MULTISENSOR MULTIBAND CROSS-ENERGY TRACKING FOR FEATURE EXTRACTION AND RECOGNITION

Stamatios Lefkimmiatis, Petros Maragos and Athanassios Katsamanis

National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece.
Email:[sleukim,maragos,nkatsam]@cs.ntua.gr

ABSTRACT

In this paper, we present a multisensor multiband energy tracking scheme for robust feature extraction in noisy environments. We introduce a multisensor feature extraction algorithm which combines both the spatial and frequency information incorporated in the speech signals captured by a microphone array. This is based on the estimation of cross-energies over multiple sensors and minimization of an error term due to noise. The relevant noise-analysis is given. Automatic Speech Recognition (ASR) experiments at various SNR levels demonstrate that the newly proposed frontend performs better than alternative schemes, especially in noisy conditions.

Index Terms— Robust Feature Extraction, ASR, Energy Tracking, Teager Energy, Microphone Array

1. INTRODUCTION

A major concern of Human-Machine Interaction (HMI) is to improve the interactions between users and computers by making computers more usable and receptive to the user's needs. Speech recognition has been one of the leading technologies to accomplish this goal. For this reason significant efforts have been made and several ASR systems have been developed and perform satisfactorily. However, the majority of the current single-channel solutions suffer from two serious drawbacks. Specifically, their efficacy degrades significantly when speech is contaminated with noise. Further, in most applications, users are required to wear head-mounted close-talking microphones. Proximity of the microphone to the speaker can ensure a high speech signal level which can partly compensate for the presence of environmental noise.

An emerging area of research which can offer a potential solution to both constraints focuses on the use of microphone arrays. The main advantage of multisensor techniques over the standard single-channel solutions is that they provide "richer" information about the acoustic environment. This is achieved by exploiting the spatial diversity of the acoustic signals to be recognized and noise, since the corresponding sources are usually physically separated in space. So, it is expected that microphone arrays can improve recognition performance, especially in the case of noisy and reverberant environments.

State of the art multisensor speech recognition systems currently apply microphone array processing as a separate noise-suppression frontend module. Input signals are filtered by a beamformer and acoustic features (typically MFCC) for ASR are then extracted from the denoised output signal. The most known and efficient beamforming algorithm is the Minimum Variance Distortionless Response (MVDR) beamformer [1]. This beamformer has the important property that maximizes the *array gain* which is a measure of the increase in signal-to-noise ratio (SNR) that is obtained by using an array rather than a single microphone. However, the MVDR beamformer suffers from a serious drawback. Its directivity factor, which is a measure describing the ability of the beamformer to suppress the noise field, is low in the lower frequency regions and thus MVDR is incapable of sufficiently removing the noise in those regions. In addition, the filtering process distorts the speech spectrum resulting in poor ASR performance.

Alternatively, we propose a multisensor feature extraction scheme. We investigate the potential of exploiting a multiband decomposition scheme for multisensor acoustic processing in noisy environments. Based on energy tracking, via the nonlinear Teager-Kaiser operator (TEO) [2], the least affected by noise subbands across the sensors of the microphone array are combined (see Fig. 1). Then the energy features from these combined subbands are extracted. This procedure is described in two steps for clarity but can be efficiently realized in a single step.

The organization of the rest of the paper is as follows: in Section 2 we review the theoretical background of the nonlinear *Teager-Kaiser* energy operator and briefly describe the Gabor filterbank that is used for the multiband decomposition. In Section 3 an analysis on how the noise affects the energy measurements at every subband is provided. In Section 4, based on the noise analysis, we propose a multisensor feature extraction method which combines the energy measurements over multiple sensors. Finally, results of multisensor ASR experiments are presented in Section 5, while conclusions and future work are found in Section 6.

2. BACKGROUND

In [2] Kaiser, based on Teager's previous work, introduced a nonlinear differential operator called *Teager-Kaiser* energy operator (TEO) Ψ. This operator can track the instantaneous energy of a source producing an oscillation. When Ψ is operating on a continuous-time signal is given by $\Psi\left(x(t)\right) = \dot{x}(t)^2 - x(t)\ddot{x}(t)$, where $\dot{x}(t)$ and $\ddot{x}(t)$ indicate the first and second time derivative of the argument. Applied to an AM-FM signal of the form $x(t) = a(t)\cos\left(\phi(t)\right)$, yields the instantaneous source energy, i.e. $\Psi\left(x(t)\right) \approx a(t)^2\omega_i(t)^2$, where the approximation error becomes negligible [3] if the instantaneous amplitude $a(t)$ and instantaneous frequency $\omega_i(t) = \dot{\phi}(t)$, do not vary too fast or too much with respect to the average value of $\omega_i(t)$. In this work instead of using the "traditional" signal energy approximation of the mean square amplitude, that only takes into account the kinetic energy of the signal's source, we will use the TEO

This work was supported by grant ΠΕΝΕΔ 2003-ΕΔ554 & 866 [co-financed by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%)], by the Greek research program ΕΗΓ10-GridNews and the EC project HIWIRE.

for computing the total source energy, which hereafter we will call *Teager* energy.

In order to apply the TEO speech or any wideband signal, it is necessary first to filter the signal and isolate specific frequency bands. This necessity comes from the fact that the operator cannot perform well in multi-component signals due to inherent limitations of the algorithm. However, this filtering process is also a common strategy followed in the majority of feature extraction algorithms, like MFCCs. In this paper, the observed signals at the outputs of the sensors, in the microphone array, are filtered through a filterbank of 35 overlapping Mel-spaced Gabor filters. These filters are chosen as an optimum candidate for being compact, smooth and minimum uncertainty filters, i.e. their rms time and frequency bandwidth product attains the minimum value in the uncertainty principle inequality [3].

3. NOISE ANALYSIS ON MULTISENSOR CROSS ENERGIES

Let us consider an M-sensor linear microphone array in a noisy environment that captures the waveform of a desired source signal. The observed signal, $y'_m(t)$, $m = 0, \ldots, M-1$, at the mth sensor corresponds to a linearly filtered version of the source signal $s(t)$, plus an additive noise component $v'_m(t)$. The additive noise component is assumed to be a zero mean, wide-sense stationary (WSS) Gaussian random process with autocorrelation function $R_m(\tau)$ and spectral density $\Phi_m(\omega)$. The noise components observed at two different sensors m, k are also considered joint WSS processes with cross-correlation function $R_{mk}(\tau)$ and cross-spectral density $\Phi_{mk}(\omega)$. The signals received by the sensors of the microphone array are fed into a time alignment module to account for the effects of propagation. In this work we do not address the problem of reverberation, thus we assume that the output signals can be denoted as :

$$y_m(t) = s(t) + v_m(t),\, m = 0, \ldots, M-1 \tag{1}$$

In the multisensor-multiband scheme, every aligned input signal is decomposed into N subband signals by the analysis filterbank. Let us denote with y_{mj} the signal observed at the output of the mth sensor and filtered by the jth filter of the filterbank. This decomposition can be expressed as:

$$y_{mj}(t) = y_m(t) * g_j(t),\, j = 0, \ldots, N-1, \tag{2}$$

where $*$ denotes convolution. Estimating the cross Teager energy [4] between a sensor pair $(m\,, k)$ of the filtered (bandpass) signals by the jth filter of the filterbank results in :

$$\Psi_c\left[y_{mj}(t), y_{kj}(t)\right] = (\dot{y}_{mj}(t)\dot{y}_{kj}(t)) - y_{mj}(t)\ddot{y}_{kj}(t)\,. \tag{3}$$

We can expand this expression using Eqs. (1) and (2) to obtain :

$$\begin{aligned}\Psi_c\left[y_{mj}(t), y_{kj}(t)\right] =& \Psi\left[s_j(t)\right] + \Psi_c\left[v_{mj}(t), v_{kj}(t)\right] + \\ & \Psi_c\left[s_j(t), v_{kj}(t)\right] + \Psi_c\left[v_{mj}(t), s_j(t)\right]\end{aligned} \tag{4}$$

The three last terms on the right side of equation (4) are error terms due to noise. If we now take the mean of Eq. (4) we will have

$$E\left\{\Psi_c\left[y_{mj}(t), y_{kj}(t)\right]\right\} = E\left\{\Psi\left[s_j(t)\right]\right\} + E\left\{\Psi_c\left[v_{mj}(t), v_{kj}(t)\right]\right\}\,, \tag{5}$$

since the last two terms on the right side of Eq. (4) are zero mean.

In order to simplify the analysis we make a fundamental assumption: the signal $s(t)$ is well approximated by an AM-FM signal, $s(t) = a(t)\cos\left[\phi(t)\right]$, with both time-varying amplitude $a(t)$ and time-varying instantaneous frequency $\omega_i(t) = \dot{\phi}(t)$. Such an approximation is well motivated for speech signals, since experimental results have produced strong evidences for the existence of amplitude and frequency modulations (AM–FM) in speech resonance signals [3]. In this case, the Teager energy of $s(t)$ will be approximately equal to : $\Psi\left[s(t)\right] \approx a(t)^2\omega_i(t)^2$. Moreover, under this assumption the bandpass signal $s_j(t)$ can be approximated as [5] :

$$\hat{s}_j(t) = a(t)\left|G_j\left[\omega_i(t)\right]\right|\cos\left\{\phi(t) + \angle G_j\left[\omega_i(t)\right]\right\} \tag{6}$$

and thus the Teager energy of the filtered signal $s_j(t)$ will equal to

$$\Psi\left[s_j(t)\right] = a(t)^2\omega_i(t)^2\left|G_j\left[\omega_i(t)\right]\right|^2\,. \tag{7}$$

At this point we will focus on the $E\left\{\Psi_c\left[v_{mj}(t), v_{kj}(t)\right]\right\}$ term of Eq. (5). Since the noise processes $v_m(t)$, $v_k(t)$ have cross spectral density $\Phi_{mk}(\omega)$, the filtered noise processes will have cross spectral density $\Phi_{(mk)j}(\omega) = \left|G_j(\omega)\right|^2\Phi_{mk}(\omega)$. In addition, since $v_{mj}(t), v_{kj}(t)$ are WSS Gaussians, the processes $\dot{v}_{mj}(t), \dot{v}_{kj}(t)$ and $\ddot{v}_{kj}(t)$ are also WSS Gaussians, and the product $\dot{v}_{mj}(t)\dot{v}_{kj}(t)$ is statistically independent of both $v_{mj}(t)$ and $\ddot{v}_{kj}(t)$ [6]. Therefore the energy operator output

$$\Psi_c\left[v_{mj}(t), v_{kj}(t)\right] = \dot{v}_{mj}(t)\dot{v}_{kj}(t) - v_{mj}(t)\ddot{v}_{kj}(t) \tag{8}$$

is the sum of two independent processes. To compute the mean of this term we have to estimate the following two quantities

$$\begin{aligned}E\left[\dot{v}_{mj}(t)\dot{v}_{kj}(t)\right] &= -R^{(2)}_{(mk)j}(0)\\ E\left[v_{mj}(t)\ddot{v}_{kj}(t)\right] &= R^{(2)}_{(mk)j}(0)\,.\end{aligned} \tag{9}$$

Of interest are the values that the second derivative of the cross-correlation, $R_{(mk)j}(\tau)$, takes at the origin

$$R^{(2)}_{(mk)j}(0) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}(j\omega)^2\left|G_j(\omega)\right|^2\Phi_{mk}(\omega)\,d\omega\,. \tag{10}$$

An approximation of this quantity similar to the one proposed in [5] can be : $R^{(2k)}_{(mk)j}(0) = \hat{R}^{(2k)}_{(mk)j}(\omega_i(t))$ where

$$\hat{R}^{(2k)}_{(mk)j}(\omega_i(t)) = (-1)^k\omega_i(t)^{2k}\left|G_j(\omega_i(t))\right|^2\Gamma_{(mk)j}\,, \tag{11}$$

with $\Gamma_{(mk)j} = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\left|\frac{G_j(\omega)}{G_j(\omega_c)}\right|^2\Phi_{mk}(\omega)\,d\omega$, the concentration of noise power within the passband of the filter $g_j(t)$. Combining Eqs. (5), (7), (8), (9), (11) we find that the mean value of the cross Teager energy $\Psi_c\left[y_{mj}(t), y_{kj}(t)\right]$ equals to

$$\begin{aligned}E\left\{\Psi_c\left[y_{mj}(t), y_{kj}(t)\right]\right\} =& E\left\{\Psi\left[s_j(t)\right]\right\}\\ & + \underbrace{2\omega_i(t)^2\left|G_j\left[\omega_i(t)\right]\right|^2\Gamma_{(mk)j}}_{\text{Error Term}}\end{aligned} \tag{12}$$

At this point of the analysis we have to note that in the case where $m = k$, all the results still hold but instead of $R_{(mk)j}$ and $\Phi_{(mk)j}$ we will have R_{mj} and Φ_{mj} respectively. Using the following inequality which appears in [6],

$$\left|\int_{-\infty}^{+\infty}\Phi_{(mk)j}(\omega)\,d\omega\right|^2 \leq \int_{-\infty}^{+\infty}\Phi_{mj}(\omega)\,d\omega\int_{-\infty}^{+\infty}\Phi_{kj}(\omega)\,d\omega \tag{13}$$

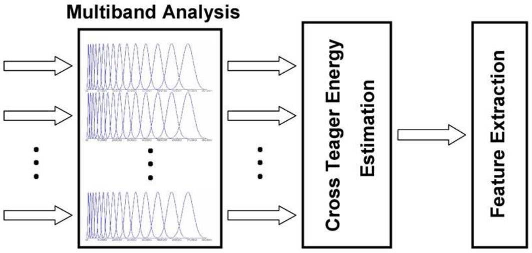

Fig. 1: Block Diagram of the Multisensor Feature Extraction Scheme.

we are led to the final inequallity

$$\left|\Gamma_{(mk)j}\right|^2 \leq \Gamma_{mj}\Gamma_{kj}\,, \tag{14}$$

which will be proved useful for the efficiency of the proposed multisensor feature extraction scheme.

4. FEATURE EXTRACTION BASED ON MINIMUM MEAN CROSS TEAGER ENERGY

Inspired by the above analysis and by the fact that the mean Teager energy of the bandpass signals has been used with success for single-channel feature extraction algorithms [7], we propose a multisensor feature extraction algorithm which combines the benefits of the "richer" acoustic information provided by the Teager energy and the benefits of the spatial information provided by the microphone array.

In the proposed scheme our goal is to minimize the error term of Eq. (12) which distorts the energy measurement of the clean bandpass signal $s_j(t)$. In order to minimize this term we have to minimize $\Gamma_{(mk)j}$ which is the concentration of the noise power within the passband of the filter. For the jth subband, since we have $2\cdot\binom{M}{2}$ possible sensor pairs[1] in the microphone array, a straightforward way to select the least distorted subband energy measurement, is to compute all the possible mean cross Teager energy measurements of the form of (12), and choose the one which corresponds to the minimum value. This energy measurement is clearly the one which is closer to the mean energy of the clean bandpass signal $s_j(t)$. For example, to select the least affected energy measurement for the jth subband we estimate the mean cross Teager energy of all the pairs $[(m,k),j]$, $m,k = 0,1,\ldots,M-1$, and choose the one with the minimum value. This approach is justified by the fact that the mean Teager energy of the bandpass source signal s_j will remain the same in all the measurements. The only term that will vary in the noisy mean cross Teager energy, $E\left\{\Psi_c\left[y_{mj}(t), y_{kj}(t)\right]\right\}$, will be the error term of Eq. (12) due to the different percentage of noise at the various sensors of the array.

Based on these findings we propose a multisensor feature extraction scheme which is summarized in Table 1 and illustrated in Fig. 1 as a block diagram. An issue that arises from the described method is its computational complexity, since for every subband frame we have to compute $2\cdot\binom{M}{2}$ results. However, if we consider the inequality in Eq. (14), then we can succeed a significant reduction of computations. According to (14), since the invoked quantities are positives and reals, the value of $\Gamma_{(mk)j}$ will be at least smaller than one of Γ_{mj} and Γ_{kj}. Therefore, it is sufficient instead of computing all the possible cross-results (step-2 of the algorithm) to compute the M mean Teager measurements $E\left\{\Psi\left[y_{mj}(t)\right]\right\}$, $m = 0,\ldots,M-1$. Then we can select the two sensors (let them be p,q) which produce the minimum values for the specific subband (let it be j), and estimate the two mean cross Teager energies of the sensor pairs (p,q), (q,p). Finally, we may choose the minimum value of the corresponding results as :

$$\begin{aligned}\min\{&E\left\{\Psi\left[y_{pj}(t)\right]\right\}, E\left\{\Psi\left[y_{qj}(t)\right]\right\},\\ &E\left\{\Psi_c\left[y_{pj}(t), y_{qj}(t)\right]\right\}, E\left\{\Psi_c\left[y_{qj}(t), y_{pj}(t)\right]\right\}\}\,.\end{aligned}$$

The selected energy measurement is guaranteed to be the least distorted from all the cross-enegies produced in the second step of the described scheme. In addition, this procedure needs only $(M+2)$ computations instead of $2\cdot\binom{M}{2}$.

Table 1 Multisensor Feature Extraction

1. Use the Gabor filterbank described in Section 2 to produce N bandpass signals for each one of the M input speech signals.
2. Estimate the short-time mean value of the cross Teager energies of all the (m,k) sensor pairs, for each one of the N bandpass signals. The short-time averaging window has duration of $30ms$ and the window shift $10ms$.
3. For every subband frame of the filterbank select the mean cross Teager energy measurement, among the $2\cdot\binom{M}{2}$ results, with the minimum value.
4. Compute via the DCT transform the cepstrum coefficients of the log short-time mean cross Teager energies.
5. Keep only the first 12 coefficients, $c_1 - c_{12}$. The zero'th-coefficient, c_0, augments the final feature vector as is also common in a typical MFCC-based frontend.

4.1. Bandpass TEO Estimation

In this section we describe a method that produces more accurate and smoother estimations of the bandpass Teager energy [8] and we provide a solution to reduce the computational complexity of the method.

The continuous-time TEO Ψ, combined with bandpass filtering and sampled at time instances $t = nT$, is given by : $\Psi\left[y(t)\right] = \dot{y}^2(t) - y(t)\ddot{y}(t)\;|_{t=nT}$, $\quad y(t) = x(t) * g(t)$, where $x(t)$ is the continuous time signal and $g(t)$ the filter's impulse response. Since convolution commutes with time-differentiation, i.e. $\frac{d^n}{dt^n}\left(x(t) * g(t)\right) = x(t) * \frac{d^n}{dt^n}g(t)$, $n = 1,2,\ldots$, operator Ψ can be written as :

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

In order to avoid the three convolutions which are time-consuming operations we can move, as proposed in [9], to the Fourier domain where the convolution becomes multiplication. Then since it holds : $\frac{d^n g(t)}{dt^n} \overset{\mathcal{F}}{\longleftrightarrow} (j\omega)^n G(\omega)$, we can compute the product $X(\omega)G(\omega)$ once and use it three times. Thus the Teager Energy of a bandpass signal can be estimated more efficiently if we express it as :

$$\begin{aligned}\Psi\left[y(t)\right] &= \left[\mathcal{F}^{-1}\left\{X(\omega)\cdot(j\omega)G(\omega)\right\}\right]^2\\ &\quad - \left[\mathcal{F}^{-1}\left\{X(\omega)\cdot G(\omega)\right\}\right]\left[\mathcal{F}^{-1}\left\{X(\omega)\cdot(j\omega)^2G(\omega)\right\}\right]\end{aligned}$$

where with $\mathcal{F}$ and $\mathcal{F}^{-1}$ we denote the *Fourier* transform and its inverse, respectively. In practice, the *Fourier* operations are implemented via FFT's.

[1] We care about the order of the sensor pairs since in general $\Psi_c\left[x(t), y(t)\right] \neq \Psi_c\left[y(t), x(t)\right]$.

5. ASR EXPERIMENTS AND RESULTS

To validate the effectiveness of the proposed feature extraction technique, we compare its performance with the single channel features TECC [7], which are also based on Teager Energies, and are extracted from the central sensor of the array . For comparison we also extract TECC features from the enhanced output after beamforming. In the last case we use two kinds of beamforming algorithms, Delay and Sum (DS) [10] and MVDR. This approach of combining the inputs by beamforming and then extracting the features is commonly used [11] and can be considered as the state of the art in multisensor ASR. In these experiments we do not use the MFCCs features in order to base our comparisons on the same type of energy and since TECCs have shown to produce similar results with MFCCs.

The speech data set, used for the ASR experiments, is a subset of the TIDIGITS database. It contains about 1000 recordings from 52 male and 52 female adult speakers contaminated by noise at various SNRs. The recordings are collected by a linear microphone array of 8 sensors with 2 cm spacing between adjacent sensors, at a sampling frequency of 16 KHz. The desired speech source is positioned directly in front of the array at a distance of 1.3 m from its center. The experiments are performed using the HTK system (context-independent, 14-state, left-right word HMMs with three Gaussian mixtures). The HMMs are trained on the clean data set (700 utterances) and tested on 5 noisy sets (5 SNRs $\times$ 300 utterances). All feature vectors are extended by their time derivatives ($\Delta, \Delta\Delta$).

Table 2 shows the average results per SNR for all the methods. With the acronym "MCTEF" standing for Multisensor Cross Teager Energy Features we refer to the features obtained by the proposed multisensor feature extraction method described in Table 1.

For the clean case and the high SNR we observe that MCTEF performs slightly worse than the Delay and Sum beamforming. However, for mid and lower SNRs the performance is consistently better than all the methods. Specifically, the improvement succeeded by MCTEF with respect to the other methods gets greater for the cases of 0-10 dBs where the noise distorts significantly the input signals. As for the clean case and the high SNR, the results are justified, since the cross-energies among the sensors are expected to have great correlation and the differences between them to be insignificant. Thus, for cases where the speech signals do not suffer greatly from noise distortions we can avoid estimating the cross-energies. Instead, we can estimate only the mean Teager energies from each sensor of the array and choose the minimum one. This can save us computational time, while the performance remains at the same level.

6. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we have introduced the idea of combining the spatial and frequency information incorporated in the speech signals, captured by a microphone array, in order to extract features that perform better in noisy conditions. This is achieved by estimating Teager cross-energies across the sensors of the array and using the one which is less distorted to produce the final feature vector. The presented results in ASR validate the theoretical analysis provided in Section 3 and indicate the proposed scheme as a promising multisensor feature extraction method.

The overall analysis and methodology presented focus on Teager energy measurements for feature extraction. However, the proposed scheme can be properly generalized to exploit the typical energy measurements as used in the widely applied MFCC feature extraction procedure. We are currently working in this direction so that the multiresolution feature extraction algorithm will also work with MFCC features.

Table 2: Speech Recognition Results (Average Word Accuracy (%) for Different Training/Testing Scenarios).

Correct Word Accuracies (%)						
SNR/ Methods	clean	20dB	15dB	10dB	5dB	0dB
MCTEF	98.37	94.69	**85.70**	**74.46**	**60.57**	**42.59**
TECC	98.06	93.46	83.45	69.66	54.03	33.91
MVDR+TECC	98.26	93.97	84.58	69.21	52.20	33.91
DS+TECC	**98.57**	**94.79**	84.03	67.93	51.69	34.53

Acknowledgements : The authors wish to thank M. Matassoni and P. Svaizer, at FBKITC-IRST, for kindly providing the multisensor speech database, and D. Dimitriadis, at NTUA, for helpful discussions about TECC features.

7. REFERENCES

[1] H. Cox, R. M. Zeskind, and M. W. Owen, "Robust Adaptive Beamforming," *IEEE Trans. Speech and Audio Processing*, vol. 35, pp. 1365–1376, 1987.

[2] J. F. Kaiser, "On A Simple Algorithm to Calculate the "Energy" of a Signal," in *ICASSP*, 1990, vol. 5, pp. 381–384.

[3] 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, pp. 3024–3051, 1993.

[4] J. F. Kaiser, "Some useful properties of Teager's energy operators," in *ICASSP*, 1993, pp. 149–152.

[5] 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*, vol. 41, pp. 3245–3265, 1993.

[6] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, International Edition, McGraw-Hill, 1991.

[7] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory Teager Energy Cepstrum Coefficients for Robust Speech Recognition," in *Interspeech*, 2005, pp. 3013–3016.

[8] D. Dimitriadis and P. Maragos, "Continuous Energy Demodulation Methods and Application to Speech Analysis,," *Speech Communication*, vol. 48, pp. 819–837, 2006.

[9] G. Evangelopoulos, I. Kokkinos, and P. Maragos, "Advances in Variational Image Segmentation using AM-FM models: Regularized Demodulation and Probabilistic Cue Integration," in *Proc. of Third Int. Workshop, VLSM*, 2005, pp. 121–136.

[10] H. Van Trees, *Optimum Array Processing, Part IV of Detection, Estimation and Modulation Theory*, Wiley, 2002.

[11] T. B. Hughes, H.-S. Kim, J. H. DiBiase, and H. F. Silverman, "Performance of an HMM Speech Recognizer Using a Real-Time Tracking Microphone Array as Input," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 346–349, 1999.