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 (MBD) 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.

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

\[ y_m(t) = s(t) + u_m(t), \quad m = 0, \ldots, M - 1 \]  

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 \( \{g_k(t)\} \):

\[ y_{mk}(t) = y_m(t) * g_k(t), \quad k = 0, \ldots, N - 1 \]

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 \( \Psi_c \) that measures...
the relative rate of change between two oscillators. More analytically:

$$\Psi_c[y_{mk}, y_{\ell k}] = \dot{y}_{mk} y_{\ell k}(t) - y_{mk} \dot{y}_{\ell k}$$  \hspace{1cm} (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:

$$\mathcal{E}\{\Psi_c[y_{mk}, y_{\ell k}]\} = \mathcal{E}\{\Psi[s_k]\} + \text{error} \hspace{1cm} (4)$$

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 $\hat{m}$ and $\hat{\ell}$ having the 1st and 2nd smallest average Teager energies:

$$\Psi_c^{min}(k) = \Psi_c^{min}[y_{\hat{m}k}, y_{\hat{\ell}k}], \hspace{1cm} \hat{m}, \hat{\ell} = \arg\min_{\hat{m}, \hat{\ell}} (\mathcal{E}\{\Psi_c^{min}[y_{mk}, y_{\ell k}]\}, \mathcal{E}\{\Psi_c^{min}[y_{\ell k}, y_{mk}]\}) \hspace{1cm} (5)$$

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 $\Psi_c^{min}(k)$ 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 $\Psi_c^{min}$ or $\Psi_c^{min}$, given two real distant recordings from a two-microphone linear array.

$^1$2 - ($M^2$) computations are needed for each band because $\Psi_c[y_{mk}, y_{\ell k}] \neq \Psi_c[y_{\ell k}, y_{mk}]$.

$$\mathcal{E}\{\Psi_c[y_{mk}, y_{\ell k}]\} = \mathcal{E}\{\Psi[s_k]\} + \text{error} \hspace{1cm} (4)$$

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 $\hat{m}$ and $\hat{\ell}$ having the 1st and 2nd smallest average Teager energies:

$$\Psi_c^{min}(k) = \Psi_c^{min}[y_{\hat{m}k}, y_{\hat{\ell}k}], \hspace{1cm} \hat{m}, \hat{\ell} = \arg\min_{\hat{m}, \hat{\ell}} (\mathcal{E}\{\Psi_c^{min}[y_{mk}, y_{\ell k}]\}, \mathcal{E}\{\Psi_c^{min}[y_{\ell k}, y_{mk}]\}) \hspace{1cm} (5)$$

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 $\Psi_c^{min}(k)$ 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 $\Psi_c^{min}$ or $\Psi_c^{min}$, given two real distant recordings from a two-microphone linear array.

$$\mathcal{E}\{\Psi_c[y_{mk}, y_{\ell k}]\} = \mathcal{E}\{\Psi[s_k]\} + \text{error} \hspace{1cm} (4)$$

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 $\hat{m}$ and $\hat{\ell}$ having the 1st and 2nd smallest average Teager energies:

$$\Psi_c^{min}(k) = \Psi_c^{min}[y_{\hat{m}k}, y_{\hat{\ell}k}], \hspace{1cm} \hat{m}, \hat{\ell} = \arg\min_{\hat{m}, \hat{\ell}} (\mathcal{E}\{\Psi_c^{min}[y_{mk}, y_{\ell k}]\}, \mathcal{E}\{\Psi_c^{min}[y_{\ell k}, y_{mk}]\}) \hspace{1cm} (5)$$

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 $\Psi_c^{min}(k)$ 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 $\Psi_c^{min}$ or $\Psi_c^{min}$, given two real distant recordings from a two-microphone linear array.

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:

$$\Psi[y_{mk}] = (y_m * \tilde{g}_k)^2 - (y_m * g_k)(y_m * \tilde{g}_k) \hspace{1cm} (8)$$

$$\Psi[\tilde{g}_{mk}] = (y_m * \tilde{g}_k)^2 - (y_m * \tilde{g}_k)(y_m * \tilde{g}_k) \hspace{1cm} (9)$$

Herein, we incorporate the “denoised” energies $\Psi_c^{min}$ and $\Psi_c^{min}$ 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 $\Psi[\tilde{g}_{mk}]$ can be substituted by two “cleaner”
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:

$$Fw_k = \sum_{n=0}^{L} a_k^2[n] f_k[n] / \sum_{n=0}^{L} a_k^2[n]$$ (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 proposed MMD scheme are assessed and compared to other multichannel processing methods like beamforming, in which

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
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 “tril” 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 CONCATENATIONS ("+") 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.

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

TABLE II
WERs (%) AFTER DELAY-AND-SUM BEAMFORMING.

<table>
<thead>
<tr>
<th>tri</th>
<th>MFCC</th>
<th>+ Fw</th>
<th>+ MIF</th>
<th>+ FMP</th>
<th>MIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>dirha-sim</td>
<td>45.1</td>
<td>36.6</td>
<td>36.3</td>
<td>37.2</td>
<td>49.2</td>
</tr>
<tr>
<td>dirha-real</td>
<td>50.2</td>
<td>42.3</td>
<td>41.4</td>
<td>43.4</td>
<td>53.1</td>
</tr>
<tr>
<td>average</td>
<td>47.7</td>
<td>39.5</td>
<td>38.9</td>
<td>40.3</td>
<td>51.2</td>
</tr>
<tr>
<td>rel. reduction (%)</td>
<td>-</td>
<td>17.2</td>
<td>18.8</td>
<td>15.4</td>
<td>-7.4</td>
</tr>
</tbody>
</table>

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