Robust AM-FM Features for Speech Recognition

Dimitris V. Dimitriadis, *Member, IEEE,* Petros Maragos, *Fellow, IEEE,*
and Alexandros Potamianos, *Member, IEEE,*

Abstract

In this paper, 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 the AM-FM 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.

EDICS Category: SPL.SPEE

Corresp. Author: D. Dimitriadis, **Email**: ddim@cs.ntua.gr, **Tel.**: +30-210-7722964, **Fax**: +30-210-7722281
Address: School of ECE, National Technical University of Athens, Athens 15773, Greece

This work was supported by the European program MUSCLE and by the NTUA research program Protagoras.

D. Dimitriadis and P. Maragos are with the National Technical University of Athens.

A. Potamianos is with the Technical University of Crete.

Robust AM-FM Features for Speech Recognition

I. INTRODUCTION

Despite voluminous recent 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 work, we focus on robust feature extraction schemes.

Motivated by strong evidence for the existence of amplitude and frequency (AM-FM) modulations in speech signals [4], a speech resonance can be modeled by an AM-FM signal,

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

and correspondingly the total speech signal as a superposition of a small number of such AM-FM signals (one for each formant). Most often, the number of observable formants does not exheed the 6, and henceforth the speech sounds will be modeled by 6 such AM-FM signals. 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 paper 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 AM-FM feature extraction can be found in [3], [7], though they deal with Speaker Recognition problems. Herein, the proposed features measure instantaneous amplitude and frequence modulation; such information is not captured by the linear source-filter model and the MFCC features. In addition, modulation features (especially FM) are expected to be resistant to noise [7]. The main contribution of this paper is to show that the proposed modulation features are robust in the presence of adverse recording conditions (and especially additive noise). The robustness of the features is demonstrated both in terms of phoneme-class discrimination (J-measure) and in terms of improvement of speech recognition performance for several speech databases, i.e. the Aurora-3, TIMIT, NTIMIT and TIMIT+Noise tasks.

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 containing acoustic information that is not visible in the linear part of the speech spectrum. In [1] preliminary ASR results have indicated that 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 MFCC features. 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 (a weighted version of the $f_i(t)$-signal deviation [6]) and F_i is the (amplitude) weighted mean frequency value of resonance i. F_i and B_i are estimated as follows from the information signals $a_i(t)$ and $f_i(t)$

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

where $i = 1, \ldots, 6$ is the speech resonance index and T the time window length. Another frequency-related feature investigated in this paper 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) feature set that is 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 behaviour of speech, e.g. the modulation pulses appearing within a single pitch period.

B. Feature Extraction Algorithm

Two significant parts of the feature extraction system are the filterbank and the single-band demodulation algorithm. The AM-FM features are computed from the instantaneous frequency and amplitude signals of each speech resonance. To extract the resonance signals $r_i(t)$ a fixed 6-filter mel-spaced Gabor filterbank is used. The Gabor filters are chosen for several reasons listed in [4], including their optimal time-frequency discriminability. The filter placing and bandwidths are dictated by the mel-scale and the need for constant-Q filterbanks. The bandwidth overlap of adjacent filters is fixed and 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 [4]. This is based on the continuous-time Teager-Kaiser energy operator (TEO) $\Psi = \dot{x}^2 - x\ddot{x}$. 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 $|a(t)| \approx \Psi[x(t)]/\sqrt{\Psi[\dot{x}(t)]}$.

There is also an ESA for discrete-time AM-FM signals [4]. In this paper, 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 Ψ and the bandpass filtering [1]:

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

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.

In order 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 produce estimates that have singularities and spikes which 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 the AM-FM feature sets to discriminate between phoneme classes is compared with that of the 'standard' MFCC feature set, both for clean and noisy conditions. For this purpose, we used the linear *Fisher Discriminant* (J-measure) [2]. The J-measure is defined as the product of the inverse inter-class and the intra-class matrices. The larger the value of this ratio the better the discrimination of the classes in the feature space. The Fisher Discriminant is defined in [2] as: $J(\mathbf{w}) = (\mathbf{w}^T\mathbf{S}_B\mathbf{w})/(\mathbf{w}^T\mathbf{S}_W\mathbf{w})$, where $\mathbf{w}$ ($||\mathbf{w}|| = 1$) is the optimal projection vector that maximizes J. The maximum J-value is $J_{max} = trace\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)$ where the within-class scatter $\mathbf{S}_W$ and the between-class scatter $\mathbf{S}_B$ are given by

$$\mathbf{S}_W = \sum_{k=1}^{N} N_k \mathbf{\Sigma}_k, \qquad \mathbf{S}_B = \sum_{k=1}^{N} N_k (\mathbf{m}_k - \mathbf{m}_0)\,(\mathbf{m}_k - \mathbf{m}_0)^T$$

where $\mathbf{m}_0$ the mean vector of the entire feature set, N the number of classes, N_k the number of samples, $\mathbf{m}_k$ and $\mathbf{\Sigma}_k$ the mean vector and the covariance matrix of the k^{th} class.

Several different combinations of phoneme classes have been tested in order to examine the discriminability of the proposed features. We present the J-measure estimates 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 SNR values was added to the clean speech signals in order 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 feature sets 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, in order to improve the feature robustness to noise. 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. In Table 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 modulation features appear to be more robust in the presence of additive noise in both cases.

TABLE I

J-MEASURE ESTIMATES FOR FEATURE SETS ON VOWELS AND FRICATIVES. (A) VOWELS AND (B) FRICATIVES.

	(a) Vowels						(b) Fricatives					
SNR / Features	-5 dB	0 dB	5 dB	10 dB	20 dB	clean	-5 dB	0 dB	5 dB	10 dB	20 dB	clean
MFCC	2.84	2.98	3.05	3.07	3.09	3.09	4.65	5.02	5.23	5.28	5.24	5.19
MFCC+IA-Mean	2.97	3.14	3.22	3.25	3.26	3.27	4.96	5.37	5.60	5.67	5.62	5.56
MFCC+IF-Mean	2.92	3.12	3.23	3.29	3.35	3.36	4.73	5.15	5.37	5.43	5.40	5.38
MFCC+FMP	2.94	3.10	3.16	3.21	3.27	3.31	4.74	5.16	5.38	5.43	5.34	5.30

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. The 'TIMIT+Noise' databases are created 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 is set equal to 10 dB. The Aurora-3 Database contains recordings, sampled at 8 kHz, from close-talking and distant microphones, at 3 driving conditions with average SNR values of 12, 9 and 5 dB, correspondingly. These recordings are mixed to create 3 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 HMM-based HTK Tools system [8]. Context-independent, 14-state left-right word HMMs were used; each state contains 16 gaussian mixtures. For the TIMIT recognition tasks, the HMM models are 3-state, left-right HMMs with 16 mixtures. The grammar used for both cases is the all-pair, unweighted grammar. Finally, for the TIMIT+Noise cases, the HMM models 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 features MFCC and the other for the modulation-based features. The data streams are assumed independent. The augmented feature vector consists of 57 coefficients, 39 samples for the 'standard' features (normalized energy, MFCCs, 1^{st} and 2^{nd} time-derivatives) and 18 for the modulation features (6 coefficients plus their 1^{st} and 2^{nd} time-derivatives). Cepstral Mean Subtraction (CMS) is applied to the standard feature stream only for the Aurora-3 database to combat convolutional mismatches (for the MM-scenario there is microphone mismatch). The frame length is set equal to 30 msec with frame-period equal to 10 msec. 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, 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 Table II

TABLE II

CORRECT WORD ACCURACIES (%) FOR MODULATION FEATURES ON THE AURORA-3 (SPANISH TASK) DATABASE.

Scenario / Features	WM	MM	HM	Average	Aver. 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

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 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 modulation features yield modest results.

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	Aver. 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

IV. CONCLUSION-DISCUSSION

In this paper, new robust modulation features have been proposed. The proposed features are mainly the 1^{st}-order statistics (mean values) of the demodulated instantaneous signals. The AM-FM features 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 MFCCs.

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 on-going research we are investigating (i) the usefulness of 2^{nd}-order statistics of the modulation signals, and (ii) ways of optimally combining linear and modulation features for ASR tasks.

REFERENCES

[1] D. Dimitriadis and P. Maragos, "Robust Energy Demodulation Based on Continuous Models with Application to Speech Recognition", in *Proc. of Eurospeech-03*, Geneva, Sept. 2003.

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

[3] H. Ezzaidi and J. Rouat, "Comparison of MFCC and Pitch Synchronous AM, FM Parameters for Speaker Identification", in *Proc. ICSLP-00*, Beijing, Oct. 2000.

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

[5] M. D. Plumpe, T. F. Quatieri and D. A. Reynolds, "Modeling of the Glotal Flow Derivative Waveform with Application to Speaker Identification", *IEEE Trans. on Speech and Audio Processing*, vol 7, no. 5, Sept. 1999.

[6] 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.

[7] T. F. Quatieri, "Discrete-Time Speech Signal Processing", *Prentice Hall*, Upper Saddle River, NJ, 2002.

[8] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, “The HTK Book (for HTK Version 3.2)”, Cambridge Research Lab: Entropics, Cambridge, England, 2002.