

# Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments

Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, and Björn Schuller

**Abstract**—Eliminating the negative effect of non-stationary environmental noise is a long-standing research topic for automatic speech recognition that stills remains an important challenge. Data-driven supervised approaches, including ones based on deep neural networks, have recently emerged as potential alternatives to traditional unsupervised approaches and with sufficient training, can alleviate the shortcomings of the unsupervised methods in various real-life acoustic environments. In this light, we review recently developed, representative deep learning approaches for tackling non-stationary additive and convolutional degradation of speech with the aim of providing guidelines for those involved in the development of environmentally robust speech recognition systems. We separately discuss single- and multi-channel techniques developed for the front-end and back-end of speech recognition systems, as well as joint front-end and back-end training frameworks.

**Index Terms**—deep learning, neural networks, robust speech recognition, non-stationary noise, multi-channel speech denoising

## I. INTRODUCTION

Recently, following years of research, Automatic Speech Recognition (ASR) has achieved major breakthroughs and greatly improved performance [1]–[3]. Plenty of speech-specific intelligent human-machine communication systems, such as smartphone assistants (e.g. Siri, Cortana, Google Now), Amazon Echo, and Kinect Xbox One, have started to become part of our daily life. However, one of the central issues that limits their performance in everyday situations is still their performance degradation due to ambient noise and reverberation that corrupt the speech as captured by microphones.

According to the spectral distribution, the noises (including reverberation) can be generally grouped into stationary noise (constant with respect to time) or non-stationary noise (i.e., varying with time, such as transient sound events, interfering speakers, and music). Provided that it is possible to reliably detect instants of the absence of the target signal (i.e., the speech signal of interest), short-term stationary additive noise can be adequately tackled with standard, unsupervised noise reduction signal processing techniques mainly developed in the 1970s and 1980s [4]. However, detecting and reducing

the effects of non-stationary ambient noise, competing non-stationary sound sources, or highly reverberant environments, is still very challenging in practice [5]–[10]. To address these issues, a new wave of research efforts has emerged over the past five years, as showcased in the robust speech recognition challenges such as REVERB and CHiME [7]–[9], [11].

In this research, *data-driven* approaches based on a supervised machine learning paradigm have received increasing attention, and have emerged as viable methods for enhancing robustness of ASR systems [12]. The primary objective of these approaches is, by means of learning from large amounts of training data, to either obtain cleaner signals and features from noisy speech audio, or directly perform recognition of noisy speech. To this end, *deep learning*, which is mainly based on *deep neural networks*, has had a central role in the recent developments [13]–[16]. Deep learning has been consistently found to be a powerful learning approach in exploiting large-scale training data to build complex and dedicated analysis systems [17], and has achieved considerable success in a variety of fields, such as gaming [18], visual recognition [19], [20], language translation [21], music information retrieval [22], and ASR [23], [24]. These achievements have encouraged increasing research efforts on deep learning with the goal of improving the robustness of ASR in noisy environments.

In this survey, we provide a systematic overview of relevant deep learning approaches that are designed to address the noise robustness problem for speech recognition. Rather than enumerating all related approaches, we aim to establish a taxonomy of the most promising approaches, which are categorised by two principles: i) according to the addressed number of channels, these approaches can be grouped into *single-channel* or *multi-channel* techniques; ii) according to the processing stages of an ASR system, in which deep learning methods are applied, these approaches can be generally classified into *front-end*, *back-end*, or *joint front- and back-end* techniques (as shown in Fig. 1). We highlight the advantages and disadvantages of the different approaches and paradigms, and establish interrelations and differences among the prominent techniques. This overview assumes that the readers have background knowledge in noise-robust ASR and deep learning. However, we provide some key concepts of the raised noise-robust speech recognition problem and neural networks, e.g., fully-connected layers, convolutional layers, and recurrent layers, for a better overview. For more detailed knowledge of noise-robust ASR systems or deep learning, the readers can refer to [25] and [26], respectively. Note that, in this overview, the term *deep neural networks* refers to

Z. Zhang, J. Pohjalainen, A. E. Mousa, and B. Schuller are with the Chair of Complex and Intelligent Systems, University of Passau, 94032 Passau, Germany (e-mail: zixing.zhang@uni-passau.de).

J. Geiger is with the European Research Centre, Huawei Technologies, Dessauerstr. 10, 80992 Munich, Germany (e-mail: juergen.geiger@huawei.com).

B. Schuller is also with the Department of Computing, Imperial College London, London SW7 2AZ, UK (e-mail: bjoern.schuller@imperial.ac.uk), and with audEERING GmbH, Gilching, Germany (email: bs@audeerling.com).

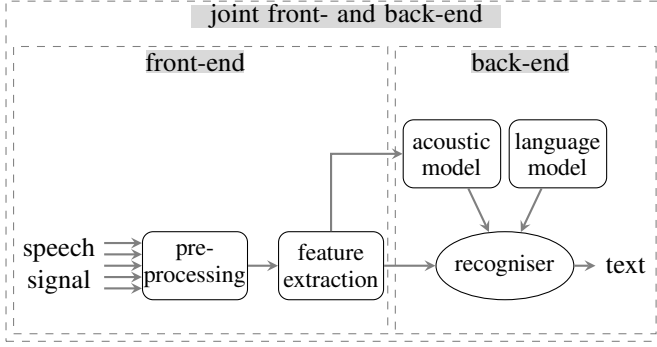


Fig. 1. General framework of a speech recognition system divided into front-end and back-end.

networks including multiple hidden layers.

Whilst several related surveys on environmentally robust speech recognition are available in the literature (e.g., [5], [25], [27]–[30]), none of these works focuses on the usage of deep learning. The emergence of deep learning is, however, deemed as one of the most significant advances in the field of speech recognition in the past decade and thus merits a dedicated survey.

The remainder of this article is organised as follows. In Section II, we briefly introduce the background of this overview. In Sections III to V, we comprehensively summarise the representative single-channel algorithms at the front-end, the back-end, and the joint front- and back-end of speech recognition systems, respectively. In Section VI, we then review promising multi-channel algorithms, before drawing our conclusions in Section VII.

## II. BACKGROUND

In this section, we briefly describe the environmental noise problem for speech recognition. We then analyse the drawbacks and limitations of traditional approaches, and introduce the opportunities for deep learning. Finally, we introduce some standard noisy speech databases and evaluation metrics for performance comparison of the following reviewed deep learning approaches.

### A. Problem Description

In real-life scenario, the *raw* speech signal  $s(t)$  is easily corrupted by convolutional noise  $r(t)$  (or Room Impulse Response [RIR]) and additive noise  $a(t)$  when transmitting through spatial channel. Thus, the observed distant-talk signal  $y(t)$  at the microphone can be written as:

$$y(t) = s(t) * r(t) + n(t). \quad (1)$$

When applying Short-Time Discrete Fourier Transform (STFFT) on the mixed/noisy speech, the length of RIR  $T_{60}$  should be considered. If it is much shorter than the analysis window size  $T$ ,  $r(t)$  only effects the speech signals within a frame (analysis window). For many applications (e.g., occurring in typical office and home environment), however, the reverberation time  $T_{60}$  ranges from 200 to 1 000 *ms* that is much longer than the analysis window size, resulting in

an undesirable influence on the following speech frames. For example, if the duration of a RIR is 1 *s* ( $T_{60}$ ) and a feature frame is extracted at every 10 *ms*, one RIR would smear across the following 100 frames. Therefore, this distorted speech in the amplitude *spectral* domain, can be formulated by (see [31] for more details):

$$Y(n, f) \approx \sum_{d=0}^{D-1} S(n-d, f)R(d, f) + A(n, f), \quad (2)$$

with an assumption that  $r(t)$  is a constant function. Particularly,  $R(d, f)$  denotes the part of  $R(f)$  (i.e., STFFT of RIR  $r(t)$ ) corresponding to a frame delay  $d$ . In this case, the channel distortion is no longer of multiplicative nature in a linear spectral domain – rather it is non-linear.

Assuming that the phases of different frames, and the speech and noise signals, are non-correlated for simplification (not the case in practise), the *power spectrum* of Eq. 2 can be approximated as

$$|Y(n, f)|^2 \approx \sum_{d=0}^{D-1} |S(n-d, f)|^2 |R(d, f)|^2 + A^2(n, f). \quad (3)$$

Then, the following relation is obtained in the *Mel spectral* domain for the  $k$ -th Mel-filter-bank output

$$Y^{mel}(n, k) \approx \sum_{d=0}^{D-1} S^{mel}(n-d, k)R^{mel}(d, k) + A^{mel}(n, k), \quad (4)$$

where  $S^{mel}(n, k) = \mathbf{B}[k] \cdot S^2(n, f)$  with  $\mathbf{B} = (b_{k,f}) \in \mathbb{R}^{K \times F}$ ,  $K$  is the number of Mel bins and  $b_{k,f}$  is the weight of the DFT bin  $f$  in the  $k$ -th Mel bin.  $R^{mel}(n, k)$  and  $A^{mel}(n, k)$  are defined similar to  $S^{mel}(n, k)$ .

To extract the Mel Frequency Cepstral Coefficients (MFCCs) in *cepstral* domain for ASR, logarithms and Discrete Cosine Transform (DCT) are further executed over the above mel spectral signals, so that

$$Y^{dct}(n, i) \approx S^{dct}(n, i) + R^{dct}(0, i) + M^{dct}(n, i), \quad (5)$$

where  $S^{dct}(n, i) = \mathbf{C}[i] \log(S^{mel}(n, k))$  with  $\mathbf{C}$  denoting a discrete cosine transformation matrix (same definition is for  $R^{dct}(0, i)$  and  $M^{dct}(n, i)$ ), and

$$M(n, i) = 1 + \frac{\sum_{d=1}^{D-1} S^{mel}(n-d, k)R^{mel}(d, k) + A^{mel}(n, k)}{S^{mel}(n, k)R^{mel}(0, k)}. \quad (6)$$

From Eq. (1) to (5), it can be found that the clean speech and the mixed/noisy speech have a highly *non-linear* correlation in either temporal, spectral, power spectral, mel spectral, log mel spectral, or cepstral domains, which results in a difficulty for noise cancellation.

Furthermore, the *time-variant* characters of RIR and additive noise (time-invariant additive noise is beyond the scope of this paper) make the task even more challenging. For RIR, many factors can lead to a change, for instance, the position of the speaker (i.e., the distance and angle between the speaker and microphone), the size, shape, and material of acoustic enclosure (such as a living room). For additive noise,

it could be abrupt sound like thunder and bark, side talking, and also music and driving noise. All these noises are almost unpredictable.

### B. Deep Learning vs Traditional Approaches

The ultimate goal of robust ASR systems is to learn well the relationship between noisy speech and the word predictions, i. e.,

$$w = f(y), \quad (7)$$

where  $y$  denotes the representation of noisy speech  $y(t)$ , and  $w$  is the target word. To simplify this process, we often divide it into two steps conducted at the system front end and back end, respectively. At the front end, speech enhancement (aka speech separation) or feature enhancement is applied to improve the quality and intelligibility of the estimated target speech on either signal level or feature level, so as to obtain the signals as clean as possible. That is,

$$s(t) \leftarrow \hat{s}(t) = f_s(y(t)). \quad (8)$$

At the back end, model updating is applied to make acoustic models adapt to the new data, i. e.,

$$w = f_m(\hat{x}), \quad (9)$$

where  $\hat{x}$  indicates the representation from enhance speech or the enhanced representation.

Traditional solutions on the front-end are mainly dominated by unsupervised signal processing approaches over the past several decades. *Spectral subtraction* [32] subtracts an averaged noise spectrum (magnitude or power spectrum) from the noisy signal spectrum, while keeping the resultant spectral magnitudes positive. It only affects the spectrum magnitudes, while the spectrum phases are obtained from the noisy signal. *Wiener filtering* [4] adopts stochastic models and is often implemented in practice using iterative approaches which base new estimates of the filter on the enhanced signal obtained by the previous iteration's estimate [33]. Another popular family of techniques comprises the *Minimum Mean Square Error (MMSE)* [34] and *log-spectral amplitude MMSE (Log-MMSE)* Short-Time Spectral Amplitude (STSA) estimators [35]. Despite that they are able to yield lower musical noise, a trade-off in reducing speech distortion and residual noise needs to be made due to the sophisticated statistical properties of the interactions between speech and noise signals [36].

Most of these unsupervised methods are based on either the additive nature of the background noise, or the statistical properties of the speech and noise signals. However, they often fail to track non-stationary noise in real-world scenarios in unexpected acoustic conditions [36]. Although some supervised machine learning approaches have been proposed, such as *Non-negative Matrix Factorisation (NMF)* [37]–[40], they struggle to obtain effective representations (aka dictionaries) of noise and speech in complex and noisy acoustic environments.

Deep learning that is mainly based on Deep Neural Networks (DNNs), however, is well suited to address such a complex *non-linear* problem [26]. The neural node, a basic unit constituting a network, is analogous to a biological neuron. The value of a node is usually computed as a weighted

sum of the inputs followed by a non-linear activation function. Theoretically, a single node can represent a huge amount of information as long as the numerical resolution allows. Practically, deep neural networks implement multiple neural network layers (each layer consists of multiple nodes). As a result, when combining many non-linear activation functions, it enables the network to learn complicated relationships between the inputs and outputs.

More specifically, typical neural layers frequently employed in deep learning include *fully-connected* layer, *convolutional* layer, and *recurrent* layer. *Fully-connected* layer is also known as dense layer and Multi-Layer Perception (MLP). In speech processing, stacking fully-connected layers on the top of extracted features (e.g., spectrogram) has already shown a great potential to extract high-level representation for speech recognition [23] via a greedy layer-wise unsupervised pre-training strategy [41].

*Convolutional* layer is a biologically inspired variant of fully-connected layer originally developed for visual perception tasks [42], and is the elementary layer to construct Convolutional Neural Networks (CNN). It employs a small size of 2D convolutional kernel ‘sweep’ over a 2D input, and delivers a representation of local activations of patterns. In image processing, convolutional layer has been frequently and clearly visualised to effectively extract the hierarchical features (see [43] for more details). This strongly encourages its applications to the speech domain, since the time-frequency representation of acoustic signals can be considered as an image. Besides, the 2D kernel can be modified into 1D kernel and directly applied to raw signals. Recent work has shown that the convolutional layer can automatically learn fundamental frequencies from raw signals [44].

In contrast to the aforementioned feed-forward layers (i. e., fully-connected layer and convolutional layer), *recurrent* layer (elementary layer for Recurrent Neural Networks [RNNs]) allows cyclical connections. These connections consequently endow the networks with the capability of accessing previously processed inputs. However, it cannot access long-term temporal contextual information since it suffers from the vanishing gradient problem when training. To overcome this limitation, Long Short-Term Memory (LSTM) [45] unit and most recently Gated Recurrent Unit (GRU) [46] were introduced, which make the recurrent layer a powerful tool for speech analysis owing to the highly time-varying character of speech and noise.

All these layer types, especially their stacked layers, provide deep neural networks the ability to deal with the raised problem of reducing noise and reverberation at the front end.

At the system back end, the Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) were widely used as acoustic models to characterise the distribution of speech a few years ago. The most common ways include Maximum A Posterior (MAP) [47] estimation and Maximum Likelihood Linear Regression (MLLR) [48]. These techniques have been successfully applied to noise adaptation. In this article, we cannot enumerate all traditional approaches, which are beyond the scope of this survey. A systematic overview of traditional approaches on the back end can be found in [25].

In spite of the success, most of these approaches suffer the significant drawbacks: (i) they are particularly designed for generative models (e.g., GMM-HMM); (ii) they assume that the adapted data match with the observed data, which is often not true in practise; (iii) they fail in modelling large-scale data and complex environments.

In recent years, the acoustic model has shifted from generative GMM to discriminative DNN owing to its powerful capability of representation learning. In this case, traditional approaches such as MAP do not work any more. New noise adaptation techniques for the DNN acoustic models need to be investigated. Besides, with the rise of big data era, it is now feasible to collect huge amounts of realistic noisy speech via the microphones that are pervasive in the world. Moreover, the advance of cloud computing makes it possible that the DNN acoustic model with millions of trainable parameters can be learnt from massive noisy data.

### C. Standard Corpora and Evaluation Metrics

To better compare the effectiveness of various deep learning approaches for noise-robust ASR, we introduce a set of widely used standard databases (see Table I) in the ASR community. Among them, the earliest and most famous databases are the Aurora series developed by the European Telecommunications Standards Institute (ETSI).

Note that all Aurora databases were artificially simulated, except Aurora 3 [49] that was recorded in a real noisy-car environment. All Aurora databases were created based on the clean and small-vocabulary database TIDigits for digit recognition, except Aurora 4 [50] that was constructed by corrupting Wall Street Journal (WSJ0) corpus and designed for Large Vocabulary Continuous Speech Recognition (LVCSR). All Aurora databases were mainly corrupted by additive noise, except Aurora 5 [51] that was developed for hand-free speech recognition and simulated by involving RIR obtained in rooms and cars.

Apart from the Aurora databases, more recently developed databases relate to CHiME series. All these CHiME databases (from the 1st to 4th) involve not only additive noise by adding various ambient noises, but also convolutional noise. More specifically, the 1st and 2nd CHiME databases [7], [11] include two tracks: one is for small vocabulary digit recognition based on Grid database, and the other is for LVCSR based on WSJ0; whereas the 3rd and 4th CHiME databases [8], [10] only include the data for LVCSR. Moreover, the 3rd and 4th CHiME databases considered more realistic noisy speech for evaluation, and applied a microphone array to obtain multi-channel signals.

Other frequently used databases include REVERB [9], AMI [52], and Voice Search [53]. Particularly, the AMI and Voice Search contain hundreds of recordings of spontaneous speech in real-life scenarios.

Overall, the standard databases were developed for scenarios from small to large vocabulary, from artificial simulation to realistic recording, from additive noise only to convolutional noise extended, and from single channel to multiple ones. All these development trends enable the ASR systems to approach a more realistic application scenario in the wild.

The *de facto* standard metric to evaluate the performance of ASR systems is *Word Error Rate* (WER) or *Word Accuracy Rate* (WAR). However, to measure the performance of the front-end techniques such as speech enhancement, other intermediate subjective and objective metrics are also frequently employed. Specifically, typical objective metrics include *segmental Signal-to-Noise Ratio* (segSNR) [55], [56], *distance measures*, *Source-to-Distortion Ratio* (SDR) [57], and *Perceptual Evaluation of Speech Quality* (PESQ) [58]. More detailed definitions and explanations of these objective metrics can be found in [59]. Although no research has proved that a good value of these intermediate metrics for enhancement techniques necessarily leads to a better WER or WAR, experimental results have frequently shown a strong correlation between them. For example, in [60] the authors conducted speech recognition on the enhanced speech, and found that SDR and WER improvements are significantly correlated with Spearman's  $\rho = 0.84$  in single-channel case, and Spearman's  $\rho = 0.92$  in two-channel case, evaluated on the CHiME-2 benchmark database.

## III. FRONT-END TECHNIQUES

The techniques at the front end often relate to speech enhancement, source separation, and feature enhancement. Both *speech enhancement* and *source separation* attempt to obtain the estimated temporal signals as clean as possible, which can certainly be used for any speech applications including ASR. *Feature enhancement*, however, mainly focuses on purifying the derived features, such as MFCCs, which are largely designed for specific intelligent tasks (i.e., ASR here). In this overview, we treat all three techniques as *enhancement* techniques, as they often share the same or similar algorithms.

When applying deep learning approaches to the environmentally robust speech recognition systems, it is particularly important to effectively and efficiently represent the information of speech signals, since training DNNs is computationally intensive. In many cases, two-dimensional representations provide speech data in an effective form, and can be obtained by applying a series of operations to the raw signals  $y(t)$ , including Short-Time Fourier Transform (STFT,  $Y(n)$ ), square magnitude ( $|Y(n)|^2$ ), Mel-frequency filterbank ( $Y^{mel}(n)$ ), log Mel-frequency filterbank ( $Y^{logMel}(n)$ ), and even Discrete Cosine Transform (DCT,  $Y^{dct}(n)$ ) (see Section II-A for more details). For a better introduction of related approaches, we separately term the data spaces after each operation as *temporal*, *magnitude-spectral*, *power-spectral*, *mel-spectral*, *log-Mel-spectral*, and *Mel-cestral* domains. Enhancement techniques can theoretically be applied to each domain, i.e., from the raw signals in the temporal domain to the MFCCs in the cepstral domain.

Deep learning-based front-end techniques are normally designed in a supervised manner. For a better review, we set the input of a learning model as  $\mathbf{y}$  that is the representation extracted from noisy speech, and the target as  $\mathbf{x}$ . Based on how the training target  $\mathbf{x}$  is obtained, the techniques can be categorised into i) *mapping-based* methods, where  $\mathbf{x}$  is the representation, straightforwardly extracted from clean speech,

TABLE I

GENERAL DESCRIPTION OF SOME STANDARD EVALUATION CORPORA FOR ENVIRONMENTALLY ROBUST SPEECH RECOGNITION. THESE CORPORA ARE EITHER *realistically* RECORDED OR ARTIFICIALLY *simulated* BASED ON CERTAIN CLEAN DATABASES. *Additive* AND/OR *convolutional* NOISES ARE COLLECTED IN VARIOUS ENVIRONMENTS.

Dataset	based on	environments	sim./real	noise types	channels
Aurora-2 [54]	TIDigits	eight conditions	sim.	add. (mainly)	single
Aurora-3 [49]	TIDigits	car	real	add. (mainly)	single
Aurora-4 [50]	WSJ0	str/tra/car/bab/res/air	sim.	add. (mainly)	dual
Aurora-5 [51]	TIDigits	rooms and cars	sim.	add. & con.	single
CHiME-1 [7]	Grid, WSJ0	home	sim	add. & con.	dual
CHiME-2 [11]	Grid, WSJ0	home	sim	add. & con.	dual
CHiME-3 [8]	WSJ0	bth/bus/caf/ped/str	real & sim	add. & con.	six
CHiME-4 [10]	WSJ0	bth/bus/caf/ped/str	real & sim	add. & con.	six
REVERB [9]	WSJCAM0	ambient noise	real & sim	add. & con.	eight
AMI [52]	-	meeting	real	con. (mainly)	four/eight
Voice Search [53]	-	voice search	sim.	add. & con.	dual

or ii) *masking-based* methods, where  $\mathbf{x}$  is a mask calculated between clean and noisy speech.

#### A. Mapping-based Deep Enhancement Methods

The mapping-based methods aim to learn a non-linear mapping function  $F$  from the observed noisy speech  $y(t)$  into the desired clean speech  $s(t)$ , as

$$y(t) \xrightarrow{F} s(t). \quad (10)$$

Owing to the fast-variation problems of raw speech signals and the high computational complexity they require, such a learning strategy is often applied to the data in the spectral and cepstral domains rather than the temporal domain.

To learn  $F$ , the neural networks are trained to reconstruct the target features  $\mathbf{x}$  (extracted from the clean speech  $s(t)$ ) from the corresponding input features  $\mathbf{y}$  (extracted from the corrupted speech  $y(t)$ ). The parameters of neural networks (models)  $\theta$  are determined by minimising the objective function of the Mean Squared Error (MSE):

$$\mathcal{J}(\theta) = \frac{1}{N} \sum_{n=1}^N \|F(\mathbf{y}_n) - \mathbf{x}_n\|^2, \quad (11)$$

where  $\|\cdot\|^2$  is the squared loss, and  $n$  denotes the frame index. After the estimated clean features  $\hat{\mathbf{x}}_n = F(\mathbf{y}_n)$  are obtained, they will be then reversed back to the time-domain signals  $\hat{s}(t)$  by using the phase information from the original noisy speech, and evaluated by the objective measures as aforementioned.

1) *Based on Stacked AutoEncoder or Deep Boltzmann Machine*: Specifically, in 2013, a Stacked AutoEncoder (SAE) was employed to map noisy speech to clean speech in the Mel-spectral domain [61]. Given an AutoEncoder (AE) that includes one non-linear encoding stage and one linear decoding stage for real valued speech as

$$\begin{aligned} h(\mathbf{y}) &= g(\mathbf{W}_1 \mathbf{y} + \mathbf{b}) \\ \hat{\mathbf{x}} &= \mathbf{W}_2 h(\mathbf{y}) + \mathbf{b}, \end{aligned} \quad (12)$$

where  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are the weight matrices of encoding and decoding,  $\mathbf{b}$  is the bias, and  $g$  denotes the activation function. The training pair for the first AE is  $\mathbf{y}$  and  $\mathbf{x}$ , and then the training pair for the next AE will be  $h(\mathbf{y})$  and  $h(\mathbf{x})$  if weight matrices of the encoder and decoder are tied, i.e.,

$\mathbf{W}_1 = \mathbf{W}_2^T = \mathbf{W}$ . The empirical results indicate that SAE-based enhancement methods notably outperform the traditional methods like MMSE for enhancing speech distorted by factory and car noises [61].

Analogous to this, another successful work has been shown in [62], where a Deep Boltzmann Machine (DBM) was utilised to estimate the complex mapping function. In the pre-training stage, noisy speech was used to train Restricted Boltzmann Machines (RBMs) layer-by-layer in a standard unsupervised greedy fashion to obtain a deep generative model [41]; whereas, in the fine-tuning process, the desired clean speech was set as the target by minimising the objective function as Eq. (22). Similar research efforts were also extensively made on the log magnitude [63] and the log-Mel-spectral domains [36], respectively.

Motivated by the fact that the same distortion in different frequency bands has different effects on speech quality, a weighted SAE was proposed in [64] and showed positive performance for denoising. In detail, a weighted reconstruction loss function is employed to train SAE on the power spectrum as

$$\mathcal{J}(\theta) = \frac{1}{N} \sum_{n=1}^N \lambda_w \|F(\mathbf{y}_n) - \mathbf{x}_n\|^2, \quad (13)$$

where  $\lambda_w$  is a weight for the  $w$ -th frequency band.

Further, related approaches were also shown in [65] and [66], where the authors utilised Stacked Denoising AutoEncoders (SDAEs) to enhance the Mel filterbank features corrupted by either additive or convolutional noise for ASR. The networks were pre-trained with multi-condition data, and fine-tuned by mapping the noisy speech to the clean speech. Experimental results indicate that the SDAE-based mapping method remarkably outperforms the spectral subtraction method in ASR.

2) *Based on LSTM-RNN*: For sequence-based pattern recognition, context information is considered to be vitally important [45]. However, the aforementioned denoising networks (i.e., SAE, DBM, and SDAE) are considered to be less capable in this respect, although certain naive solutions for context-dependent processing have been applied, such as expanding several sequential frames as a long vector input [62]. RNNs, especially the LSTM-RNNs, have been frequently

demonstrated to be highly capable of capturing the context information in a long sequence [67], [68].

In this light, Maas et al. [12] introduced RNNs to clean distorted input features (i.e., MFCCs). Specifically, the model was trained to predict clean features when presented with a noisy input frame by frame. This enhancement model has been shown to be competitive with other DNN-based mapping models at various levels of SNR when evaluated by ASR systems. Following from this work, Wöllmer et al. [69] further proposed to use LSTM-RNNs to handle highly non-stationary additive noise, which was then extended to coping with reverberation in [70]–[74]. With the help of LSTM-RNN, the speech recognition systems perform much better than the ones without LSTM-RNN when decoding noisy speech [70]–[73].

3) *Based on CNN*: Owing to the capability to capture the inherent representations embedded in the spectro-temporal feature space or in the raw signals, CNNs have attracted increasing interest in recent years [2], [75]. For image restoration and further image processing tasks, *deep convolutional encoder-decoder* networks were proposed in [76] and delivered promising performance. This network was further introduced for speech enhancement [77], where the time-frequency spectrum (spectrogram) is viewed as an image. Specifically, the encoder network includes multiple convolutional layers in order to discover the primary information from the spectrum, and the decoder network is composed of a hierarchy of decoders, one corresponding to each encoder, for compensating the details. In order to have suitable error back-propagation to the bottom layers and to pass important information to the top layers, symmetrical links between convolutional and de-convolutional layers are added by employing skip-layer connections.

However, one main drawback of the widely used spectral or cepstral representations is discarding of potentially valuable information, such as phase. When recovering the speech, the noisy phase spectrum is straightforwardly applied in constructing the enhanced speech, even though it may suffer from distortion.

Most recently, a novel network structure, namely WaveNet [78], was announced to synthesise natural speech. It takes a series of small causal and dilated convolutional layers with exponentially increasing dilation factors, which contributes to a receptive field growth that is exponential with respect to depth, and a significant reduction of the computational complexity. This provides an opportunity to directly map the noisy speech to clean speech in temporal domain, which is supposed to retain the complete speech information (including phase). Two exemplary works are shown in [79] and [80]. Particularly in [79], an explicit prior model that learns the conditional distribution of speech samples for clean speech is further incorporated with WaveNet, to regularise the enhanced speech to be more speech-like.

To further refine the model enhancement performance, adversarial training has recently attracted increasing attention. This training algorithm implements two networks, i.e., one generative network (G) and one discriminative network (D), in a cascaded network structure. The generative network tries to map the noisy speech into the clean speech so as to fool the

discriminative network, whereas the discriminative network aims to distinguish whether its inputs come from the enhanced speech (False) or the clean speech (True). Therefore, the two networks play a minimax game, and are optimised by

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})} [\log(D(\mathbf{x}))] + \mathbb{E}_{\tilde{\mathbf{x}} \sim p_{data}(\tilde{\mathbf{x}})} [\log(1 - D(G(\mathbf{y})))] \quad (14)$$

The adversarial training strategy was examined in [81] and [82], and was found to perform superior to other traditional approaches, such as Wiener filtering.

4) *Brief Discussion*: The above reviewed works reflect a trend that the employed representations for enhancement have gradually moved from cepstral domain into temporal domain, mainly thanks to i) the powerful capability of deep learning to automatically extracted effective representations from raw data which ideally retain the complete information compared with the manually extracted features like MFCCs; ii) the advance of novel architecture of neural networks (e.g., dilated CNN [78]) which dramatically reduce the computational load; iii) the development of cloud computing which makes it possible to handle such a task.

They also reflect another trend relating to the network training strategy, which starts to shift from traditional way with a single network to an adversarial way with two networks [83]. The adversarial training strategy regards the enhancement process as an image generation process, with the aid of a discriminative network to enhance the generative quality of the generative network. With recent rapid development of GAN in machine learning [84], it can be expected that further improvements will be achieved in speech/feature enhancement in the future.

However, while many works simply regard the spectrogram as a traditional visual image, few works specifically take their differences into account. Traditional visual images are locally correlated, i.e., nearby pixels are likely to have similar intensities and colours; whereas the spectrograms often include harmonic correlations which spread along frequency axis while local correlation may be weaker. Therefore, more efforts are required towards this direction.

## B. Masking-based Deep Enhancement Methods

Different from the mapping-based methods, masking-based methods aim to learn a regression function from a noisy speech spectrum  $Y(n, f)$  to a Time-Frequency (T-F) mask  $M(n, f)$ . That is,

$$Y(n, f) \xrightarrow{F} M(n, f). \quad (15)$$

1) *Masks*: Two most commonly used masks in the literature include: *binary*-based mask [85] and *ratio*-based mask [86]. Typical binary-based mask often refers to *Ideal Binary Mask* (IBM), where a T-F mask unit is set to 1 if the local SNR is greater than a threshold  $R$  (indicating clean speech domination), or 0 if otherwise (indicating noise domination). That is,

$$M^b(n, f) = \begin{cases} 1, & \text{if } SNR(n, f) > R, \\ 0, & \text{otherwise,} \end{cases} \quad (16)$$

where  $SNR(n, f)$  denotes the local SNR within the T-F unit at the frame index  $n$  and the frequency bin  $f$ . Hence, the IBM is a binary matrix. Typical ratio-based mask often indicates the so-called *Ideal Ratio Mask* (IRM), where a T-F mask unit is assigned by the soft ratio of the clean speech and the noisy (mixture) speech, as follows:

$$M^r(n, f) = \frac{S^\alpha(n, f)}{S^\alpha(n, f) + N^\alpha(n, f)}, \quad (17)$$

where  $S(n, f)$  and  $N(n, f)$  are the magnitudes of clean speech and noise in the T-F domain, respectively, and  $\alpha$  is a warping factor of the magnitudes in order to differentially affect the sharpness of the mask or the dynamic ranges of the features. Specifically, for example, if  $\alpha=2/3$ , 1, or 2, the IRM is calculated from an ‘auditory’, magnitude, or power spectrum, respectively. When  $\alpha = 2$ , the IRM is closely related to the Wiener filter, and can be viewed as its instantaneous version. From Eq. (16) and (17), it can be seen that IRM-based approaches could deliver a less distorted enhanced speech, while it could potentially involve much interference [87].

Wang and Wang [88] first introduced DNNs to perform IBM estimation for speech separation, and reported large performance improvement over non-DNN-based methods. Subsequently, Wang et al. [89] compared a variety of masks and indicated that ratio masking (e.g., IRM) is superior to binary masking (e.g., IBM) in terms of objective intelligibility and quality metrics. This conclusion was further supported by the work in [90], where the obtained results suggested that IRM achieves better ASR performance than IBM. Further, motivated by the advantages and disadvantages of IBM and IRM, Grais et al. [87] combined the IBM- and the IRM-based enhanced (separated) speech by another neural network, to exploit the compensation between two approaches.

Rather than estimating the masks in the T-F domain, the masking-based approaches were also successfully applied to a reduced feature space – Mel-frequency domain [90], [91] and log-Mel-frequency domain [92] that have frequently been proven to be effective for ASR in deep learning. The experimental results in [91] showed that the masking-based approaches in the Mel-frequency domain perform better than the ones in the T-F domain in terms of SDR.

Further, another trend in masking-based approaches is replacing DNNs with LSTM-RNNs as the mask learning model [60], [91], [92], since LSTM-RNNs have shown to be capable of learning the speech and noise context information in a long temporal range thus also often being able to model events that appear non-stationary in the short term. The research efforts [91] have demonstrated that LSTM-RNNs can notably outperform DBM/SAE alternatives in the mask estimation for source separation.

However, both the IBM and IRM-based approaches for calculating the target masks simply ignore the distorted phase information, even though it has been shown to be helpful for speech enhancement [93]. For this reason, Erdogan et al. [94] proposed a *Phase-Sensitive Mask* (PSM) that is calculated by

$$M^p(n, f) = \frac{|S(n, f)|}{|Y(n, f)|} \cos(\theta), \quad (18)$$

where  $\theta$  is the difference between the clean speech phase  $\theta^s$  and the noisy speech phase  $\theta^n$ , i.e.,  $\theta = \theta^s - \theta^n$ . The experimental results on CHiME-2 database show that it outperforms the phase-nonsensitive approaches.

Note that, PSM does not completely enhance reverberant speech, since it cannot completely restore the phase. For this reason, Williamson and Wang [95] further developed this approach, naming it *complex IRM* (cIRM). It is defined as

$$M^c(n, f) = \frac{|S(n, f)|}{|Y(n, f)|} e^{j(\theta^s - \theta^n)}. \quad (19)$$

Therefore, cIRM can be regarded as the IRM in the complex domain, while PSM corresponds to the real component of the cIRM. Both two phase-based masks were demonstrated to be more effective than normal IRMs in suppressing the reverberated noise in [95].

2) *Objective Functions and Training Strategies*: In the neural network training stage, given the input  $\mathbf{y}$  from the T-F domain of mixed noisy signals  $Y(n, f)$  and the target  $\mathbf{x}$  from the calculated T-F mask  $M(n, f)$ , the parameters of neural networks  $\theta$  are determined by the so called *Mask Approximation* (MA) objective function. That is, it attempts to minimise the MSE between the estimated mask and the target mask as follows

$$\mathcal{J}(\theta) = \frac{1}{N} \sum_{n=1}^N \|F(\mathbf{y}_n) - M(n, f)\|^2, \quad (20)$$

where  $\|\cdot\|^2$  is the squared loss,  $n$  denotes the frame index, and  $F(\mathbf{y}_n)$  is restricted to the range [0,1].

In the test stage, to filter out the noise, the estimated mask  $\hat{M}(n, f) = F(\mathbf{y}_n)$  is sequentially applied to the spectrum of the mixed noisy signal  $\mathbf{y}$  by

$$\hat{\mathbf{x}}_n = \mathbf{y}_n \otimes \hat{M}(n, f), \quad (21)$$

where  $\otimes$  denotes the elementwise multiplication. After that, it transforms the estimated clean spectrum  $\hat{\mathbf{x}}$  back to the time-domain signal  $\hat{s}(t)$  by an inverse STFT.

Apart from the MA-based objective function, more and more studies have recently started to use *Signal Approximation* (SA) objective functions [91], [96], [97]. Such an alternative straightforwardly targets minimising the MSE between the estimated clean spectrum  $\hat{\mathbf{x}} = \mathbf{y} \otimes \hat{M}(n, f)$  and the target clean spectrum  $\mathbf{x}$  by

$$\mathcal{J}(\theta) = \frac{1}{N} \sum_{n=1}^N \|\mathbf{y}_n \otimes \hat{M}(n, f) - \mathbf{x}_n\|^2. \quad (22)$$

This is indeed similar to the objective function used for the mapping-based methods (cf. Section III-A). Employing the SA based objective function was empirically examined to perform better than the MA-based one for source separation [91]. Furthermore, the conclusions found in [91] and [98] indicate that combining the two objective functions (i.e., MA and SA) can further improve the speech enhancement performance in both the magnitude and the Mel-spectral domains.

Due to the importance of phase information as aforementioned, Weninger et al. [60] took the phase information in the objective function, which is called *Phase-sensitive SA*

(PSA). Specifically, the network does not predict the phase, but still predicts a masking. However, in the objective function (cf. Eq. (22)), the terms of  $\mathbf{y}_n$  and  $\mathbf{x}_n$  are in the complex domain, making the network learn to shrink the mask estimates when the noise is high [60].

Additionally, a multi-task learning framework was proposed in [96], [97] to jointly learn multiple sources (i.e., speech and noise) and the mask simultaneously. The assumption behind this idea is that the relationship between noise and its caused speech distortion could be learnt and help for estimating the clean speech. The experimental results have shown that such a joint training framework is superior to the isolated training way [96].

Although the masking-based approaches were initially designed for removing additive noise, recent research has showed that they are capable of eliminating convolutional noise as well [60], [92], [94].

#### IV. BACK-END TECHNIQUES

The back-end techniques are also known as *model-based* techniques. They leave the noisy observation unchanged, and instead let the neural networks automatically find out the relationship between the observed noisy speech and the phonetic targets. Compared with the aforementioned front-end techniques, a drawback of the back-end techniques is that they have to change the parameters, or even structures, of previous trained Acoustic Model (AM).

Early works focus on the improvement of the model structure in order to make it more robust to recognise noisy speech. The most popular approaches involve with a combination of DNNs and HMMs, such that they take advantage of neural networks for discriminative classification and HMM for context learning. A *tandem* structure is one typical approach proposed in [101]. It utilises the neural networks to predict phonemes, and explicitly considers the phoneme prediction as a discriminative feature and combines it with original features for HMM to make a final prediction.

*Multi-stream* HMM architecture [15] is another popular approach to incorporate DNN with traditional GMM-HMM model. Specifically, given the HMM emission state  $s$  and the input vector  $\mathbf{y}$ , at every time frame  $n$ , the double-stream HMM has access to two independent information sources,  $p_G(\mathbf{y}_n|s_n)$  and  $p_L(\mathbf{y}_n|s_n)$ , the acoustic likelihoods of the GMM and the RNN predictions, respectively. In particular, the RNN-based AM is discriminatively trained to generate frame-wise phone predictions. The double-stream emission probability is computed as

$$p(\mathbf{y}_n|s_n) = p_G(\mathbf{y}_n|s_n)^\lambda p_L(\mathbf{y}_n|s_n)^{1-\lambda}, \quad (23)$$

where the variable  $\lambda \in [0, 1]$  denotes the stream weight. This approach combines GMM and DNN to leverage the reliable adaptation performance of the former and high quality classification performance of the latter. In particular, when setting  $\lambda$  to 0, the structure is known as *hybrid* NN/HMM model, where only the likelihoods of the NN predictions are employed for HMMs.

Recently, owing to the capability of LSTM-RNN in learning long-term dependence, the frequently used fully-connected

layers in DNN have been shifted to LSTM layers in either tandem [102], double-stream [15], or hybrid structures [103].

One main drawback of the combined NN/HMM structures is that it highly depends on the HMM model that, nevertheless, is gradually losing its ground in speech recognition, and being replaced by the rapidly developed DNN model-only [2]. Therefore, HMM-independent approaches are more than necessary than ever before. The widely used approach comes to *multi-condition* training [104]. In doing this, various acoustic variations caused by different noises are provided in the training process, reducing the acoustic distribution mismatch between the training and the test speech. However, it requires a large amount of data in various noisy conditions, which is rarely the case in practise.

To release the large-data-size requirement and make the model become flexible, another common way is *model adaptation*, which aims to modify the parameters of a pre-trained AM to compensate the acoustic distribution mismatch. However, modifying the entire weights of the neural networks (AM) with small adaptation data easily leads to overfitting and results in noise-dependent parameter sets [105]. Alternatively, a part of neural network parameters can be modified. For example, the authors of the work [105] added an extra layer with linear activations to the input layer, the hidden layers, or the output layer of neural networks, for model adaptation, which contributes to a considerable system robustness in environmentally noisy conditions.

Rather than forcing the pre-trained AM to adapt to various noisy conditions, an alternative approach aims to let the network-based AM be informed about the noise information (or acoustic space information [109]) when training, which is often termed as *Noise-Aware Training* (NAT) [104]. In this case, a noise estimation  $\hat{\mathbf{n}}$  presented in the signal serves as an auxiliary input and is incorporated with the original observation input  $\mathbf{y}$ , i.e.,  $[\mathbf{y}, \hat{\mathbf{n}}]$ . In this way, the DNN is being given additional cues in order to automatically learn the relationship between noisy speech and noise, which is beneficial to predict phonetic targets [104]. Experimental results on the ‘Aurora-4’ database show that the NAT-based AM is quite noise robust.

Therefore, the key point is changed to how to represent the noise information. Early works implement traditional signal processing approaches, such as MMSE, and estimate the noise over each sentence. Recently, a more general way to represent noise is employing *i-vectors* [117], which were originally developed for speaker recognition. The i-vector can be calculated either from the hand-crafted features such as commonly used MFCCs [106], or from the automatically extracted bottleneck representations by DNN [107]; and either from the raw noisy features, or from the enhanced features, e.g., Vector Taylor Series (VTS) [107].

Most of these studies assume that the noise is stationary within an utterance, so that the obtained noise estimation or i-vector can be applied to the whole utterance. Nevertheless, this is not always the case in practise. To address this issue, dynamic NAT was introduced in [108], where the authors used masking-based approaches (cf. Section III-B) to estimate the noise that varies along time. This approach performs more efficient especially for non-stationary noise, whereas it



TABLE II

A SUMMARY OF REPRESENTATIVE *single-channel* APPROACHES BASED ON DEEP LEARNING FOR ENVIRONMENTALLY ROBUST SPEECH RECOGNITION. THOSE METHODS ARE SUMMARISED AT DIFFERENT ASR PROCESSING STAGES (*front-END*, *back-END*, AND *JOINT FRONT- AND BACK-END*).

stage	approaches	typical publications	advantages	disadvantages
front (mapping)	MFCC	[12], [66], [69]–[71]	low dimension, require less computational load	lose much information, (almost) irreversible to raw signals
	(log) Mel (log/power) mag. temporal	[61], [65], [72], [73] [36], [63], [77], [99] [79], [80]	low dimension, require less computational load invertible to the audio signal retain the complete information	lose some information, (almost) irreversible to raw signals high dimension, require high computational load, each element is equally important large data size, require heavy computational load
front (masking)	IBM	[87]–[89]	little interference	much magnitude distortion
	IRM	[87], [89]–[91], [97]	little magnitude distortion	much interference
	PSM	[94]	less phase distortion and little interference	do not restore the complete phase information
	cIRM	[100]	learn a complete relationship between noisy and clean speech for both magnitude and phase	relative complex to compute
	MA SA PSA	[88], [90], [91] [87], [94], [97] [60]	most straightforward way directly optimise the objective considered phase information when predicting mask	no enhanced phase no enhanced phase
back	tandem	[101], [102]	} explicitly make use of discriminative features	} HMM model dependent
	double-stream	[15], [73]		
	hybrid	[103]	} straightforward and very efficient	} require many data in different noisy scenarios
	multi-condit. train.	[88], [104]		
	model adapt.	[105]	flexible to different noisy environments	require a relatively large amount of adaptation data, otherwise easy overfitted
back	NAT	[104], [106], [107]	easy to be implemented	require another disassociated model to estimate noise and cannot be optimised jointly
	dynamic NAT multi-task train.	[108] [109], [110]	more efficient to deal with non-stationary noise exploit the clean-sensitive speech	more efforts to estimate noise could lose some discriminative features
joint	re-training	[70], [72]	no need to change the structure of acoustic model	do not guarantee a better speech recogniser since the two nets are optimised by different metrics
	joint	[111]–[116]	exploit the complementary of enhancement networks and speech recognition networks	tricky to combine the two networks
	end-to-end	[2], [16]	automatically distil the salient features for speech recognition from raw noisy speech (or low-level features), so it reduces the information loss	require a large amount of training data and heavy computational load

requires an extra DNN for noise estimation.

Apart from these approaches, a *multi-task learning* based AM has attracted increasing attention. For example, the work done in [109] and [110] respectively introduced similar multi-task learning architectures but different network types (i.e., one is a feed-forward DNN and the other one is a LSTM-RNN) for noisy speech recognition, where the primary task is the senone classification and the augmented task is reconstructing the clean speech features. In these architectures, the objective function is calculated by

$$\mathcal{J}(\theta) = \lambda E_c + (1 - \lambda) E_r, \quad (24)$$

where  $E_c$  and  $E_r$  indicate the senone classification error and the clean feature reconstruction error, respectively. The underlying assumption of this approach is that the representations that are good for producing clean speech should be easier to be classified.

## V. JOINT FRONT- AND BACK-END TRAINING TECHNIQUES

Most research efforts on fighting with the environmental noise in the past few year were separately made on the system front end or back end. That is, speech/feature enhancement and speech recognition are often designed independently and, in many cases, the enhancement part is tuned according to the metrics such as segSNR, SDR, and PESQ, which are not directly correlated with the final recognition performance.

To address this issue, a straightforward way is employing the enhanced speech obtained in the front end to *re-train* the pre-trained AM in the back-end [70]. This simply remains everything unchanged but a further re-training process on the AM.

A more sophisticated *joint* DNN structure was proposed in [111], [112], where the authors concatenated two independent pre-trained DNNs. The first DNN performs the reconstruction of the clean features from noisy features augmented by a noise estimation. The second DNN attempts to learn the mapping between the reconstructed features and the phonetic targets [111]. Then, joint the two individual networks as one and further fine-tune the network parameters together. Compared with the re-training strategy, the joint neural networks could learn more discriminative representations for speech recognition when reconstructing the clean features from the noisy ones by feature enhancement in the front-end.

Furthermore, the work done in [118] even left out the pre-training process, and directly concatenated a DNN-based speech separation front-end and a DNN-based AM back-end to build a large neural network, and jointly adjusted the whole weights. In doing this, the enhancement front-end is able to provide enhanced speech desired by the acoustic model and the acoustic model can guide the enhancement front-end to produce more discriminative enhancement. In other words, the linguistic information contained in the acoustic models can flow back to influence the enhancement front-end at the training stage. Similar work was further done in [113]–[115].

Despite the considerable effectiveness of such joint training frameworks, the enhancement process and the speech recognition process suffer from a uni-directional communication. To this end, a novel architecture was proposed in [116]. It jointly optimises the enhancement network and speech recognition network in a parallel way, rather than a cascaded way; the activations of the hidden layer of each network will

be mutually concatenated as new inputs of their next hidden layer. Thus, all the components of two networks are jointly trained and better cooperate with each other.

More recently, an *end-to-end* architecture has attracted dramatic attention, and have shown great promise in latest ASR systems [2], [75]. Its central idea is to jointly optimise the parameters of the networks at the front-end which automatically learn the inherent representations from low-level features/signals for the task at hand, and the networks at the back-end which provide final predictions. For noisy speech recognition, a quite recent and well-developed framework has been reported in [16], where two tasks were evaluated: the Aurora-4 task with multiple additive noise types and channel mismatch, and the ‘AMI’ meeting transcription task with significant reverberation. In this framework, a variety of very deep CNNs with many convolutional layers were implemented, and each of them is followed by four fully-connected layers and one softmax output layer for senone prediction. Compared with DBMs, the CNNs have the advantages [16]: 1) they are well suited to model the local correlations in both time and frequency in speech spectrogram; and 2) translational invariance, such as the frequency shift due to speaker or speaking style variations, can be more easily captured by CNNs. The reported results on the AMI corpus by using the proposed end-to-end framework is much higher than the results of traditional DBMs and are competitive to the LSTM-RNN-based AM; and the results on Aurora-4 beat any other published results on this database, even without performing any speech and feature enhancement approaches.

## VI. MULTI-CHANNEL TECHNIQUES

Microphone arrays and *multi-channel* processing techniques have recently played an increasingly significant role in the development of robust ASR [8], [9]. A central approach is *acoustical beamforming*, i.e., spatio-temporal filtering that operates on the outputs of microphone arrays and converts them to a single-channel signal while amplifying the speech from the desired direction and attenuating the noise coming from other directions. The beamformer output is often further enhanced by a *microphone array post-filter* [119], [120]. After that, the back-end techniques for single-channel speech can be applied to this enhanced data for speech recognition.

With the rapid development, deep learning has emerged as a powerful tool to evolve the traditional methods. In the following, we separately discuss the latest deep learning approaches either in a *supportive* way to assist traditional beamforming methods (a well known survey can be found in [121]) and post-filtering methods in the front end, or an *independent* way to address the multi-channel speech recognition in a joint front and back end. Note that we do not summarise the back-end techniques in this section since it shares the same techniques with the ones for single channel as aforementioned.

### A. Front End: NN-Supported Beamformers and Post-Filters

Beamformers in general require a Direction-Of-Arrival (DOA) estimate for the target signal. In *Delay-and-Sum (DS) beamforming*, which is one of the simplest approaches and

applies a fixed delay operation to align the signals of the different microphones before summing them, so as to focus on the desired target direction. In contrast, *adaptive* beamformers update the filter coefficients based on estimates of the noise and signal statistics, and have now become the dominate approaches to address the non-stationary noise due to its time-varying attribute. Among them, the *Minimum Variance Distortionless Response (MVDR)* approach and the *Generalised EigenValue (GEV)* approach have shown to be particularly promising recently [8], [10].

Specifically, the MVDR beamforming works in the frequency domain and aims to minimise the energy at the beamformer output, while simultaneously keeping the gain in the direction of the target signal fixed at unity. The complex-valued signal model is  $\mathbf{Y}(n) = S(n)\mathbf{d} + \mathbf{A}(n)$ , where the vector  $\mathbf{Y}(n) = (Y_1(n), \dots, Y_M(n))^T$  contains the instantaneous noisy observations at the  $n$ -th time instant on a given discrete frequency bin as registered by the  $M$  microphones,  $S(n)$  is the corresponding complex frequency bin of the unknown transmitted signal, the steering vector  $\mathbf{d}$  is the desired signal spatial signature encoding its direction of arrival and  $\mathbf{A}(n)$  is a  $(M \times 1)$  vector containing the noise and interference contributions. Both the signal and the noise are assumed to have zero mean. In operation, the beamformer computes a linear combination of a complex weight vector  $\mathbf{w}$  and the observation vector  $\mathbf{Y}(n)$  as  $\mathbf{x}(n) = \mathbf{w}^H \mathbf{Y}(n)$ , where  $(\cdot)^H$  denotes the Hermitian transpose. In determining  $\mathbf{w}$  using the MVDR criterion, the spatial covariance matrix representing the covariance of the noise plus interference will be needed. It is generally unknown but can be estimated as a sample covariance matrix of a suitable segment of  $N$  observations as  $\mathbf{R}_{VV} = (1/N) \sum_n \mathbf{Y}(n) \mathbf{Y}^H(n)$  [136]. By then minimising  $\mathbf{w}^H \mathbf{R}_{VV} \mathbf{w}$  with respect to  $\mathbf{w}$  subject to the constraint  $\mathbf{w}^H \mathbf{d} = 1$ , as mentioned above, the MVDR beamformer filter coefficients are given by [137]

$$\hat{\mathbf{w}}_{MVDR} = \frac{\mathbf{R}_{VV}^{-1} \mathbf{d}}{\mathbf{d}^H \mathbf{R}_{VV}^{-1} \mathbf{d}}. \quad (25)$$

The MVDR beamformer is not robust against an inaccurately estimated steering vector  $\mathbf{d}$  [138]. In contrast, GEV beamformer requires no DOA estimate and is based on maximising the output signal-to-noise ratio [139]. The beamformer filter coefficients for a given frequency bin are found as the principal eigenvector of a generalised eigenvalue problem as required by [139]

$$\hat{\mathbf{w}}_{GEV} = \arg \max_{\mathbf{w}} \frac{\mathbf{w}^H \mathbf{R}_{SS} \mathbf{w}}{\mathbf{w}^H \mathbf{R}_{VV} \mathbf{w}}, \quad (26)$$

where  $\mathbf{R}_{SS}$  and  $\mathbf{R}_{VV}$  are the required estimates of the spatial covariance matrices of the target speech and noise/interference, respectively.

Recently reported NN-supported beamformers can be generally categorised into two types [135]: (i) beamformers with a *mask estimation* network [122], [123], [125], [140]; and (ii) beamformers with a *filter estimation* network [126]–[128]. Both approaches aim to obtain an enhanced signal based on the formalisation of the conventional filter-and-sum beamformer

TABLE III

A SUMMARY OF REPRESENTATIVE *multi-channel* APPROACHES BASED ON DEEP LEARNING FOR ENVIRONMENTALLY ROBUST SPEECH RECOGNITION. THOSE METHODS ARE SUMMARISED AT DIFFERENT ASR PROCESSING STAGES (*front*-END AND *joint* FRONT- AND BACK-END).

stage	approaches	typical publica- tions	advantages	disadvantages
front	mask estimation	[122]–[125]	avoid relying on a DOA estimation	require large-scale training data
	filter coefficients estimation	[126]–[128]	easily to be integrated with DNN AM as a joint network	based on the simulated data in all possible scenarios
	post-filter estimation	[129]	do not require explicit estimates of the signal and noise statistics	require large-scale simulated data
joint	channel concatenation	[130], [131]	require no knowledge of microphone array geometry and signal information	unclear on a severe mismatch among multiple channels
	cross-channel max-pooling	[132]	able to pick the most informative channel	unable to make use of the spatial information found in multi-channel signals
	factoring spatial & spectral filtering	[133]	robust to varying target speaker direction of arrival	additional computational cost
	end-to-end	[134], [135]	automatically extracted the underlying and salient representations over multiple channels	heavy parameters tuning and computational load

in the time-frequency domain. The difference between them is how the filter coefficients are generated by neural networks. The former approach uses neural networks to estimate noise or speech masks (cf. Section III-B), which are then applied to calculate the spatial covariance matrix further followed by a calculation of filter coefficients by Eq. (25) and (26). On contrary, the later approach skips a series of interval process. It directly utilises neural networks to estimate the filter coefficients. In both approaches, the estimated filter coefficients are then applied to the multi-channel noisy signals to obtain the enhanced speech signals.

Specifically, the mask estimation-based beamformer was firstly investigated in [122], where LSTM-RNNs were used to estimate two IBMs for each microphone channel: The two IBMs receptively indicate for each T-F bins whether they are presumably dominated by speech or noise. To train neural networks, the authors further used a multi-task learning framework; the inputs are noisy speech, and the targets are two IBMs. These obtained masks are then condensed to a single speech and a single noise mask by a median filter, which are sequentially used for estimating the spatial covariance matrices  $\mathbf{R}_{SS}$  and  $\mathbf{R}_{VV}$  and in turn the beamformer coefficients  $\hat{\mathbf{w}}_{GEV}$ . However, this approach requires *both* speech and noise counterparts of the noisy speech for each microphone channel. In this case, only simulated data is possible to be employed for the network training. To relax this requirement to some extent, a follow-up work has been presented in [123], where only the clean speech was employed for the mask estimation. This slight improvement enables one to utilise more realistic noisy and clean speech pairs, which can be recorded simultaneously by a close microphone (for clean speech) and a distant microphone array (for noisy speech). The experimental results shown in [123] were competitive with the ones in previous work [122]. Apart from the mask estimation for GEV, similar approach was also applied to MVDR, where the steering vector is calculated by the principal component of the estimated spatial covariance matrix of speech, i.e.,  $\mathbf{d} = \mathcal{P}(\mathbf{R}_{VV})$ . The effectiveness of all these mentioned approaches has been demonstrated in the fourth CHiME Challenge [125], [140].

A typical filter coefficients-based approach was evaluated in [126], where the networks were trained with generalised cross correlation from simulated multi-channel data from a given array geometry using all possible DOA angles. As conventional neural networks are not able to handle complex

values directly, the real and imaginary parts of each complex weight are predicted independently [126]. Similar investigation was also shown in [127], [128].

As for post-filtering, very few recent papers appear to have used neural networks for this purpose. One such study evaluated a non-deep MLP network in predicting the post-filter parameters for a circular microphone array [129].

### B. Joint Front- and Back-End Multi-Channel Techniques

Rather than using neural networks to support traditional beamformers and post-filters for speech enhancement, joint front- and back-end multi-channel ASR systems have recently attracted considerable attention with a goal of decreasing the WER directly [131], [132], [134]. In [130], the individual features extracted from each microphone channel are concatenated as a long single feature vector and fed into a DNN for AM. Whilst such a feature concatenation operation is simple, it was still found to be effective for dereverberation on the AMI dataset [130], and was further verified in [131].

A more sophisticated approach was proposed in [132]. In this work, the authors utilised a joint network structure of several individual convolutional layers followed by a shared fully-connected feedforward network. In more detail, each individual convolutional layer was operated on each channel independently with the magnitude spectrum as input, and a max pooling was proceeded across channels to choose the channel with the largest response in each node. This algorithm performs better than the one by applying a CNN after a DS beamformer [132].

Encouraged by this work as well as the research trend of end-to-end ASR systems, this work was extended to handle raw speech directly and without the operation of cross-layer max pooling [134]. The advantage of these extensions is that the system can automatically exploit the spatial information found in the fine time structure, which primarily lies in the previously discarded FFT phase value, of the multichannel signals [134]. A follow-up work was reported in [133], where the authors employed two convolutional layers, instead of one layer, at the front-end. The assumption is that the spatial and spectral filtering operations can be separately processed by two convolutional layers. That is, the first layer is designed to be spatially selective, and the second layer is implemented to decompose frequencies that are shared across all spatial filters.

TABLE IV

BENCHMARKS FOR FOUR SELECTED STANDARD CORPORA (I. E., AURORA-4, CHiME-2, CHiME-4, AND AMI). NOTE THAT ONLY THE DEEP LEARNING RELATED APPROACHES WERE INDICATED FOR EACH PRESENT SYSTEM. THAT IS, MANY OTHER TRADITIONAL APPROACHES MIGHT ALSO BE UTILISED.

FURTHER NOTE THAT DNN MENTIONED IN THE TABLE GENERALLY REFERS TO DEEP BOLTZMANN MACHINE (DBM) OR DEEP BELIEF NETWORK (DBN). MCT: MULTI-CONDITION TRAINING; MTL: MULTI-TASK LEARNING; WRN: WIDE RESIDUAL NETWORK [141]; VDCNN: VERY DEEP CNN.

pub.	single channel			multiple channels		model	WER (or SDR)
	front-end	back-end	joint	front	joint		
Aurora-4							
[90]	IBM/IRM (MA; Mel)	MCT	end-to-end			DNN	16.50%
[104]		NAT; MCT; re-training				DNN	12.40%
[142]		NAT, MTL					8.80%
[16]						VDCNN	8.81%
CHiME-2							
[70]	mapping: MFCC	Re-training	joint			BLSTM	26.73%
[143]		multi-stream, re-training				BLSTM	41.42%
[91]	masking: IRM (MA, SA; Mel)					LSTM	17.68 (SDR)
[103]		hybrid				BLSTM	22.20%
[72]	mapping: log Mel	re-training				BLSTM	22.16%
[94]	masking: PSM (SA; log Mel)					BLSTM	14.76 (SDR)
[63]	mapping: log mag	hybrid				DNN	≈25%
[110]	masking: IRM (SA; log Mel)	MTL; hybrid				BLSTM	16.04%
[144]	masking: IRM (MA; Mel)					DNN	15.40%
[60]	masking: IRM (PSA; Mel)					BLSTM	13.76%
[114]	masking: IRM (power spec.)	multi-stream; model adapt.; MCT				DNN	10.63%
CHiME-4 <sup>a</sup>							
[125]				mask est.		BLSTM	/; 4.0%, 5.2%
[140]				mask est.		WRN & LSTM	1.7%, 9.9%; 3.1%, 3.9%
[145]				filter coeff. est.; mask est.		LSTM	21.4%, 20.9%; 5.0%, 6.4%
[146]		re-training; hybrid		mask est.		BLSTM	/; 3.4%, 4.4%
[147]		NAT			end-to-end	VDCNN & LSTM	12.9%, 13.9%; 6.3%, 6.4%
AMI <sup>b</sup>							
[130]		MCT			channel concatenation	DNN-HMM	57.30%
[132]					cross-channel max-pooling	CNN	49.40%
[131]					channel concatenation	DNN	44.80%
[16]			end-to-end			VDCNN	46.90%
[126]				filter coeff. est.	end-to-end	DNN	42.20%
[135]				mask est.	end-to-end		39.00%

[a] results are provided for the one and six channel signals (separated by ‘;’) on the simulated and real subsets of the evaluation dataset (separated by ‘,’);

[b] only the results for MDM subset are provided.

By factoring the spatial and the spectral filters as separate layers in the network, the performance of the investigated system was notably improved in terms of WER [133].

## VII. CONCLUSIONS

In this survey, we have attempted to provide a comprehensive overview on the state-of-the-art and most promising deep learning approaches with the goal of improving the environmental robustness of speech recognition systems. These technologies are mainly introduced from the viewpoint of single-channel and multi-channel processing at different stages of the ASR processing chain, i. e., the front-end, the back-end, or the joint front- and back-end.

To intuitively compare the performance of different approaches, we selected four benchmark databases from Table I, i. e., Aurora, CHiME-2, CHiME-4, and AMI. The rationale behind this choice is that Aurora-4 includes a large vocabulary among the Aurora series corpora; CHiME-2 is more frequently used in the past few years in comparison with CHiME-1 in single and two channels; CHiME-4 is the most recently employed standard database compared with CHiME-3 to address both additive and convolutional noise in multiple channels; and AMI has large-scale data compared with all aforementioned

databases and is public available compared with Voice Search. The experimental results for each benchmark database are shown in Table IV.

From the table, we can find that i) the deep learning-based robust ASR systems are shifting from taking conventional hand-crafted features (e. g., MFCCs) as the input, to automatically extracting the representative and discriminative features directly from noisy raw speech, mainly due to the fact that raw speech signals keep the entire information (e. g., phase) related to the targets (i. e., phoneme or word); ii) separate front-end and back-end systems are gradually defeated by joint, even end-to-end training systems, owing to the powerful non-linear learning capability of deep neural networks which can optimise all processing stages simultaneously; iii) the importance of multi-channel approaches is more striking considering the promising performance they offer.

Due to the growth in popularity of microphones embedded in smartphones for example, more realistic and large size of data are increasingly utilised to train speech recognition model for fighting with the diverse and severe acoustic environments. This consequently requires more complex and deep neural network structures and high computing resources.

Despite great achievements that deep learning has accom-

published in the fast few years, as shown in the literature, an obvious performance gap still remains between the state-of-the-art noise-robust system and the one evaluated in a degradation-free, clean environment. Therefore, further efforts are still required for speech recognition to overcome the adverse effect of environmental noises [8]–[10]. We hope that this review could help researchers and developers to stand on the frontier of the developments in this field and to make greater breakthroughs.

#### ACKNOWLEDGEMENTS

This work was supported by Huawei Technologies Co. Ltd.

#### REFERENCES

- [1] G. Saon, T. Sercu, S. Rennie, and H.-K. J. Kuo, "The IBM 2016 english conversational telephone speech recognition system," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 7–11.
- [2] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. International Conference on Machine Learning (ICML)*, New York City, NY, 2016, 173–182.
- [3] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," Microsoft Research, Tech. Rep. MSR-TR-2016-71, Oct. 2016.
- [4] P. C. Loizou, *Speech enhancement: theory and practice*. Abingdon, UK: Taylor Francis, 2013.
- [5] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012.
- [6] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, pp. 40:1–40:30, May 2012.
- [7] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, May 2013.
- [8] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, 2015, pp. 504–511.
- [9] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj *et al.*, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, Dec. 2016.
- [10] 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," *Computer Speech & Language*, Dec. 2016, in press.
- [11] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matasoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 126–130.
- [12] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. INTERSPEECH*, Portland, OR, 2012, pp. 22–25.
- [13] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Improving keyword spotting with a tandem BLSTM-DBN architecture," in *Proc. Advances in Non-Linear Speech Processing: International Conference on Nonlinear Speech Processing (NOLISP)*, Vic, Spain, 2010, pp. 68–75.
- [14] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Speech Processing for Natural Interaction with Intelligent Environments*, vol. 4, no. 5, pp. 867–881, Oct. 2010.
- [15] J. Geiger, F. Weninger, J. F. Gemmeke, M. Wöllmer, B. Schuller, and G. Rigoll, "Memory-enhanced neural networks and NMF for robust ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1037–1046, June 2014.
- [16] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, Dec. 2016.
- [17] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation for speech analysis – an overview," *IEEE Signal Processing Magazine*, vol. 34, July 2017, 24 pages.
- [18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
- [19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [20] Y. Liu, Y. Liu, S. Zhong, and S. Wu, "Implicit visual learning: Image recognition via dissipative learning model," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 2, pp. 31:1–31:24, Jan. 2017.
- [21] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, Oct. 2016.
- [22] M. Schedl, Y.-H. Yang, and P. Herrera-Boyer, "Introduction to intelligent music systems and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 2, pp. 17:1–17:8, Oct. 2016.
- [23] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
- [24] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
- [25] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, Apr. 2014.
- [26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
- [27] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, Apr. 1995.
- [28] L. Deng, "Front-end, back-end, and hybrid techniques for noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data*. Berlin/Heidelberg, Germany: Springer, 2011, pp. 67–99.
- [29] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Berlin, Germany: Springer Science & Business Media, 2012, vol. 201.
- [30] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*. Hoboken, NJ: John Wiley & Sons, 2012.
- [31] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1305–1319, Mar. 2007.
- [32] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
- [33] J. H. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Transaction on Signal Processing*, vol. 39, no. 4, pp. 795–805, Apr. 1991.
- [34] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [35] —, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 2, pp. 443–445, Apr. 1985.
- [36] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions*

- on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, Jan. 2015.
- [37] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
  - [38] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, “Non-negative matrix factorization as noise-robust feature extractor for speech recognition,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, TX, 2010, pp. 4562–4565.
  - [39] F. Weninger, J. Feliu, and B. Schuller, “Supervised and semi-supervised suppression of background music in monaural speech recordings,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 61–64.
  - [40] J. Geiger, J. F. Gemmeke, B. Schuller, and G. Rigoll, “Investigating NMF speech enhancement for neural network based acoustic models,” in *Proc. INTERSPEECH*, Singapore, 2014, pp. 2405–2409.
  - [41] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.
  - [42] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
  - [43] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proc. European conference on computer vision (ECCV)*, Zurich, Switzerland, 2014, pp. 818–833.
  - [44] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5200–5204.
  - [45] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
  - [46] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder-decoder approaches,” in *Proc. Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, Doha, Qatar, 2014, pp. 103–111.
  - [47] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains,” *IEEE transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.
  - [48] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.
  - [49] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, “SPEECHDAT-CAR. A large speech database for automotive environments,” in *Proc. the Second International Conference on Language Resources and Evaluation LREC*, Athens, Greece, 2000, 6 pages.
  - [50] D. Pearce and J. Picone, “Aurora working group: DSR front end LVCSR evaluation AU/384/02,” *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep.*, 2002.
  - [51] H.-G. Hirsch, “Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments,” *Niederrhein Univ. of Applied Sciences*, 2007.
  - [52] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, “The ami meeting corpus: A pre-announcement,” in *International Workshop on Machine Learning for Multimodal Interaction*, 2005, pp. 28–39.
  - [53] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strophe, “‘your word is my command’: Google search by voice: a case study,” in *Advances in Speech Recognition*. Springer, 2010, pp. 61–90.
  - [54] D. Pearce and H. Hirsch, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proc. INTERSPEECH*, Beijing, China, 2000, pp. 29–32.
  - [55] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech quality*. Upper Saddle River, NJ: Prentice-Hall, 1988.
  - [56] J. H. Hansen and B. L. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” in *Proc. International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998, pp. 2819–2822.
  - [57] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.
  - [58] I.-T. R. P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” 2001.
  - [59] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
  - [60] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in *Proc. International Conference on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, 2015, pp. 91–99.
  - [61] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 436–440.
  - [62] Y. Xu, J. Du, L. Dai, and C. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
  - [63] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, “Learning spectral mapping for speech dereverberation and denoising,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, Apr. 2015.
  - [64] B. Xia and C. Bao, “Speech enhancement with weighted denoising auto-encoder,” in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 3444–3448.
  - [65] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, “Reverberant speech recognition based on denoising autoencoder,” in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 3512–3516.
  - [66] X. Feng, Y. Zhang, and J. Glass, “Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1759–1763.
  - [67] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, “Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, Oct. 2010.
  - [68] A. Graves, “Generating sequences with recurrent neural networks,” *arXiv preprint arXiv:1308.0850*, Aug. 2013.
  - [69] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, “Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6822–6826.
  - [70] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, “The munich feature enhancement approach to the 2nd CHiME challenge using BLSTM recurrent neural networks,” in *Proc. 2nd CHiME workshop on machine listening in multisource environments*, Vancouver, Canada, 2013, pp. 86–90.
  - [71] Z. Zhang, J. Pinto, C. Plahl, B. Schuller, and D. Willett, “Channel mapping using bidirectional long short-term memory for dereverberation in hand-free voice controlled devices,” *IEEE Transactions on Consumer Electronics*, vol. 60, no. 3, pp. 525–533, Aug. 2014.
  - [72] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, “Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments,” *Computer Speech & Language*, vol. 28, no. 4, pp. 888–902, July 2014.
  - [73] F. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachikawa, J. Geiger, B. Schuller, and G. Rigoll, “The MERL / MELCO / TUM system for the REVERB challenge using deep recurrent neural network feature enhancement,” in *Proc. REVERB Workshop, held in conjunction with ICASSP 2014 and HSCMA 2014*, Florence, Italy, 2014, pp. 1–8.
  - [74] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, “Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks,” in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 3593–3597.
  - [75] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, “Convolutional, long short-term memory, fully connected deep neural networks,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 4580–4584.
  - [76] X. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in *Proc. Advances In Neural Information Processing Systems (NIPS)*, Barcelona Spain, 2016, pp. 2802–2810.

- [77] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, Sep. 2016.
- [78] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, Sep. 2016.
- [79] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian WaveNet," *Proc. Interspeech*, pp. 2013–2017, 2017.
- [80] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," *arXiv preprint arXiv:1706.07162*, June 2017.
- [81] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, Mar. 2017.
- [82] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2008–2012.
- [83] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Montréal, Canada, 2014, pp. 2672–2680.
- [84] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, 2017, 14 pages, to appear.
- [85] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Boston, MA: Springer US, 2005, pp. 181–197.
- [86] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, Nov. 2006.
- [87] E. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley, "Combining mask estimates for single channel audio source separation using deep neural networks," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 3339–3343.
- [88] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, July 2013.
- [89] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
- [90] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7092–7096.
- [91] F. Weninger, J. R. Hershey, J. L. Roux, and B. W. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, 2014, pp. 577–581.
- [92] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 3709–3713.
- [93] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, Apr. 2011.
- [94] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 708–712.
- [95] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, July 2017.
- [96] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1562–1566.
- [97] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.
- [98] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 4390–4394.
- [99] Y. Xu, J. Du, L. Dai, and C. Lee, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
- [100] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *Proc. IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5590–5594.
- [101] S. Sharma, D. Ellis, S. S. Kajari, P. Jain, and H. Hermansky, "Feature extraction using non-linear transformation for robust speech recognition on the Aurora database," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000, pp. 1117–1120.
- [102] M. Wöllmer, F. Eyben, B. W. Schuller, Y. Sun, T. Moosmayr, and N. Nguyen-Thien, "Robust in-car spelling recognition – a tandem BLSTM-HMM approach," in *Proc. INTERSPEECH*, Brighton, United Kingdom, 2009, pp. 2507–2510.
- [103] J. Geiger, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 631–635.
- [104] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7398–7402.
- [105] S. Mirsamadi and J. H. Hansen, "A study on deep neural network acoustic model adaptation for robust far-field speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2430–2434.
- [106] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 2180–2184.
- [107] C. Yu, A. Ogawa, M. Delcroix, T. Yoshioka, T. Nakatani, and J. H. Hansen, "Robust i-vector extraction for neural network adaptation in noisy environment," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2854–2857.
- [108] Y. Xu, J. Du, L. Dai, and C. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 2670–2674.
- [109] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 5014–5018.
- [110] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 1–5.
- [111] K. H. Lee, S. J. Kang, W. H. Kang, and N. S. Kim, "Two-stage noise aware training using asymmetric deep denoising autoencoder," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5765–5769.
- [112] K. H. Lee, W. H. Kang, T. G. Kang, and N. S. Kim, "Integrated DNN-based model adaptation technique for noise-robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5245–5249.
- [113] T. Gao, J. Du, L. Dai, and C. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Australia, 2015, pp. 4375–4379.
- [114] Z. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, Apr. 2016.
- [115] M. Mimura, S. Sakai, and T. Kawahara, "Joint optimization of denoising autoencoder and DNN acoustic model based on multi-target learning for noisy speech recognition," in *Proc. INTERSPEECH*, San Francisco, CA 2016, pp. 3803–3807.
- [116] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "A network of deep neural networks for distant speech recognition," in *Proc. IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 4880–4884.
- [117] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.
- [118] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 2504–2508.



- [119] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post-filtering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.
- [120] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, Nov. 2003.
- [121] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
- [122] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, 2015, pp. 444–451.
- [123] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 196–200.
- [124] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 1981–1985.
- [125] T. Menne, J. Heymann, A. Alexandridis, K. Irie, A. Zeyer, M. Kitzka, P. Golik, K. Ilia, L. Durde, R. Schlatter, H. Ney, R. Haeb-Umbach, and A. Mouchtaris, "The RWTH/UPB/FORTH system combination for the 4th CHiME challenge evaluation," in *Proc. 4th International Workshop on Speech Processing in Everyday Environments (CHiME)*, San Francisco, CA, USA, 2016, pp. 49–51.
- [126] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5745–5749.
- [127] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 1976–1980.
- [128] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 271–275.
- [129] P. Pertilä and J. Nikunen, "Microphone array post-filtering using supervised machine learning for speech enhancement," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 2675–2679.
- [130] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, 2013, pp. 285–290.
- [131] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 5542–5546.
- [132] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, Sep. 2014.
- [133] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. W. Senior, K. K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, May 2017.
- [134] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 4624–4628.
- [135] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 2632–2641.
- [136] X. Mestre and M. A. Lagunas, "On diagonal loading for minimum variance beamformers," in *Proc. 3rd IEEE International Symposium on Signal Processing and Information Technology*, Darmstadt, Germany, 2003, pp. 459–462.
- [137] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.
- [138] A. Khabbazi-basmenj, S. A. Vorobyov, and A. Hassanien, "Robust adaptive beamforming based on steering vector estimation with as little as possible prior information," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2974–2987, June 2012.
- [139] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1529–1539, July 2007.
- [140] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proc. 4th International Workshop on Speech Processing in Everyday Environments (CHiME)*, San Francisco, CA, 2016, pp. 12–17.
- [141] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, May 2016.
- [142] S. Kundu, G. Mantena, Y. Qian, T. Tan, M. Delcroix, and K. C. Sim, "Joint acoustic factor learning for robust deep neural network based automatic speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5025–5029.
- [143] J. Geiger, E. Marchi, F. Weninger, B. Schuller, and G. Rigoll, "The TUM system for the REVERB challenge: Recognition of reverberated speech using multi-channel correlation shaping dereverberation and BLSTM recurrent neural networks," in *Proc. REVERB Workshop, held in conjunction with ICASSP 2014 and HSCMA 2014*, Florence, Italy, 2014, pp. 1–8.
- [144] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 92–101, Jan. 2015.
- [145] X. Xiao, C. Xu, Z. Zhang, S. Zhao, S. Sun, and S. Watanabe, "A study of learning based beamforming methods for speech recognition," in *Proc. CHiME Workshop*, San Francisco, CA, 2016, pp. 26–31.
- [146] H. Erdogan, T. Hayashi, J. R. Hershey, T. Hori, C. Hori, W.-N. Hsu, S. Kim, J. Le Roux, Z. Meng, and S. Watanabe, "Multi-channel speech recognition: LSTMs all the way through," in *Proc. CHiME-4 Workshop*, San Francisco, CA, 2016.
- [147] Y. Qian and T. Tan, "The SJTU CHiME-4 system: Acoustic noise robustness for real single or multiple microphone scenarios," in *Proc. CHiME-4 Workshop*, San Francisco, CA, 2016.