# Robust AM-FM Features for Speech Recognition

Dimitrios Dimitriadis, *Member, IEEE*, Petros Maragos, *Fellow, IEEE*, and Alexandros Potamianos, *Member, IEEE*

***Abstract*—In this letter, a nonlinear AM-FM speech model is used to extract robust features for speech recognition. The proposed features measure the amount of amplitude and frequency modulation that exists in speech resonances and attempt to model aspects of the speech acoustic information that the commonly used linear source-filter model fails to capture. The robustness and discriminability of the AM-FM features is investigated in combination with mel cepstrum coefficients (MFCCs). It is shown that these hybrid features perform well in the presence of noise, both in terms of phoneme-discrimination (J-measure) and in terms of speech recognition performance in several different tasks. Average relative error rate reduction up to 11% for clean and 46% for mismatched noisy conditions is achieved when AM-FM features are combined with MFCCs.**

***Index Terms*—AM-FM, ASR, features, nonlinear, speech.**

## I. Introduction

DESPITE the intense research activity, automatic speech recognition (ASR) systems do not yet exhibit acceptable performance in many real-life environments. Robust ASR is an active research field, and a variety of algorithms can be used to improve speech recognition performance under adverse conditions, including speech enhancement techniques, robust feature extraction, and model compensation. In this letter, we focus on robust feature extraction schemes.

Motivated by strong evidence for the existence of amplitude and frequency (AM-FM) modulations in speech signals [5], a speech resonance can be modeled by an AM-FM signal

$$r_i(t) = a_i(t)\cos\left(2\pi\int_0^t f_i(\tau)d\tau\right) \quad (1)$$

and correspondingly the total speech signal as a superposition of a small number of such AM-FM signals [9]. We have found that speech sounds are better modeled by six such AM-FM signals (roughly corresponding one to each resonance). The estimation of their instantaneous frequencies $f_i(t)$ and amplitude envelopes $|a_i(t)|$ is referred to as the "Demodulation Problem" and is significant for speech applications.

This letter deals with both extracting speech features inspired by the AM-FM model and applying them to speech classification and recognition tasks. Other work related to Teager energy or AM-FM feature extraction can be found in [3] and [10] dealing with the speaker recognition tasks and in [4], [6], and [11] for the ASR problem. The proposed features measure instantaneous amplitude and frequency modulation; such information is not captured by the linear source-filter model and the mel cepstrum coefficients (MFCCs). In addition, modulation features are expected to be resistant to noise (see Tables I–III). The main contribution of this letter is to show that the proposed modulation features, when combined with the MFCCs, are robust to noise. The feature robustness is demonstrated both in terms of phoneme-class discrimination (J-measure) and in terms of improvement of ASR performance for several speech databases, i.e., the Aurora-3 and the TIMIT-based tasks.

TABLE I
J-Measure Estimates for Feature Sets on Vowels and Fricatives. (a) Vowels and (b) Fricatives

| Vowels | | | | | | |
|---|---|---|---|---|---|---|
| Features \ SNR | -5 dB | 0 dB | 5 dB | 10 dB | 20 dB | clean |
| MFCC | 2.84 | 2.98 | 3.05 | 3.07 | 3.09 | 3.09 |
| MFCC+IA-Mean | 2.97 | 3.14 | 3.22 | 3.25 | 3.26 | 3.27 |
| MFCC+IF-Mean | 2.92 | 3.12 | 3.23 | 3.29 | 3.35 | 3.36 |
| MFCC+FMP | 2.94 | 3.10 | 3.16 | 3.21 | 3.27 | 3.31 |

(A)

| Fricatives | | | | | | |
|---|---|---|---|---|---|---|
| Features \ SNR | -5 dB | 0 dB | 5 dB | 10 dB | 20 dB | clean |
| MFCC | 4.65 | 5.02 | 5.23 | 5.28 | 5.24 | 5.19 |
| MFCC+IA-Mean | 4.96 | 5.37 | 5.60 | 5.67 | 5.62 | 5.56 |
| MFCC+IF-Mean | 4.73 | 5.15 | 5.37 | 5.43 | 5.40 | 5.38 |
| MFCC+FMP | 4.74 | 5.16 | 5.38 | 5.43 | 5.34 | 5.30 |

(B)

TABLE II
Correct Word Accuracies (%) for Modulation Features on the Aurora-3 (Spanish Task) Database

| Features \ Scenario | WM | MM | HM | Average | Av. Rel. Improv. |
|---|---|---|---|---|---|
| Aurora Frontend (WI007) | 92.94 | 80.31 | 51.55 | 74.93 | - |
| MFCC+CMS (Baseline) | 93.68 | 92.73 | 65.18 | 83.86 | 35.62 |
| MFCC+CMS+IA-Mean | 93.22 | 91.35 | 71.35 | 85.31 | 41.40 |
| MFCC+CMS+IF-Mean | 90.71 | 89.52 | 72.36 | 84.20 | 36.98 |
| MFCC+CMS+FMP | 94.38 | 92.46 | 72.79 | 86.54 | 46.31 |

TABLE III
Correct Phoneme Accuracies (%) for Modulation Features on the TIMIT Tasks

| Phoneme Accuracy for the TIMIT Tasks (%) for SNR=10 dB | | | | | | | |
|---|---|---|---|---|---|---|---|
| | TIMIT | NTIMIT | TIMIT +Babble | TIMIT +White | TIMIT +Pink | TIMIT +Car | Av. Rel. Improv. |
| MFCC | 58.40 | 42.42 | 27.71 | 17.72 | 18.60 | 52.75 | - |
| MFCC+ IA-Mean | 59.61 | 43.53 | 39.25 | 26.03 | 31.05 | 56.50 | 17.62 |
| MFCC+ IF-Mean | 59.41 | 43.70 | 38.56 | 26.05 | 32.81 | 56.75 | 19.13 |
| MFCC+ FMP | 59.92 | 43.69 | 38.60 | 26.15 | 32.84 | 55.97 | 18.17 |

Manuscript received December 30, 2004; revised March 29, 2005. This work was supported by the FP6 European programs HIWIRE and MUSCLE and by the NTUA research program Protagoras. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Bastiaan Kleijn.

D. Dimitriadis and P. Maragos are with the School of Electrical and Computer Engineering, National Technical University of Athens, Athens 15773, Greece (e-mail: ddim@cs.ntua.gr; maragos@cs.ntua.gr).

A. Potamianos is with the Department of Electronics and Computer Engineering, Technical University of Crete, Chania, GR 73100, Crete, Greece (e-mail: potam@telecom.tuc.gr).

Digital Object Identifier 10.1109/LSP.2005.853050

## II. Feature Extraction

The AM-FM model suggests the decomposition of speech signals into a series of a few instantaneous frequency and amplitude signals. These signals can be considered as time-frequency distributions [7], containing acoustic information that is not captured by the linear speech model. In [1], preliminary ASR results have indicated that a significant part of the acoustic information cannot be modeled by the linear source-filter acoustic model, and thus, the need for nonlinear features becomes apparent. These features, which are based on either the FM or the AM part, provide additional acoustic information. The modulation features have two major advantages compared to the linear MFCCs. They can model the dynamic nature of speech and capture some of its fine structure and its rapid fluctuations. Second, they appear to be relatively noise resistant and, thus, yield improved results, especially for speech recognition in noise, when a mismatch in the training and testing conditions is present.

### A. Modulation Features

The AM-FM model suggests that the formant frequencies are not constant during a single pitch period, but they can vary around a center frequency. These variations are partly captured by the *Frequency Modulation Percentages* (FMP) features defined as $FMP_i = B_i/F_i$ for each speech resonance $i$, where $B_i$ is the mean bandwidth (an amplitude-weighted version of the $f_i(t)$-signal deviation), and $F_i$ is the weighted mean frequency value of resonance $i$. $F_i$ and $B_i$ provide more accurate and smooth estimates [8], and they are derived as follows from the information signals $a_i(t)$ and $f_i(t)$:

$$F_i = \frac{\int_0^T f_i(t) a_i^2(t)dt}{\int_0^T a_i^2(t)dt},$$
$$B_i = \frac{\int_0^T [\dot{a}_i^2(t) + (f_i(t) - F_i)^2 a_i^2(t)]dt}{\int_0^T a_i^2(t)dt} \quad (2)$$

where $i = 1, \ldots, 6$ is the speech resonance index, and $T$ is the time window length. Another frequency-related feature investigated in this letter is the short-time weighted mean of the instantaneous frequency signal $f_i(t)$, i.e., the *Instantaneous Frequency Mean* (IF-Mean). The proposed features provide information about the speech formant fine structure taking advantage of the excellent time resolution of the ESA. Transitional phenomena and instantaneous formant variations are mapped onto these FM features.

Next, we attempt to model the fine structure of the amplitude envelope signal (AM) with the *Mean Instantaneous Amplitude* (IA-Mean) features that are defined as the short-time mean of the instantaneous amplitude signal $|a_i(t)|$ for each speech resonance $i$. The IA-Mean features parametrize the resonance amplitudes and capture part of the nonlinear behavior of speech, e.g., the modulation pulses appearing within a single pitch period.

### B. Feature Extraction Algorithm

The AM-FM features are computed from the instantaneous frequency and amplitude signals of each speech resonance. To extract the resonance signals $r_i(t)$, we used a fixed six-filter mel-spaced Gabor filterbank. The Gabor filters were chosen for several reasons listed in [5], including their optimal time-frequency discriminability. The filter placing and bandwidths were dictated by the mel-scale and the need for fixed overlap. We used a bandwidth overlap of adjacent filters equal to 50%. Once the resonance signals $r_i(t)$ are extracted, they are demodulated and the $f_i(t)$, $|a_i(t)|$ are obtained.

Among the various demodulation approaches to estimate the model parameters of a single resonance, we use the Energy Separation Algorithm (ESA), due to its excellent time resolution and low complexity [5]. The ESA estimates of the instantaneous frequency and amplitude signals are given by $f(t) \approx (1/2\pi)\sqrt{\Psi[\dot{x}(t)]/\Psi[x(t)]}$ and $|a(t)| \approx \Psi[x(t)]/\sqrt{\Psi[\dot{x}(t)]}$, where $\Psi[x] = \dot{x}^2 - x\ddot{x}$. There is also an ESA for discrete-time AM-FM signals [5]. In this letter, we use a more robust ESA, where the discrete-time signal is expanded over the continuous-time domain, and then, the continuous-time ESA is applied upon. This approach combines differentiation and Gabor filtering of the signal into convolutions of the signal with time derivatives of the filter impulse response. The advantages of such an approach is that one can avoid the noisy one-sample discrete-time approximations of the derivatives and achieve smoother estimates of the signal's derivatives, even in the presence of noise. Since the convolution operation commutes with time differentiation, we can combine the operator $\Psi$ and the bandpass filtering [1]

$$\Psi[x(t) * g(t)] = \left[x(t) * \frac{dg(t)}{dt}\right]^2 - (x(t) * g(t))\left[x(t) * \frac{d^2 g(t)}{dt^2}\right] \quad (3)$$

where $x(t)$ is the input signal, and $g(t)$ is the Gabor impulse response. In this approach, the necessary processes of bandpass filtering and the subsequent differentiations are combined into a single convolution of the speech signals with the time derivatives of the Gabor impulse response. So, a filtering/demodulation algorithm is obtained, which we call *Gabor ESA*. This algorithm exhibits important advantages compared to the original discrete demodulation algorithm DESA [1], most notably smoother instantaneous estimates.

To obtain robust AM-FM features, it is crucial that the demodulation algorithm can provide smooth and accurate estimates for $f_i(t)$ and $|a_i(t)|$. There are cases when the demodulation algorithms presented above provide estimates that have singularities and spikes that should be eliminated before the feature measurement process. For this purpose, a binomial smoothing of the energy signals is done to smooth out the highpass modeling error of the ESA. Also, a post-processing scheme is applied upon the demodulated instantaneous signals that employs a median filter with a short window.

## III. Classification and Recognition

### A. Classification Results

In this section, we investigate the phoneme classification properties of the proposed features. The ability of each of these hybrid features to discriminate between phoneme classes is compared to that of the "standard" MFCCs, both for clean and noisy conditions. For this purpose, we have used the linear

*Fisher Discriminant* [2] and the corresponding J-measure, which is the ratio of the interclass scatter divided by the intraclass scatter. The larger the value of this ratio, the better the discrimination of the classes in the feature space. We have used the maximum J-value of the Fisher Discriminant, which equals $J = \mathrm{trace}\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)$, where $\mathbf{S}_W$ and $\mathbf{S}_B$ are the within-class and the between-class scatter matrix, respectively.

Several different combinations of phoneme classes have been tested to examine the discriminability of the proposed features. Herein, we present the J-measure estimates only for two different groups of phonemes, the vowels ($/iy/$, $/ih/$, $/eh/$, $/ae/$, $/aa/$, $/uh/$, $/uw/$, $/ah/$, and $/er/$) and the fricatives ($/f/$, $/s/$, $/sh/$, $/v/$, $/th/$, and $/z/$). The J-measure was computed for both clean and noise-corrupted instances of the phonemes. Specifically, white Gaussian noise of different signal-to-noise ratio (SNR) values was added to the clean speech signals to test the degradation of class discrimination under adverse conditions. For the estimation of the J-measures, smoothed-out versions of both the proposed modulation features and the MFCC features are used. Instances of the phonemes are extracted from the TIMIT database according to the given transcriptions. The steady-state part of the phoneme is extracted (middle one third) and the features are estimated over this segment (i.e., the steady state of the phoneme is assumed as one large speech frame). In this case, the dynamic time-varying phenomena present in speech are not taken under consideration, and they are partly smoothed out. However, the proposed scheme can clearly demonstrate the ability of the various features to discriminate among phonemes in the presence of noise.

In the case of ASR tasks, we augment the linear feature vectors with the robust nonlinear features to improve the feature robustness to noise [6], [9]. Therefore, we have tested the discriminability of the augmented features according to the J-measures. The input vectors are derived by the concatenation of the MFCCs with the modulation features. InTable I, the J-measure estimates are shown for the augmented feature vectors, for both clean and noisy conditions. The MFCCs are strongly affected by noise, and their classification properties deteriorate rapidly as the SNR level decreases. The J-measure values of the proposed hybrid features appear to be more robust in the presence of additive noise in both cases.

### B. Recognition Results

We have applied the proposed features to the Aurora-3 Speech Database (Spanish task), which is a word-level recognition task, and to the TIMIT-based tasks, which are phoneme-based recognition tasks. We created the "TIMIT + Noise" databases by adding babble, white, pink, and car noise to the test set of the TIMIT database, which is sampled at 16 kHz; the SNR level was set equal to 10 dB. The Aurora-3 Database contains recordings, sampled at 8 kHz, from two different microphones, at three driving conditions. These recordings are mixed to create three different training/testing scenarios: the *Well-Matched* scenario, the *Medium-Mismatch* scenario, where the mismatch is mainly due to the usage of different microphones, and the *High-Mismatch* scenario with different noise levels in the training and the testing sets. The ASR experiments have been performed using the hidden Markov model (HMM)-based HTK Tools system [12]. Context-independent, 14-state, left-right word HMMs with 16 Gaussian mixtures are used. For the TIMIT-based tasks, the phoneme HMMs are three-state, left-right with 16 mixtures. The grammar used for both cases is the all-pair, unweighted grammar. Finally, for the TIMIT + Noise cases, the HMMs are trained in the clean speech training set and tested in the noise-corrupted versions of the testing set.

The input vectors are split into two different data streams: one for the standard MFCCs and the other for the modulation features. The data streams are assumed independent. The augmented feature vector consists of 57 coefficients, 39 samples for the "standard" features (normalized energy, MFCCs, first and second time derivatives), and 18 for the modulation features (six coefficients plus their first and second time derivatives). For the Aurora-3 database, Cepstral Mean Subtraction (CMS) is applied to the MFCCs to combat convolutional mismatches. The frame length is set equal to 30 ms with frame period equal to 10 ms. The weights of the two independent data streams are optimized on held-out data. In practice, the stream weight for the AM-FM features decreases with the SNR level, which is another indication of the robustness of the proposed features. More specifically, for the clean case, i.e., the TIMIT task, the stream weights are set $s_1 = 1.00$ and $s_2 = 0.20$ for the MFCCs and the modulation features, correspondingly. For the low SNR cases, the stream weights are $s_1 = 1.00$ and $s_2 = 0.50$ or $s_2 = 1.00$, depending on the noise level. In Tables II and III, the recognition results are presented for the Aurora-3 and the TIMIT tasks, respectively. By combining MFCCs with AM-FM features, we achieve a performance improvement for the clean and especially for the noisy conditions. The improvement is larger for the HM-scenario of the Aurora-3 database and the TIMIT + Noise tasks, where additive noise is the main source of degradation. On the other hand, for the NTIMIT and the Aurora-3 WM-, MM-scenarios, where the convolutional noise is dominant, the hybrid features yield modest results.

## IV. Conclusion/discussion

In this letter, new robust modulation features have been proposed. These features are mainly the first-order statistics (mean values) of the demodulated instantaneous signals, and they provide robustness to additive noise tasks but less so for convolutional noise. Relative error rate reduction up to 46% for mismatched noisy conditions is achieved when these features are combined with the MFCCs.

The modulation features can model more accurately the voiced speech signals, but they appear to have good results even for the unvoiced speech case. We have presented strong indications that modulation features can model and classify different phoneme classes better and more efficiently than the classic MFCC features, especially in the presence of additive noise. In our ongoing research, we are investigating 1) the usefulness of second-order statistics of the modulation signals, 2) ways of optimally combining linear and modulation features for ASR tasks, and 3) the relation of unvoiced speech with the AM-FM model.

## REFERENCES

[1] D. Dimitriadis and P. Maragos, "Robust energy demodulation based on continuous models with application to speech recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 2853–2856.

[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 2001.

[3] H. Ezzaidi and J. Rouat, "Comparison of MFCC and pitch synchronous AM, FM parameters for speaker identification," in *Proc. ICSLP*, vol. 2, Beijing, China, Oct. 2000, pp. 318–321.

[4] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," *IEEE Signal Process. Lett.*, vol. 6, no. 10, pp. 259–261, Oct. 1999.

[5] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.

[6] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. ICASSP*, Seattle, WA, May 1998, pp. 617–620.

[7] J. W. Pitton, K. Wang, and B. H. Juang, "Time-frequency analysis and auditory modeling for automatic recognition of speech," *Proc. IEEE*, vol. 84, no. 9, pp. 1199–1215, Sep. 1996.

[8] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3795–3806, Jun. 1996.

[9] ——, "Time-frequency distributions for automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 196–200, Mar. 2001.

[10] T. F. Quatieri, C. R. Jankowski Jr, and D. A. Reynolds, "Energy onset times for speaker identification," *IEEE Signal Process. Lett.*, vol. 1, no. 11, pp. 160–162, Nov. 1994.

[11] H. Tolba and D. O'Shaughnessy, "Automatic speech recognition based on cepstral coefficients and a mel-based discrete energy operator," in *Proc. ICASSP*, vol. 2, Seattle, WA, May 1998, pp. 973–976.

[12] S. Y. Young *et al*. The HTK Book (for HTK Version 3.2) [Online]. Available: http://htk.eng.cam.ac.uk