MODULATION FEATURES FOR SPEECH RECOGNITION

Dimitrios Dimitriadis[†], *Petros Maragos*[†] *and Alexandros Potamianos*[‡]

[†] Dept. ECE, National Technical University of Athens, Zografou, 15773 Athens, Greece
[‡] Bell Labs, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.
ddim@cs.ntua.gr, maragos@cs.ntua.gr, potam@research.bell-labs.com

ABSTRACT

Automatic speech recognition (ASR) systems can benefit from including into their acoustic processing part new features that account for various nonlinear and time-varying phenomena during speech production. In this paper, we develop robust methods to extract novel acoustic features from speech signals of the modulation type based on time-varying models for speech analysis. Further, we integrate the new speech features with the standard linear ones (mel-frequency cesptrum) to develop a augmented set of acoustic features and demonstrate its efficacy by showing significant improvements in HMM-based word recognition over the TIMIT database.

1. INTRODUCTION

The traditional approach to speech modelling has been the linear model where the true nonlinear physics of speech production have not been taken under consideration. Therefore, the most common speech features used in ASR are based on short-time smoothed cepstra stemming from the linear model. This representation ignores the nonlinear aspects of speech and is sensitive to small signal durations. In this paper, we focus on improving the acoustic processing part of ASR systems by developing robust instantaneous features based on modulation models for speech production and by using these features to increase the recognition performance of ASR systems whose pattern classification part is based on Hidden Markov Models (HMM).

Our motivations for this research work include the following: (i) Adding new information to the feature set such as instantaneous information can model better the dynamics and time evolution of speech features. (ii) Robustness to large speaker population or large vocabularies with confusable words can be achieved by using speech processing models motivated by the physics of speech production and auditory perception.

In Section 2 of this paper, the use of modulation models for speech resonances is reviewed and a robust demodulation algorithm for extracting the parameters of such models is developed. In Section 3, we describe how to extract novel short-time feature vectors from speech signals that contain modulation information and use them as additional (to the cepstrum) features to develop a generalized set of acoustic features for improving HMM-based recognition.

2. SPEECH MODULATION MODEL AND DEMODULATION ALGORITHM

There is much experimental and theoretical evidence for the existence of amplitude and frequency modulation (AM-FM) in speech resonance signals, which make the amplitude and frequency of the resonance (formant) vary instantaneously within a pitch period.

Motivated by this evidence, Maragos, Quatieri and Kaiser [4] model each speech resonance with an AM-FM signal,

$$x(t) = a(t)\cos[2\pi \int_0^t f(\tau)d\tau] \qquad (1)$$

and the total speech signal as a superposition of such AM-FM signals, one for each formant. Here $a(t)$ and $f(t)$ are the instantaneous amplitude and frequency which represent the time-varying formant signal. The short-time formant frequency average $f_c = (1/T)\int_0^T f(t)dt$, where T is in the order of a pitch period, is said to be the carrier frequency of the AM-FM signal. The classical linear model of speech views a formant frequency as constant, i.e., equal to f_c, over a short time (10-30 ms) frame. However, the AM-FM model can both yield the average f_c and provide additional information about the formant's instantaneous frequency deviation $f(t) - f_c$ and its amplitude intensity $|a(t)|$. To isolate a single resonance from the original speech signal, band-pass filtering is first applied around estimates of formant center frequencies. Then for demodulating a resonance signal, Maragos et al. [4] used the nonlinear Teager-Kaiser

This research work was supported by the Greek Secretariat for Research and Technology and by the European Union under the program ΕΠΕΤ-98 with Grant # 98ΓΤ26. It was also partially supported by the basic research program ARCHIMEDES of the NTUA Institute of Communication and Computer Systems.

0-7803-7402-9/02/$17.00 ©2002 IEEE

energy-tracking operator

$$\Psi[x(t)] \equiv \left[\frac{dx(t)}{dt}\right]^2 - x(t)\frac{d^2x(t)}{dt^2} \qquad (2)$$

to develop the following nonlinear algorithm

$$\frac{1}{2\pi}\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \approx f(t) \ , \ \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \approx |a(t)| \qquad (3)$$

This is the energy separation algorithm (ESA) and provides AM-FM demodulation by tracking the physical energy implicit in the source, producing the observed acoustic resonance signal and separating it into its amplitude and frequency components. It yields very good estimates of the instantaneous frequency signal $f(t) \geq 0$ and of the amplitude envelope $|a(t)|$ of an AM-FM signal, assuming that $a(t), f(t)$ do not vary too fast (small bandwidths) or too greatly compared with the carrier frequency f_c. A very efficient and computationally simple discrete version of the ESA also exists, called DESA [4], which is obtained by using a discrete energy operator on discrete-time nonstationary sinusoids. Extensive experiments on speech demodulation using the DESA in [4, 5] indicate that these amplitude/frequency modulations exist in real speech resonances and are necessary for its naturalness.

The main disadvantage of the DESA is a moderate sensitivity to noise. Thus, we describe next an alternative approach [2] where we first interpolate the discrete-time signal using smoothing splines [7], and then apply the continuous-time ESA (3). *Splines* are piecewise polynomial functions constructed as a linear combination of B-Splines. A spline function of order ν has continuous derivatives up to order $\nu - 1$, which is important when using the energy operator Ψ. At first we used exact splines to improve the performance of the ESA, tested on noisy AM-FM signals with different levels of SNR. The results were disappointing as the exact fitting of the signal representation curve, due to the presence of noise, was creating large estimation errors. The problem of noise led us to optimally interpolate signal samples with *smoothing splines*, whose main advantage is that the interpolating polynomial does not pass "precisely" through the signal samples but "close enough". The smooth spline interpolating function is defined as the function s_ν that minimizes the mean square error criterion

$$E = \underbrace{\sum_{n=-\infty}^{+\infty} (x[n] - s_\nu(n))^2}_{E_d} + \lambda \underbrace{\int_{-\infty}^{+\infty} \left(\frac{\partial^r s_\nu(t)}{\partial t^r}\right)^2 dt}_{E_s}$$

where E_d is the data fitting error and E_s quantifies the non-smoothness of the interpolant by the mean square value of its derivative. The positive design parameter λ controls the trade-off between how smooth the interpolating curve will be and how close to the data points the interpolant will pass. (For $\lambda = 0$ we obtain exact splines with no data smoothing.) Given the initial signal samples $x[n]$, $n = 1, \ldots, N$, the interpolating spline function of order $\nu = 2r - 1$ is given by [7]

$$s_\nu(t) = \sum_{n=-\infty}^{+\infty} c[n]\beta_\nu(t - n) \qquad (4)$$

where $\beta_\nu(t)$ is the B-spline of order ν, and the coefficients $c[n]$ depend only on the data $x[n]$, the parameter λ and the analytic expression of the B-spline. The coefficient sequence $c[n]$ can be determined recursively by using the sequence $x[n]$ as input to excite an IIR filter. This IIR filter has a symmetric impulse response, and all its poles are always inside the unit circle. Thus, the spline coefficients $c[n]$ can be determined stably via a few recursive equations [2, 7].

The above spline interpolation leads us to a new approach for ESA-based Demodulation Algorithm whose basic steps are the following: (i) By using smoothing splines, the original discrete-time signal $x[n]$ is interpolated to create a continuous-time expansion $s_\nu(t)$. (ii) The continuous-time energy operator Ψ and the continuous ESA are applied to the continuous-time signal $s_\nu(t)$. This requires computing $\Psi[s_\nu(t)]$ and $\Psi[\partial s_\nu(t)/\partial t]$, which in turn require the derivatives $\partial^r s_\nu(t)/\partial t^r$ for $r = 1, 2, 3$. We can obtain closed-form expressions for these derivatives that involve only the coefficients $c[n]$ and the B-spline functions [2]. The continuous ESA (3) estimates the instantaneous amplitude $a(t)$ and frequency $\omega(t)$ of the continuous signal $s_\nu(t)$. (iii) The information-bearing signals $a(t), \omega(t)$ are sampled to obtain estimates of the instantaneous amplitude $A[n] = a(nT)$ and frequency $\Omega[n] = T\omega(nT)$ of the original discrete signal $x[n]$. This whole approach above is called the **Spline-ESA**.

By setting $\nu = 5$, the time-window (i.e., the number of input samples required to produce one output sample) of Spline-ESA becomes the same with that of the DESA. Extensive comparisons [2] between the Spline-ESA (with $\nu = 5$ and λ fixed to a constant value in the order of 0.5) versus the DESA have demonstrated that, while both algorithms perform well in signal-plus-noise environments with high SNRs, the Spline-ESA outperforms the DESA in low SNRs. This robustness in the presence of noise, is the main advantage of the Spline-ESA.

The ESAs are efficient demodulation algorithms only when they are used on narrowband AM-FM signals [1]. This constraint makes the use of *filterbanks* (i.e., parallel arrays of bandpass filters) inevitable for wideband signals like speech. Thus, each short-time segment (analysis frame) of a speech signal is simultaneously filtered by all the bandpass filters of the filterbank, and then each filter output is demodulated using the ESA. In our on-going research on speech analysis and recognition [5, 6] we have been using filter-

banks with Gabor bandpass filters whose center frequencies are spaced on a mel-frequency scale. Figure 1 shows an example of demodulating three bands of a speech phoneme into their instantaneous amplitude and frequency signals.

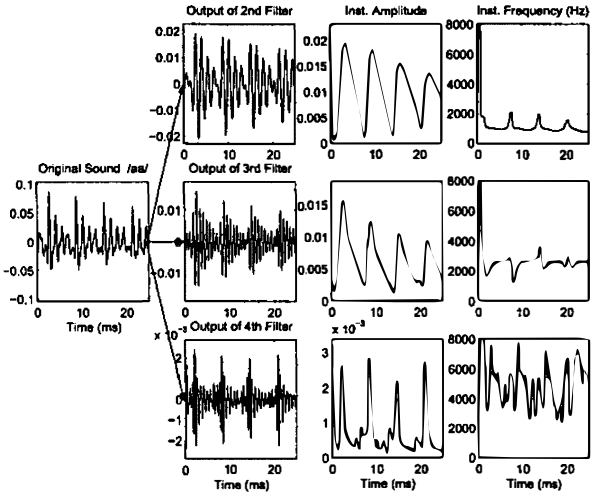

Fig. 1. Demodulating a speech phoneme using a Gabor filterbank and the Spline-ESA.

3. MODULATION FEATURE EXTRACTION AND PHONEME RECOGNITION

The feature vectors used in speech recognition, are typically computed over a 20-30 ms window and are updated every 5-10 ms. The 'standard' feature set consists of the first twelve *mel-frequency cepstrum coefficients (MFCC)*, the mean-square amplitude(i.e. energy) of the signal and their first and second time derivatives.

We shall augment the 'standard' feature vector and thus create a *hybrid feature vector* by incorporating information from the nonlinear structure of speech of the modulation type as additional features. We use feature vectors that contain information both from the smoothed cepstrum of the linear model, as well as from the speech modulations.

We have used the hybrid feature vector as input to a hidden Markov model (HMM)–based speech recognizer. The HMM recognizer is the HTK system [8]. In the experiments presented below, context-independent, 3-state, left-right HMMs and unweighted word-pair grammar were used. The input vectors are split into two different data streams, one for the standard features (MFCC) and the other for the modulation features. The modulation features are assumed to be independent of the linear features and to belong to separate probability 'streams'. Each one of these streams has an independent probability distribution. These distributions are modelled by a certain number of Gaussian mixture probability densities, called mixture components [8].

We have experimented with a broad range for the number of Gaussian mixture densities, but we are presenting the recognition results only for the cases of 8 and 16 mixtures, since these values are the most representative. Stream-weights affect directly the recognition process. Our experiments have shown that the best recognition results are obtained when the data-stream weights are equal and sum up to one, so we set the weights each equal to 0.5.

The experiments were made over the TIMIT database. This database consists of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of US. All speech signals in TIMIT are sampled at 16 kHz. The training set consists of 3696 sentences and the test set by 1344 sentences. Each one of these sentences was segmented into 25-ms speech frames, whose update period was 10 ms. The (linear and nonlinear) feature sets were extracted from each such frame.

We have automated the extraction of modulation features from speech signals in the following way: First, we use a parallel filterbank of overlapping Gabor bandpass filters whose center frequencies are spaced on a mel-frequency scale. Second, the output signals from each Gabor bandpass filter are demodulated via the Spline-ESA into their instantaneous amplitude $a(t)$ and frequency $f(t)$ component signals. For each such short-time analysis frame and for each band, the weighted mean F_w and standard deviation B_w of the instantaneous frequency signal are estimated as in [5]:

$$F_w \equiv \frac{\int_{t_0}^{t_0+T} f(t)a^2(t)dt}{\int_{t_0}^{t_0+T} a^2(t)dt} \quad (5)$$

$$B_w \equiv \frac{\int_{t_0}^{t_0+T} [(\dot{a}(t)/2\pi)^2 + (f(t)-F_w)^2 a^2(t)]dt}{\int_{t_0}^{t_0+T} a^2(t)dt} \quad (6)$$

where t_0 and T are the start and duration of the analysis frame, respectively. Next, we compute the *frequency modulation (FM) percentage* in each band, as the ratio $K = B_w/F_w$. For each analysis frame, the FM percentages K_i, $i = 1, ..., L$, are computed, one for each narrowband speech component, where L is the number of filters in the filterbank. The modulation feature set consists of the sequence of the FM percentages K_i and their first and second time derivatives. This is a total of $3L$ numbers per frame. We have experimented with mel-spaced filterbanks consisting of $L = 12$ and $L = 6$ Gabor filters, spanning the whole frequency range and overlapping by 50%. We have used these modulation feature vectors to augment the standard feature vectors employed in speech recognition tasks. The augmented hybrid feature set consists of the standard and the modulation feature set. The two different feature subsets are treated as separate streams by the HTK system with independent probability distributions.

Word Percent Correct [1]		
# Gaussian Mixtures	MFCC	MFCC+FM
5	67.15	79.12
8	73.95	84.31
10	77.33	85.51
12	77.95	85.76
16	78.76	86.83

Table 1. Recognition Results

Table 1 reports the word recognition results over the TIMIT database using either only the standard features (column MFCC) or the augmented standard-plus-modulation features (column MFCC+FM). Clearly, our experiments on word-recognition by augmenting the standard feature set with modulation information show a significant improvement over using only the standard features. This relative error rate reduction approaches 40% when using 8 Gaussian mixtures. Thus, the FM modulation percentage features provide an improvement to the recognition performance with a moderate increase in the size of the feature vector.

The results in Table 1 refer to the case of a 6-channel filterbank (i.e. 18 modulation features); hence, the augmented feature set has a size of 57. We have experimentally found that measuring the modulations in the outputs of only 6 Gabor filters, yields better recognition results than using 12 filters. For example, the correct word recognition for the 12-channel filterbank was 80.96% (using 8 Gaussian mixtures) compared to 84.3% for 6 channels. Note that the 12-channel case employs a larger feature vector of size 75 despite its inferior recognition performance. This difference in the recognition rates, can be explained based on the modulation model for speech resonances. In the 12-channel case, the large number of filters causes each bandpass filter to have a narrower bandwidth and hence pass a smaller part of the AM-FM modulation structure of the neighbor speech resonances. In contrast, the filters in the 6-channel filterbank, have a wider bandwidth and hence keep a richer part of the modulation information.

4. CONCLUSIONS

In this paper, we have described how to apply an efficient nonlinear DSP algorithm to speech signals in order to extract novel acoustic features related to their nonstationary dynamics of the modulation type. Further we have developed a hybrid feature set for speech recognition that includes both the standard linear features as well as the new nonlinear features and applied this new feature set to HMM-based word recognition. Our experimental results have shown a significant improvement in recognition over the TIMIT database.

Given the relation of the underlying nonstationary models to the physics and true dynamics of speech production and given the efficiency of the nonlinear DSP algorithm we developed to extract the corresponding nonlinear features, we believe that the modulation models and related nonlinear algorithm have a strong potential in speech recognition.

In the near future, we intend to apply the modulation features for speech recognition in noisy environments and for large vocabulary speech recognition. Regarding these modulation features, the Spline-ESA can offer robustness in the speech demodulation problem.

[1] The percentage number of words correctly recognized is given by the ratio of the number of correct labels to the total number of words in the defining transcription files.

5. REFERENCES

[1] A. C. Bovik, P. Maragos, and T.F. Quatieri, "AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators", *IEEE Trans. Signal Processing*, vol. 41, Dec. 1993.

[2] D. Dimitriadis and P. Maragos, "An Improved Energy Demodulation Algorithm Using Splines", *Proc. ICASSP-01*, Salt Lake, Utah, May 2001.

[3] D. Dimitriadis, P. Maragos, V. Pitsikalis and A. Potamianos, "Modulation and Chaotic Acoustic Features for Speech Recognition", *J. Control and Intelligent Systems*, Invited Paper, accepted for publication, 2002.

[4] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.

[5] A. Potamianos and P. Maragos, "Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Demodulation", *J. Acoust. Soc. Amer.*, 99 (6), pp.3795–3806, June 1996.

[6] A. Potamianos and P. Maragos, "Time-Frequency Distributions for Automatic Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol.9, pp.196-200, Mar. 2001.

[7] M. Unser, A. Aldroubi and M. Eden, "B-Spline signal processing: Part I–Theory. Part II–Efficient design and applications" *IEEE Trans. Signal Processing*, vol. 41, pp. 821–848, Feb. 1993.

[8] S. Young, *The HTK Book*, Cambridge Research Lab: Entropics, Cambridge, England, 1995.