Multiband, Multisensor Robust Features for Noisy Speech Recognition

Dimitrios Dimitriadis, Petros Maragos and Stamatios Lefkimmiatis

National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece.

Email:[ddim, maragos, sleukim]@cs.ntua.gr

Abstract

This paper presents a novel feature extraction scheme taking advantage of both the nonlinear modulation speech model and the spatial diversity of speech and noise signals in a multisensor environment. Herein, we propose applying robust features to speech signals captured by a multisensor array minimizing a noise energy criterion over multiple frequency bands. We show that we can achieve improved recognition performance by minimizing the Teager-Kaiser energy of the noise-corrupted signals in different frequency bands. These Multiband, Multisensor Cepstral (MBSC) features are inspired by similar ones already been applied to single-microphone noisy Speech Recognition tasks with significantly improved results. The recognition results show that the proposed features can perform better than the widely-used MFCC features.

Index Terms: Speech recognition, modulations, robust features, multisensor array, multiband processing.

1. Introduction

Nowadays, significant research interest is focused on the use of multisensor systems. Their improved results in speech enhancement and robust speech recognition tasks seem very promising. The main advantage is that the microphone array can simultaneously exploit the spatial diversity of speech and noise, so both spectral and spatial characteristics of the speech signals are considered. Usually, the spatial discrimination of the array is examined by beamforming algorithms, like those proposed in [1]. In most cases though, the obtainable noise reduction performance is not sufficient and post-filtering techniques, using Wiener filters, are applied to further enhance the output of the beamformer. In general, these techniques accomplish higher noise reduction than the *Minimum Variance Distortionless Response* (MVDR) beamformer alone. Despite its theoretical optimality, Wiener filtering is difficult to be realized due to its need for estimating the 2^{nd} order statistics for both the clean speech and noise signals. A variety of post-filtering techniques, trying to address this issue, have been proposed [2, 3, 4].

One of the early methods for post-filtering is due to Zelinski [2], and was further studied by Marro et al. [5]. The generalized version of Zelinski's algorithm is based on the assumption of a spatially uncorrelated noise field. A more accurate noise field model has been introduced by McCowan et al. [4], where a known noise field coherence function is proposed improving the overall performance. Both methods are characterized by a certain drawback as the noise power spectrum at the beamformer's output is over-estimated, [4, 6]. In addition, the post-filtering scheme distorts the speech spectrum, especially in the lower bands of the speech spectrum. Some of the frequency content of the speech signal is mistakenly taken for noise and thus smoothed out. Therefore, these methods do not perform well in the feature extraction process of noisy ASR tasks.

Herein, we are proposing a novel feature extraction algorithm minimizing the interference of the corrupting noise signal in different frequency bands. These features are called *Multiband, Multisensor Cepstral Coefficients* (MBSC). The single-sensor version of the proposed features, called *Teager-Energy Cepstral Coefficients* (TECCs), has been presented in [7] with promising improvements over the recognition rates. Now, we are expanding the notion of the TECCs in a multisensor environment.

2. Background on the Feature Extraction Process of TECCs

The typical MFCCs are estimated over a filterbank of triangular filters with fixed overlap as the log mean squared amplitudes of the bandpassed signals. Herein, we propose incorporating information about the time-varying and the dynamic nature of speech when using the *Teager-Kaiser Energy* (TK-Energy) instead of the typical squared-amplitude approach, [8]. In this way, the features' acoustic information is 'richer' as information concerning the instantaneous frequency is, also, incorporated. In addition, we propose using an auditory-inspired filterbank instead of the usual triangular filterbank considering the advantages of the human hearing process. The proposed nonlinear features are called *TECCs* in direct proportion to the widely-used MFCCs.

Recent studies of the human hearing physiology, [9, 10, 11], have shown that the human physiology dictates that the auditory filter bandwidths should be given by the $ERB(f)$ function

$$ERB(f_c) = 6.23(f_c/1000)^2 + 93.39(f_c/1000) + 28.52 \quad (1)$$

where f_c is the filter center frequency in Hz. Moreover, the filter placing is equidistant in the *Critical Band* (Bark) frequency scale

$$Bark(f_c) = \frac{26.81 f_c}{f_c + 3920} - 0.53 \quad (2)$$

and $0 \leq f_c \leq F_s/2$, where F_s the sampling frequency. Finally, a good approximation of the auditory filters are the asymmetrical Gammatone filters, [9],

$$g(t) = A t^{n-1} \exp\left(-2\pi b ERB(f_c) t\right) \cos(2\pi f_c t) \quad (3)$$

where $A,\ b,\ n$ are the Gammatone filter design parameters and f_c its center frequency. In [9] it is proposed that the auditory filters should have $b = 1.019$ and $n = 4$.

The proposed features are proved to be more robust in additive noise and provide additional acoustic information when compared to the typical MFCCs, [7]. The Gammatone filters are smoother and broader than the usual triangular filters due to the increased filter overlap dictated by the ERB-curve, providing additional robustness to noise, [12].

The *TECC* extraction algorithm can be summarized as follows:

i. Use of a Gammatone filterbank, as mentioned above, to create a set of bandpass speech signals. The number N of filters can vary from 30 to 200 filters depending on the SNR levels. In general, more filters provide additional robustness to noise and better frequency resolution, Fig 1,

ii. Estimation of the mean TK-energy for each one of the framed and bandpassed signals,

iii. Cepstrum coefficient computation of the log mean energies with the DCT, and

iv. Truncation of these Cepstrum coefficients by keeping the first 12 coefficients, $c_1 - c_{12}$. The 0^{th}-coefficient, c_0, augments the final feature vector similarly to the MFCC extraction scheme.

The first two steps combine the auditory filtering process with the more 'natural' approach of the speech energy notion. These steps differentiate the proposed algorithm from the widely-spread MFCC extraction algorithm. A mean TK-energy coefficient corresponds to each one of the frequency bands. These energy coefficients are highly dependant on the spectral content of the speech signal at any given time.

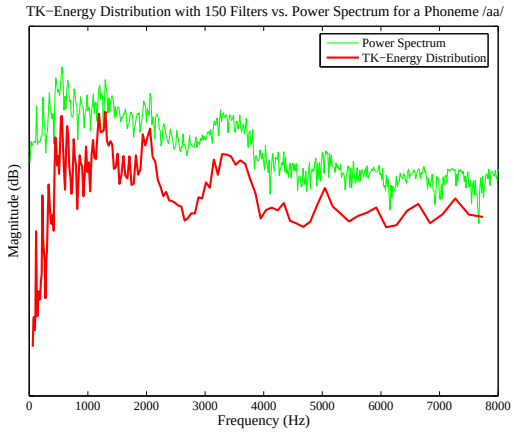

Figure 1: Normalized Power Spectrum and TK-Energy Time-Frequency Distribution, with 150 Filters, for a Phoneme */aa/*.

As shown in Fig. 1 these coefficients are closely related to the corresponding speech spectrum but they provide additional information concerning the instantaneous frequency content of the speech bands. For the estimation of the TK-energy Time-Frequency distribution, a filterbank with 150 filters is used, similar to the one mentioned above. Some of the presented differences, between the common Spectral Envelope and the TK-energy distributions, are due to the filter spacing. For the case of linear spacing, we would obtain a similar distribution to the common, linear Spectral Envelope curve. Finally, note the increased detail of the proposed distribution in the lower bands of the speech spectrum due to the fact that the filters are placed closer in this part of the speech spectrum than those placed in the higher part of it, as the filters are placed according to the Critical Band spacing, Eq. (2). The larger part of the acoustic information is located exactly in this part of the spectrum.

The TK-energy coefficients are highly correlated due to the increased filter overlap in the frequency axis. A certain level of decorrelation can be achieved through the DCT. The estimation of the Cepstrum coefficients and their truncation process remain unaltered, similarly to that of the MFCCs. Though, the ASR results show significant improvement, especially when applied to noisy recognition tasks. According to the experimental results, the use of Gammatone filters and the TK-energy appears to provide the much wanted additional robustness to noise.

3. Noise Analysis

Going one step further, we consider the case of a M-sensor linear microphone array in a noisy environment capturing a speech waveform. The observed signal $\hat{y}_m(n)$ on the m^{th} sensor, $m = 0, \ldots, M-1$, corresponds to a linearly filtered version of the arriving source signal $s(n)$, plus an additive noise component $\hat{v}_m(n)$. This additive noise component is assumed to be a zero mean, wide-sense stationary (WSS) Gaussian random process with an autocorrelation function $R_m(\tau)$ and a spectral density $\Phi_m(\omega)$. The signals, received by the sensors, have to be scaled and time-aligned (T.A.) accounting for the spatial propagation effects. The obtained, time-aligned, signals are denoted as

$$y_m(n) = s(n) + v_m(n) \tag{4}$$

where $m = 0, \ldots, M-1$ the number of the aligned signals.

Herein, we are proposing a multisensor-multiband processing scheme where every aligned input signal is decomposed into N bandpass signals using an auditory-based analysis filterbank, as in Section 2. Let us denote with y_{mk} each of the signals observed at the output of the m^{th} sensor and filtered by the k^{th} filter. As in Section 2, the TK-energy of each passband signal is estimated

$$\Psi\left[y_{mk}(t)\right] = \dot{y}_{mk}^2(t) - y_{mk}(t)\ddot{y}_{mk}(t) \tag{5}$$

Expanding this expression, we shall obtain

$$\begin{aligned}\Psi\left[y_{mk}(t)\right] &= \Psi\left[s_k(t)\right] + \Psi\left[v_{mk}(t)\right] + \Psi_c\left[s_k(t), v_{mk}(t)\right] \\ &+ \Psi_c\left[v_{mk}(t), s_k(t)\right]\end{aligned} \tag{6}$$

where Ψ_c is the cross-Teager energy between the source signal and the noise, as defined in [13].

The last three terms are related to the noise signal. Assuming that the signal $s(t)$ can be approximated by an AM-FM signal, we are able to simplify further Eq. (6). Such an approximation is well-motivated for speech signals since experimental results have produced strong evidence for the existence of amplitude and frequency modulations (AM–FM) in speech resonance signals. Thus, the corresponding TK-energy is given, [14], by $\Psi\left[s_k(t)\right] \approx a_k^2(t)\omega_k^2(t)$, where $a_k(t)$, $\omega_k(t)$ are the instantaneous amplitude and frequency signals. Additionally, the passband signal $s_k(t)$ can be approximated, [15], by

$$\hat{s}_k(t) \approx a_k(t)\left|G_k\left[\omega_k(t)\right]\right|\cos\left\{\phi(t) + \angle G_k\left[\omega_k(t)\right]\right\} \tag{7}$$

Finally, the TK-energy of the filtered signal $s_k(t)$ is

$$\Psi\left[s_k(t)\right] \approx a_k^2(t)\omega_k^2(t)\left|G_k\left[\omega_k(t)\right]\right|^2 \tag{8}$$

Since the noise process $v_m(t)$ has a spectral density given by $\Phi_m(\omega)$, the spectral density $\Phi_{mk}(\omega)$ of its filtered version, $v_{mk}(t) = v_m(t) * g_k(t)$, is

$$\Phi_{mk}(\omega) = \left|G_k(\omega)\right|^2 \Phi_m(\omega). \tag{9}$$

Furthermore, the $v_{mk}(t)$, $\dot{v}_{mk}(t)$ and $\ddot{v}_{mk}(t)$ processes are WSS and Gaussians. In addition, $\dot{v}_{mk}(t)$ is statistically independent of both $v_{mk}(t)$ and $\ddot{v}_{mk}(t)$, [16]. Therefore the energy operator output is the sum of two independent processes, i.e.

$$\Psi\left[v_{mk}(t)\right] = \dot{v}_{mk}^2(t) - v_{mk}(t)\ddot{v}_{mk}(t) \tag{10}$$

Computing its mean value, only two quantities are necessary

$$\begin{aligned} E\left[\dot{v}_{mk}^2(t)\right] &= -R_{mk}^{(2)}(0) \\ E\left[v_{mk}(t)\ddot{v}_{mk}(t)\right] &= R_{mk}^{(2)}(0) \end{aligned} \tag{11}$$

The second time-derivative of the autocorrelation function $R_{mk}^{(2)}(\tau)$ of the filtered noise process, at $\tau = 0$, is given by

$$R_{mk}^{(2)}(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (j\omega)^2 \left|G_k(\omega)\right|^2 \Phi_m(\omega)\, d\omega \tag{12}$$

It can be approximated, as in [15], by

$$R_{mk}^{(2)}(0) \approx R_{mk}^{(2)}(\omega_k(t)) = \omega_k^2(t) \left|G_k(\omega_k(t))\right|^2 \Gamma_{mk} \tag{13}$$

where

$$\Gamma_{mk} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left| \frac{G_k(\omega)}{G_k(\omega_c)} \right|^2 \Phi_m(\omega)\, d\omega$$

is the concentration of noise power within the passband of the filter $g_k(t)$.

Finally, from Eqs. (6, 8, 10, 11, 13) and noting that the last two terms on the right side of Eq. (6) are zero mean, the mean value of the TK-energy $\Psi\left[y_{mk}(t)\right]$ is

$$\begin{aligned} E\left[\Psi\left[y_{mk}(t)\right]\right] =& a_k^2(t)\omega_k^2(t) \left|G_k\left[\omega_k(t)\right]\right|^2 \\ &+ \underbrace{2\omega_k^2(t) \left|G_k\left[\omega_k(t)\right]\right|^2 \Gamma_{mk}}_{\text{Error Term}} \end{aligned} \tag{14}$$

The last term at the right side of Eq. (14) is an 'error' term due to the presence of noise. Therefore, the estimation of the mean value of the TK-energy of the passband source signal contains an additional term that corresponds to the mean TK-energy of the noise component that contaminates the specific subband.

4. Band selection based on Minimum Mean Teager-Kaiser Energy

It is simple to show that we should minimize the $E\left[\Psi\left[y_{mk}(t)\right]\right]$ quantity, which is closely related to the noise power within the passband of the filter, to minimize the error term of Eq. (14). It is, also, straightforward to conclude that the least affected, by noise, passband signal is the one corresponding to the minimum mean TK-energy. Due to the M sensors of the microphone array, M signals are available. So, we can choose among these M signals and their N bandpassed signal components those that appear to have the minimum mean TK-energy, one for each frequency band. For example, selecting the least affected k^{th} bandpass signal, where $k = 0, 1, \ldots, N-1$, we should estimate the mean TK-energy of all the (m, k), $m = 0, 1, \ldots, M-1$ bandpass signals and choose that one with the minimum value. This is explained by the fact that the mean TK-energy of the source speech signal's content, in a certain frequency band, is the same across all sensors. The only term varying in the noisy mean TK-energy estimation, $E\left[\Psi\left[y_{mk}(t)\right]\right]$, is the error term in Eq. (14) due to different amounts of noise reaching the various sensors of the array. Thus, selecting the signals $y_{mk}(t)$ with the minimum mean TK-Energies, we, in fact, select those passband signals that are less affected by noise.

Based on this analysis, we propose the following feature extraction algorithm:

i. Time-alignment of the M recorded signals,

ii. Filtering the time-aligned signals through an auditory-inspired filterbank with N filters, similarly to the process mentioned in Section 2, and thus creating MxN bandpass signals,

iii. Framing these MxN signals and computation of their mean TK-energy coefficients,

iv. Selection among each M subband components the one that appears to have the minimum mean TK-energy coefficient,

v. Estimation of the TECC coefficients considering only the selected subband mean TK-Energies, as in Section 2.

Having selected those subbands with the minimum mean TK-energy quantity *Multi-Band, Multi-Sensor Cepstral Coefficients* (MBSC-Min), a single output vector is obtained where all its coefficients are the least affected by noise. The estimating feature algorithm is formulated, as

$$\text{TK-Coefficient}_k(\tau) = \min_{\text{M}} E\left[\Psi\left[y_{mk}(t)\right]\right] \tag{15}$$

where τ is time (in frames), $y_{mk}(t)$ is the subband signal produced as the output of the k^{th} filter of the filterbank and the m^{th} sensor of the array. The rest are the same as in Section 2.

A second approach was examined setting the TK-energy coefficients as the mean values of the mean subband TK-Energies (*MBSC-Mean*)

$$\text{TK-Coefficient}_k(\tau) = \frac{1}{M} \sum_{m=0}^{M-1} E\left[\Psi\left[y_{mk}(t)\right]\right] \tag{16}$$

Some additional features have been examined based on different selection criteria, such as the median value or the A-Trimmed mean values of the sorted vector of the corresponding subbands TK-energies. The experimental results on ASR tasks have shown some improvements when compared to the original noisy speech task. Though, the best results are obtained when the minimum TK-energy coefficients (MBSC-Min) are selected, as presented in Table 1.

5. ASR Experiments and Results

The speech data set, used for the experiments, is a subset of the TIDIGITS database recorded in a room with diffuse noise. This data set contains about 10 recordings from each one of 52 male and 52 female adult speakers. These recordings are collected by a linear microphone array consisting of 16 sensors with a 2 cm spacing between adjacent sensors. The desired speech source is positioned directly in front of the array at a distance of $1.3\ m$ from its center. The diffuse noise field is created by several loudspeakers emitting noise with average $SNR = 0\ dB$. All of the recordings are sampled at 16 kHz.

Several algorithms, originally proposed for speech enhancement tasks, are used for the ASR feature extraction process. These algorithms are the MVDR, Zelinski and McCowan algorithms, as in Section 1 and [2, 4]. The extracted features are processed by the HTK Toolkit to examine their performance in ASR tasks. For the needs of the ASR tasks, the database is divided into 2 non-overlapping sets; 700 of the recorded sentences are used for training and the rest 300 are used for testing. The database was recorded by broadcasting the *Source Speech* set through a loudspeaker placed in the middle of the room. The *Clean Speech* set consists of those recordings held without the existence of any noise field. For the case of the *Clean and Noisy Speech* feature sets, as appeared in Table 1, we keep only the

center microphone recordings, ignoring the rest of the recordings. This way, we can apply single-microphone feature extraction algorithms to these signals. Finally, context-independent, 12-state, left-right word HMMs with 3 gaussian mixtures are used. The grammar used is the all-pair, unweighted grammar.

The features, extracted by the speech enhancing algorithms, are the common MFCCs plus their 1^{st} and 2^{nd}-order time-derivatives (D+DD). It is, also, examined if the Cepstral Mean Subtraction (CMS) scheme can improve their performance. The MFCCs are extracted directly from the frequency versions of the enhanced signals to avoid inserting additional modeling errors. For the single-microphone case, the MFCCs and TECCs have been extracted using only the center-microphone recordings, as described above. The *Clean Signal TECCs* present improved performance when compared to the *Clean Signal MFCCs* and the **MBSC-Min** to the *Noisy Signal TECCs*, correspondingly. Finally, the corresponding time-derivatives and CMS are estimated for the *Multi-Band, Multi-Sensor Cepstral Coefficients* (MBSC-Mean and -Min) feature sets, too.

Correct Word Accuracies (%)				
	Input Signal - Features	D+DD	D+DD +CMS	
Single Mic. Input	Source - MFCC	95.61	96.82	Clean Speech
	Clean - MFCC	95.48	96.37	
	Clean - TECC	96.50	96.63	
	Noisy - MFCC	94.02	94.98	Noisy Speech
	Noisy - TECC	93.38	95.36	
Multi Mic. Input	McCowan - MFCC	93.51	93.83	
	Zelinski - MFCC	94.34	95.48	
	MVDR - MFCC	94.78	95.67	
	MBSC-Mean	**95.67**	**95.80**	
	MBSC-Min	**96.12**	**96.12**	

Table 1: Speech Recognition Results (Correct Word Accuracies %) for source, clean and noisy input speech signals. The results correspond to single- and multi-sensor speech recordings and MFCC and TECC features.

This speech database was not originally designed for ASR tasks so it lacks of training and testing variability in the speaker and sentence fields. The recording conditions are considered matched, as training/testing recording conditions are the same.

6. Discussion – Conclusions

In this paper, we are presenting a novel feature extraction algorithm. We propose minimizing the noise interference over multiple bands of the speech spectrum across multisensor signals. The application of such multisensor, multiband approach yields improved recognition results, as presented in Table 1. The results obtained by the enhancing algorithms are far worse than those of the clean speech results. Most of the speech enhancing algorithms tend to create artifacts in the lower part of the speech spectrum destroying the corresponding formant structure of the specific band. This is the reason why these algorithms appear to have such a poor performance in the ASR tasks. On the contrary, the proposed algorithm succeed in preserving the formant structure and suppress most of the noise over multiple subbands. In addition, the proposed features do not need any prior knowledge of the noise field model. The recognition results are very encouraging for most of the presented features (MBSC-Min and -Mean sets). Though, further research is needed. The proposed methodology is, also, applied to speech enhancement tasks with comparable improvements of its performance.

7. Acknowledgements

We wish to gratefully acknowledge help from M. Matassoni and P. Svaizer, at FBKITC-IRST, for kindly providing the multisensor speech database. This research work was supported in part by the FP6-IST STREP program 'HIWIRE' of the European Union and in part by the Greek research programs 'ΕΗΓ10-GridNews' and 'ΠΕΝΕΔ 2003-ΕΔ554' of the General Secretariat for Research and Technology.

8. References

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, 1988.

[2] R. Zelinski, "A Microphone Array With Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *ICASSP*, 1988.

[3] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction," in *ICASSP*, 1997.

[4] I. A. McCowan and H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. Speech and Audio Processing*, 2003.

[5] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of Noise Reduction Techniques Based on Microphone Arrays with Postfiltering," *IEEE Trans. Speech and Audio Processing*, 1988.

[6] S. Fischer and K. D. Kammeyer, "Broadband Beamforming With Adaptive Postfiltering for Speech Acquisition in Noisy Environments," in *ICASSP*, 1997.

[7] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition," in *Eurospeech*, 2005.

[8] J. F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal," in *ICASSP*, 1990.

[9] T. Irino and R. D. Patterson, "A Time-Domain, Level-Dependent Auditory Filter: The Gammachirp," *Journ. Acoust. Society of America*, 1997.

[10] O. Ghitza, "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition," *IEEE Trans. Speech and Audio Processing*, 1994.

[11] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, 1990.

[12] M. D. Skowronski and J. G. Harris, "Increased MFCC Filter Bandwidth for Noise-Robust Phoneme Recognition," in *ICASSP*, 2002.

[13] J. F. Kaiser, "Some Useful Properties of Teager's Energy Operators," in *ICASSP*, 1993.

[14] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Processing*, 1993.

[15] 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*, 1993.

[16] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1991.