Improved Frequency Modulation Features for Multichannel Distant Speech Recognition

Isidoros Rodomagoulakis and Petros Maragos

***Abstract*—Frequency modulation features capture the fine structure of speech formants that constitute beneficial to the traditional energy-based cepstral features by carrying supplementary information. Improvements have been demonstrated mainly in Gaussian mixture model (GMM)–hidden Markov model (HMM) systems for small and large vocabulary tasks. Yet, they have limited applications in deep neural network (DNN)–HMM systems and distant speech recognition (DSR) tasks. Herein, we elaborate on their integration within state-of-the-art front-end schemes that include post-processing of MFCCs resulting in discriminant and speaker-adapted features of large temporal contexts. We explore: 1) multichannel demodulation schemes for multi-microphone setups; 2) richer descriptors of frequency modulations; and 3) feature transformation and combination via hierarchical deep networks. We present results for tandem and hybrid recognition with GMM and DNN acoustic models, respectively. The improved modulation features are combined efficiently with MFCCs yielding modest and consistent improvements in multichannel DSR tasks on reverberant and noisy environments, where recognition rates are far from human performance.**

***Index Terms*—Frequency modulation features, demodulation, deep bottleneck features, distant speech recognition.**

I. Introduction

MODULATION features stemming from the AM-FM speech model were originally conceived for ASR [1] as capturing the second-order non-linear structure of speech formants, providing complementary information to the traditional energy-based cepstral features (e.g., MFCCs and PLPs). Their fusion presents robustness in noisy and mismatched conditions as indicated in recent works [2], [3]. However, only a few works [4], [5] examine their performance in DSR tasks with reverberation. Recently, bottleneck Multilayer Perceptons (MLPs) have been proposed in [2] to combine frequency micro-modulation features with PLPs using network's non-linear transformations instead of Linear Discriminant Analysis (LDA) which is suboptimal for non-Gaussian features. Following the tandem approach [6], improved and deeper nets were proposed in [7], [8], while hierarchical architectures [9] were beneficial for feature combination.

Deep Neural Networks (DNNs) have resulted in innovative ways to improve feature extraction and acoustic modeling in speech recognition [10]. Recently, end-to-end systems [11], [12] have been developed to combine all recognition stages into Recurrent Neural Networks (RNNs) of long memory in order to transform unsegmented sequences of raw speech signals into sequences of phone labels, outperforming in many cases the hybrid DNN-HMM state-of-the-art systems. However, they require large amounts of data and processing capacity while poor performance persists due to high levels of noise and reverberation in many DSR scenarios [13], [14] commonly found in modern applications.

Although DNNs can learn many types of variation depending on the training data, they can be sensitive to data mismatches, while feature transformations learned in a data-driven way may not generalize well for out-of-domain data. Model adaptation with regularization mechanisms [15] and iVector based adaptation [16] have been proposed for coping with unseen acoustic data. However, robust acoustic features are typically used to improve acoustic models when dealing with noisy and channel-degraded acoustic data. A comprehensive survey on robust feature extraction strategies and features for DNN-based recognition can be found in [17].

Multi-microphone setups with array processing [18] offer flexibility on multi-source and noisy acoustic scenes by capturing the spatial diversity of speech and non-speech sources, allowing more sophisticated front-ends with channel combination [19], beamforming [20] and speech enhancement [21], which were recently revised and solved with DNNs. However, the most significant improvements have been achieved with multi-style training on multichannel data [22], [23], while incorporating deep learning in traditional array processing methods is still under investigation.

Our goal in this work is to increase the robustness of frequency modulation features in noise and reverberation in order to combine them efficiently with standard MFCC-based front-ends for state-of-the-art speech recognition with GMM and DNN acoustic models. First, we propose a Multichannel, Multiband Demodulation (MMD) scheme that utilizes the noise diversity across microphone array signals aiming at improved demodulation of speech resonances and more accurate estimations of instantaneous modulations [24]. Secondly, we explore richer representations of the estimated modulations either by applying

Manuscript received November 17, 2018; revised March 20, 2019 and May 8, 2019; accepted May 24, 2019. Date of publication June 24, 2019; date of current version July 25, 2019. This work was supported in part by the European Regional Development Fund of the EU and in part by Greek National Funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call "Research–Create–Innovate" through Project T1EDK-01248, "i-Walk." The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Michiel Bacchiani. *(Corresponding author: Isidoros Rodomagoulakis.)*

The authors are with the School of Electrical and Computer Engineering, National Technical University of Athens, 15773 Athens, Greece (e-mail: irodoma@cs.ntua.gr; maragos@cs.ntua.gr).

Digital Object Identifier 10.1109/JSTSP.2019.2923372

signal compression on the raw signals, or by transforming mid-duration temporal contexts of their first-order statistics into hierarchical deep bottleneck networks, which are able to combine both non-linear transformation and fusion of heterogeneous features. Finally, we incorporate the proposed features combined with MFCCs in standard recognition recipes leveraging multi-style training and beamforming. Experiments are conducted in simulated and real data with strong background noise and reverberation.

Sections II and III include theoretical background on single- and multi-channel energy tracking, respectively; Section IV presents the proposed MMD approach with indicative results on the demodulation of speech phonemes; Section V describes the extraction of frequency modulation features along with the proposed hierarchical bottleneck DNN scheme; The experimental framework and the employed DSR corpora are described in Section VI, while Sections VII and VIII present the obtained results and the conclusions of the paper.

II. Background on Single-Channel Energy Tracking

Kaiser in [25] continued Teager's pioneering work to introduce a nonlinear operator called Teager-Kaiser energy operator (TEO) Ψ. The operator is given by

$$\Psi[x(t)] = \dot{x}(t)^2 - x(t)\ddot{x}(t) \tag{1}$$

in the continuous time domain, where $\dot{x}(t)$ and $\ddot{x}(t)$ correspond to the first and second time derivatives of the argument. It can track very accurately and fast the instantaneous energy of an oscillator-like source.

When Ψ is applied to an AM-FM signal of the form $x(t) = a(t)\cos(\phi(t))$, the instantaneous energy of the source is given by $\Psi[x(t)] \approx a(t)^2\omega(t)^2$, where the approximation error becomes negligible if the instantaneous amplitude $a(t)$ and instantaneous frequency $\omega(t) = \dot{\phi}(t)$ do not vary rapidly and significantly compared to the average value of the carrier frequency $\omega(t)$. As a result, Ψ is the main ingredient of the first Energy Separation Algorithm (ESA)

$$\omega(t) \approx \sqrt{\Psi[\dot{x}(t)]/\Psi[x(t)]},\ \alpha(t) \approx \Psi[x(t)]/\sqrt{\Psi[\dot{x}(t)]} \tag{2}$$

developed in [26] and used in speech demodulation, which was motivated by the strong evidence of the existence of AM-FM modulation in speech resonance signals.

Modeling speech signals as a composition of superimposed bandlimited AM-FM signals drove a plethora of applications on speech analysis [26]–[28] and speech recognition [2], [30] areas, in which TEO is extensively used for robust energy estimation and demodulation of speech components, yielding energy features which are robust to noise [31], [32], and modulation features [1] that carry complementary information about the nonlinear nature of speech.

To represent the interaction between two real time functions, an energy-like function, called the cross-Teager energy operator (cross-TEO) Ψ_c has also been defined [33], [34]. This function can been viewed as cross-energy between two real signals $x(t)$ and $y(t)$ and is defined by

$$\Psi_c[x(t), y(t)] = \dot{x}(t)\dot{y}(t) - y(t)\ddot{x}(t) \tag{3}$$

It measures the instantaneous differences in the relative rate of change between $x(t)$ and $\dot{y}(t)$. In the general case, if x and y represent displacements in some generalized motions, it has dimensions of energy (per unit mass). This energy-like quantity was used in [33] to analyze the output $\Psi_c[x(t), y(t)]$ of the energy operator applied to $x(t) + y(t)$. It was also used in [34] as cross-energy between a signal $x(t)$ and its higher order derivatives for the development of higher order energy measures useful for AM-FM demodulation. Other applications can be found in signal detection problems (e.g., Difference in time of arrival estimation and detection of transients) [35] alternatively to cross correlation. Moreover, in [36], cross-energy features between adjacent microphones were found more robust to noise compared to other energy based features. Based on the above theoretical and practical results, we are motivated to use cross-Teager in order to extract speech modulation features in a multi-microphone framework by proposing a multichannel demodulation robust in noise.

III. Multichannel Energy Tracking

The employed energy tracking scheme exploits the spatial diversity of noise $u_m(t)$ exhibited across the M noisy recordings

$$y_m(t) = s(t) + u_m(t),\ m = 1, \ldots, M \tag{4}$$

of a microphone array capturing the clean source speech signal $s(t)$ in the continuous time domain t. Note that, reverberation effects and time alignment issues between y_m are not taken into account in the following analysis. Following a multiband analysis for energy estimation, the recordings are decomposed into N frequency bands for the derivation of their bandpass components $y_{mk}(t),\ k = 1, \ldots, N$, which correspond to speech resonances. The kth resonance of the recording y_m can be modeled by an AM–FM signal as

$$y_{mk}(t) = a_{mk}(t)\cos\left(\int_0^t \omega_{mk}(\tau)\, d\tau\right), \tag{5}$$

where $a_{mk}(t)$ and $\omega_{mk}(t)$ are its instantaneous amplitudes and angular frequencies. Speech components can be obtained by decomposing $y_m(t)$ with a Mel-spaced Gabor filterbank $\{g_k(t)\}$:

$$y_{mk}(t) = y_m(t) * g_k(t),\ k = 0, \ldots, N-1 \tag{6}$$

where $g_k(t)$ corresponds to the impulse response of the bandbass Gabor filter centered at frequency $\omega_c(k)$ in band k

$$g_k(t) = \exp(-\beta^2 t^2)\cos[\omega_c(k)t] \tag{7}$$

The constants β and $\omega_c(k)$ are the filter parameters. The Gabor filters were chosen for several reasons listed in [26], including their optimal time-frequency discriminability. Thus, we can estimate the energy of the source in band k by applying TEO directly on y_{mk} components, or alternatively, an energy-like quantity is given by the cross-Teager energy between the components of adjacent microphones. Given the correlated kth bandpass signals from any two microphones m and ℓ, their interaction can be described by the cross-Teager energy operator:

$$\Psi_c[y_{mk}, y_{\ell k}](t) = \dot{y}_{mk}(t)\dot{y}_{\ell k}(t) - y_{mk}(t)\ddot{y}_{\ell k}(t) \tag{8}$$

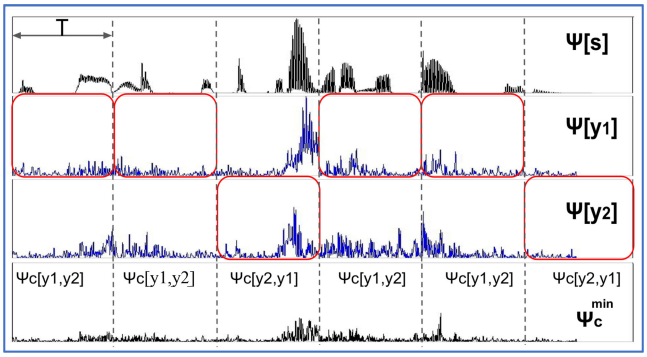

Fig. 1. Multichannel energy tracking for band $k = 3$: Given the noisy recordings $y_1(t), y_2(t)$ (2nd and 3rd row) of an array, their average Teager energies are computed in non-overlapping frames of duration T. The most robust cross-Teager energy $\Psi_c^{\min}(k)$ is found in 4th row between the channels having the 1st (red rectangulars) and 2nd smaller average energies.

which in general measures the relative rate of change between two oscillators. As discussed in [36] and [37], two useful properties of the operator can be derived:

1) On averaging over time, noise $u_m(t)$ contributes as an additive error term to the average Teager energy $\Psi[s_k]$ of the kth resonance of the source signal:

$$\mathcal{E}\{\Psi_c[y_{mk}, y_{\ell k}](t)\} = \mathcal{E}\{\Psi[s_k](t)\} + \text{error} \quad (9)$$

where $\mathcal{E}\{.\}$ denotes expectation. The above stands assuming that the additive noise component is a zero mean, wide-sense stationary Gaussian random process. Consequently, the energy with the minimum average

$$\Psi_c^{\min}(k) = \Psi_c[y_{\hat{m}k}, y_{\hat{\ell}k}], \quad (10)$$

which is formed by microphones $(\hat{m}, \hat{\ell})$, is expected to lie closer to $\Psi[s_k(t)]$.

2) We could search for the best pair $(\hat{m}, \hat{\ell})$ among all pairs of microphones (m, ℓ), $m, \ell \in [1, \ldots, M]$, $m \neq \ell$ as follows:

$$(\hat{m}, \hat{\ell}) = \arg\min_{m,\ell} \big(\mathcal{E}\{\Psi_c[y_{mk}, y_{\ell k}](t)\}\big) \quad (11)$$

However the search is computationally intensive.[1] Thus, it suffices to search only between those microphones having the 1st and 2nd smallest average Teager energies.

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 the best pair of microphones $(\hat{m}, \hat{\ell})$ that yield $\Psi_c^{\min}(k)$, in each band k, is realized in medium-duration non-overlapping frames 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 by $\Psi_c^{\min}$, given two real distant recordings from a two-microphone linear array.

[1]The number of computations needed for each band is $2 \cdot \binom{M}{2}$ because the operator Ψ_c is not commutative, e.g., $\Psi_c[y_{mk}, y_{\ell k}]\} \neq \Psi_c[y_{mk}, y_{\ell k}]$.

IV. Multichannel, Multiband Demodulation

The kth AM-FM resonance component y_{mk} is demodulated on its instantaneous amplitudes $a_k(t)$ and angular frequencies $\omega_k(t)$ based on the ESA formulas

$$\omega_{mk}(t) \approx \sqrt{\Psi[\dot{y}_{mk}]/\Psi[y_{mk}]},\ \alpha_{mk}(t) \approx \Psi[y_{mk}]/\sqrt{\Psi[\dot{y}_{mk}]} \quad (12)$$

In this work, we modify ESA by replacing single-channel energies $\Psi[y_{mk}]$ with the cross-channel energy $\Psi_c^{\min}(k)$ from (10), which is the best estimation of source's energy as evidenced in (9):

$$\omega_k(t) \approx \sqrt{\Psi_c[\dot{y}_{\hat{m}k}, \dot{y}_{\hat{\ell}k}]/\Psi_c[y_{\hat{m}k}, y_{\hat{\ell}k}]} \quad (13)$$

$$\alpha_k(t) \approx \Psi_c[y_{\hat{m}k}, y_{\hat{\ell}k}]/\sqrt{\Psi_c[\dot{y}_{\hat{m}k}, \dot{y}_{\hat{\ell}k}]} \quad (14)$$

Smoother approximations and more robust to noise are achieved by Gabor-ESA [28], which combines time-differentiation and filtering of the signal with a Gabor filterbank into convolutions of the signal with the time-derivatives of the Gabor filter's impulse response. The advantages of such approaches are that we can both avoid having the noisy one-sample discrete-time approximations of the derivatives and also succeed in having smoother estimates of the signal's time-derivatives in the presence of noise. In our case, bandpass filtering is applied within the cross-Teager operator:

$$\Psi_c[y_{\hat{m}k}, y_{\hat{\ell}k}] = (y_{\hat{m}} * \dot{g}_k)(y_{\hat{\ell}} * \dot{g}_k) - (y_{\hat{m}} * g_k)(y_{\hat{\ell}} * \ddot{g}_k) \quad (15)$$

$$\Psi_c[\dot{y}_{\hat{m}k}, \dot{y}_{\hat{\ell}k}] = (y_{\hat{m}} * \ddot{g}_k)(y_{\hat{\ell}} * \ddot{g}_k) - (y_{\hat{m}} * \dot{g}_k)(y_{\hat{\ell}} * \dddot{g}_k) \quad (16)$$

Note that in Eqns. (15) and (16) we have replaced the derivatives of the Gabor bandpass filtered signals with convolutions of the original signals with corresponding derivatives of the Gabor filter's impulse response. This has the effect of regularizing the combination of the Teager-Kaiser energy operator and bandbass filtering, as shown in [28]. From now on, we call this multichannel version of Gabor-ESA with the cross-Teager energies as Multichannel Multiband Demodulation (MMD) scheme.

A. Single- vs. Multi-Channel Demodulation

The goal of the following analysis is to compare the proposed MMD method with the single-channel Gabor-ESA [28], targeting robust demodulation of far-field speech. The focus is on frequency demodulation for the accurate estimation of instantaneous frequencies, since we use them to extract the proposed features for DSR. Single- and multi-channel Gabor-ESA are compared in terms of instantaneous frequency estimation error, which is computed between the instantaneous frequencies of the clean speech source and the estimated ones on the far-field version of the speech signal captured by the microphones. For the comparison, the Root Mean Square (RMS) error is computed in each band. We report the average relative reduction of

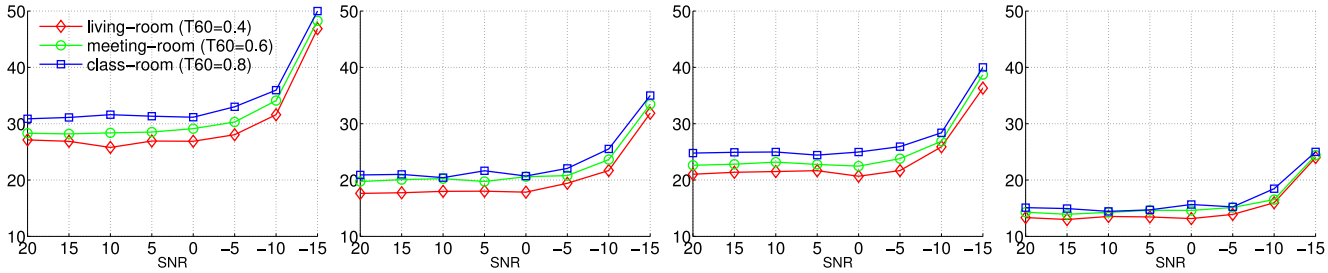

Fig. 2. Relative reduction (%) of Root Mean Square (RMS) error on the instantaneous frequency estimations per phoneme category achieved by the proposed MMD approach compared to single-channel Gabor-ESA demodulation.

the RMS error across phoneme instances and bands that MMD yields compared to the single-channel approach.

Without losing generalization on continuous speech, the comparison is applied on phoneme instances from all phoneme classes, in which AM-FM modulations are expected to differ. For instance, according to [28], consonants present stronger frequency modulations than vowels. Moreover, we need to have access in both clean and far-field versions of a speech source in order to compute the demodulation RMS error. Based on the above, we conduct the analysis on simulated far-field data for phonemes. The simulations are realized on the clean phonemes of the well known TIMIT corpus including:

- convolution with simulated room impulse responses using the Image-Source Method (ISM) [38] to match three environments of various reverberation times (T_{60}): a livingroom ($T_{60} = 0.4$ s), a meeting room ($T_{60} = 0.6$ s), and a class room ($T_{60} = 0.8$ s).
- addition of randomly selected noises (also reverberated) from the RWCP sound scene database [39] in order to simulate noisy domestic backgrounds of SNRs in $[20, 15, \ldots, -15]$ dB.

We configure ISM to simulate a microphone setup of three microphones arranged in a 30-cm equidistant linear array located in the center of each room, where a moving source is assumed to form a small spiral trajectory three meters away from the array. Overall, 100 instances of each phoneme are simulated for each of the 24 SNR-T_{60} combinations, resulting in 2400 signals. As evidenced in Fig. 2, the relative improvements gained by using MMD increase as conditions get more difficult, especially for low SNR values. It also appears that the robustness of Teager energy in low-frequency bands [40] benefits vowels the most compared to nasals, stops, and fricatives, in which the improvements are modest.

V. Frequency Modulation Features

First order statistics over the frequency micro-modulations $f_k(t) = \omega_k(t)/2\pi$ have yielded improved results combined with MFCCs in noisy LVCSR tasks [2]. Experimental evidence in [37] and [2], showed that instantaneous frequencies are more beneficial than amplitudes as supplementary features to MFCCs. A possible explanation is that although the amplitudes capture part of the nonlinear behavior of speech, e.g., the modulation pulses appearing within a single pitch period, they are expected to be correlated with the energy-based MFCCs. A basic descriptor over instantaneous frequencies is their temporal mean across the samples of a short-time frame, e.g., $1/L \times \sum_L f_k[n]$, where k is the filter index and L the frame length in samples. The so-called Mean Instantaneous Frequency (MIF) features are extensively used in our previous works. Herein, we compare MIF with two new and richer descriptors, designed for DNN acoustic modeling: 1) the Compressed Instantaneous Frequencies (CIF) and 2) the bottleneck features derived from hierarchical deep networks. All features are extracted over instantaneous frequencies that are estimated following the single-channel demodulation or the proposed MMD approach. Feature extraction is realized in frames of 32 ms ($L = 512$ for 16 kHz signals) across steps of 10 ms, similarly to the short-time analysis used in MFCCs.

A. Compressed Instantaneous Frequencies (CIF)

As depicted in Fig. 4, the estimated instantaneous frequencies $f_k(t)$ of a 32 ms short-time frame of 16 kHz speech contain periodic patterns that can be described compactly with a few of its Discrete Cosine Transform (DCT) coefficients. The exact number of the selected coefficients is a trade-off between the reconstruction error they achieve and their dimensionality, taking into account the fidelity of the employed network in which they are fed for the extraction of bottleneck features from larger temporal contexts.

Searching the optimum number of DCT coefficients to use in each band is realized on the set of corrupted TIMIT phones. Fig. 3 depicts the percentage of variance (energy) covered on instantaneous frequencies signals by $[5 \ldots 15]$ coefficients in each band. As observed, more coefficients are needed in higher bands where modulations are stronger and more noisy due to the wider filters that are applied. However, we decided to use 10 coefficients in each band, from which the percentages start converging. An example of reconstructing the instantaneous frequencies by using 10 DCT coefficients is depicted in Fig. 4.

B. Hierarchical Deep Bottleneck Features

The complementary MFCC and frequency modulation features are transformed and combined through a hierarchical network of bottleneck DNNs for the extraction of long-term deep features, which are augmented with speaker adapted features.

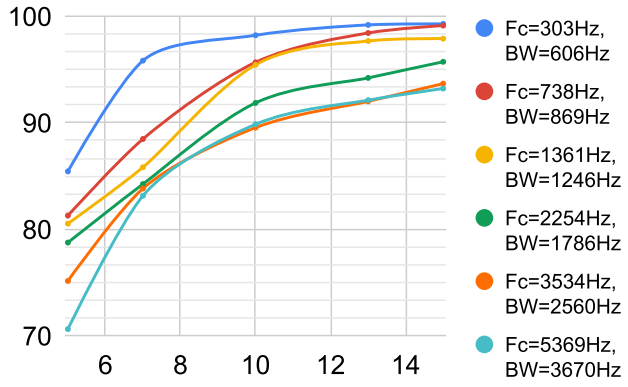

Fig. 3. Variance percentage covered on the instantaneous frequencies computed over the corrupted TIMIT phones after reconstructing them with a varying number of DCT coefficients. $[5 \ldots 15]$ coefficients are tested in each band, where Fc and BW correspond to its central frequency and bandwidth.

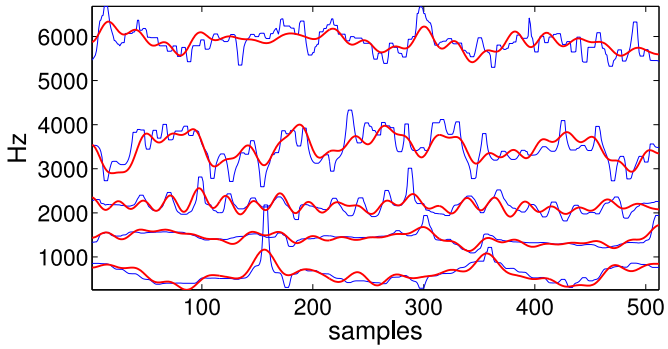

Fig. 4. Extraction of Compressed Instantaneous Frequency (CIF) features: Instantaneous frequency modulations in six Mel-spaced bands of phoneme "ah" and their reconstructions (red thick lines) using 10 DCT coefficients per band.

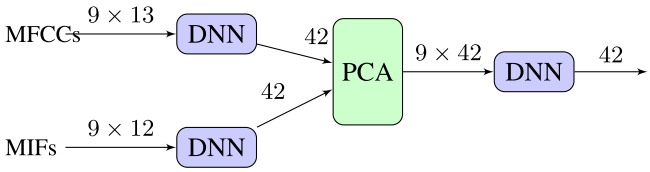

Fig. 5. Extraction of 42 deep hierarchical bottleneck features after transforming and combining MFCCs with mean instantaneous modulation frequencies (MIFs) spanning contexts of approximately 800 ms ($9 \times 9 = 81$ frames).

As shown in Fig. 5, first, compression of 9-frame temporal contexts is realized for each feature set through bottleneck networks. Subsequently, the activations of their bottleneck layers are concatenated and given in 9-frame vectors to the combination network after reducing their dimensionality by applying Principal Component Analysis (PCA), retaining 95% of the total variability. The final feature vector is formed after augmenting the bottleneck features of the combination network with the initial MFCCs transformed using feature-space Maximum Likelihood Linear Regression (fMLLR).

VI. Experimental Framework

We evaluate specific combinations (concatenations), indicated as '+', of MFCC and frequency modulation features for three recognition schemes with GMM, DNN, and TDNN acoustic models. All features are extracted either on a noisy recording from a single channel of a microphone array or on its denoised version after delay and sum beamforming, indicated as ('_dsb') in the results. The frequency modulation features (MIF and CIF) are extracted on instantaneous frequencies that are estimated either by single-channel Gabor ESA or after the proposed MMD demodulation ('_mmd'). Note that our intent is not to compare directly beamforming and MMD results because the latter is proposed for speech demodulation method and not for speech denoising. However, we compare modulation features in DSR tasks from single- and multi-channel demodulation, and additionally, we test those features on beamformed signals as well. LDA, MLLT and fMMLR transformations are applied to features according to [41]. More specifically, the three considered recognition schemes are:

1) *GMM-HMM recognition* with triphones and speaker adaptive training on top of features of Fig. 6 for GMM.
2) *Tandem recognition* with subspace GMMs on hierarchical deep bottleneck features of Fig. 5.
3) *Hybrid recognition* with DNN and TDNN acoustic models on top of features of Fig. 6 for DNN.

The schemes are tested in three DSR corpora according to their latest and optimized recipes we found in Kaldi repository. We modified the front-end part of the recipes to support the proposed combinations of features without changing the rest of the scripts for LM, AM and decoding. More specifically, GMM-, DNN-based, and tandem recognition is tested on the DIRHA-English corpus, while hybrid recognition with TDNNs is tested on the well known AMI and CHiME-4 corpora. Separate AMs are trained for each corpus as described in the following paragraphs.

A. Feature Extraction Configuration

Multiband speech demodulation is realized with a Mel-spaced Gabor filterbank spanning the interval $[0, F_s/2]$ Hz. The filter placing and bandwidths are dictated by the Mel-scale and the need for fixed overlap. In the case of MIF features, 12 filters are used for the extraction of 12 features in each short-time frame. Filters' bandwidths are overlapped by 70% for better formant localization. In the case of CIF features and in order to keep dimensionality in controllable levels, six filters are used for the extraction of 60 CIF (10 DCT coefficients per filter) features. In this case, the frequency overlap between filters is reduced to 50% because the filters are wider, offering a sufficient coverage over speech formants. Instantaneous frequencies are smoothed with a 7-sample median filter in order to eliminate possible singularities that are caused by instabilities of the Teager-Kaiser energy operator in small amplitude values. Features are mean and variance normalized to cope with long-term effects. Standardization is applied per filter in utterance level before extracting the features in frames. Multichannel demodulation is realized by using the same channels which are employed for beamforming according to the setup of each database. Finally, modulation features are spliced in the same way as MFCCs and both sets are concatenated to the input of the employed networks. Note that

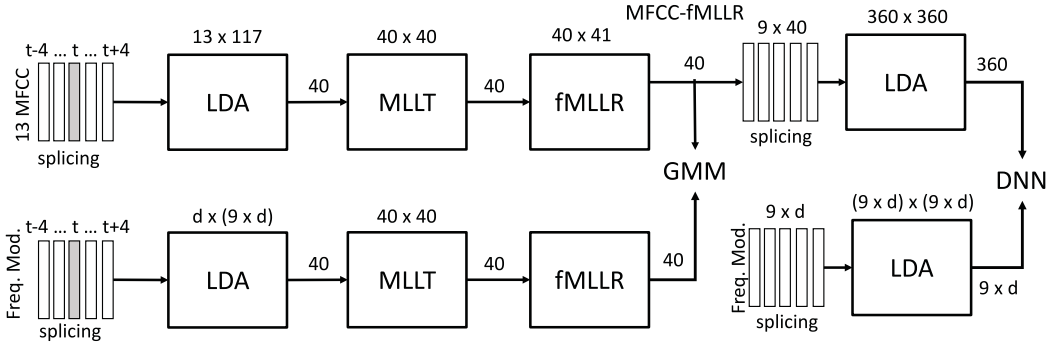

Fig. 6. Extraction and combination of MFCC-fMLLR [41] features with MIF ($d = 12$) and CIF ($d = 60$) frequency modulation features for GMM and DNN acoustic modeling.

LDA and fMMLR transformations, where they are referred, are applied separately on top of the two feature streams, as depicted in Fig. 6.

B. Delay-and-Sum Beamforming

Speech denoising is also used in the front-end stage, in which the available multichannel data are beamformed by using the BeamformIt tool [42] based on the setup of each database, as described in VI-C. The BeamformIt tool a state-of-the-art delay-and-sum beamformer that is extensively used in several multichannel DSR systems and supports blind reference-channel selection and two-step time delay of arrival Viterbi postprocessing.

C. Mutli-Microphone DSR Corpora

DIRHA-English corpus: The corpus [43] includes one-minute sequences simulating real-life scenarios of voice-based domestic control. Real far-field speech was recorded in a Kitchen-Livingroom space by 21 condenser microphones arranged on pairs and triplets on the walls, and pentagon arrays on the ceilings. 12 US and 12 UK English native speakers were recorded on WSJ, phonetically-rich, and home automation sentences. Moreover, 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. In our experiments, beamforming is applied on the six channels (LA1-LA6) of the pentagon ceiling-array located in the livingroom.

AMI corpus: The proposed features are also tested on the three tasks of the AMI meeting corpus [44] which consists of 100 hours of meeting recordings captured, transcribed and organized for DSR benchmarking according to three microphone setups: a) individual headset microphones (IHM), b) single distant microphone (SDM), and c) multiple distant microphones (MDM). The three tasks offer us the opportunity to test the robustness of the proposed features on various setups. For the MDM scenario, the eight channels of the 10 cm radius circular table array are combined via beamforming. Overlapping speech segments are excluded from our experiments. Additionally, the employed trigram language model is trained only on the transcriptions of the *train* set, without using the Fisher transcriptions as the standard Kaldi recipe supports. We report results on the *eval* set.

CHIME-4 corpus: The CHiME-4 task [45] is a far-field speech recognition challenge for single- and multi-microphone tablet device recordings in everyday scenarios under four noisy environments: street (STR), pedestrian area (PED), cafe (CAF) and bus (BUS). For training, 1600 utterances were recorded in the four environments from four speakers, and additional 7138 noisy utterances were simulated from WSJ0 by adding noises from the four noisy environments. The challenge setup consists of three tracks in which recognition is realized by using one (1ch), two (2ch), or six (6ch) channels from the tablet array. Multichannel recognition (2ch, 6ch) is based on beamforming. We report results for the three tracks on the 2640 utterances of the evaluation set, which consists of 330 utterances in each of the same eight conditions. Our recognition setup, as described in the following paragraphs, is based on the latest baseline Kaldi recipe in which TDNN acoustic models are trained on beamformed signals, while no RNNLM rescoring is applied. Separate TDNNs are trained for each task by using the corresponding data and alignments, while the networks are cross-entropy (xent) regularized using the clean speech part.

D. Data Augmentation for DIRHA-English

In the absence of sufficient training data for environments with distant microphones, a practical and widely used approach for acoustic modeling is to generate artificial training data by simulating the expected acoustic conditions of the target environment. The simulation process involves convolution of studio-quality speech with room impulse responses and noise addition in several SNR levels. We follow a slightly different approach for the case of the DIRHA-English database, where in order to increase robustness and reduce the training-testing mismatch, we generate beamformed signals for training, like the ones we intent to recognize. Thus, the ceiling-array recordings for beamforming are simulated by using RIRs measured from various positions in the room.

E. Recognition Schemes

1) Baseline GMM-HMM System: A baseline GMM-HMM recognizer is built based on the standard Kaldi recipe. First, tied-state triphones are trained on 13 MFCCs with their first- and second-order derivatives and then, LDA, MLLT and fMLLR transformations [41] are applied to train speaker independent models (*tri6*). Gaussian subspace acoustic models (*sgmm*) are also developed in which the universal background model (UBM) is trained on the *tri6* GMMs. Regarding language modeling, trigrams are trained on the transcriptions of the training sets.

2) Tandem Recognition: A GMM-HMM system is trained on top of the deep bottleneck features extracted by the proposed hierarchical scheme of bottleneck DNNs that is developed using TensorFlow. Each DNN consists of 6 hidden layers ($[1048, 1048, 1048, 42, 1048, 5000]$) of tanh nonlinearities. The bottleneck layer has 42 nodes while the last hidden layer acts like mixture components of the pdf in the softmax layer, comparable to GMMs. Nine frames are spliced and given to the input of the network which is trained to classify $2.5\,M$ frames from the DIRHA-English corpus to one of each 3405 nodes of the softmax layer corresponding to the senones of the baseline GMM-HMM system that provides the frame-state alignments. The network weights are trained layer-wise in 20 epochs by following the iterative stochastic gradient descent training using minibatches of 256 vectors. To prevent overfitting and for adjusting the learning rate parameter, 10% of the training corpus (chosen randomly) is used as cross-validation set.

3) Hybrid Recognition: Neural networks are trained to provide pseudo log-likelihood scores for HMM decoding. Herein, DNNs [46] and Time Delay Neural Networks (TDNNs) [47] are considered on spliced frames of MFCCs appended with modulation features. Substantially, their first layers act as feature transformation and fusion units on the combined feature sets similarly to the already described bottleneck networks. DNNs of six fully-connected hidden layers of 2048 sigmoid nonlinearities are trained on mini-batches of 512 samples in which 9-frame splices of 40 fMLLR-transformed MFCCs are included. Training is realized in three stages: 1) DBN pre-training, 2) frame cross-entropy training, and 3) sequence-training optimizing the sequential Minimum Bayes Risk criterion. The developed TDNNs, capable of tackling long-term interactions between speech and corrupting sources in reverberant environments, consist of five time-delay layers modeling multi-scale contexts of $\{[-2, 2], [-1, 1], [-1, 1], [-3, 3], [-6, 0]\}$ compared to the running frame in time t. Their input features are 11-frame splices of 40-dimensional hi-resolution MFCCs appended with 100-dimensional i-vectors extracted using a 512-Gaussian UBM. The training data are augmented by applying 3-way speed perturbation using factors of $[0.9, 1.0, 1.1]$ and rate perturbations by picking uniformly random values in [0.125,2].

VII. Results

Baseline recognition results on the DIRHA-English corpus are presented in Table I where is evident how MIFs benefit MFCCs in simulated and real data mostly a) when extracting the features from a single channel, b) after using multichannel demodulation, and c) after beamforming that yields the lowest WER. On the other hand, as shown in Table II, although linear transformations (LDA, MLLT and fMLLR) benefit MFCCs, they deteriorate the performance of the combined features because they are not uncorrelated and Gaussian like MFCCs. However, better combinations are accomplished after using DNN-based non-linear transformations. As shown in Table III, hierarchical deep bottleneck features with subspace GMMs yield significantly better results over the SAT system. Additionally, the contribution of modulation features is increased after applying multichannel demodulation compared to beamforming. In hybrid recognition results of Table IV, the proposed features achieve modest improvements on MFCC-fMLLR for DNNs. Accordingly, they also benefit hi-resolution MFCCs with i-vectors for TDNNs, yielding relative improvements up to 15% over the baseline Kaldi recipes, as Tables V, and VI show. It also seems that CIF features benefit MFCCs mostly in hybrid recognition compared to MIF. Moreover, it is worth mentioning that in the results of Table VI, MMD appears improved when the

TABLE I
GMM-HMM Recognition WERs (%) With Triphones on Combinations ("+") of MFCC and Modulation Features Extracted After MMD ("_mmd") or Beamforming ("_dsb")

Feat. Combinations	dirha-sim	dirha-real	average
MFCC	62.9	67.9	65.4
MFCC + MIF	47.7	52.9	50.3
MFCC + MIF_mmd	45.1	51.6	48.4
MFCC_dsb	36.8	40.5	38.7
MFCC_dsb + MIF_dsb	37.2	38.8	**38.0**

TABLE II
GMM-HMM Recognition WERs (%) After f MLLR-Based Speaker Adaptive Training (SAT)

DIRHA	features	mono	tri	-LDA-MLLT	-fMLLR
dirha-sim	MFCC_dsb	57.8	36.8	31.8	24.3
	+ MIF_dsb	54.7	37.2	32.8	26.6
dirha-real	MFCC_dsb	61.5	40.5	30.9	29.5
	+ MIF_dsb	52.3	38.8	33.4	31.2
average	MFCC_dsb	59.7	38.7	31.4	**26.9**
	+ MIF_dsb	53.5	38.0	33.1	28.9

TABLE III
Tandem Recognition WERs (%) With Subspace GMMs on Hierarchical DNN Bottleneck Features Appended With f MLLR-Transformed MFCCs

Feat. combinations	dirha-sim	dirha-real	average
MFCC_dsb	23.4	29.1	26.25
MFCC_dsb + MIF_dsb	22.8	28.8	25.8
MFCC_dsb + MIF_mmd	22.3	28.5	25.4
MFCC_dsb + CIF_mmd	21.6	27.8	**24.7**

TABLE IV
Hybrid Recognition WERs (%) Using DNN Acoustic Models Trained on Multiple-Frame Combined Features

Feat. combinations	dirha-sim	dirha-real	average
MFCC_dsb-fmllr	19.0	25.0	22.0
MFCC_dsb-fmllr + MIF_dsb	18.8	24.6	22.0
MFCC_dsb-fmllr + MIF_mmd	18.3	24.3	21.3
MFCC_dsb-fmllr + CIF_mmd	18.0	23.9	**20.9**

TABLE V
WERs (%) ON AMI CORPUS USING CROSS ENTROPY (XENT) REGULARIZED TDNN WITH CLEANED DATA AND SEPARATE ALIGNMENTS PER TASK

Feat. combinations	IHM	SDM	MDM
MFCC_dsb + ivector	25.7	50.1	43.9
MFCC_dsb + ivector + MIF_dsb	25.8	48.2	41.1
MFCC_dsb + ivector + MIF_mmd	25.8	48.2	40.9
MFCC_dsb + ivector + CIF_mmd	**25.6**	**46.8**	**40.3**

TABLE VI
WERs (%) ON CHIME-4 SIM/REAL TEST SETS FOLLOWING THE BASELINE KALDI RECIPE FOR TDNNS ON DELAY-AND-SUM BEAMFORMED SIGNALS WITHOUT USING RNNLM RESCORING

	simulated			real		
Feat. combinations	1ch	2ch	6ch	1ch	2ch	6ch
MFCC_dsb + ivector	16.6	13.2	10.3	16.4	13.5	9.7
+ MIF_dsb	15.9	12.9	10.1	16.3	13.3	9.4
+ MIF_mmd	15.9	13.1	9.8	16.3	13.4	9.2
+ CIF_mmd	15.5	12.3	**9.3**	15.8	12.9	**9.1**

employed channels are increased from two to six. Finally, the proposed features yield improvements in both real and simulated data, and DSR performance is improved without degradation in clean speech as indicated by the results on the AMI IHM task.

VIII. CONCLUSION

A new approach is presented for robust demodulation of the frequency micro-modulations of speech based on multichannel speech energy tracking over the signals of a microphone array. Better estimations of instantaneous frequencies enable the extraction of improved modulation features which are combined efficiently with standard feature sets in state-of-the-art recognition setups. Modest and consistent improvements are achieved in three challenging DSR corpora.

REFERENCES

[1] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 621–624, Sep. 2005.

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

[3] 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. Speech Commun. Technol.*, 2014, pp. 895–899.

[4] I. Rodomagoulakis, G. Potamianos, and P. Maragos, "Advances in large vocabulary continuous speech recognition in Greek: Modeling and nonlinear features," in *Proc. Eur. Signal Process. Conf.*, 2013, pp. 1–5.

[5] V. Mitra *et al.*, "Improving robustness against reverberation for automatic speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 525–532.

[6] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 3, pp. 1635–1638.

[7] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4153–4156.

[8] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Proc. Int. Conf. Speech Commun. Technol.*, 2011, pp. 237–240.

[9] Z. Tüske, R. Schlüter, and H. Ney, "Deep hierarchical bottleneck MRASTA features for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6970–6974.

[10] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. New York, NY, USA: Springer, 2014.

[11] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4835–4839.

[12] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv:1412.5567*.

[13] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 7, 2016.

[14] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 547–554.

[15] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7893–7897.

[16] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2013, pp. 55–59.

[17] V. Mitra *et al.*, "Robust features in deep-learning-based speech recognition," in *New Era for Robust Speech Recognition*. New York, NY, USA: Springer, 2017, pp. 187–217.

[18] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York, NY, USA: Springer, 2013.

[19] 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. Acoust., Speech, Signal Process.*, 2014, pp. 5542–5546.

[20] X. Xiao *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 5745–5749.

[21] M. Delcroix *et al.*, "Multichannel speech enhancement approaches to DNN-based far-field speech recognition," in *New Era for Robust Speech Recognition*. New York, NY, USA: Springer, 2017, pp. 21–49.

[22] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2013, pp. 285–290.

[23] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 5220–5224.

[24] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Instantaneous frequency and bandwidth estimation using filterbank arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8032–8036.

[25] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 381–384.

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

[27] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3795–3806, 1996.

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

[29] "Speech detection and speaker localization in domestic environments." 2014. [Online]. Available: http://dirha.fbk.eu/hscma

[30] I. Rodomagoulakis, P. Giannoulis, Z.-I. Skordilis, P. Maragos, and G. Potamianos, "Experiments on far-field multichannel speech processing in smart homes," in *Proc. 18th Int. Conf. Digit. Signal Process.*, 2013, pp. 1–6.

[31] F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," *IEEE Signal Process. Lett.*, vol. 6, no. 10, pp. 259–261, Oct. 1999.

[32] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory teager energy cepstrum coefficients for robust speech recognition," in *Proc. Int. Conf. Speech Commun. Technol.*, 2005, pp. 3013–3016.

[33] J. F. Kaiser, "Some useful properties of Teager's energy operators," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1993, vol. 3, pp. 149–152.

[34] P. Maragos and A. Potamianos, "Higher order differential energy operators," *IEEE Signal Process. Lett.*, vol. 2, no. 8, pp. 152–154, Aug. 1995.

[35] A.-O. Boudraa, J.-C. Cexus, and K. Abed-Meraim, "Cross ψ(B)-energy operator-based signal detection," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4283–4289, 2008.

[36] S. Lefkimmiatis, P. Maragos, and A. Katsamanis, "Multisensor multiband cross-energy tracking for feature extraction and recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4741–4744.

[37] I. Rodomagoulakis and P. Maragos, "On the improvement of modulation features using multi-microphone energy tracking for robust distant speech recognition," in *Proc. Eur. Signal Process. Conf.*, 2017, pp. 558–562.

[38] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[39] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," presented at the Int. Lang. Resources and Eval. Conf. (LREC), 2000.

[40] D. Dimitriadis, A. Potamianos, and P. Maragos, "A comparison of the squared energy and Teager-Kaiser operators for short-term energy estimation in additive noise," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2569–2581, Jul. 2009.

[41] S. P. Rath, D. Povey, K. Veselỳ, and J. Cernockỳ, "Improved feature processing for deep neural networks," in *Proc. Int. Conf. Speech Commun. Technol.*, 2013, pp. 109–113.

[42] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep. 2007.

[43] 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 Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 275–282.

[44] J. Carletta *et al.*, "The AMI meeting corpus: A pre-announcement," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, 2006, pp. 28–39.

[45] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.

[46] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Int. Conf. Speech Commun. Technol.*, 2013, pp. 2345–2349.

[47] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts." in *Proc. Int. Conf. Speech Commun. Technol.*, 2015, pp. 3214–3218.