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

Article *in* ACM Transactions on Intelligent Systems and Technology · May 2017

DOI: 10.1145/3178115

5 authors, including:

Zixing Zhang
Technische Universität München
84 PUBLICATIONS **1,758** CITATIONS
SEE PROFILE

Jouni Pohjalainen
Aalto University
43 PUBLICATIONS **824** CITATIONS
SEE PROFILE

Amr Mousa
Apple Inc.
29 PUBLICATIONS **467** CITATIONS
SEE PROFILE

Björn Schuller
Imperial College London
959 PUBLICATIONS **25,925** CITATIONS
SEE PROFILE

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

Zixing Zhang[a], Jürgen Geiger[b], Jouni Pohjalainen[a], Amr El-Desoky Mousa[a], Björn Schuller[a,c]

[a] *Chair of Complex and Intelligent Systems, University of Passau, Innstr. 41, 94032 Passau, Germany*
[b] *European Research Center, Huawei Technologies, Dessauerstr. 10, 80992 Munich, Germany*
[c] *Department of Computing, Imperial College London, 180 Queens' Gate, London SW7 2AZ, UK*

Abstract

Eliminating the negative effect of highly non-stationary environmental noise is a long-standing research topic for speech recognition but remains an important challenge nowadays. To address this issue, traditional unsupervised signal processing methods seem to have touched the ceiling. However, data-driven based supervised approaches, particularly the ones designed with deep learning, have recently emerged as potential alternatives. In this light, we are going to comprehensively summarise the recently developed and most representative deep learning approaches to deal with the raised problem in this article, with the aim of providing guidelines for those who are going deeply into the field of environmentally robust speech recognition. To better introduce these approaches, we categorise them into single- and multi-channel techniques, each of which is specifically described at the front-end, the back-end, and the joint framework of speech recognition systems. In the meanwhile, we describe the pros and cons of these approaches as well as the relationships among them, which can probably benefit future research.

1. Introduction

Recently, Automatic Speech Recognition (ASR) has achieved tremendous success in both academic and industrial areas [1, 2, 3]. A large amount of relevant applications have started to become as an important part of our daily life, including various smartphone assistants (e. g., Siri, Cortana, Google Now), Amazon echo, kinect Xbox, and the like. However, one of the vital issues which severely degrades their performance in real life relates to the corruption effect of various environmental noises on speech.

The ambient noises that widely corrupt the clean speech $s(t)$ in the realistic scenarios mainly include i) various *additive* noises $a(t)$, such as *stationary* white noise (e. g., constant hum from machinery), and *non-stationary* interfering talking, music, and natural sounds; and ii) non-stationary *convolutional* noise $r(t)$ that is largely caused by the room reverberation and the propagation channels. Hence, the distorted speech $y(t)$ can be expressed as

$$y(t) = s(t) * r(t) + a(t), \tag{1}$$

where t denotes the time index. Provided that it is possible to reliably detect instants of the absence of the target signal (i. e., the speech of interest), short-term stationary additive noise can be adequately tackled with standard, unsupervised noise reduction signal processing techniques developed in the 1970s and 1980s [4]. However, detecting and reducing the impact of unknown non-stationary noise, or competing non-stationary sound sources, is still very challenging in practice, owing to the unstable characteristics of non-stationary noise [5, 6, 7, 8, 9].

To overcome this problem, a new wave of research efforts have been made over the past few years, for example, the organisations of REVERB and a series of CHiME challenges [7, 10, 8, 6]. In such a new research stage, *data-driven* based approaches in a supervised machine learning paradigm has received increased attention, and have emerged as influential alternatives for enhancing the noise-robustness (i. e., non-stationary noise) of ASR systems [11]. The primary objective of these approaches is, via learning from large amounts of training data, to either obtain cleaner speech signals or features from noisy speech, or directly explore the discriminating phoneme representations from noisy speech. To this end, a revolutionary technique called *deep learning* has particularly played a central role in the recent developments [12, 13, 14]. The essential idea of deep learning is to extract high-level and complex abstractions of data by using neural networks in a deep structure (i. e., multiple layers). Deep learning has been repeatedly verified to be powerful in making use of massive training data to build complex and dedicated analysis systems, and has achieved considerable success in a variety of fields, such as gaming [15], visual recognition [16], language translation [17], and ASR [18, 19]. All these achievements have leveraged increasing research efforts on deep learning to efficiently improve 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 endeavour to establish a taxonomy of the most promising approaches, which are categorised by the two principles: i) According to the addressed channel number, the approaches can be grouped into *single-channel* and *multi-channel* techniques. ii) According to the processing stages where deep learning is applied to, these ap-

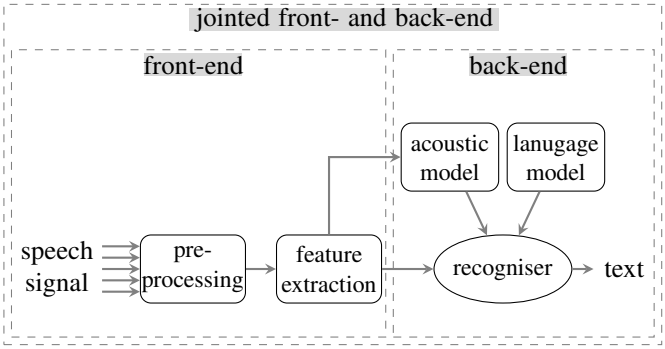

Figure 1: Framework of environmentally robust speech recognition system.

proaches can be generally classified into *front-end*, *back-end*, and *joint front- and back-end* techniques (as shown in Fig. 1). We then highlight their advantages and disadvantages, and establish their interrelations and distinctions among the prominent techniques.

Whilst several related surveys are available in the literature (e. g., [20, 21, 22, 23, 24, 5]), all these works have no focus on the usage of deep learning. This point, however, is deemed as one of the most considerable advances for the addressed problem in recent years, and forms the key concentration of this survey.

The remainder of this article is organised as follows. In Section 2, we briefly introduce the principle of some basic neural networks. In Sections 3 to 5, 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 6, we then review the most promising multi-channel algorithms, before drawing the conclusions in Section 7.

2. Related Neural Networks

Before going deeply into the review, we shortly describe the widely used neural networks for deep learning as priori knowledge. These neural networks mainly include Deep Boltzmann Machines, Stacked Autoencoders, Recurrent Neural Networks, and Convolutional Neural Networks. For more details about these networks, the reader is referred to [25] and [26].

2.1. Deep Boltzmann Machines and Stacked Autoencoders

Two of the most popular architectures at the early development stage of deep learning are *Deep Boltzmann Machines* (DBMs) [27] and *Stacked AutoEncoders* (SAEs) [28]. Each of these is obtained by stacking multiple layers of *Restricted Boltzmann Machines* (RBMs) or feedforward autoencoders, respectively. The essential idea of these networks is to utilise *deep* architectures to model the complex and potentially non-linear functions that represent *high-level* abstractions, via an unsupervised pre-training in a greedy layerwise fashion [27, 29]. Each layer is trained with an encoder $f(\cdot)$ and a decoder $g(\cdot)$ by minimising the reconstruction error for its input $\mathbf{y}$:

$$g(f(\mathbf{y})) \approx \mathbf{y}. \tag{2}$$

The output of the encoder $f(\mathbf{y})$ forms an alternative representation of the input $\mathbf{y}$, and is then fed into a successive layer as input. This procedure is repeated layer-by-layer until all predefined layers are initialised. Training the stacked layers in this manner allows a deep network to incrementally learn a representation that is more robust than the one learnt by training the whole network in ensemble from a random initialisation of weights. For a detailed description as to why pre-training autoencoders yields better performance, the reader is referred to [30]. An extension of SAE is *Stacked Denoising AutoEncoder* (SDAE) [31, 32], where the initial inputs $\mathbf{y}$ are corrupted into a partially destroyed version $\widetilde{\mathbf{y}}$ by means of stochastic mapping $\widetilde{\mathbf{y}} \sim q_d(\widetilde{\mathbf{y}}|\mathbf{y})$. By doing this, the generative capability of the high-level representations is improved [31, 32].

Further, a variation of DBMs is Deep Belief Network (DBNs). Different from DBMs of which all connections are undirected, the top two layers of DBNs form an undirected graph and the remaining layers form a belief net with directed top-down connections [33]. In this article, unless otherwise stated the term Deep Neural Networks (DNNs) particularly refers to as DBMs or DBNs.

2.2. Recurrent Neural Networks

In contrast to the aforementioned neural networks, where connections are only available between two adjacent layers, *Recurrent Neural Networks* (RNNs) allow cyclical connections. These connections consequently endow the RNNs with the capability of accessing previously processed inputs. That is, the hidden state $\mathbf{s}_t$ at time step t is obtained by the previous hidden state $\mathbf{s}_{t-1}$ at time step $t-1$ and the input $\mathbf{y}_t$ at the time step t:

$$\mathbf{s}_t = f(U\mathbf{y}_t + W\mathbf{s}_{t-1}), \tag{3}$$

where U and W are the associated weight matrices; the function f is usually a nonlinearity, such as tanh, sigmoid, or ReLU.

The standard RNNs, however, cannot access long-range context since the backpropagated error when training either blows up or decays over time (the vanishing gradient problem) [34]. To overcome this limitation, Hochreiter and Schmidhuber introduced *Long Short-Term Memory* (LSTM) RNNs [34], which are able to store the information in memory cells over a long period. The LSTM-RNNs replace the traditional neurons of RNNs by so-called *memory blocks*. Analogous to the cyclic connections in RNNs, these memory blocks are recurrently connected. Every memory block consists of self-connected linear memory cells and three multiplicative gate units: *input*, *output*, and *forget* gates. The input and output gates scale the input and output of the cell, respectively, while the forget gate scales the internal state. In other words, the input, output, and forget gates are responsible for writing, reading, and resetting the memory cell values, respectively. Therefore, LSTM-RNNs could also be regarded as a natural extension of deep neural networks for temporal sequential data, where the deepness comes from layers through time. Furthermore, similar to the DBM and the SAE, stacking RNN blocks into a deep structure has attracted increased attention, leading to a Deep RNN (DRNN) [35].

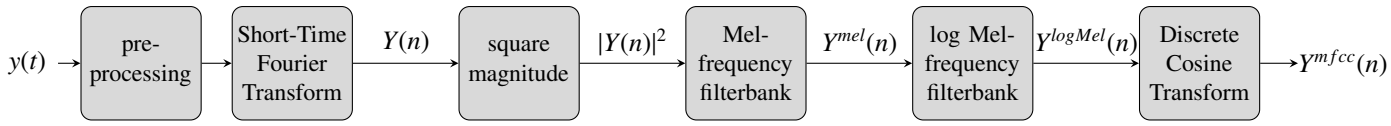

Figure 2: Block diagram of single-channel based speech recognition front-end. y and Y indicates the noisy speech in the temporal and spectral/cepstral domains, respectively; t and n denotes the time and frame indexes, respectively. For ease of notation, the (Mel-)frequency bin f or the MFCC index k is left out.

2.3. Convolutional Neural Networks

Convolutional Neural Network (CNN) is a biologically inspired variant of Multilayer Perceptron (MLP) originally developed for visual perception tasks [36, 37, 38]. It consists of one, or more, convolutional layers (often with a pooling layer), and then followed by one or more fully connected layers.

Typically, convolutional layer is referred to as feature extraction layer; it includes multiple filters (aka kernels), the size of which is normally much smaller than the dimension of the input. Each neuron of the filter is connected with a local receptive field of previous layers and extracts a local feature representation. Each of the filter is convolved with the input, giving rise to produce multiple corresponding feature maps. The feature maps are then subsampled, typically with mean or max pooling, over other small regions to reduce the feature dimension. Several such convolutional and subsample layers are concatenated in a deep structure, finally followed by fully connected layers for classification or regression. One of the advantages of CNN is that it can extract efficient features by automatic learning. For speech processing, as the time-series property of speech, 1D filter instead of 2D filter is usually considered to take input samples.

3. Front-End Techniques

The block diagram of single-channel based ASR front-end is illustrated in Fig. 2. It is given the noisy speech signal $y(t)$, which is obtained from the clean speech $s(t)$ corrupted by the convolutional noise $r(t)$ and the additive noise $a(t)$ (as shown in Eq. (1)). Then, the mixed noisy signal $y(t)$ is preprocessed for framing, etc. After that, a series of processes is executed to extract the features of Mel-Frequency Cepstral Coefficients (MFCCs) for speech recognition, including Short-Time Fourier Transform (STFT), square magnitude, Mel-frequency filterbank, log Mel-frequency filterbank, and Discrete Cosine Transform (DCT). For better introduction of related approaches, we separately term the data spaces after each processing stage as temporal, Time-Frequency (T-F) spectral, power spectral, Mel spectral, logMel spectral, and cepstral domains.

In such a set of processes, enhancement techniques can be theoretically applied to each domain, i. e., from the raw signal in the temporal domain to the MFCCs in the cepstral domain. The enhancement techniques often relate to speech enhancement, source separation, and feature enhancement. Specifically, the objective of *speech enhancement* is to improve the quality and intelligibility of the estimated target speech $\hat{s}(t)$; whilst *source separation* aims to split multiple mixed sources so as to acquire multiple independent speech sources (e. g., speakers). Thus, source separation can be viewed as a specific case of speech enhancement, where all other interfering sound sources are regarded as noise. 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 like teleconferencing. *Feature enhancement*, however, mainly focuses on purifying the derived distorted features, such as MFCCs, which are largely designed for specific intelligent tasks (i. e., ASR here). In this survey, we unified all three terms as *enhancement* techniques, as they often share the same or similar algorithms.

Before discussing the deep-learning based enhancement methods at length, we introduce the most common objective performance measures for speech enhancement and source separation, and briefly review the central traditional enhancement techniques as priori knowledge.

3.1. Objective Measures

The *de facto* standard metric to evaluate the ASR system performance is the Word Error Rate (WER). However, for speech enhancement and source separation, there are several independent ways available for evaluating the quality and intelligibility of enhanced speech. The most general and reliable way is performing *subjective* listening tests [39]. This method, however, is notoriously expensive and time consuming. For this reason, the research community established a variety of *objective* measurements to quantify the similarity of the enhanced and the original speech [40, 41, 42, 4].

Specifically, *segmental Signal-to-Noise Ratio* (segSNR) is one of the most popular objective measures [40, 41]. Rather than working on the whole signal, segSNR averages the SNRs over speech segments calculates the average of the SNR values over segments of the speech signals. Compared with SNR, segSNR gives better results for waveform encoders. In addition, *distance measures* are used to evaluate the difference between the clean and enhanced (or noisy) signal in different representation domains (e. g., spectral or cepstral); and *Source-to-Distortion Ratio* (SDR) is also frequently used in evaluating the performance of source separation algorithms [43]. Further, *Perceptual Evaluation of Speech Quality* (PESQ) [44, 45] is a test methodology developed for quality assessment of telephone-band speech. It makes use of psychoacoustic models and has been widely adopted for speech quality assessment in place of subjective listening tests. It has also shown some potential in predicting speech intelligibility but this appears to depend on the noise conditions [46].

3.2. Traditional Enhancement Techniques

Automatic removal of non-stationary noise is a central task in speech enhancement and many unsupervised signal processing

methods have been developed for it [4]. The advantage of these methods over supervised approaches is that they do not require training with noisy speech and an exactly corresponding clean target. In the context of this survey, however, we generally assume that such training material with the expected type of noise is available.

In general, these unsupervised algorithms model the noise based on the same noisy inputs that they aim to enhance. The simplest approach to automatically constructing a noise model is by means of a Voice Activity Detector (VAD) such that the noise model is updated during non-active segments. As the performance of the unsupervised noise reduction methods is strongly affected by the noise modelling approach, more sophisticated noise estimation techniques have also been developed, that track spectral minima also during speech activity [47, 48].

Whilst an instantaneous representation of the noise is obtained, noise reduction methods typically use it to construct an enhancement filter in order to remove the noise in each short-term analysis frame, and then combine the enhanced frames by the overlap-add method [49] in order to resynthesise the (enhanced) signal. Classical single-channel noise reduction methods include variants of *spectral subtraction* [50, 51, 52], where an averaged noise spectrum (magnitude or power spectrum) is subtracted from the noisy signal spectrum, while keeping the resultant spectral magnitudes positive. Spectral subtraction only affects the spectrum magnitudes, while the spectrum phases are obtained from the noisy signal. While spectral subtractive methods rely on deterministic noise models, *Wiener filtering* [4] adopts stochastic models and is often implemented in practice using iterative approaches which base new estimates of the Wiener filter on the enhanced signal obtained by the previous iteration's Wiener filter estimate [53]. Noniterative approaches based on *a priori* SNR estimation have also been developed for implementing Wiener filtering [54, 55, 4]. Another popular family of techniques comprises the *minimum mean square error (MMSE)* [56] and *log-spectral amplitude MMSE (Log-MMSE)* short-time spectral amplitude (STSA) estimators [57]. Their performance is still considered to be among the best of the published general-purpose speech noise reduction methods [58] and in comparison to many other such methods, they exhibit lower amounts of musical noise (spectral components left as residual after enhancement) [59]. In part, this can be attributed to the temporal smoothing of their "decision-directed" estimation approach, where the spectral estimate of each frame is based partially on the estimates from previous frames via the *a priori* SNR estimate updated by using a memory coefficient [59, 54, 55]. More recently, spectral subtraction [60] and the decision-directed MMSE [58] methods have also been applied in the spectral modulation domain in order to better handle non-stationary noise.

Particularly, a supervised learning-based method, namely *Non-negative Matrix Factorisation* (NMF), was recently frequently utilised [61]. The main idea is to find a dictionary (basis) matrix $\mathbf{W}$ and an activation matrix $\mathbf{H}$, so that their multiplications can be used to represent the original matrix $\mathbf{V}$ of noisy speech spectrum, i. e., $\mathbf{V} \approx \mathbf{WH}$, where both $\mathbf{W}$ and $\mathbf{H}$ are with a constraint of non-negativity. The columns of $\mathbf{W}$ are dictionary atoms $\mathbf{w}$, representing the spectra of acoustic events. To enhance the target speech, the dictionaries of speech and noise are separately trained on clean speech and noise only [62, 63]. Nevertheless, it is noted that it highly relies on the assumption that the training speech or noise are in the same acoustic distribution with the unseen speech and noise data. To improve the generalisation ability, an *exemplar-based* approach has been proposed to obtain larger dictionaries [64], which is designed to cover multiple speakers and pronunciation variants in a speech dictionary. In realistic scenarios, the applied environments or the speakers may vary and are not easy to be predicted. A compromise solution to address this issue is *semi-supervised NMF* [65]. That is, either the speech dictionary or the noise dictionary is pre-defined in the training stage. Then, in the test time, it estimates the unknown dictionary $\mathbf{W}^i$ for the source i alongside with the activations $\mathbf{H}$ of all dictionary atoms. This method has been examined in [66, 67], and has shown its efficiency.

3.3. Deep-Learning-based Enhancement Techniques

For better reviewing, we set the input of learning model (i. e., deep neural networks) as $\mathbf{y}$ that is extracted from noisy speech, and the target as $\mathbf{x}$. Based on the types of training target $\mathbf{x}$, the enhancement methods can be categorised into i) *mapping-based* methods and ii) *masking-based* methods.

3.3.1. Mapping-based Methods

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

$$y(t) \xrightarrow{F} s(t), \tag{4}$$

where F is the learnt mapping function. Owing to the high-dimension and fast-variation problems of raw speech signals, such a learning strategy is often applied to the spectral and cepstral domains rather than the temporal domain. That is, the data used for network training can be absolute magnitude $Y(n)$, power magnitude $|Y(n)|^2$, Mel-frequency filterbank $Y^{mel}(n)$, log Mel-frequency filterbank $Y^{logMel}(n)$, or MFCC $Y^{mfcc}(n)$, as shown in Fig. 2. In doing this, the data dimensions can be significantly reduced, leading to less computational complexity.

To learn F, the neural network is 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 the model θ are determined by minimising the objective function of Mean Squared Error (MSE):

$$\mathcal{J}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\|F(\mathbf{y}_n) - \mathbf{x}_n\|^2, \tag{5}$$

where $\|\cdot\|^2$ is the squared loss, and n denotes the frame index. After the estimated clean features $\hat{\mathbf{x}}_n = F(\mathbf{y}_n)$ obtained, they will be then inversed back to the time-domain signal $\hat{s}(t)$ by using the phase information from original noisy speech, and evaluated by the objective measures as mentioned in Section 3.1.

1) Based on SAE or DBM: Specifically, in 2013 Lu et al. [68] employed SAE to map the noisy speech to the clean speech in the Mel spectral domain. Given an AE that includes one non-linear encoding stage and one linear decoding stage for real valued speech as

$$\begin{aligned} h(\mathbf{y}) &= g(\mathbf{W}_1\mathbf{y} + \mathbf{b}) \\ \hat{\mathbf{x}} &= \mathbf{W}_2 h(\mathbf{y}) + \mathbf{b}, \end{aligned} \tag{6}$$

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 encoder and decoder are tied, i. e., $\mathbf{W}_1 = \mathbf{W}_2^T = \mathbf{W}$. The empirical results indicate that SAE-based enhancement method notablely outperforms the traditional method like MMSE for enhancing the speech distorted by factory and car noises [68].

Analogous to this, another successful work has been done in [69], where DBM was utilised to estimate the complex mapping function. In the pre-training stage, noisy speech was used to train RBMs layer-by-layer in a standard unsupervised greedy fashion to obtain a deep generative model [29]; whereas in the fine-tuning process, the desired clean speech was set as to the target by minimising the objective function as Eq. (13). Similar research efforts were also extensively made on the T-F [70] and the logMel spectrum domains [71], 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 [72] which has shown positive performance for denoising. In detail, a weighted reconstruction loss function is employed to train SAE on the power spectrum as

$$\mathcal{J}(\theta) = \frac{1}{N}\sum_{n=1}^{N} \lambda_w \|F(\mathbf{y}_n) - \mathbf{x}_n\|^2, \tag{7}$$

where λ is the weighting function.

Further, similar work was also done in [73] and [74], where the authors utilised SDAE 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. The experiments conducted on CENSREC-4 database indicate that the SDAE-based mapping method remarkably outperforms the spectral subtraction method [73] in ASR.

2) Based on LSTM-RNN: For the mapping-based speech enhancement, context information of speech and noise is considered to be important [75]. However, the aforementioned networks (i. e., SAE, DBM, and SDAE) are evaluated to be less capable in this aspect, although certain naive solutions were performed, such as expanding several sequential frames as a long vector input [69]. RNN, espcially the ones equiped with LSTM blocks (namely LSTM-RNN), has been frequently verified to be highly capable of capturing the context information in long sequence [76, 77], as mentioned in Section 2.2.

In this light, Maas et al. [11] introduced RNN to purify the input features (i. e., MFCCs). Specifically, the model was trained to predict clean features given noisy input frame by frame. Evaluated by ASR systems, this enhancement model is shown to be competitive with other DNN-based mapping models at various levels of SNR [11]. Following this work, Wöllmer et al. [78] proposed to use LSTM-RNN to handle highly non-stationary additive noise, which was then extended to cope with reverberation in [13, 79, 80, 81, 82]. With the help of LSTM-RNN, the speech recognition systems perform much better than the one without LSTM-RNN when decoding the noisy speech.

3) Based on CNN: More recently, an interesting research topic has emerged towards automatically learning robust phoneme representations from original noisy speech. For example, Chang et al. [83] proposed a novel feature extraction framework that concatenates a CNN and a fully connected deep neural network. Specifically, Gabor filter kernels are integrated with CNN, which aims to capture the spectro-temporal receptive field as it is designed to model human auditory processing. The bottleneck features generated from the deep neural networks are then used for training traditional Hidden Markov Model (HMM), providing superior performance on two standard noisy speech databases (i. e., Aurora-4 and noisy WSJ) [83].

3.3.2. *Masking-based Methods*

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

$$Y(n, f) \xrightarrow{F} M(n, f). \tag{8}$$

Generally, there are two typical masks available in the literature: *Ideal Binary Mask* (IBM) [84] and *Ideal Ratio Mask* (IRM) [85]. For the IBM, a T-F mask unit is set to as 1 if the the local SNR is greater than a threshold R (indicating clean speech domination), or 0 if otherwise (indicating noise domination). That is,

$$M^b(n, f) = \begin{cases} 1, & \text{if } SNR(n, f) > R, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $SNR(n, f)$ denotes the local SNR within the T-F unit at frame index n and frequency bin f. Hence, the IBM is a binary matrix.

The IBM is indeed a binary approximation to IRM. For the IRM, a T-F mask unit is assigned by the soft ratio of the clean speech energy and the noisy (mixture) speech energy, as follows:

$$M^s(n, f) = \frac{|S(n, f)|^2}{|S(n, f)|^2 + |N(n, f)|^2}, \tag{10}$$

where $S(n, f)$ and $N(n, f)$ are the clean speech and noise in the T-F spectral domain, respectively. Therefore, the IRM is closely related to the Wiener filter, and can be viewed as its instantaneous version.

In the network training stage, given the input $\mathbf{y}$ from the T-F spectrum 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 θ are determined by the called *Mask Approximation* (MA) objective function. That is, it attempts to minimise the

MSE between the estimated mask and the target mask as follows

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

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

$$\hat{\mathbf{x}}_n = \mathbf{y}_n \otimes \hat{M}(n,f), \tag{12}$$

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 inverse STFT.

Specifically, Wang and Wang [86] first introduced DNNs to perform IBM estimation for speech separation, which significantly outperforms other methods without deep learning. Subsequently, Wang et al. [87] compared a variety of masks and indicated that the IRM is superior to the IBM in terms of objective intelligibility and quality metrics. Such a conclusion was further verified by the work in [88], where the obtained results suggested that IRM achieves better ASR performance than IBM. Nevertheless, the work done by Grais et al. [89] showed that combining IBM and IRM can deliver better performance than each of them used independently for source separation.

Rather than estimating the masks in the T-F spectral domain, the masking-based approaches were also successfully applied to a reduced feature space – Mel frequency domain [90, 88] and its logarithmic scale [91] that have frequently been proven to be effective for ASR in deep learning. The experimental results in [90] showed that the masking-based approaches in the Mel frequency domain perform better than the ones in the T-F spectral domain in terms of SDR.

Further, another trend of the masking-based approaches towards replacing DNNs with LSTM-RNNs as the mask learning model [91, 92, 90], since LSTM-RNNs have shown to be capable of learning the speech and noise context information in a long temporal range [76, 77] which is vitally important for such a sequence learning task. The research efforts made in [90] have demonstrated the LSTM-RNNs can notably outperform the DNNs in mask estimation, as well as NMF, for source separation.

Apart from employing MA-based objective function to optimise the model, more and more studies have recently started to use *Signal Approximation* (SA) objective function [93, 94, 90]. Such an alternative straightforwardly targets at minimising the MSE between the estimated clean spectrum $\hat{\mathbf{x}} = \mathbf{y} \otimes F(\mathbf{y})$ and the target clean spectrum $\mathbf{x}$ by

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

This is indeed similar to the objective function used for the mapping-based methods (cf. Section 3.3.1). Empirically, employing the objective function based on SA performs better than the one based on MA for source separation [90]. Moreover, the conclusions found in [90] and [95] indicated that combining the two objective functions (i. e., MA and SA) can further improve the speech enhancement performance in both the T-F and the Mel domains.

Most of these methods, however, have only considered the magnitude or power spectrum when calculating the target ideal mask. The distorted phase information is completely ignored, even though it was proved to be helpful for speech enhancement [96]. To this end, the studies done by Erdogan et al. [97] and Weninger et al. [92] have taken the phase information into account, contributing to better performance in terms of SDR. Similarly, Williamson et al. [98] proposed complex ratio masking for DNN-based monaural speech separation, which learns the real and imaginary components of complex spectrograms jointly in the Cartesian coordinate system instead of learning magnitude spectrograms only in the traditional polar coordinate system. Notable performance improvement was observed as well.

Additionally, a multi-task learning framework has proposed in [93, 94] 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 [93].

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 [91, 97, 92].

4. 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 speech and the phonetic targets. The most widely used approach in robust ASR often involves with *multi-condition* training [99]. In doing this, various acoustic variations caused by different noises are provided in the training process, reducing the acoustic distribution gap between the training and the test data.

Another common way to compensate the acoustic distribution mismatch is *model adaptation*, which is performed on the pre-trained Acoustic Modelling (AM). However, modifying the entire weights of the neural networks with small adaptation data easily leads to overfitting and results in extremely large noise dependent parameter sets. Alternatively, small subsets of neural network weights can be modified. For example, the authors of the work [100] added extra layer with linear activation to the network of input, hidden layers, or output, for model adaptation, which contributes to a considerable system robustness in noisy conditions.

Rather than forcing the pre-trained AM to adapt to various noisy conditions, an alternative way aims to let the network-

based AM be informed about the noise information when training, which is often termed as *Noise-Aware Training* (NAT) [99]. In this case, a noise estimate present in the signal serves as an augmented input and is incorporate with the original observation input, 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 in a way that is beneficial to predict phonetic target [99]. Experimental results on Aurora-4 for speech recognition showed that the DNN-based AM trained with NAT has remarkable noise robustness. A similar work was also done in [101], where the noise estimation was replaced by the room information as an augmented input for dereverberation.

Further, a more general way to involve with the noise information in the network input comes to employing *i-vector*, which was originally developed for speaker recognition. In [102], Karanasou et al. introduced an i-vector to represent acoustic environment, except another conventional one for speaker. Then, the i-vectors are concatenated to the original acoustic features for every frames of the data. These expanded features form the input for neural network training and decoding. Analogous to this, Yu et al. [103] used the extracted i-vector to represent the noisy environment; however, the i-vectors are calculated from the Vector Taylor Series (VTS) enhanced features and bottleneck features rather than MFCCs.

Recently, a quite successful work was reported by [12], where a *double-stream HMM* architecture was used for fusing two AMs. One is traditional Gaussian Mixture Model (GMM) based AM, and other one is LSTM-RNN-based AM. 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 LSTM predictions, respectively. Particularly, the LSTM-RNN-based AM was discriminatively trained to generate frame-wise phone prediction. The double-stream emission probability is computed as

$$p(\mathbf{y}_n|s_n) = p_G(\mathbf{y}_n|s_n)^{\lambda} p_L(\mathbf{y}_n|s_n)^{1-\lambda}, \tag{14}$$

where the variable $\lambda \in [0, 1]$ denotes the stream weight. With the help of LSTM-RNN-based AM, the performance of traditional GMM-HMM system is boosted for noisy speech recognition. Moreover, the effectiveness of such a system has already been shown as a winner of the second CHiME Challenge.

Following research was introduced in [104] and found that the GMM-based AM is even unnecessary when the LSTM-RNN-based AM was trained to predict HMM states (i. e., senone, rather than phonemes), resulting in a *hybrid* LSTM/HMM model. That is, only the likelihoods of the LSTM predictions are employed for HMM, i. e., $\lambda = 0.0$ in Eq. (14). The work also revealed that combining the LSTM-RNN-based AM for phoneme or state predictions can further improve the performance.

Recently, a *multi-task learning* based AM has attracted increasing attention. For example, the work of [101] and [105] respectively introduced similar multi-task learning architectures but different network types (i. e., one is DNN and the other one is LSTM-RNN) for noisy speech recognition, where the primary task is senone classification and the augmented task is reconstructing the clean speech features. In these architectures, the objective function is calculated by

$$\mathcal{J}(\theta) = \lambda E_c + (1 - \lambda) E_r, \tag{15}$$

where E_c and E_r indicate the senone classification error and the reconstruction error, respectively. The underlying assumption of this kind of multi-task learning is that the representations that are good for producing clean speech should be easier to classify.

5. Jointed Front- and Back-End Techniques

To better take advantage of the techniques in both the front-end and the back-end, more and more interests are focusing on jointing the two. The straightforward way to do this is employing the enhanced features in the front-end for *re-training* the AM in the back-end [81].

More sophisticatedly, Lee et al. [106] proposed a *cascaded* DNN structure, which concatenated two independent fine-tuned DNNs. The first DNN performs the reconstruction of the clean features from noisy features augmented by the noise estimation. The second DNN attempts to learn the mapping between the reconstructed features and the phonetic targets [106]. The assumption of such a cascaded structure is that it could learn a most discriminative representation for speech recognition when reconstructing the clean feature from the noisy one by feature enhancement in the front-end.

A similar work has also been done by Wang et al. [107], who concatenated a DNN-based speech separation front-end and a DNN-based AM back-end to build a larger neural network, and jointly adjust the weights in each model. In doing this, the separation front-end is able to provide enhanced speech desired by the AM back-end, and the AM back-end can guide the separation front-end to produce more discriminative enhancement [107].

To utmost explore the potential of deep neural networks at different stages in the speech recognition chain, *end-to-end* systems have attracted increasing interests in recent years and have achieved tremendous achievement in ASR [2, 108]. The central idea is to jointly optimise the parameters of the networks at the front-end which automatically learn the inherent representations for the task at hand, and the networks at the back-end which provide final predictions.

A quite recent and well-developed framework was reported in [14], 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 the feedforward DNN, the CNNs have these advantages [14]: 1) It is well suited to model the local correlations in both time and frequency in

Table 1: A summary of representative *single-channel* methods based on deep learning for environmentally robust speech recognition surveyed in this paper. Those methods involve with different processing stages (*front*-end, *back*-end, and *jointed* front- and back-end), specific approaches, applied neural networks (NNs), *addtive* (add.) and/or *convolutional* (con.) noise types, and used databases.

published papers	stages	approaches	applied NNs	noise types	databases
Lu et al. 13 [68]	front	mapping-based	DSAE	add.	Japan. speech
Xu et al. 14 [69]	front	mapping-based	DBM	add.	TIMIT
Han et al. 15 [70]	front	mapping-based	DBM	add. & con.	2nd ChiME
Xu et al. 15 [71]	front	mapping-based	DBM	add.	Aurora-2
Ishii et al. 13 [73]	front	mapping-based	DBM	con.	CENSREC-4
Feng et al. 14 [74]	front	mapping-based	SDAE	add. & con.	2nd ChiME
Maas et al. 12 [11]	front	mapping-based	DRNN	add.	Aurora-2
Wöllmer et al. 13 [78]	front	mapping-based	LSTM-RNN	add.	Buckeye
Zhang et al. 14 [13]	front	mapping-based	LSTM-RNN	con.	TV control
Weninger et al. 14 [79]	front	mapping-based	LSTM-DRNN	add. & con.	2nd ChiME
Weninger et al. 14 [80]	front	mapping-based	LSTM-DRNN	con.	REVERB
Weninger et al. 13 [81]	front	mapping-based	LSTM-RNN	add. & con.	2nd ChiME
Chang & Morgen 14 [83]	front	mapping-based	CNN	add.	Aurora-4, WSJ
Wang & Wang 13 [86]	front	masking-based	DNN-SVM	add.	TIMIT
Narayanan & Wang 13 [88]	front	masking-based	DNN	add.	Aurora-4
Wang et al. 14 [87]	front	masking-based	DNN	add.	TIMIT
Weninger et al. 14 [90]	front	masking-based	LSTM-DRNN	add. & con.	2nd ChiME
Erdogan et al. 15 [97]	front	masking-based	DRNN	add. & con.	2nd ChiME
Weninger et al. 15 [92]	front	masking-based	LSTM-RNN	add. & con.	2nd ChiME
Huang et al. 15 [94]	front	masking-based	DRNN	add.	TIMIT, etc.
Grais et al. 16 [89]	front	masking-based	DNN	add.	SiSEC2015
Seltzer et al. 13 [99]	back	multi-condition, NAT	DNN	add.	Aurora-4
Mirsamadi et al. 15 [100]	back	model adaptation	DNN	add. & con.	Aspire
Giri et al. 15 [101]	back	multi-task, NAT	DNN	con.	REVERB
Karanasou et al. 14 [102]	back	factorised i-vector	DNN	add.	WSJ
Geiger et al. 14 [12]	back	multi-stream	LSTM-RNN	add. & con.	2nd ChiME
Geiger et al. 14 [104]	back	hybrid	LSTM-RNN	add. & con.	2nd ChiME
Chen et al. 15 [105]	back	multi-task	LSTM-RNN	add. & con.	2nd ChiME
Lee et al. 16 [106]	jointed	cascaded, NAT	DNN	add. & con.	Aurora-5
Wang & Wang 16 [107]	jointed	cascaded	DNN	add. & con.	2nd ChiME
Amodei et al. 16 [2]	jointed	end-to-end	CNN+DNN	add.	multiple
Qian et al. 16 [14]	jointed	end-to-end	Very Deep CNN	add. & con.	Aurora-4, AMI

speech spectrogram; 2) Translational invariance, such as frequency shift due to speaking styles or speaker variations, can be more easily captured by CNNs. The reported results on AMI by using the proposed end-to-end framework is much higher than the traditional DNN and is competitive to the LSTM-RNN-based AM, and the results on Aurora-4 achieve the best compared with any other published results on this database, even without performing any speech and feature enhancement approaches.

6. Multi-Channel Techniques

Recent advance in robust ASR has also highly acknowledged the practical importance of using microphone arrays and *multi-channel* processing (cf. the recent CHiME and REVERB challenges [8, 6] and their successful contributions which introduced improvements on the multi-channel techniques over the baseline, e. g., [109, 110, 111, 112]). A central technique is *acoustical beamforming*, i. e., spatio-temporal filtering that operates on the outputs of the microphone array and converts it to a single-channel signal while focusing on the desired speech and attenuating the noise coming from other directions. Beamforming is a long-standing research topic in array signal processing with many applications, not limited to spatial audio processing [113, 114]. Beamformer output is often further enhanced by a *microphone array post-filter* [115, 116]. In this section, we first review the standard microphone array enhancement techniques, and then discuss the deep learning in a supportive way or an independent way.

6.1. Beamforming and Post-Filtering

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, the signals received by the different microphones are considered to have been delayed according to the target signal DOA. Fixed delay operators that compensate for the arrival delays are applied to the signals of the different microphones before summing them, so as to focus on the desired target direction; signal statistics are not considered by the DS beamformer (apart from a possible DOA estimation step). In contrast, *adaptive beamforming* techniques estimate statistics from the observed signal and determine the beamformer filter coefficients based on these statistics according to some criterion. In particular, the *Minimum Variance Distortionless Response (MVDR)* or *Capon beamformer* 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 $\mathbf{Y}(n) = (Y_1(n), \ldots, Y_M(n))^T$ contains the instantaneous noisy observations at the nth 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

$$x(n) = \mathbf{w}^H \mathbf{Y}(n), \tag{16}$$

where $(\cdot)^H$ denotes the Hermitian transpose. The general approach of Eq. (16) is known as *filter-and-sum beamforming*. 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 $\mathbf{R}_{VV} = (1/N)\sum_n \mathbf{Y}(n)\mathbf{Y}^H(n)$ [117]. 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 weight vector for the discrete frequency bin in question is given by [118]

$$\hat{\mathbf{w}}_{MVDR} = \frac{\mathbf{R}_{VV}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_{VV}^{-1}\mathbf{d}}. \tag{17}$$

In multi-channel audio enhancement, the enhancement and noise reduction provided by the beamformer alone is typically not sufficient, as reducing noise and reverberation at low frequencies would require very large arrays [115]. Therefore, the output of a beamformer is often further enhanced by *post-filtering* relying on a single-channel Wiener filtering approach with statistics estimated from the multi-channel observations [119, 116, 120]. An optimal multi-channel Wiener filter formulation according to the MMSE criterion gives rise to the MVDR beamformer followed by a single-channel Wiener post-filter [115, 116, 120]. The Zelinski post-filter [119, 115] shows reasonable performance but is based on simplified assumptions including a perfectly incoherent noise field with zero correlation between the noise on different channels, an assumption that does not hold with low noise frequencies and closely spaced microphones. In order to address this issue, McCowan and Bourlard [116] proposed a generalisation of the Zelinski post-filter that assumes prior knowledge of the complex coherence of the noise field and demonstrated it with a diffuse noise model.

The MVDR beamformer is not robust against an inaccurately estimated steering vector $\mathbf{d}$ [121]. In contrast, *Generalised EigenValue (GEV) beamforming* requires no DOA estimate and is based on maximising the output signal-to-noise ratio [122]. The beamformer filter coefficients for a given frequency bin are found as the principal eigenvector of a generalised eigenvalue problem as required by [122]

$$\hat{\mathbf{w}}_{GEV} = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^H\mathbf{R}_{SS}\mathbf{w}}{\mathbf{w}^H\mathbf{R}_{VV}\mathbf{w}}, \tag{18}$$

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. When applied to broadband signals such as audio, this principle leads to considerable distortion of the speech due to independent SNR optimisation of each frequency bin. In [122], this is compensated by single-channel post-filtering in order to achieve a performance comparable to the MVDR beamformer.

6.2. Neural-Network-Supported Beamformers

To calculate the spatial covariance matrices $\mathbf{R}_{SS}$ and $\mathbf{R}_{VV}$ of the GEV beamformer (see Eq. (18)), Heymann et al. [123] proposed in the third CHiME challenge (CHiME-3) to use LSTM-RNNs to firstly estimate two Ideal Binary Masks (IBMs) for each microphone channel: one to indicate which T-F bins are presumably dominated by speech, and another to indicate which T-F bins are dominated by noise. This idea is similar to the one in [88], where a DNN was employed to predict the noise mask. To train the neural networks, the authors in [123] further used a multi-task learning framework with the input of noisy speech, and two corresponding IBM targets respectively from clean speech and noise in separate output layers. After the masks for each channel are obtained, they 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 $\hat{\mathbf{w}}_{GEV}$.

The aforementioned supervised neural network training process, however, requires *both* speech and noise counterparts of the noisy speech for each microphone channel. In this case, only the simulated data is possible to be employed for network training. To relax this requirement to some extent, a follow-up work was presented in [124], where only the clean speech was employed for mask estimation. Such a slight improvement enables it to utilise more realistic noisy and clean speech pair, which can be recorded simultaneously by close microphone (for clean speech) and distant microphone array (for noisy speech). The experimental results shown in [124] were competitive with

Table 2: A summary of representative *multi-channel* methods based on deep learning for environmentally robust speech recognition surveyed in this paper. Those methods involve with different processing stages (*front*-end, *back*-end, and *jointed* front- and back-end), specific approaches, applied neural networks (NNs), *addtive* (add.) and/or *convolutional* (con.) noise types, and used databases. WRBR: wide residual BLSTM-RNN; NIN: network in network; CLDNN: jointed convolutional, LSTM, and DNN networks.

published papers	stages	approaches	applied NNs	noise types	databases
Heymann et al. 15 [123]	front	NN-GEV	LSTM-RNN	add. & con.	3rd ChiME
Heymann et al. 16 [124]	front	NN-MVDR/GEV	LSTM-RNN	add. & con.	3rd ChiME
Menne et al. 16 [112]	front/back	superdirective, MVDR, NN-GEV; WRBR AM	LSTM-RNN	add. & con.	4th ChiME
Heymann et al. 16 [125]	front/back	NN-GEV; WRBR AM	WRBR	add. & con.	4th ChiME
Sivasankaran et al. 15 [126]	front	PSD esti.	DNN	add. & con.	3rd ChiME
Yoshioka et al. 15 [109]	back	NIN-CNN	NIN-CNN	add. & con.	3rd ChiME
Xiao et al. 15 [127]	front	DOA esti.	MLP	add. & con.	WSJ
Xiao et al. 16 [128]	front	beamform. weights esti.	DNN	add. & con.	AMI
Pertila et al. 14 [129]	front	post-filtering esti.	MLP	add. & con.	TIMIT
Liu & Nikunen 14 [130]	jointed	channel concatenation	DNN	add. & con.	AMI
Swietojanski et al. 14 [131]	jointed	max-pooling across mulitple channels	CNN	add. & con.	AMI
Hoshen et al. 15 [132]	jointed	end-to-end	CNN-DNN	add. & con.	voice search
Swietojanski et al. 13 [133]	jointed	concatenated features, hybrid AM	DNN	add. & con.	AMI
Sainath et al. 16 [134]	jointed	factoring spatial & spectral filterings	CLDNN	add. & con.	voice search
Li et al. 16 [135]	jointed	adaptive spatial filtering	LSTM-RNN, CLDNN	add. & con.	voice search

previous work. Furthermore, the effectiveness of the neural-network-supported GEV beamformer has been demonstrated in the recent 4th CHiME Challenge [112, 125].

In CHiME-3, Sivasankaran et al. [126] used a DS beamformer to obtain single-channel noisy spectra, and then used a DNN to map them into speech and noise spectra, which were then employed in constructing a multi-channel Wiener filter according to [136], in order to replace the basic MVDR beamformer used in the challenge.

While deep learning approaches have shown moderate success in CHiME-3 in estimating speech and noise statistics for multi-channel Wiener filtering and beamforming, the winning contribution was still able to rely on the baseline MVDR beamformer complemented with improved DOA estimate (the steering vector) based on T-F masks estimated using complex GMMs [109, 137]. Recent studies have, however, started to investigate the use of DNNs from the earliest stages of multi-channel analysis, such as in DOA and steering vector estimation. DOA estimation using DNNs was studied in [127] and a follow-up study investigated the implementation of DS beamforming in an ASR system directly using a feedforward DNN that is used to predict the beamformer's weight vector [128]. In both cases, the network was trained with generalised cross correlation from simulated multi-channel data from a given array geometry using all possible DOA angles. In the latter paper, the future work envisioned by the authors includes the use of spatial covariance matrices as input to the network, so that they could also use speech and noise statistics in predicting (adaptive) beamformer parameters.

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].

6.3. End-to-End Multi-Channel Approaches

Rather than using neural networks to support traditional beamformers and post-filters for speech enhancement, end-to-end multichannel ASR systems have recently attracted more attention with a straightforward target of WER decrease [130, 131, 132]. In [133], the individual features extracted from each microphone channel are concatenated as a long single feature vector and fed into DNN for AM. Whilst such a feature concatenation operation is simple, it was still found to be effective on the AMI dataset [133], and was further verified in [130].

Recently, a more sophisticated approach was proposed in [131]. In this work, the authors utilised a joint network structure of single convolutional layer followed by several fully connected DNN layers. In more detail, the convolutional layer was operated on each channel independently with the magnitude spectrum as input, while max pooling was proceeded across channels to choose the channel with the largest respond in each node. This algorithm was found to perform better than the one by applying CNN after a DS beamformer output [131].

Motivated by the success of this work as well as the research trend of end-to-end ASR systems, a study [132] extended the work of [131] on the raw signals and without the operation of cross-layer max pooling. The advantage of this extended work is that it can automatically exploit the spatial information found in the fine time structure, which greatly lies in the previously discarded FFT phase value, of the multichannel signals.

A follow-up work was reported in [134], where the authors employed two convolutional layers, instead of one layer, at the front-end. The assumption is that the spatial and spectral filtering process can be separately processed by two convolutional layers. That is, the first layer is designed to be spatially selective, and second layer is implemented to decompose frequency that are shared across all spatial filters. 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 [134].

7. Conclusion

In this survey, we provided a comprehensive review 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 in the viewpoint of single-channel and multi-channel, representative works of which are respectively summarised in Table 1 and Table 2. Both single and multi-channel techniques include the approaches that can be carried out in different ASR processing stages, i. e., the front-end, the back-end, or the jointed front- and back-end.

Thanks to the advance of deep learning, major achievements have been accomplished in the past four years, as shown in the literature. However, an obvious performance gap still remains between the enhanced system and the one evaluated in the clean environment. Thus, further efforts are still required for speech recognition to overcome the adverse effect of environmental noise [8, 6, 9].

We hope that this review could provide researchers and developers an opportunity to stand on the frontier of development in this field and to make greater breakthroughs.

Acknowledgements

This work was supported by the 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, 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, 10 pages.

[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] 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.

[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] 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*, 2016, in press.

[10] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "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.

[11] 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.

[12] J. T. 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, Jun 2014.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. 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.

[20] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, Apr 1995.

[21] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Berlin, Germany: Springer Science & Business Media, 2012, vol. 201.

[22] L. Deng, "Front-end, back-end, and hybrid techniques for noise-robust speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data*. Berlin/Heidelburg, Germany: Springer, 2011, pp. 67–99.

[23] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*. Hoboken, NJ: John Wiley & Sons, 2012.

[24] 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.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[27] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, Jan 2006.

[28] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle *et al.*, "Greedy layer-wise training of deep networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2006, pp. 153–160.

[29] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul 2006.

[30] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, vol. 11, pp. 625–660, Mar 2010.

[31] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. International Conference on Machine Learning (ICML)*, Helsinki, Finland, 2008, pp. 1096–1103.

[32] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, Dec 2010.

[33] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," in *Proc. International Conference on Artificial Intelligence and Statistics Conference (AISTATS)*, Clearwater Beach, FL, 2009, pp. 448–455.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 6645–6649.

[36] 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.

[37] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, Nov 1968.

[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[39] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, Apr 2000.

[40] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech quality*. Prentice-Hall, 1988.

[41] J. Hansen and B. 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.

[42] 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.

[43] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul 2006.

[44] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, UT, 2001, pp. 749–752.

[45] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. part ii – psychoacoustic model," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, Oct 2002.

[46] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[47] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul 2001.

[48] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on speech and audio processing*, vol. 11, no. 5, pp. 466–475, Sep 2003.

[49] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[50] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Washington, D.C., 1979, pp. 208–211.

[51] 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.

[52] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[53] 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.

[54] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, GA, 1996, pp. 629–632.

[55] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 59–67, Jan 2004.

[56] 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.

[57] ——, "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.

[58] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 2, pp. 282–305, Feb 2012.

[59] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transaction on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, Apr 1994.

[60] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010.

[61] 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.

[62] B. Schuller, F. Weninger, M. Wllmer, 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.

[63] J. T. 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.

[64] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, Sep 2011.

[65] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. International Conference on Independent Component Analysis and Signal Separation (ICA)*, London, UK, 2007, pp. 414–421.

[66] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proc. International Conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel, 2012, pp. 322–329.

[67] 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.

[68] 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.

[69] Y. Xu, J. Du, L. R. Dai, and C. H. 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.

[70] 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.

[71] Y. Xu, J. Du, L. R. Dai, and C. H. 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.

[72] B. Xia and C. Bao, "Speech enhancement with weighted denoising auto-encoder." in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 3444–3448.

[73] 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.

[74] 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.

[75] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[76] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, Aug 2013.

[77] M. Wollmer, 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.

[78] 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.

[79] 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, Jul 2014.

[80] F. Weninger, S. Watanabe, J. Le Roux, J. Hershey, Y. Tachioka, 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.

[81] 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.

[82] 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, 3593–3597.

[83] S.-Y. Chang and N. Morgan, "Robust CNN-based speech recognition with gabor filter kernels." in *Proc. INTERSPEECH*, Singapore, 2014, pp. 905–909.

[84] D. Wang, *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Boston, MA: Springer US, 2005, pp. 181–197.

[85] 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.

[86] 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, Jul 2013.

[87] 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.

[88] 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.

[89] 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.

[90] F. Weninger, J. R. Hershey, J. Le Roux, and B. 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.

[91] 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.

[92] 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.

[93] P. S. 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar 2016.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] J. T. 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.

[105] Z. Chen, S. Watanabe, H. Erdoğan, 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.

[106] 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.

[107] Z. Q. 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.

[108] 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.

[109] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, 2015, pp. 436–443.

[110] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe, "The MERL / SRI system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, 2015, pp. 475–481.

[111] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The USTC-iFlytek system for CHiME-4 challenge," in *Proc. 4th International Workshop on Speech Processing in Everyday Environments (CHiME)*, San Francisco, CA, 2016, pp. 36–38.

[112] T. Menne, J. Heymann, A. Alexandridis, K. Irie, A. Zeyer, M. Kitza, P. Golik, I. Kulikov, L. Drude, R. Schlüter, 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.

[113] B. D. V. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr 1988.

[114] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, July 1996.

[115] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[116] I. A. 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.

[117] 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.

[118] 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.

[119] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New York, NY, USA, 1988, pp. 2578–2581.

[120] S. Lefkimmiatis and P. Maragos, "A generalized estimation approach for linear and nonlinear microphone array post-filters," *Speech Communication*, vol. 49, no. 7, pp. 657–666, Aug 2007.

[121] A. Khabbazibasmenj, 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, Jun 2012.

[122] 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, Jul 2007.

[123] 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.

[124] 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.

[125] ——, "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.

[126] S. Sivasankaran, A. A. Nugraha, E. Vincent, J. A. Morales-Cordovilla, S. Dalmia, I. Illina, and A. Liutkus, "Robust ASR using neural network based speech enhancement and feature simulation," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, 2015, pp. 482–489.

[127] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 2814–2818.

[128] 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.

[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] 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.

[131] 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.

[132] 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.

[133] 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.

[134] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform CLDNNs," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5075–5079.

[135] 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.

[136] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sep 2010.

[137] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5210–5214.