Proc. 25th European Signal Processing Conference (EUSIPCO-2017), Kos, Greece, Aug.2017.

On the Improvement of Modulation Features Using Multi-Microphone Energy Tracking for Robust Distant Speech Recognition

Isidoros Rodomagoulakis and Petros Maragos
School of ECE, National Technical University of Athens, 15773 Athens, Greece
irodoma@cs.ntua.gr, maragos@cs.ntua.gr

***Abstract*—In this work, we investigate robust speech energy estimation and tracking schemes aiming at improved energy-based multiband speech demodulation and feature extraction for multi-microphone distant speech recognition. Based on the spatial diversity of the speech and noise recordings of a multi-microphone setup, the proposed Multichannel, Multiband Demodulation (MMD) scheme includes: 1) energy selection across the microphones that are less affected by noise and 2) cross-signal energy estimation based on the cross-Teager energy operator. Instantaneous modulations of speech resonances are estimated on the denoised energies. Second-order frequency modulation features are measured and combined with MFCCs achieving improved distant speech recognition on simulated and real data recorded in noisy and reverberant domestic environments.**

I. Introduction

Several scientific projects [18], [7] and challenges [8], [10] have been launched during the last decade targeting intelligent interfaces for indoors smart environments. Distant speech recognition (DSR) via distributed microphones is examined in most of them. State-of-the-art developments in acoustic modeling for speech recognition [21] have demonstrated high levels of recognition performance under clean conditions or high signal-to-noise ratios (SNRs), making voice-enabled user interfaces practically usable in a variety of everyday environments. However, untethered, far-field, and always-listening operation, robust to noise and reverberation, still constitutes a challenge that limits their universal applicability.

Multi-microphone setups offer flexibility on multi-source and noisy acoustic scenes by capturing the spatial diversity of speech and non-speech sources. Richer multichannel observations may be potentially exploited and fused in many stages of the recognition pipeline. To name a few established approaches in the literature from early to late fusion: channel selection, beamforming, feature enhancement, and rescoring have brought notable improvements to recognition rates. More recently, some of these approaches were revised in the framework of Deep Neural Networks (DNNs) where non-linear modeling is feasible. Networks are trained to extract bottleneck features [5], and combine channels [12], achieving similar or better results compared to beamforming. However, training DNNs on multi-style and multi-channel data [20] is the main focus, while incorporating traditional array processing methods remains unexploited.

This research work was supported by the EU under the project I-SUPPORT with grant H2020-643666.

Non-linear features stemming from the AM-FM speech model were originally conceived for ASR in [4] as capturing the second-order non-linear structure of speech formants, whereas the linear speech model and its corresponding features (e.g., MFCCs) capture the first-order linear structure of speech. Their fusion exhibits robustness in noise and mismatch training/testing conditions (e.g., in Aurora-4 task), as indicated by the single-channel ASR results in recent works [5], [16]. However, only a few works [19], [15] examine their performance in reverberant environments.

Herein, we extend our previous work [19] on modulation features for DSR by proposing a multi-channel scheme for energy tracking that is robust to noise and applicable in the workflow of multiband speech demodulation for improved estimation of the AM–FM speech model parameters. Noise is minimized across the available bands and channels by selecting the "cleanest" in terms of Teager-Kaiser Energy (TKE) or by estimating cross-channel energies using the cross-TKE (CTKE) operator. A similar approach has been followed in [11] for the extraction of multisensor, multiband energy features. Although the robustness of cross-energy operators have been analyzed in early studies [14], only a few works [3] employ them.

II. Multichannel Energy Tracking

Let us denote with

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

the noisy speech recordings captured by M microphones of an array, where s is the source signal and u_m is the microphone-dependent noise. Note that reverberation effects and time alignment issues between y_m are not taken into account in the following analysis. Bandlimited speech components are obtained by decomposing y_m with a Mel-spaced Gabor filterbank $\{g_k(t)\}$:

$$y_{mk}(t) = y_m(t) * g_k(t),\ k = 0, \dots, N-1 \tag{2}$$

The signals recorded by adjacent microphones are expected to be correlated. A measure of their interaction can be given by the cross-Teager energy [9] operator Ψ_c that measures

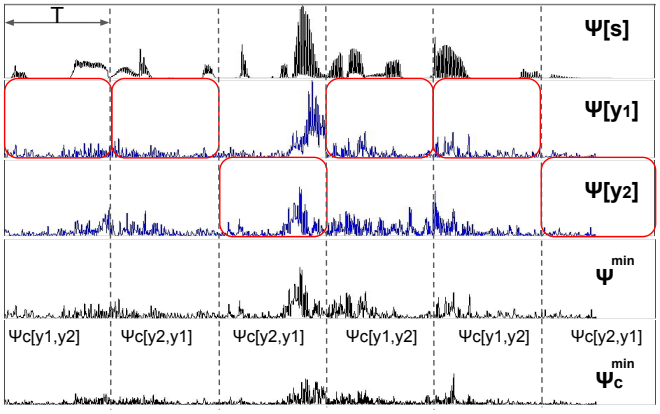


Fig. 1. Multichannel energy tracking: Given the noisy recordings y_1, y_2 (2nd and 3rd row) of an array, the minimum Teager energy Ψ^{min} is selected among them (in red rectangulars) after averaging in non-overlapping frames of duration T. The minimum cross-Teager energy $\Psi_c^{min}[y_{\hat{m}}, y_{\hat{\ell}}]$ is found between the channels $\hat{m}$ and $\hat{\ell}$ having the 1st and 2nd smaller energies.

the relative rate of change between two oscillators. More analytically:

$$\Psi_c[y_{mk}, y_{\ell k}] = \dot{y}_{mk}\dot{y}_{\ell k}(t) - y_{mk}\ddot{y}_{\ell k} \tag{3}$$

where dots and double dots correspond to the first- and second-order derivatives, respectively. Based on the analysis of [11], noise $u(t)$ contributes as an additive error term on averaging:

$$\mathcal{E}\{\Psi_c[y_{mk}, y_{\ell k}]\} = \mathcal{E}\{\Psi[s_k]\} + error \tag{4}$$

Consequently, the energy $\Psi_c^{min}[y_{\hat{m}k}, y_{\hat{\ell}k}]$ with the minimum average, formed by microphones $(\hat{m}, \hat{\ell})$, is expected to lie closer to $\Psi[s_k(t)]$. Another outcome of [11] was that instead of searching $(\hat{m}, \hat{\ell})$ among all pairs of microphones, which is computationally intensive [1], it suffices to search between microphones $\bar{m}$ and $\bar{\ell}$ having the 1st and 2nd smallest average Teager energies:

$$\Psi_c^{min}(k) = \Psi_c^{min}[y_{\hat{m}k}, y_{\hat{\ell}k}], \tag{5}$$
$$(\hat{m}, \hat{\ell}) = \arg\min_{\bar{m},\bar{\ell}} \big(\mathcal{E}\{\Psi_c^{min}[y_{\bar{m}k}, y_{\bar{\ell}k}]\}, \mathcal{E}\{\Psi_c^{min}[y_{\bar{\ell}k}, y_{\bar{m}k}]\}\big)$$

As a result, based on the fact that noise contributes as an additive term in both Teager and cross-Teager energies of the bandpass microphone signals, taking the minimum among them yields the most robust energy for demodulation. Tracking of $\Psi^{min}(k)$ and $\Psi_c^{min}(k)$, in each band k, is realized in medium-duration non-overlapping frames of T sec for fine temporal resolution against the instantaneous changes of the acoustic conditions due to noise changes and speaker's motion. An example is shown in Fig. 1, where the energy of the 3rd ($k = 3$) bandlimited component of $s(t)$ is approximated with Ψ^{min} or Ψ_c^{min}, given two real distant recordings from a two-microphone linear array.

[1] $2 \cdot \binom{M}{2}$ computations are needed for each band because $\Psi_c[y_{mk}, y_{\ell k}]\} \neq \Psi_c[y_{mk}, y_{\ell k}]$

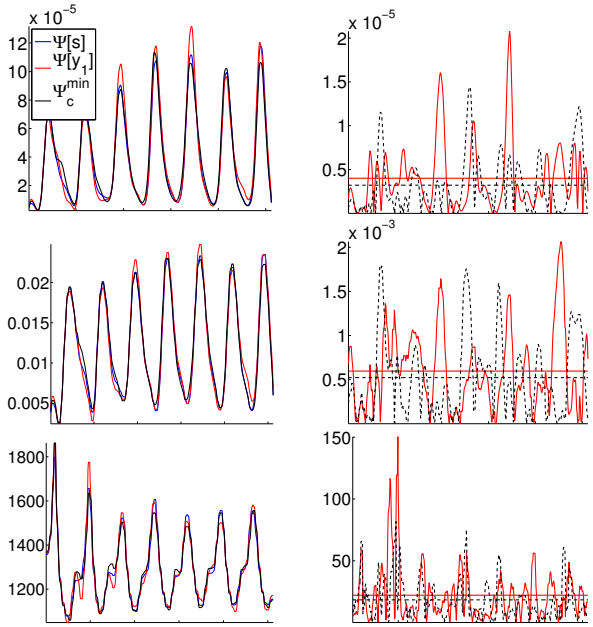


Fig. 2. Teager energies (top row), instantaneous amplitudes (middle row) and instantaneous frequencies in Hertz (bottom row) of a 32 ms long frame from the steady state of an instance of phoneme "ah". Demodulation of the 3rd speech component ($k = 3$) is realized using: a) the clean source $s(t)$ (blue lines), b) the 1st channel $y_1(t)$ of a three-microphone linear array whose signals are simulated using the Image Source Method (ISM) with noise ($SNR = 5dB$) (red lines), and c) all the simulated channels (y_1, y_2, y_3) using the proposed MMD scheme (black lines). The figures on the right column show the estimation errors, with the flat lines showing their averages.

III. Multichannel, multiband demodulation

The kth resonance of a speech signal $s(t)$ can be modeled by an AM–FM signal as

$$r_k(t) \;=\; a_k(t)\cos\Big(\int_0^t \omega_k(\tau)\,d\tau\Big)\,, \tag{6}$$

where $a_k(t)$ and $\omega_k(t)$ are its instantaneous amplitudes and angular frequencies. Given a noisy observation y_m for $s(t)$, demodulation is realized based on the widely known Energy Separation Algorithm (ESA) [13] formulas

$$\omega_k(t) \approx \sqrt{\frac{\Psi[\dot{y}_{mk}]}{\Psi[y_{mk}]}},\;\; a_k(t) \approx \frac{\Psi[y_{mk}]}{\sqrt{\Psi[\dot{y}_{mk}]}} \tag{7}$$

Smoother approximations that are more robust to noise are achieved by Gabor-ESA [6], which combines bandpass filtering in the Teager energy operator as convolution with the corresponding bandpass Gabor filter:

$$\Psi[y_{mk}] = (y_m * \dot{g}_k)^2 - (y_m * g_k)(y_m * \ddot{g}_k) \tag{8}$$
$$\Psi[\dot{y}_{mk}] = (y_m * \ddot{g}_k)^2 - (y_m * \dot{g}_k)(y_m * \dddot{g}_k) \tag{9}$$

Herein, we incorporate the "denoised" energies Ψ_c^{min} and Ψ^{min} within the Gabor-ESA framework for improved speech demodulation. The energies are tracked with the proposed multichannel scheme based on the M microphone array signals. $\Psi[y_{mk}]$ and $\Psi[\dot{y}_{mk}]$ can be substituted by two "cleaner"

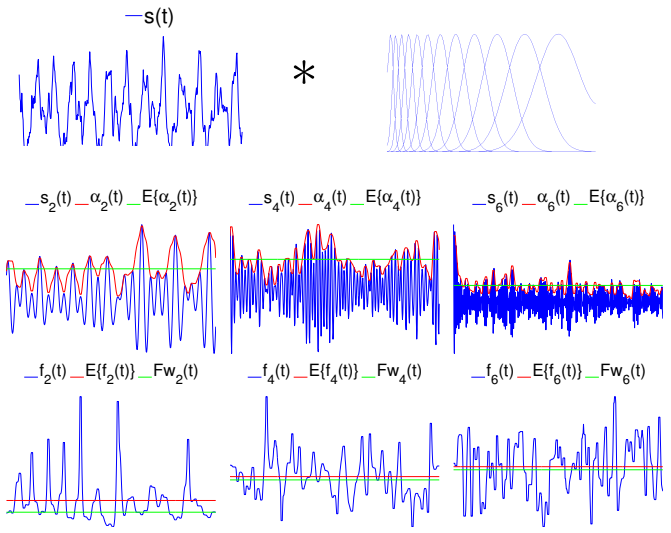


Fig. 3. Extraction of MIA, MIF, and Fw modulation features on a noisy 32 ms segment $s(t)$. Gabor-ESA with 12 filters is employed for the demodulation of each bandpass speech resonance $s_k(t)$ to its instantaneous AM–FM parameters $a_k(t)$ and $f_k(t)$.

versions:

$$1.\ \Psi[y_{\hat{m}k}, y_{\hat{\ell}k}],\ \ \Psi[\dot{y}_{\hat{m}k}, \dot{y}_{\hat{\ell}k}] \quad \text{or} \quad 2.\ \Psi[y_{\bar{m}k}],\ \ \Psi[\dot{y}_{\bar{m}k}] \tag{10}$$

In response to (8) and (9) the cross energies are:

$$\Psi[y_{\hat{m}k}, y_{\hat{\ell}k}] = (y_{\hat{m}} * \dot{g}_k)(y_{\hat{\ell}} * \dot{g}_k) - (y_{\hat{m}} * g_k)(y_{\hat{\ell}} * \ddot{g}_k) \tag{11}$$

$$\Psi[\dot{y}_{\hat{m}k}, \dot{y}_{\hat{\ell}k}] = (y_{\hat{m}} * \ddot{g}_k)(y_{\hat{\ell}} * \ddot{g}_k) - (y_{\hat{m}} * \dot{g}_k)(y_{\hat{\ell}} * \dddot{g}_k) \tag{12}$$

Figure 2 demonstrates an example of how the energy of a bandlimited component of a clean utterance recorded by a close-talk microphone is better approximated by the multichannel energy Ψ_c^{min} compared to $\Psi[y_1]$ given the noisy recordings (y_1, y_2, y_3) of a distant three-microphone linear array. Better estimations of the instantaneous amplitudes and frequencies are also evident after applying the proposed Multichannel, Multiband Demodulation (MMD) scheme.

IV. Improved Modulation Features

The estimation of instantaneous amplitudes $a_k[n]$ and frequencies[2] $f_k[n]$ is realized following short-time processing in frames of length L. As depicted in Fig. 3, first, each recording y_m is convolved with a Gabor filterbank $\{g_k(t)\},\ k \in [1, K]$. Then, for each frequency band k, the corresponding multichannel energy is found, and based on that, the instantaneous AM–FM parameters $a_k[n]$ and $f_k[n]$ are estimated using ESA. To cope with singularities caused by small energies, the instantaneous signals are smoothed by a median filter. In this work, second-order modulation features are extracted by measuring statistics over $a_k[n]$ and $f_k[n]$, namely (a) Mean Instantaneous Amplitudes (MIAs) (b) Mean Instantaneous Frequencies (MIFs), (c) Weighted Frequencies (Fw), and (d) Frequency Modulation Percentages (FMPs). MIAs and MIFs are the short-time means of $a_k[n]$ and $f_k[n]$. Motivated by the non-linear human perception of speech, MIAs are transformed using a logarithm. MIFs are only scaled from the frequency domain to the [0,1] range by dividing with $f_s/2$. Fw features are the micro-fluctuations of the instantaneous frequencies around the center frequency of filter k, estimated as:

$$Fw_k = \sum_{n=0}^{L} a_k^2[n] f_k[n] \ / \ \sum_{n=0}^{L} a_k^2[n] \tag{13}$$

Finally, $FMP_k = B_k / Fw_k$, where B_k is the mean bandwidth of $f_k[n]$ in band k, an amplitude-weighted deviation [4]. All features are mean and variance normalized to cope with long-term effects. Standardization is applied per utterance, across filters for MIA in order to keep the relative information that exists between the coefficients, and per filter for the rest.

To test the robustness of the improved modulation features against their single-channel version, we simulate noisy far-field speech by creating distorted versions of a sample of clean TIMIT phonemes. Clean speech is convolved with room impulse responses simulated using the Image-Source Method (ISM) [1] to match the environment of a small room, while white Gaussian noise is added to simulate the noisy background. Three microphones, arranged in a 30-cm equidistant linear array, were assumed in the center of the room, three meters away from the speaker. Figure 4 shows the relative improvements gained for a selection of features. For each phoneme and frequency band, estimation errors correspond to the amount of mismatch of the features extracted on the noisy signals against the features extracted on the clean source.

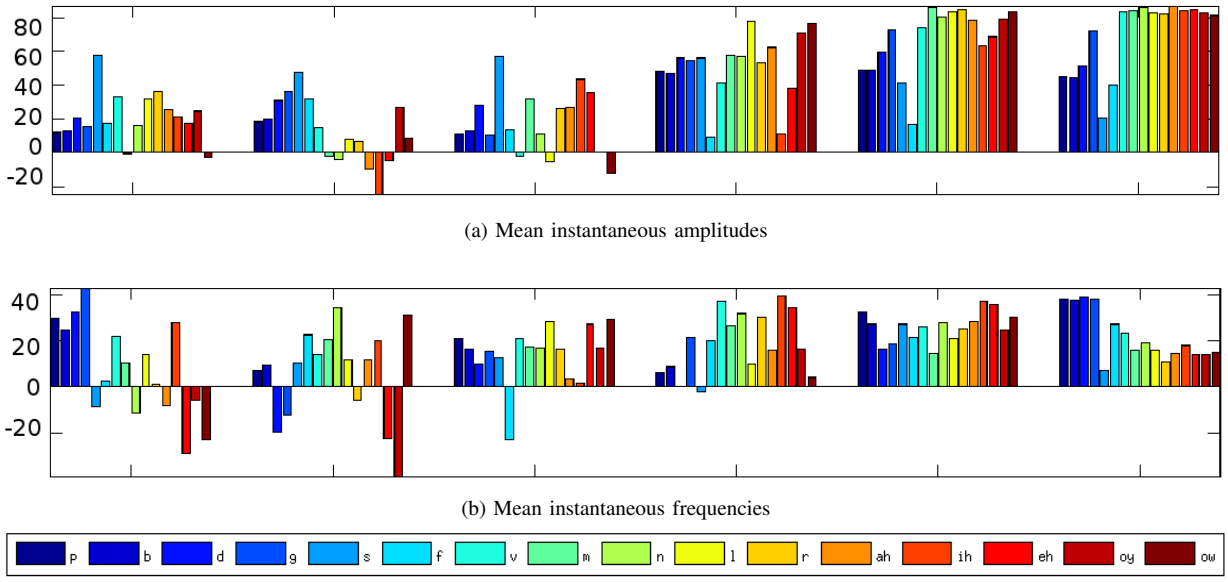
V. DSR on Simulated and Real Data

Several hybrid feature vectors are tested by combining frequency modulation features (e.g., MIFs, Fw, and FMPs) with the traditional MFCCs targeting improved performance in challenging conditions. Any improvements gained by the proposed MMD scheme are assessed and compared to other multichannel processing methods like beamforming, in which features are extracted on denoised signals.

A. DIRHA-English corpus

The employed DSR corpus [18] includes a large set of one-minute sequences simulating real-life scenarios of speech-based domestic control. The sequences were generated by mixing real and simulated far-field speech with typical domestic background noise. Real far-field speech was recorded in a Kitchen-Livingroom space by 21 condenser microphones arranged in pairs and triplets on the walls, and pentagon arrays on the ceilings. 12 US and 12 UK English native speakers were recorded on Wall Street Journal, phonetically-rich, and home automation sentences. Clean speech was recorded in a studio by the same speakers on the same material and convolved with the corresponding room impulse responses to produce simulated far-field speech. Overall, 1000 noisy and reverberant

[2]Instantaneous frequencies $f_k[n] = \omega_k[n]/2\pi,\ k \in [1, K]$ are measured in Hz.



(a) Mean instantaneous amplitudes

(b) Mean instantaneous frequencies

Fig. 4. Relative reduction (%) of demodulation error after using cross-Teager energy in Gabor-ESA. Root-mean-square errors are between: (a) MIA and (b) MIF features on clean and noisy far-field speech. Clean speech corresponds to the central frames of 50 randomly selected instances for each of 16 TIMIT phonemes uniformly selected from each phoneme category, while their far-field version have been simulated using the Image Source Method (ISM) for a linear array with three microphones, in which Gaussian noise ($SNR = 5\ dB$) was added.

utterances of real (dirha-real) and simulated (dirha-sim) far-field multichannel speech were extracted by the sequences and used for experimentation.

B. Experimental framework

13 MFCCs are derived from 40 Mel-spaced triangular filters spanning the interval $[0, f_s/2]$. Short-time analysis is applied every 10 ms over 25 ms long speech frames that are Hamming filtered and pre-emphasized. Cepstral mean normalization is applied per utterance in order to cope with channel distortions. A Mel-spaced filterbank of 12 Gabor filters with 70% overlap is used for the extraction of AM–FM features in 32 ms long mean and variance normalized frames shifted in 10 ms steps. Both feature sets are appended with their first- and second-order derivatives before their concatenation. MMD-based modulation features are extracted using the channels (LA1-LA6) of the six-microphone pentagon array located in the center of the Livingroom, while MFCC and single-channel modulation features are extracted on the signals of the central microphone (LA6) of the array.

State-of-the-art delay-and-sum beamforming is employed for speech denoising. The array channels (LA1-LA6) are beamformed using the BeamformIt tool [2], which is extensively used in several works for multichannel DSR and provides reliable results based on blind reference-channel selection and two-step time delay of arrival Viterbi postprocessing.

An HMM-GMM recognizer is built using the Kaldi toolkit [17]. Since our goal is to compare the different feature sets, eliminating as much as possible other factors, we are presenting results using "tri1" acoustic models, that is triphone modeling with no further feature transformation (e.g., LDA, MLLT, and SAT). GMM acoustic models are trained on matched conditions using microphone-dependent contaminated data produced by convolving clean utterances with various room impulse responses. The same microphones are used for training and testing.

A trigram language model is used for decoding, trained on the transcriptions of the training set of the corpus. Note that training and testing are based on the scripts provided with the database.

C. Results

Recognition experiments are conducted on the dirha-sim and dirha-real datasets. Amplitude modulation features (MIAs) are tested individually and compared to MFCCs as both of them are energy-based features and expected to be correlated. The results of Table I show that the combined features yield significant improvements over MFCCs, for both simulated and real data, with MIFs performing slightly better than Fw and FMPs. The MMD scheme achieves improvements of 1%-3% to all modulation features. "MFCC+Fw_mmd" yields 26% relative improvement compared to MFCCs, achieving 48.4% Word Error Rate (WER), which is the best score on average across the datasets.

Notable improvements are observed after using beamforming. As presented in Table II, recognition with MFCCs is improved by 17%, while modulation features keep contributing positively by reaching relative improvement of 18.8%. The results show that beamforming may lead to better modulation features for recognition rather than multichannel demodulation. However, note that the latter lacks a signal alignment stage in contrast with beamforming. Moreover, beamforming

TABLE I

WER (%) USING TRIPHONE ACOUSTIC MODELS (TRI1) ON CONCATANATIONS ("+") OF MFCCS WITH FREQUENCY MODULATION FEATURES (FW, MIF, FMP) AND ALTERNATIVELY WITH THEIR IMPROVED VERSIONS DERIVED BY THE PROPOSED MMD ("_MMD") SCHEME. AMPLITUDE MODULATION FEATURES (MIA), WHICH ARE DESIGNED TO WORK SIMILARLY TO MFCCS, ARE TESTED SEPARATELY.

tri1	MFCC	+ Fw	+ Fw_mmd	+ MIF	+ MIF_mmd	+ FMP	+ FMP_mmd	MIA	MIA_mmd
dirha-sim	62.9	48.1	46	47.7	45.1	47.3	46.2	62.1	61.9
dirha-real	67.9	55.3	51.3	52.9	51.6	54.6	52.8	68.5	68.3
average	65.4	51.7	48.7	50.3	**48.4**	51.0	49.5	65.3	65.1
rel. reduction (%)	-	21.0	25.61	23.1	26.1	22.1	24.3	0.2	0.5

TABLE II

WERS (%) AFTER DELAY-AND-SUM BEAMFORMING.

tri1	MFCC	+ Fw	+ MIF	+ FMP	MIA
dirha-sim	45.1	36.6	36.3	37.2	49.2
dirha-real	50.2	42.3	41.4	43.4	53.1
average	47.7	39.5	**38.9**	40.3	51.2
rel. reduction (%)	-	17.2	18.8	15.4	-7.4

is expected to reduce some reverberation effects, which are avoided in the analysis of the current work. Overall, the moderate performance in both simulated and real data is mainly due to lack of feature transformations for speaker and environment adaptation. Improved results are expected by employing non-linear transformations for modulation features.

VI. CONCLUSIONS

We have introduced a multi-channel energy tracking scheme for energy-based demodulation targeting noise minimization across the channels of a microphone array by selecting the minimum Teager and cross-Teager energies. The latter is a measure of interaction between two oscillators, used herein as a multi-channel energy estimator. The obtained results are promising: demodulation errors due to noise are decreased, leading to improved AM-FM features that exhibit robustness in DSR when combined with the complementary MFCCs.

ACKNOWLEDGMENT

The authors wish to thank M. Omologo, M. Ravanelli, and L. Cristoforetti of Fondazione Bruno Kessler Italy, for providing the DIRHA-English corpus and their Kaldi scripts for training and testing.

REFERENCES

[1] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2021, September 2007.

[3] A.-O. Boudraa, J.-C. Cexus, and K. Abed-Meraim, "Cross ψ b-energy operator-based signal detection," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4283–4289, 2008.

[4] 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, 2005.

[5] D. Dimitriadis and E. Bocchieri, "Use of micro-modulation features in large vocabulary continuous speech recognition tasks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1348–1357, 2015.

[6] D. Dimitriadis and P. Maragos, "Continuous energy demodulation methods and application to speech analysis," *Speech Communication*, vol. 48, no. 7, pp. 819–837, 2006.

[7] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, D. van Leeuwen, M. Lincoln, and V. Wan, "The 2007 AMI(DA) system for meeting transcription," in *Multimodal Technologies for Perception of Humans*. Springer, 2008, vol. LNCS-4625, pp. 414–428.

[8] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 547–554.

[9] J. F. Kaiser, "Some useful properties of teager's energy operators," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP)*, vol. 3, 1993, pp. 149–152.

[10] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[11] S. Lefkimmiatis, P. Maragos, and A. Katsamanis, "Multisensor multiband cross-energy tracking for feature extraction and recognition," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP)*, 2008, pp. 4741–4744.

[12] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing (ICASSP)*, 2014, pp. 5542–5546.

[13] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Signal Processing Letters*, vol. 41, no. 10, pp. 3024–3051, 1993.

[14] P. Maragos and A. Potamianos, "Higher order differential energy operators," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 152–154, 1995.

[15] V. Mitra, J. Van Hout, W. Wang, M. Graciarena, M. McLaren, H. Franco, and D. Vergyri, "Improving robustness against reverberation for automatic speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 525–532.

[16] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Graciarena, "Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions." in *Proc. Int. Conf. on Speech Communication and Technology (Interspeech)*, 2014, pp. 895–899.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," 2011.

[18] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015, pp. 275–282.

[19] I. Rodomagoulakis, G. Potamianos, and P. Maragos, "Advances in large vocabulary continuous speech recognition in Greek: Modeling and nonlinear features," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2013, pp. 1–5.

[20] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013, pp. 285–290.

[21] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2015.