Short-time Instantaneous Frequency and Bandwidth Features for Speech Recognition

Pirros Tsiakoulis [#1], Alexandros Potamianos [*2], Dimitrios Dimitriadis [#3]

[#] *School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece*
[1] ptsiak@ilsp.gr
[3] ddim@cs.ntua.gr

[*] *Department of Electronics and Computer Engineering, Technical University of Crete, Chania, Greece*
[2] potam@telecom.tuc.gr

***Abstract*—In this paper, we investigate the performance of modulation related features and normalized spectral moments for automatic speech recognition. We focus on the short-time averages of the amplitude weighted instantaneous frequencies and bandwidths, computed at each subband of a mel-spaced filterbank. Similar features have been proposed in previous studies, and have been successfully combined with MFCCs for speech and speaker recognition. Our goal is to investigate the stand-alone performance of these features. First, it is experimentally shown that the proposed features are only moderately correlated in the frequency domain, and, unlike MFCCs, they do not require a transformation to the cepstral domain. Next, the filterbank parameters (number of filters and filter overlap) are investigated for the proposed features and compared with those of MFCCs. Results show that frequency related features perform at least as well as MFCCs for clean conditions, and yield superior results for noisy conditions; up to 50% relative error rate reduction for the AURORA3 Spanish task.**

***Index Terms*—AM–FM, speech recognition, instantaneous frequency, instantaneous bandwidth, filterbank overlap**

I. Introduction

Time-frequency distributions and non-linear speech models have been successfully used as feature extraction tools for robust speech recognition [1], [2], [3]. In this paper we examine modulation related features extracted via the nonlinear AM–FM speech model, using a mel-spaced Gabor filterbank. Short-time amplitude, frequency, and bandwidth related features are estimated and evaluated in the context of both clean and noisy speech recognition.

The AM–FM model has been successfully applied in various areas of signal processing including speech, music and image processing. Specifically in speech processing, the AM–FM model has been used for speech analysis and modeling [4], [5], speech synthesis [4], speech recognition [2], and speaker identification [6], [7]. Significant improvement in speech recognition accuracy has been shown in [2], where amplitude and frequency modulation related features are included in the speech recognition front-end, especially for noisy conditions. A frequency domain alternative of instantaneous frequency, namely the first normalized spectral moment, has also been explored for speech recognition [1], [3], whereas bandwidth related features are considered to carry less beneficial phonetic information.

In this paper, we investigate the stand-alone performance of short-time averages of the amplitude weighted instantaneous frequencies and bandwidths. As far as the bandwidth is concerned we examine the recognition performance of both its amplitude and frequency components [8], jointly as well as independently. We also investigate the filterbank parametrization as well as decorrelation techniques for the frequency and bandwidth front-ends.

II. Amplitude, Frequency and Bandwidth Estimates

The AM–FM model is a nonlinear model that describes a speech resonance as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure [9]

$$r(t) = a(t)\cos(2\pi[f_c t + \int_0^t q(\tau)d\tau] + \theta) \qquad (1)$$

where f_c is the "center value" of the formant frequency, $q(t)$ is the frequency modulating signal, and $a(t)$ is the time-varying amplitude. The instantaneous frequency signal is defined as $f(t) = f_c + q(t)$. The speech signal $s(t)$ is modeled as the sum $s(t) = \sum_{k=1}^{K} r_k(t)$ of K such AM–FM signals.

The estimation of the amplitude and frequency components, namely the demodulation of each resonant signal, can be done with the *energy separation algorithm* (ESA), or utilizing the Hilbert transform demodulation (HTD) algorithm. ESA exploits the differential Teager–Kaiser Energy Operator (TEO), in order to estimate the amplitude envelope $|a(t)|$ and instantaneous frequency $f(t)$ signals of the speech resonance signal $r(t)$ [9]. The energy operator tracks the energy of the source producing an oscillation signal $r(t)$ and is defined as $\Psi[r(t)] = [\dot{r}(t)]^2 - r(t)\ddot{r}(t)$ where $\dot{r}(t) = dr/dt$[1].

According to the ESA the frequency and amplitude estimates are respectively [9]

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

Usually the discrete time (DESA[2]) counterparts are used,

[1] A detailed study of the behavior of the TEO can be found in [10]

[2] DESA is actually a family of efficient algorithms that use various discrete time approximations of the continuous TEO [9].

which are defined by similar equations, using the discrete energy operator $\Psi^d[r[n]] = r^2[n] - r[n+1]r[n-1]$.

For the purpose of the feature extraction process, a multi-band demodulation analysis (MDA) is performed [8]. The speech signal is decomposed into resonant signals using a mel-spaced Gabor filterbank. The raw instantaneous frequency ($f(t)$) and amplitude ($|a(t)|$) signals are estimated by demodulating each resonant signal. Next a short-time analysis is performed, where the instantaneous envelope is averaged and log compressed $A = \log(\int_{t_0}^{t_0+T}[a(t)]^2 dt)$, whereas for the frequency estimation an amplitude weighting is performed [8]

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)[a(t)]^2 dt}{\int_{t_0}^{t_0+T}[a(t)]^2 dt} \tag{3}$$

For the bandwidth estimation both frequency and amplitude components are considered, and similar weighting with the squared amplitude is also applied

$$[B_w]^2 = \frac{\int_{t_0}^{t_0+T}\left[(\dot{a}(t)/2\pi)^2 + (f(t)-F_u)^2[a(t)]^2\right] dt}{\int_{t_0}^{t_0+T}[a(t)]^2 dt} \tag{4}$$

The amplitude component is considered by the term $(\dot{a}(t)/2\pi)^2$, which describes the rate of decay of the amplitude envelope, and is closely related to the formants' bandwidths. In order to explore the relative importance of the two components, we consider these two components separately as follows

$$[B_w^f]^2 = \frac{\int_{t_0}^{t_0+T}\left[(f(t)-F_u)^2[a(t)]^2\right] dt}{\int_{t_0}^{t_0+T}[a(t)]^2 dt} \tag{5}$$

$$[B_w^a]^2 = \frac{\int_{t_0}^{t_0+T}(\dot{a}(t)/2\pi)^2}{\int_{t_0}^{t_0+T}[a(t)]^2 dt} \tag{6}$$

where B_w^f is the frequency component of bandwidth, and B_w^a is the amplitude related one. We also introduce the positive decay amplitude related component – B_w^{a+}, which is calculated as B_w^a, but only in the decaying amplitude regions $\dot{a}(t) < 0$ (i.e., in the $\dot{a}(t) > 0$ regions $a(t)$ and $\dot{a}(t)$ are set to 0 prior to bandwidth estimation).

There is a close relationship of the above frequency and bandwidth estimates, with the first and second normalized spectral moments. More specifically under certain conditions they are considered to be equivalent [1]. The general n-th spectral moment of a short-time resonant signal $r_k(t)$, corresponding to the output of the k-th filter in a filterbank analysis, is defined as

$$S_k^n = \int_0^\pi |R_k(\omega)|^\gamma \omega^n d\omega \tag{7}$$

where $R_k(\omega)$ is the fourier transform of $r_k(t)$. The n-th normalized spectral moment is defined as

$$N_k^n = S_k^n / S_k^0 \tag{8}$$

The standard MFCC features can be considered as the DCT of the (log) zero order spectral moment (S^0, $\gamma = 2$). The first normalized spectral moment (N^1) has also been used for speech recognition, termed as Spectral Subband Centroids [3].

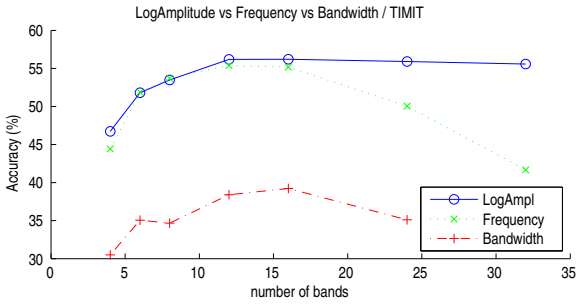

Fig. 1. Comparison of phone recognition rates for the TIMIT task, for the (log) amplitude A (without DCT), frequency F_w, and bandwidth B_w features, as a function of the number of bands. A 50% overlap is used in all filterbanks.

III. FREQUENCY AND BANDWIDTH-BASED FRONT-ENDS

Next we investigate the design of a speech recognition front-end that uses stand-alone frequency- and bandwidth-based features. Two important issues are investigated, namely, the selection of the filterbank parameters and the decorrelation of the feature vector. The analysis is grounded with the "standard" MFCC front-end.

A. Filterbank Parametrization

There are four main filterbank parameters for consideration in the MDA analysis: (1) the number of analysis bands (filters), (2) the type of the filters used, (3) the filter bandwidth (or the filter overlap), and (4) the distribution of the filters in the frequency scale. Previous studies have also examined the aforementioned parameters for frequency-based feature sets, however we address here some new findings.

Considering the number of analysis bands, previous studies (e.g. [3]) have reported an optimal number of analysis bands, beyond which there is a degradation in the recognition performance. This result is replicated in Fig. 1, where the short-time frequency and bandwidth feature recognition rates are plotted in relation to the number of analysis bands (TIMIT phone recognition task, see also the next section). For comparison, the recognition rates using energy-based features are also plotted (log amplitude without DCT).

We can see that energy- and frequency-based features have similar performance until around 12 to 16 filters. Further increase in the number of filters gives no improvement for the energy features, whereas for the frequency features a serious degradation is observed. Bandwidth features, in general, have lower performance, and similar behavior to frequencies. For the frequency and bandwidth features, the degradation is due to the narrowing of the filters, since their overlap is kept at 50%. The bandwidth reduction results in a high influence from the harmonics of the fundamental frequency. This is especially pronounced in the lower and phonetically critical bands if a log scale is used (such as in our case), where the bandwidth becomes comparable to the interharmonic distance. This is probably one of the reasons that some previous efforts involving frequency features (spectral moment based

estimation) use a filterbank with frequencies linearly spaced [3]. In order to overcome the harmonic interplay, for the frequency and bandwidth estimation, we increase the overlap between filters by widening their bandwidths. This results in significant improvement of the performance of both frequency and bandwidth features (see Table I). This is also mirrored in the filter type, where we have observed that Gabor filters (both frequency and time domain) have in general better performance than the standard frequency domain triangular filterbank, since they are wider in the central frequency region. A direct definition of the frequency overlap for the Gabor filterbank does not exist, instead we derive an *equivalent overlap* based on the energy overlap[3], which for the triangular filterbank is 0.25 (i.e. 25%). The equivalent overlap is derived as the square root of the energy overlap. Table I shows the TIMIT phone recognition rates for amplitude, frequency and bandwidth based features extracted using filterbanks with equivalent overlap of 50%, 60%, 70% and 80% (16 mel-spaced Gabor filters up to 8 kHz). Best recognition rates are obtained with equivalent overlap of 70% for frequency, 80% for bandwidth, whereas no significant improvement is observed for amplitude features (with or without DCT).

TABLE I
TIMIT PHONE RECOGNITION RATES (%) FOR AMPLITUDE, FREQUENCY AND BANDWIDTH FEATURES FOR DIFFERENT FILTERBANK OVERLAPS.

Overlap / Features	50%	60%	70%	80%
A	**56.76**	55.35	53.77	51.67
A_{DCT}	**60.09**	**60.38**	59.95	58.86
F_w	49.57	59.40	**61.21**	60.86
B_w	37.37	46.51	51.14	**53.03**

B. Decorrelation of feature vector

A common technique used for decorrelating the feature vector for speech recognition is the discrete cosine transform (DCT). Decorrelation is a necessary step, since the HMM framework used for recognition usually assumes independence between the feature vector components, i.e., diagonal covariance matrices. Although for energy-based features the DCT is beneficial, we have experimentally found that for frequency- and bandwidth-related features only moderate correlation exists between coefficient of adjacent filters. This can be seen in Fig. 2(c), where the Pearson correlation coefficient matrix has been computed for the frequency F_w feature vector. For reference, the correlation matrix for the amplitude A feature vector is shown in (a). Also the DCT's of the two feature vectors are shown in (b), (d). Clearly, the frequency-based features are only moderately correlated, and the correlation increases after the application of the DCT. Similar results can be obtained for bandwidth-based features. Overall, the DCT is used only for energy-based features, whereas frequency- and bandwidth-based features are used as are, without any transformation. Retaining the frequency domain representation

[3]Computed as the overlap ratio of the magnitude frequency responses of adjacent filters.

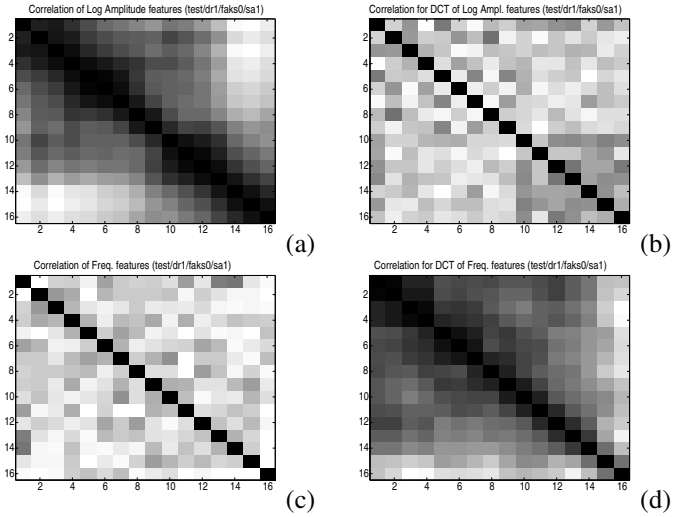

Fig. 2. The correlation coefficient matrix is shown for energy- and frequency-based feature vectors, computed for a TIMIT utterance using 16 mel-spaced filters. The correlation matrix is shown for: (a) the (log) amplitude A feature vector, (c) the frequency F_w feature vector, (b),(d) DCT transformed vectors for (a),(c), respectively. Absolute correlation values are shown in greyscale; black corresponds to 1 (fully correlated) and white to 0 (uncorrelated).

of the feature vector (instead of transforming them, e.g., to the cepstrum domain) is advantageous for a variety of robust speech recognition algorithms, e.g., frequency warping, spectral mask estimation.

IV. EXPERIMENTAL RESULTS

The following feature sets are examined: MFCC, the standard features (without C0 or energy term), A_{DCT}, short-time log amplitude (13 DCT coef., no C0), F_w, short-time instantaneous frequency estimated by (3), N^1, the first normalized spectral moment estimated by (8) using the same frequency-domain Gabor filterbank used in the MDA analysis, B_w, short-time bandwidth (4), B_w^f, the frequency component of bandwidth (5), B_w^a, the amplitude component (6), B_w^{a+}, the decaying amplitude component. All feature vectors are augmented by their first and second time-derivatives. A Gabor filterbank is used for all features, with the exception of MFCCs where the standard triangular filterbank is used. 50% is overlap is used for the amplitude features and an equivalent of 70% overlap is used for frequency and bandwidth based features.

A. Clean recording conditions

Performance was evaluated for the phone recognition task on the TIMIT database. Using the HTK framework, 3-state phonemic HMMs with a mixture of 16 Gaussians per state were trained using 4 reestimation iterations. Three different filterbanks were used, having 16, 20 and 26 mel-spaced filters up to 8 kHz. The results are summarized in Table II.

The recognition performance of the short-time frequency (F_w) features is better than the standard MFCC features in the cases of filterbanks with 16 and 20 filters, but slightly worse in the 26 filters case. This suggests that in the case of 26 filters a wider filterbank should probably be used. Furthermore the

TABLE II
PHONE RECOGNITION RATES (%) ON THE TIMIT DATABASE.

#Filters / Features	16	20	26
MFCC	60.20	60.58	**60.66**
A_{DCT}	60.09	60.68	**61.16**
F_w	61.21	**61.34**	59.88
N^1	60.54	**61.02**	60.38
B_w	51.14	**51.22**	49.05
B_w^f	**48.17**	47.67	44.14
B_w^a	48.06	**49.37**	48.15
B_w^{a+}	50.49	**51.31**	50.95

spectral moment estimation (N^1) also benefits from the use of a wider filterbank, having similar performance to the F_w. The performance of bandwidth related features is also noteworthy, since it exceeds 50%. The amplitude related component seems to be a better estimate than the frequency counterpart, and more specifically the decaying amplitude estimation.

Furthermore, we have augmented the frequency features with the log energy coefficient (E), and the zeroth cepstral coefficient (C0), and compared it with the corresponding MFCC features. The results are summarized in Table IV. The performance of frequency feature vector plus energy (or C0) compares well with the MFCC vector plus energy (or C0).

TABLE III
PHONE RECOGNITION RATES (%) ON THE TIMIT DATABASE USING AUGMENTED VECTORS

#Filters / Features	16	20	26
MFCC+E	64.06	**64.28**	64.10
MFCC+C0	64.16	**64.29**	64.24
F_w+E	63.78	**63.99**	62.55
F_w+C0	**64.28**	64.11	62.73

The results shown in Table II, as well as in Table IV, suggest that frequency related features can be used as an alternative ASR front-end, with very good performance. This has been verified also on digit and word-recognition tasks [11]. This can be achieved with the use of wider filterbanks in order to overcome the harmonic influence. Moreover the bandwidth estimates carry significant phonetic information.

B. Noisy conditions

Frequency-based features have been shown to be robust in additive noise [2], [3]. We performed a preliminary study of noisy speech recognition with the new feature extraction technique, on the Spanish Task of the Aurora 3 database. The recognition experiments were performed on the 8 kHz dataset, which was analyzed with a filterbank of 12 Gabor filters equally spaced in the Mel frequency scale up to 4 kHz. The results are summarized bellow, for three different noise situations: well-matched (WM), medium-mismatched (MM), and high-mismatched (HM). It is clear that the frequency features perform significantly better in all noise situations. Moreover the recognition improvement increases as the noise situation gets worse.

TABLE IV
WORD RECOGNITION RATES (%) ON THE AURORA 3 SPANISH TASK

	WM	MM	HM
MFCC+E	**86.88**	**73.72**	**42.23**
F_w+E	**92.22**	**84.53**	**73.56**

V. CONCLUSIONS

We investigated the use of short-time amplitude weighted instantaneous frequencies and bandwidths as a stand alone ASR front-end. Our investigation showed that a frequency front-end can be superior to a power spectral based front-end, especially in noisy situations. We also found that bandwidth features can carry substantial phonetic information that can be exploited for speech recognition. Designing the appropriate filterbank for frequency- and bandwidth-based features was essential to achieving this high performance, in order to avoid the influence of the pitch harmonics. In this study, we used a Gabor filterbank with mel-spaced center frequencies, and 70% filter overlap. However a more extensive research is needed to determine the optimal filterbank setup. Complementary information to the averages of instantaneous frequency and bandwidth must also be investigated in the ASR context.

ACKNOWLEDGMENT

This research was co-financed partially by E.U.-European Social Fund (80%) and the Greek Ministry of Development-GSRT (20%) under Grant PENED-2003-ED866.

REFERENCES

[1] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 196–200, Mar 2001.

[2] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM–FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, September 2005.

[3] J. Chen, Y. A. Huang, Q. Li, and K. K. Paliwal, "Recognition in noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 258–261, February 2004.

[4] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM–FM modulation model," *Speech Communication*, vol. 28, pp. 195–209, July 1999.

[5] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, September 1999.

[6] C. R. Jankowski Jr., T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *ICASSP-95*, Detroit, USA, May 1995.

[7] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 6, pp. 1097–1111, August 2008.

[8] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *Journal of Acoustical Society of America*, vol. 99, pp. 3795–3806, June 1996.

[9] 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, no. 10, pp. 3024–3051, October 1993.

[10] D. Dimitriadis, A. Potamianos, and P. Maragos, "A comparison of the squared energy and teager-kaiser operators for short-term energy estimation in additive noise," *IEEE Trans. Signal Processing*, vol. 57, no. 7, pp. 2569–2581, July 2009.

[11] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral moment augmented by low order cepstral coefficients for robust ASR front-end," *IEEE Signal Processing Letters*, submitted 2009.