Spectral Moment Features Augmented by Low Order Cepstral Coefficients for Robust ASR

Pirros Tsiakoulis, *Member, IEEE,* Alexandros Potamianos, *Member, IEEE,* and Dimitrios Dimitriadis, *Member, IEEE*

***Abstract*—We propose a novel ASR front-end, that consists of the first central Spectral Moment time-frequency distribution Augmented by low order Cepstral coefficients (SMAC). We prove that the first central spectral moment is proportional to the spectral derivative with respect to the filter's central frequency. Consequently, the spectral moment is an estimate of the frequency domain derivative of the speech spectrum. However information related to the entire speech spectrum, such as the energy and the spectral tilt, is not adequately modeled. We propose adding this information with few cepstral coefficients. Furthermore, we use a mel-spaced Gabor filterbank with 70% frequency overlap in order to overcome the sensitivity to pitch harmonics. The novel SMAC front-end was evaluated for the speech recognition task for a variety of recording conditions. The experimental results have shown that SMAC performs at least as well as the standard MFCC front-end in clean conditions, and significantly outperforms MFCCs in noisy conditions.**

***Index Terms*—First Spectral Moment, Low Order Cepstral Coefficients, SMAC, Robust Speech Recognition**

I. Introduction

MOST of the features used for automatic speech recognition, try to capture the time-frequency information from the speech signal, either by employing a filterbank analysis, or through parametric modeling. Indeed the widely used Mel Frequency Cepstral Coefficients (MFCC) front-end is based on a time-frequency energy distribution. It employs discrete cosine transformation (DCT) in the frequency domain for decorrelation of the feature vector. The first and second order time domain derivatives are usually included.

A theoretical analysis of the time-frequency distributions for automatic speech recognition can be found in [1], where short-time averages from various time-frequency distributions are shown to be equivalent under certain conditions. The first three spectral moments, namely the zeroth, first, and second, are respectively metrics for energy, frequency and bandwidth [2]. Energy related features, such as the established MFCCs, are dominant in speech applications. Frequency related features have also been considered in various speech applications. Frequency estimates are distributed densely in spectral peak and sparsely in spectral valley regions. This time-frequency distribution, is known as *pyknogram*, and was introduced for the speech formant estimation task [3]. Pyknogram based features were recently proposed for speaker identification [4]. Spectral moment based frequency features have been proposed for robust speech recognition [5]. Time domain frequency and bandwidth related features were also exploited for robust speech recognition as additional features to the standard feature vector [6]. Gaussian modeling of the smooth spectral envelope was also investigated for speech recognition [7].

In this letter, we propose a novel time-frequency driven front-end for automatic speech recognition. The proposed front-end retains the frequency domain representation and provides a zero mean feature vector, facilitating a variety of robust speech recognition algorithms, e.g., frequency warping, spectral mask application, multi-band analysis, vocal tract normalization etc. It is based on the first spectral moment and augmented by few low order cepstral coefficients – SMAC. The spectral moment component captures information about the resonances of the speech signal under the notion of the pyknogram. However, solely the pyknogram does not model the relative importance of each resonance. This could be why previous attempts to use frequency only based front-ends, usually have worse recognition performance for the clean speech and well matched cases [5], [8]. For this reason SMAC also incorporates low order cepstral coefficients, to capture the rough spectral envelope. The spectral moment components are sampled in the standard Mel frequency scale, with a Gabor filterbank. Moreover, we investigate the number of necessary coefficients for the spectral envelope estimation, as well as the filterbank parametrization to overcome the sensitivity to the pitch harmonics reported in previous studies [5].

II. Central Spectral Moment Estimation

Assume that the discrete- short-time speech signal $x(n)$ is filtered by a bank of K band-passed filters with center frequencies ω_k. The resulting band-passed signals $x_k(n),\ k = 1 \dots K$ in time and frequency domain are given by

$$x_k(n) = x(n) * h_k(n) \leftrightarrow X_k(\omega) = X(\omega) H_k(\omega) \qquad (1)$$

where $h_k(n)$ is the impulse response and $H_k(\omega)$ the frequency response of the k-th filter. The generalized m-th spectral moment, and central spectral moment of each signal $x_k(n)$, for an arbitrary constant γ, are respectively defined as

Manuscript received November 14, 2009; revised February 22, 2009. This research was co-financed partially by E.U.-European Social Fund (80%) and Greek Ministry of Development-GSRT (20%) - Grant PENED-2003-ED866.

P. Tsiakoulis is with the School of Electrical & Computer Engineering, National Technical University of Athens, Zografou, Athens, GR–15773, Greece; email: ptsiak@ilsp.gr.

A. Potamianos is with the Department of Electronics & Computer Engineering, Technical University of Crete, Chania, GR–73100, Greece; email: potam@telecom.tuc.gr.

D. Dimitriadis is now with the AT&T Research Labs Inc, Florham Park, NJ 07932, USA; email: ddim@research.att.com.

Digital Object Identifier XX.XXXX/XXX.XXXX.XXXXXXX

$$S^m(k) = \int_0^\pi |X_k(\omega)|^\gamma \omega^m d\omega \quad (2)$$

$$S_c^m(k) = \int_0^\pi |X_k(\omega)|^\gamma (\omega - \omega_k)^m d\omega \quad (3)$$

The respective normalized moments are defined as

$$N^m(k) = S^m(k)/S^0(k) \quad (4)$$

$$N_c^m(k) = S_c^m(k)/S_c^0(k) \quad (5)$$

From (2)-(5) stems $S_c^0(k) \triangleq S^0(k)$ and $N_c^1(k) = N^1(k) - \omega_k$.

The zero order spectral moment S^0 for $\gamma = 2$ yields the spectral filterbank energies, which are usually used in the MFCC derivation. The first spectral moment (N^1) tracks the weighted average formant frequency in each band, and it has also been used for speech recognition [1], [5]. Time domain equivalent estimations have been also used for speech recognition [1], [6], [9].

A. Spectral Moment Estimation with a Gabor Filterbank

Assuming $h_k(n)$ is the impulse response of the real Gabor filter, the frequency response can be expressed as

$$H_k(\omega) = (\sqrt{\pi}/2\alpha)(e^{-(\omega-\omega_k)^2/4\alpha^2} + e^{-(\omega+\omega_k)^2/4\alpha^2}) \quad (6)$$

where α is a parameter controlling the filter's bandwidth. The spectral moment estimation usually considers only the positive frequency component, since the integration is performed in the positive frequencies

$$H_k^+(\omega) = (\sqrt{\pi}/2\alpha)e^{-(\omega-\omega_k)^2/4\alpha^2} \quad (7)$$

We prove that under the above assumption the first central spectral moment $S_c^1(k)$ is proportional to the derivative of the zero order spectral moment $S^0(k)$ with respect to the filter's central frequency ω_k. Starting from the definition of $S^0(k)$ in (2), and taking the derivative with respect to ω_k we have

$$\frac{dS^0(k)}{d\omega_k} = \frac{d}{d\omega_k}\int_0^\pi |X_k(\omega)|^\gamma d\omega = \int_0^\pi \frac{d|X_k(\omega)|^\gamma}{d\omega_k} d\omega$$
$$\simeq \int_0^\pi |X(\omega)|^\gamma \frac{d|H_k^+(\omega)|^\gamma}{d\omega_k} d\omega \quad (8)$$

The derivative of $H_k^+(\omega)$ with respect to ω_k is

$$\frac{d|H_k^+(\omega)|^\gamma}{d\omega_k} = (\sqrt{\pi}/2\alpha)^\gamma \frac{de^{-\gamma(\omega-\omega_k)^2/4\alpha^2}}{d\omega_k}$$
$$= (\sqrt{\pi}/2\alpha)^\gamma 2(\gamma/4\alpha^2)(\omega-\omega_k)e^{-\gamma(\omega-\omega_k)^2/4\alpha^2}$$
$$= (\gamma/2\alpha^2)(\omega-\omega_k)|H_k^+(\omega)|^\gamma \quad (9)$$

Replacing (9) in (8) we get

$$\frac{dS^0(k)}{d\omega_k} \simeq \frac{\gamma}{2\alpha^2}\int_0^\pi |X(\omega)|^\gamma |H_k^+(\omega)|^\gamma (\omega-\omega_k) d\omega$$
$$\simeq \frac{\gamma}{2\alpha^2}\int_0^\pi |X_k(\omega)|^\gamma (\omega-\omega_k) d\omega = \frac{\gamma}{2\alpha^2} S_c^1(k) \quad (10)$$

It simply follows that

$$S_c^1(k) \simeq \frac{2\alpha^2}{\gamma}\frac{dS^0(k)}{d\omega_k} \quad (11)$$

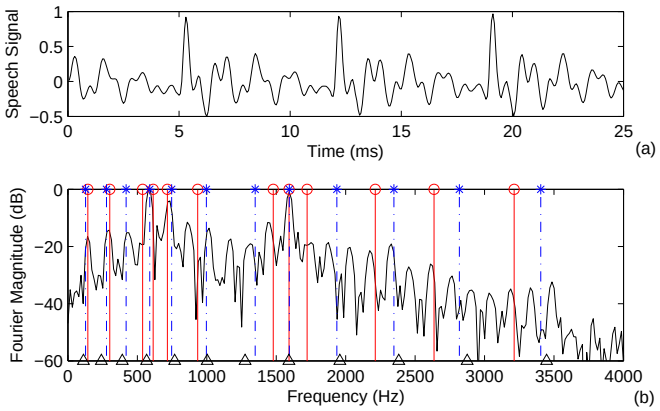

Fig. 1. (a): A 25-ms speech frame (phoneme /ae/, male speaker). (b) The corresponding DFT spectrum (up to 4kHz), and superimposed the first spectral moment estimations for two mel-spaced Gabor filterbanks with constant bandwidth of 118 and 236 Mels respectively. The spectral moment estimates are shown with stars / dashed-dotted lines for the narrow filterbank, and with circles / vertical solid lines for the wide one. The filterbank center frequencies are shown with triangles on the x axis.

Furthermore, the first normalized central spectral moment is proportional to the derivative of the logarithm of the zero order spectral moment (log power spectrum) with respect to ω_k:

$$N_c^1(k) \simeq \frac{2\alpha^2}{\gamma S^0(k)}\frac{dS^0(k)}{d\omega_k} = \frac{2\alpha^2}{\gamma}\frac{d\log(S^0(k))}{d\omega_k} \quad (12)$$

Eq. (12) identifies the close relationship between $N_c^1(k)$ and the log power spectrum, which is used in the standard MFCC front-end. This result also explains the pyknogram structure: if the filter frequency is centered before (after) the spectral peak, the derivative in (12) is positive (negative), so the estimate moves towards the spectral peak (see Fig. 1(b)). Eq. (12) also reveals the importance of the filter's bandwidth, i.e. the smaller the parameter α, the closer the estimate $N^1(k)$ will be to the center frequency ω_k ($\alpha \to 0 \Rightarrow N_c^1(k) \to 0 \Rightarrow N^1(k) \to \omega_k$).

B. Sensitivity to the pitch harmonics

Eq. (12) implies sensitivity of the spectral moment estimation to the harmonics of the fundamental frequency, if the filter's bandwidth is narrow. In such a case the estimation will select the strongest harmonic within the filter's bandwidth. This can be seen in Fig. 1(b), where we plot the first spectral moment estimates ($N^1(k)$) for two mel-spaced Gabor filterbanks with different bandwidths. Both filterbanks have constant bandwidth on the Mel scale. The bandwidth of the first one is set to 118 Mels, which is roughly equivalent to the standard triangular filterbank with 50% overlap. While the second one, the bandwidth is set to 236 Mels (equivalent to a 70% overlap). One can clearly see, that the frequency estimations for the narrow filterbank are placed at the pitch harmonics closest to the filters' center frequency. This is more pronounced in the lower filters that are narrower. On the other hand, the frequency estimations for the wide filterbank are more biased towards the formant frequency. This is mirrored in the distribution of the estimates, which are more dense around formants for the wider filterbank. Wideband analysis could be a potential solution to the problem. However factors such as,

a) the reduced frequency resolution, b) modulation effects in the voiced regions, and c) increased number of frames for the calculation of derivative features, introduce further complexity.

C. Estimation under Noise

The ability of the first spectral moment to track the local spectral peaks makes it a perfect candidate for estimation in noisy conditions, where the spectral peaks are not seriously affected. Fig. 2(d) shows a pyknogram example for an utterance from TIMIT corrupted with additive babble noise at 5dB (the noise signal was extracted from the NoiseX92 database). The spectrogram of the noise corrupted speech signal is shown Fig. 2(b). For comparison we also show the spectrogram and pyknogram of the original speech signal in Fig. 2(a),(c) respectively. For visualization purposes, the pyknograms were constructed using a Gabor filterbank with 64 linearly spaced filters up to 4kHz, and constant bandwidth of 400Hz. One can see in Fig. 2, that the spectral moment estimation is noise robust. The frequency estimation is not seriously affected, provided that the local spectral peak stays above the noise.

III. The SMAC front-end

The SMAC front-end consists of the first normalized central spectral moment (N_c^1, $\gamma = 2$) and few low order cepstral coefficients. The spectral moment component captures the resonance structure of the speech signal under the notion of the pyknogram (Fig. 2(c)). This is equivalent with a flat spectral estimation, since the information for the relative importance of each resonance is lost. For this reason the pyknogram is augmented by few low order cepstral coefficients, to capture the rough spectral envelope. Alternative estimations of both signal energy and rough envelope can also be used. No further transformation such as DCT is performed, since the spectral moment components are mostly uncorrelated [9].

As we saw in the previous section, the estimation of the first spectral moment is sensitive to the pitch harmonics if the filter is narrow, which is not a desired behavior for the speech recognition task. For this reason, for the computation of spectral moments we propose to increase the frequency overlap between adjacent filters of the filterbank. Previous studies address this problem by either reducing the number of filters used, or by using linearly spaced filterbanks [5], [6], [8]. Filterbanks with large frequency overlap have also been used for formant tracking [3], and speaker identification [4].

In the SMAC feature extraction process we employ a mel-spaced Gabor filterbank[1], having 12 filters up to 4kHz in the narrow-band speech case (8kHz), and 16 filters up to 8kHz in the wide-band case (16kHz). Although, similar or better performance could be achieved with more filters, we constrain the number of filters to retain a low feature vector dimension. The overlap between adjacent filters is controlled by adjusting their bandwidth. A constant bandwidth of 236 Mels was experimentally found to be close to the optimal value [9]. For simplicity, the same filterbank is used for extracting the low order cepstral coefficients that augment the feature vector. We have experimentally determined that adding the zero and first order cepstral coefficients (C0, C1) improves performance. C0 and C1 incorporate information about the signal energy and spectral tilt respectively. The addition of more coefficients offers little improvement or even degradation. Finally, the standard delta and delta-delta features are appended to the SMAC vector. Henceforth, we use the term SMAC to refer to the front-end that uses 16 filters up to 8kHz or 12 up to 4kHz, and includes only C0 and C1, unless indicated otherwise.

[1]Different filter types with proper bandwidths would perform similarly.

IV. Experimental Results

A. Clean Recording Conditions

Performance was evaluated for the TIMIT phone recognition task (16kHz). 3-state context-independent phonemic HMM with a mixture of 16 Gaussians per state were trained using 4 reestimation iterations, using the HTK framework. The full train and test sets of the TIMIT database were used. The original phoneme set (61 phonemes) was used for the training process, which for the test process was mapped to the standard phoneme set with 39 phonemes. In this experiment we evaluated also the number of filters, as well as the number of additional cepstral coefficients in the SMAC feature vector (noted by the trailing number). For comparison purposes we also include experimental results for the spectral moment vector without any additional cepstral coefficient - SM, as well as for the MFCC feature vector with and without C0. The results, summarized in Table I, show that the SMAC features perform slightly better than the MFCC features (some differences are not statistically significant). We conclude that adding only C0 and C1 to the SMAC feature vector is sufficient.

TABLE I
Phone Recognition Rates (%) on the TIMIT Database.

	16 filters	20 filters	26 filters
MFCC (no C0)	64.09	64.38	64.48
SM	65.00	65.25	64.66
MFCC	67.61	67.59	67.73
SMAC0 (SM+C0)	68.00	68.22	67.52
SMAC1 (SM+C0-C1)	**68.66**	68.63	68.40
SMAC2 (SM+C0-C2)	68.73	**68.96**	68.68
SMAC3 (SM+C0-C3)	68.66	**69.06**	**68.84**

B. Additive Noise Conditions

The SMAC front-end was also evaluated in additive noise conditions, and compared to the MFCC front-end, the PLP front-end [10] and the RASTA-PLP front-end [11]. The MFCC and SMAC front-ends were also compared when noise suppression via Wiener Filtering (WF) was employed [12].

We conducted word recognition experiments on the AURORA 2 database (8 kHz), which contains artificially corrupted speech with various types of noise at different levels. Context-independent, 16-state, left-right word HMMs with 3 Gaussian mixtures per state were used (no grammar was used). The models were trained on the *clean training set* and tested for all *noisy test sets*. The recognition rates averaged per noise level across all noise types are shown in Table II. The SMAC front-end performs significantly better than the MFCC and PLP front-ends across all noise levels, and slightly better than

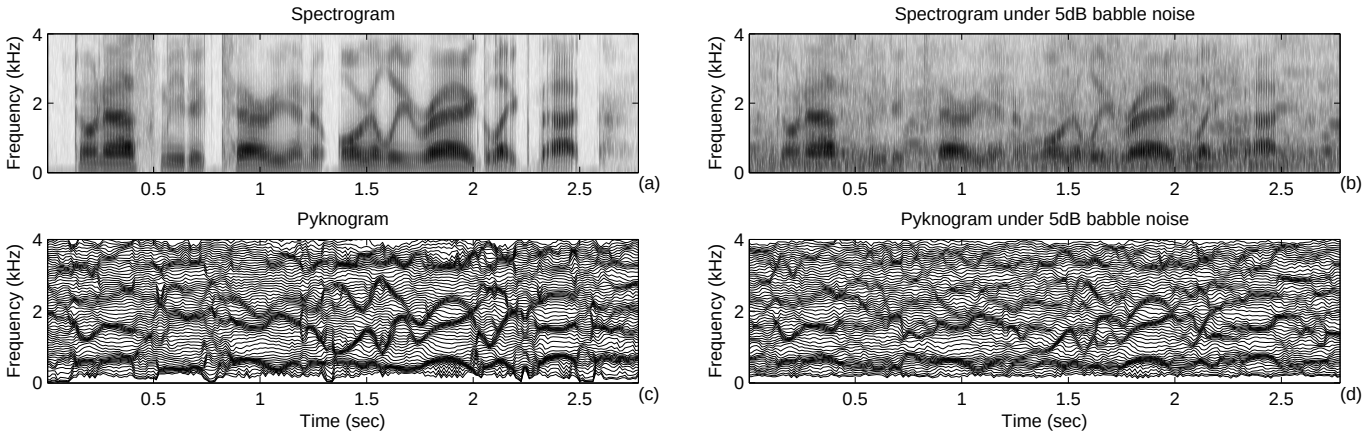

Fig. 2. Spectrogram vs Pyknogram. (a) the spectrogram of TIMIT utterance, (c) its pyknogram, (b),(d) spectrogram and pyknogram under 5dB babble noise.

RASTA-PLP. SMAC also outperforms MFCC when WF noise suppression is applied (on both train and test data sets).

TABLE II
AURORA 2 WORD RECOGNITION RATES

	20 dB	15 dB	10 dB	5 dB
MFCC (39)*	94.07	85.04	65.51	38.45
PLP (39)	94.16	85.30	67.14	41.18
RASTA-PLP (39)	96.66	92.04	77.13	45.18
SMAC (42)	**97.05**	**93.05**	**78.78**	**46.48**
WF+MFCC (39)	97.70	95.31	89.13	74.37
WF+SMAC (42)	**97.52**	**95.69**	**90.75**	**77.46**

*The feature vector size is shown next to each front-end.

Furthermore, we present recognition results on the connected digit recognition task of the AURORA 3 database (8 kHz), using the same configuration as in the AURORA 2 experiment. Table III summarizes the results for the Spanish and Italian tasks, for well-matched (WM), medium-mismatched (MM), and high-mismatched (HM) noise conditions.

TABLE III
AURORA 3 WORD RECOGNITION RATES

	Spanish Task			Italian Task		
	WM	MM	HM	WM	MM	HM
MFCC (39)	86.88	73.72	42.23	**93.64**	82.02	39.84
PLP (39)	92.04	83.84	52.72	88.24	72.51	38.98
RASTA-PLP (39)	93.94	88.25	72.93	83.76	75.27	**63.33**
SMAC (42)	**94.25**	**89.21**	**77.68**	88.14	**82.30**	51.63
WF+MFCC (39)	94.84	88.31	78.32	**95.89**	**89.81**	**73.52**
WF+SMAC (42)	**94.87**	**91.09**	**81.65**	91.43	86.42	62.23

The SMAC features perform significantly better in all noise conditions for the Spanish task, for both the baseline and the WF case. The results on the Italian task show a mixed behavior in the baseline case; MFCC is better in the WM, SMAC in the MM, and RASTA-PLP in the HM case. Including WF, the MFCC front-end outperforms SMAC. However, the Italian task results for SMAC and RASTA-PLP front-ends were greatly affected by unbalanced insertion/deletion ratios.

In general, the SMAC front-end improves over the MFCCs as the SNR decreases. Similar conclusions can be drawn from preliminary large vocabulary recognition experiments on the Wall Street Journal Aurora 4 database.

V. CONCLUSION

We proposed the use of the first central Spectral Moment Augmented by low order Cepstral coefficients (SMAC), as an alternative ASR front-end. The innovation introduced by the SMAC front-end is twofold. First, the augmentation of the spectral moment with few low order cepstral coefficients, reintroduces the coarse spectral envelope information in the feature vector. Second, the use of a Gabor filterbank with larger bandwidth alleviates the sensitivity of the spectral moment estimation to the pitch harmonics. We evaluated the proposed front-end in clean and noisy speech recognition tasks. The results have shown that the SMAC front-end performs similarly to the standard front-end in clean recording conditions, and outperforms it for a wide range of noisy tasks.

REFERENCES

[1] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 3, pp. 196–200, March 2001.
[2] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal–part 1: Fundamentals," *Proceedings of the IEEE*, vol. 80, pp. 520–538, 1992.
[3] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. of Acoust. Soc. of America*, vol. 99, pp. 3795–3806, June 1996.
[4] 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.
[5] 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.
[6] 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.
[7] M. N. Stuttle and M. J. F. Gales, "A mixture of Gaussians front end for speech recognition," in *EUROSPEECH*, 2001.
[8] K. K. Paliwal and B. S. Atal, "Frequency-related representation of speech," in *EUROSPEECH*, 2003.
[9] P. Tsiakoulis, A. Potamianos, D. Dimitriadis, "Short-time instantaneous frequency and bandwidth features for speech recognition," in *ASRU*, 2009.
[10] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. of Acoust. Soc. of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
[11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Process.*, vol. 2, no. 4, pp. 578–589, 1994.
[12] D. Dimitriadis, J. C. Segura, L. Garcia, A. Potamianos, P. Maragos, and V. Pitsikalis, "Advanced front-end for robust speech recognition in extremely adverse environments," in *INTERSPEECH*, 2007.