speech recognition course as an online tutorial.

- Evaluate the effectiveness of each algorithm on a real corpus of speech data.

IV. EVALUATION

While the preferred method of evaluation would be to study the effects of the front-end algorithms on the overall Speech-to-Text word error rate (WER), this is not a plausible course of action due to the current state of development of the ISIP recognition system. The ISIP recognizer does not currently support a training mode, which means it cannot use feature vectors generated by our front-end to train the acoustic models. The only acoustic models available to the ISIP recognizer are based on external software, so any WER experiments would suffer greatly from mismatched acoustic information. Instead, a state-of-the-art phone classification system will be used to evaluate the effectiveness of each feature extraction algorithm.

Our evaluation will do frame-level classification experiments on a reduced phone set of the

SWITCHBOARD (SWB) Corpus and OGI Alphadigits Corpus. In this work, we will use a Support Vector Machine (SVM) as the classifier. The core component in this paradigm is SVMLite, an SVM toolkit which is available as freeware. This SVM package can be applied to large datasets and is capable of handling classification tasks with tens of thousands of support vectors. For our evaluation, it will classify at the frame level the feature vectors generated by the front-end into several phonetic classes. The training data consists of two types of data: first, the Deterding Vowel data, which was collected at 10 kHz sampling rate and low pass filtered at 4.7 kHz. A window duration of 50 msec. was used for generating the features. Second, a subset of the SWB corpus consisting of 16 phones extracted from continuous speech. The phones were chosen to represent vowels, the fricatives 's' and 'f' and the liquids 'l' and 'r'. The segmentation was based on a 44 phone context-independent system. Feature vectors were generated by computing 12 mel-scaled cepstra along with energy. A frame duration of 10 msec. and a window duration of 25 msec. was used for data generation. By comparing the output of this classifier and the reference information, which can be obtained by the state-level forced alignments, of the input speech data, we can evaluate the performance of each algorithm implemented by our front-end.

V. SCHEDULE

A schedule for the major tasks in this project is shown in Figure 4.

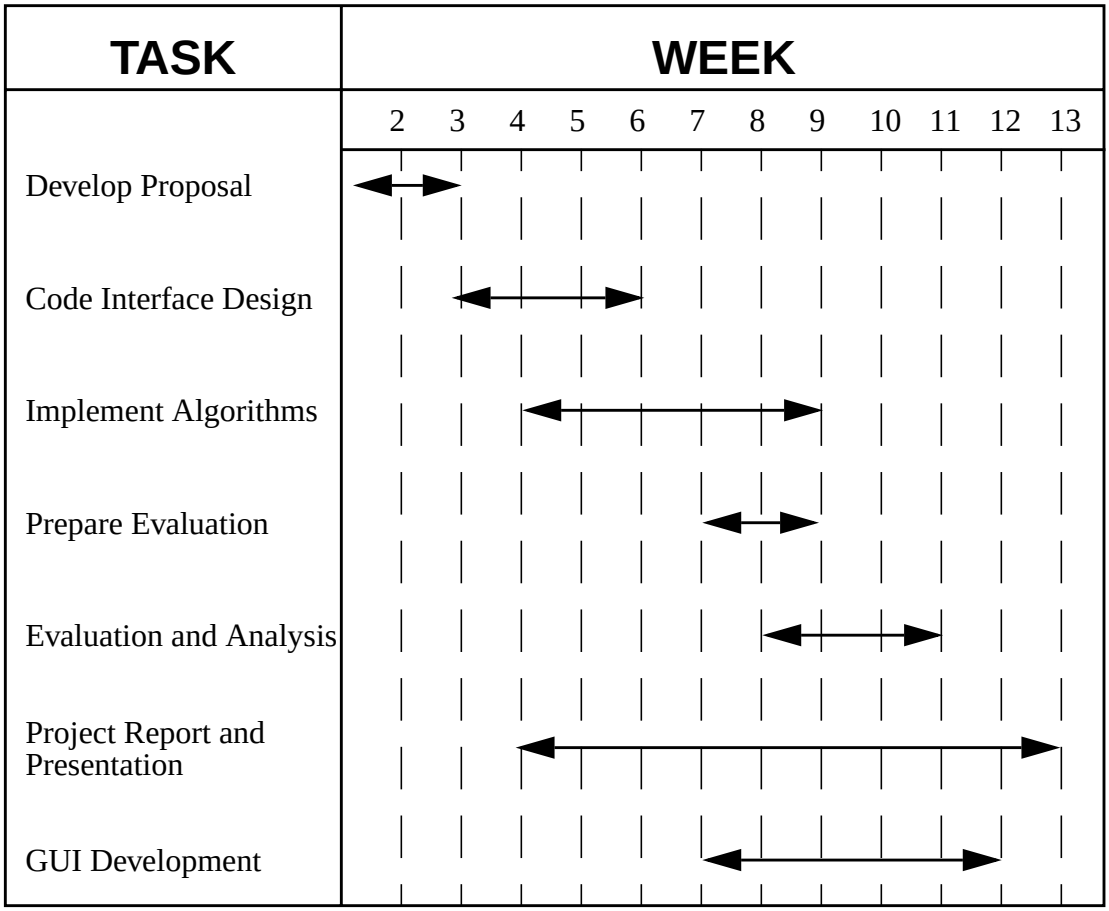

Figure 4. Schedule of Key Tasks.

A few clarifications should be made as to the reasoning behind the above schedule. The interface design is a crucial step towards the success of the project since this code must be written in such a way that it lives beyond the scope of this course. Also, the code must be exceedingly well documented, to the point of a direct correlation between lines of code and equations in the technical paper, hence the report writing progresses simultaneously with the code development. Algorithm implementation is scheduled to allow ample time for software debugging. The GUI also cannot wait until the end of the project as it must go through interface design reviews as well to assure a high quality product at the end of the term.

VI. REFERENCES

[1] S. Balakrishnama, A. Ganapathiraju, "Linear Discriminant Analysis - A Brief Tutorial," *Institute for Signal and Information Processing*, March 1998.

[2] S.B. Davis, P. Mermelstien, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 4, pp. 357-366, 1980.

[3] J. Deller, J. Proakis, J. Hansen, *Discrete-Time Processing of Speech Signals*, MacMillan Publishing Co., New York, New York, USA, 1993.

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

[5] C.R. Jankowski, H. Hoang-Doan, L.P. Lippmann, "A Comparison of Signal Processing Front Ends for Automatic Word Recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 286-292, 1995.

[6] D. O'Shaughnessy, "Speech Technology," *Applied Speech Technology*, ed. A. Syrdal, R. Bennett, and S. Greenspan, CRC press, Boca Raton, pp. 47-98, 1995.

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

[8] J. Proakis, D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1996.

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

[10] L.R. Rabiner, R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1978.

[11] A.J. Robinson, "Dynamic Error Propagation Networks," Ph.D. Thesis, Cambridge Univ. Eng. Dept. February 1989.

[12] J. Godfrey, E. Holliman and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development", *Proceedings of the IEEE ICASSP*, vol. 1, pp. 517-520, San Francisco, CA, USA, March 1992.