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

ZIXING ZHANG, Imperial College London
JÜRGEN GEIGER, Huawei Technologies Duesseldorf GmbH
JOUNI POHJALAINEN and AMR EL-DESOKY MOUSA, University of Passau
WENYU JIN, Huawei Technologies Duesseldorf GmbH
BJÖRN SCHULLER, Imperial College London

Eliminating the negative effect of non-stationary environmental noise is a long-standing research topic for automatic speech recognition but still remains an important challenge. Data-driven supervised approaches, especially the 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. In the meanwhile, we discuss the pros and cons of these approaches and provide their experimental results on benchmark databases. We expect that this overview can facilitate the development of the robustness of speech recognition systems in acoustic noisy environments.

CCS Concepts: • **Computing methodologies** → **Speech recognition**; • **Human-centered computing** → **Human computer interaction (HCI)**;

Additional Key Words and Phrases: Robust speech recognition, deep learning, neural networks, non-stationary noise, multi-channel speech recognition

1 INTRODUCTION

Recently, following years of research, Automatic Speech Recognition (ASR) has achieved major breakthroughs and greatly improved performance (Amodei et al. 2016; Saon et al. 2016; Xiong et al. 2016). 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 (Loizou 2013). 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 (Barker et al. 2013, 2015; Kinoshita et al. 2016; Vincent et al. 2016; Yang and Chen 2012; Yoshioka et al. 2012). 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 (Barker et al. 2013, 2015; Kinoshita et al. 2016; Vincent et al. 2013).

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 (Maas et al. 2012). 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 (Geiger et al. 2014c; Qian et al. 2016; Wöllmer et al. 2010a, 2010b). 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 (Zhang et al. 2017), and has achieved considerable success in a variety of fields, such as gaming (Mnih et al. 2015), visual recognition (Liu et al. 2017; Russakovsky et al. 2015), language translation (Wu et al. 2016), music information retrieval (Schedl et al. 2016), and ASR (Dahl et al. 2012; Hinton et al. 2012). 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, and (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 Figure 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 Li et al. (2014) and Goodfellow et al. (2016), respectively. Note that, in this overview, the term *deep neural networks* refers to networks including multiple hidden layers.
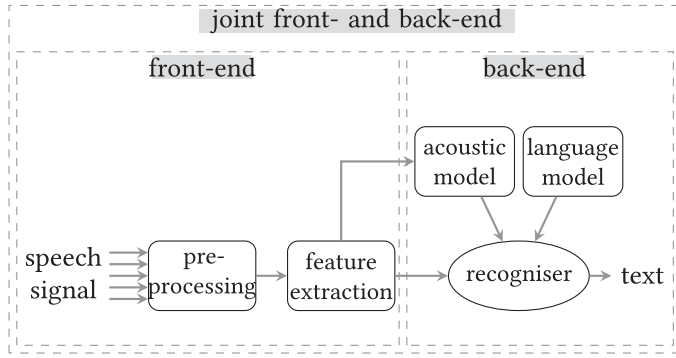
Fig. 1. General framework of a speech recognition system divided into front-end and back-end.

Whilst several related surveys on environmentally robust speech recognition are available in the literature (e.g., Acero (2012), Deng (2011), Gong (1995), Li et al. (2014), Virtanen et al. (2012), and Yoshioka et al. (2012)), 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 2, we briefly introduce the background of this overview. 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 promising multi-channel algorithms, before drawing our conclusions in Section 7.

2 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.

2.1 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

$$y(t) = s(t) * r(t) + n(t). \tag{1}$$

When applying Short-Time Discrete Fourier Transform (STDFT) 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, then $r(t)$ 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 1,000ms 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 1s (T_{60}) and a feature frame is extracted at every 10ms, one RIR would smear across the following 100 frames. Therefore, this distorted speech in the amplitude *spectral* domain, can be formulated by

(see Avargel and Cohen (2007) for more details):

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

with an assumption that $r(t)$ is a constant function. Particularly, $R(d, f)$ denotes the part of $R(f)$ (i.e., STDFT 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 Equation (2) can be approximated as

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

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

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

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

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

From Equations (1) to (5), it can be found that the clean speech and the mixed/noisy speech have a highly *non-linear* correlation in 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 article) 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 an abrupt sound like thunder and barks, side talking, and also music and driving noise. All these noises are almost unpredictable.

2.2 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.,

$$w = f(\mathbf{y}),$$ (7)

where $\mathbf{y}$ 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,

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

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

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

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* (Boll 1979) 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* (Loizou 2013) adopts stochastic models and is often implemented in practice using iterative approaches that base new estimates of the filter on the enhanced signal obtained by the previous iteration's estimate (Hansen and Clements 1991). Another popular family of techniques comprises the *Minimum Mean Square Error (MMSE)* (Ephraim and Malah 1984) and *log-spectral amplitude MMSE (Log-MMSE)* Short-Time Spectral Amplitude (STSA) estimators (Ephraim and Malah 1985). Despite that they are able to yield lower musical noise, a tradeoff 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 (Xu et al. 2015).

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 (Xu et al. 2015). Although some supervised machine-learning approaches have been proposed, such as *Non-negative Matrix Factorisation* (NMF) (Geiger et al. 2014a; Lee and Seung 1999; Schuller et al. 2010; Weninger et al. 2012), 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 (Goodfellow et al. 2016). 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 conected* layer is also know 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 (Dahl et al. 2012) via a greedy layerwise unsupervised pre-training strategy (Hinton and Salakhutdinov 2006).

Convolutional layer is a biologically inspired variant of fully connected layer originally developed for visual perception tasks (LeCun et al. 1989) and is the elementary layer to construct Convolutional Neural Networks (CNN). It employs a small size of two-dimensional (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 Zeiler and Fergus (2014) 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 a 1D kernel and directly applied to raw signals. Recent work has shown that the convolutional layer can automatically learn fundamental frequencies from raw signals (Trigeorgis et al. 2016).

In contrast to the aforementioned feed-forward layers (i.e., fully connected layer and convolutional layer), a *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, the Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) unit and, most recently, Gated Recurrent Unit (GRU) (Cho et al. 2014) 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) (Gauvain and Lee 1994) estimation and Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland 1995). 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 Li et al. (2014).

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; and (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 learned from massive noisy data.

2.3 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 1) 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 (Moreno et al. 2000), which 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 (Pearce and Picone 2002), which was constructed by corrupting the Wall Street Journal (WSJ0) corpus and designing it for Large Vocabulary Continuous Speech Recognition (LVCSR). All Aurora databases were mainly corrupted by additive noise, except Aurora 5 (Hirsch and Finster 2005), which 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 the CHiME series. All these CHiME databases (from the first to fourth) involve not only additive noise by adding

Table 1. General Description of Some Standard Evaluation Corpora for Environmentally Robust Speech Recognition

Dataset	based on	environments	sim./real	noise types	channels
Aurora-2 (Pearce and Hirsch 2000)	TIDigits	eight conditions	sim.	add. (mainly)	single
Aurora-3 (Moreno et al. 2000)	TIDigits	car	real	add. (mainly)	single
Aurora-4 (Pearce and Picone 2002)	WSJ0	str/tra/car/bab/res/air	sim.	add. (mainly)	dual
Aurora-5 (Hirsch and Finster 2005)	TIDigits	rooms and cars	sim.	add. & con.	single
CHiME-1 (Barker et al. 2013)	Grid, WSJ0	home	sim	add. & con.	dual
CHiME-2 (Vincent et al. 2013)	Grid, WSJ0	home	sim	add. & con.	dual
CHiME-3 (Barker et al. 2015)	WSJ0	bth/bus/caf/ped/str	real & sim	add. & con.	six
CHiME-4 (Vincent et al. 2016)	WSJ0	bth/bus/caf/ped/str	real & sim	add. & con.	six
REVERB (Kinoshita et al. 2016)	WSJCAM0	ambient noise	real & sim	add. & con.	eight
AMI (Carletta et al. 2005)	-	meeting	real	con. (mainly)	four/eight
Voice Search (Schalkwyk et al. 2010)	-	voice search	sim.	add. & con.	dual

These corpora are either *real*istically recorded or artificially *sim*ulated based on certain clean databases. *Add*itive and/or *con*volutional noises are collected in various environments.

various ambient noises but also convolutional noise. More specifically, the first and second CHiME databases (Barker et al. 2013; Vincent et al. 2013) 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 third and fourth CHiME databases (Barker et al. 2015; Vincent et al. 2016) only include the data for LVCSR. Moreover, the third and fourth 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 (Kinoshita et al. 2016), AMI (Carletta et al. 2005), and Voice Search (Schalkwyk et al. 2010). 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 vocabularies, 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) (Hansen and Pellom 1998; Quackenbush et al. 1988), *distance measures*, *Source-to-Distortion Ratio* (SDR) (Vincent et al. 2006), and *Perceptual Evaluation of Speech Quality* (PESQ) (P.862 2001). More detailed definitions and explanations of these objective metrics can be found in Hu and Loizou (2008). 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 Weninger et al. (2015) 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.

3 FRONT-END TECHNIQUES

From Section 3 to 5, we review some key techniques, which are concisely summarised and compared in Table 2. 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 2.1 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-cepstral* 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 **y** that is the representation extracted from noisy speech and the target as **x**. Based on how the training target **x** is obtained, the techniques can be categorised into (i) *mapping-based* methods, where **x** is the representation, straightforwardly extracted from clean speech, or (ii) *masking-based* methods, where **x** is a mask calculated between clean and noisy speech.

3.1 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

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

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 **x** (extracted from the clean speech $s(t)$) from the corresponding input features **y** (extracted from the corrupted speech $y(t)$). The parameters of neural networks (models) θ are determined by minimising the objective function of the Mean Squared Error (MSE):

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

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.

3.1.1 Based on Stacked AutoEncoder or Deep Bolzmann Machine. Specifically, in 2013, a Stacked AutoEncoder (SAE) was employed to map noisy speech to clean speech in the Mel-spectral domain (Lu et al. 2013). Given an AutoEncoder (AE) that includes one non-linear encoding stage and

one linear decoding stage for real valued speech as

$$h(\mathbf{y}) = g(\mathbf{W}_1\mathbf{y} + \mathbf{b})$$
$$\hat{\mathbf{x}} = \mathbf{W}_2 h(\mathbf{y}) + \mathbf{b},$$

(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 (Lu et al. 2013).

Analogously to this, another successful work has been shown in Xu et al. (2014b), where a Deep Bolzmann Machine (DBM) was utilised to estimate the complex mapping function. In the pre-training stage, noisy speech was used to train Restricted Bolzmann Machines (RBMs) layer by layer in a standard unsupervised greedy fashion to obtain a deep generative model (Hinton and Salakhutdinov 2006); whereas, in the fine-tuning process, the desired clean speech was set as the target by minimising the objective function as Equation (22). Similar research efforts were also extensively made on the log magnitude (Han et al. 2015) and the log-Mel-spectral domains (Xu et al. 2015), 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 Xia and Bao (2013) and showed positive performance for denoising. In detail, a weighted reconstruction loss function is employed to train SAE on the power spectrum as

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

(13)

where λ_w is a weight for the wth frequency band.

Further, related approaches were also shown in Ishii et al. (2013) and Feng et al. (2014), 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.

3.1.2 Based on LSTM-RNN. For sequence-based pattern recognition, context information is considered to be vitally important (Hochreiter and Schmidhuber 1997). 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 (Xu et al. 2014b). RNNs, especially the LSTM-RNNs, have been frequently demonstrated to be highly capable of capturing the context information in a long sequence (Graves 2013; Wöllmer et al. 2010a).

In this light, Maas et al. (2012) 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. (2013) further proposed to use LSTM-RNNs to handle highly non-stationary additive noise, which was then extended to coping with reverberation in Weninger et al. (2013, 2014a, 2014c) and Zhang et al. (2014, 2016). With the help of LSTM-RNN, the speech recognition systems perform much better than the ones without LSTM-RNN when decoding noisy speech (Weninger et al. 2013, 2014a, 2014c; Zhang et al. 2014).

3.1.3   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 (Amodei et al. 2016; Sainath et al. 2015). For image restoration and further image processing tasks, *deep convolutional encoder-decoder* networks were proposed in Mao et al. (2016) and delivered promising performance. This network was further introduced for speech enhancement (Park and Lee 2016), where the time-frequency spectrum (spectrogram) is viewed as an image. Specifically, the encoder network includes multiple convolutional layers 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. 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 (Oord et al. 2016), 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 Qian et al. (2017) and Rethage et al. (2017). Particularly in Qian et al. (2017), 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

$$\min_{G} \max_{D} V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log(D(\mathbf{x}))] + \mathbb{E}_{\hat{\mathbf{x}} \sim p_{data}(\hat{\mathbf{x}})}[\log(1 - D(G(\mathbf{y})))]. \qquad (14)$$

The adversarial training strategy was examined in Pascual et al. (2017) and Michelsanti and Tan (2017) and was found to perform superior to other traditional approaches, such as Wiener filtering.

3.1.4   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 that 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 (Oord et al. 2016)) that dramatically reduce the computational load, and (iii) the development of cloud computing that makes it possible to handle such a task.

They also reflect another trend relating to the network training strategy, which starts to shift from a traditional way with a single network to an adversarial way with two networks (Goodfellow et al. 2014). 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 (Creswell et al. 2017), 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 that spread along frequency axis while local correlation may be weaker. Therefore, more efforts are required towards this direction.

3.2 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,

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

3.2.1 Masks. Two most commonly used masks in the literature include *binary*-based masks (Wang 2005) and *ratio*-based masks (Srinivasan et al. 2006). Typical binary-based masks often refer 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,

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

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 masks often indicate 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:

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

where $S(n, f)$ and $N(n, f)$ are the magnitudes of clean speech and noise in the T-F domain, respectively, and α is a warping factor of the magnitudes 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 Equations (16) and (17), it can be seen that IRM-based approaches could deliver a less distorted enhanced speech, while it could potentially involves much interference (Grais et al. 2016).

Wang and Wang (2013) first introduced DNNs to perform IBM estimation for speech separation, and reported large performance improvement over non-DNN-based methods. Subsequently, Wang et al. (2014) 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 Narayanan and Wang (2013), 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. (2016) 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 (Narayanan and Wang 2013; Weninger et al. 2014b) and log-Mel-frequency domain (Weninger et al. 2014) that have frequently been proven to be effective for ASR in deep learning. The experimental results in Weninger

et al. (2014b) 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 (Weninger et al. 2014, 2014b, 2015), 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 (Weninger et al. 2014b) 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 (Paliwal et al. 2011). For this reason, Erdogan et al. (2015) proposed a *Phase-Sensitive Mask* (PSM) that is calculated by

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

where θ is the difference between the clean speech phase θ^s and the noisy speech phase θ^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 (2017b) further developed this approach, naming it *complex IRM* (cIRM). It is defined as

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

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 Williamson and Wang (2017b).

3.2.2 Objective Functions and Training Strategies. In the neural network training stage, given the input **y** from the T-F domain of mixed noisy signals $Y(n, f)$ and the target **x** from the calculated T-F mask $M(n, f)$, the parameters of neural networks θ 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:

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

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 **y** by

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

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 (Huang et al. 2014, 2015; Weninger et al. 2014b). 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

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

This is indeed similar to the objective function used for the mapping-based methods (cf. Section 3.1). Employing the SA-based objective function was empirically examined to perform better than the MA-based one for source separation (Weninger et al. 2014b). Furthermore, the conclusions found in Weninger et al. (2014b) and Wang and Wang (2015) 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 previously mentioned, Weninger et al. (2015) 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. Equation (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 (Weninger et al. 2015).

Additionally, a multi-task learning framework was proposed in Huang et al. (2014, 2015) 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 (Huang et al. 2014).

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 (Erdogan et al. 2015; Weninger et al. 2014, 2015).

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 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 a previously trained Acoustic Model (AM).

Early works focus on the improvement of the model structure 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 Sharma et al. (2000). 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 (Geiger et al. 2014c) 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 framewise phone predictions. The double-stream emission probability is computed as

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

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

Table 2. A Summary of Representative *Single-Channel* Approaches Based on Deep Learning for Environmentally Robust Speech Recognition

stage	approaches	typical publications	advantages	disadvantages
front (mapping-based)	MFCC	(Feng et al. 2014; Maas et al. 2012; Weninger et al. 2013; Wöllmer et al. 2013; Zhang et al. 2014)	low dimension, require less computational load	lose much information, (almost) irreversible to raw signals
	(log) Mel	(Ishii et al. 2013; Lu et al. 2013; Weninger et al. 2014a, 2014c)	low dimension, require less computational load	lose some information, (almost) irreversible to raw signals
	(log/power) mag.	(Han et al. 2015; Park and Lee 2016; Xu et al. 2014c, 2015)	invertible to the audio signal	high dimension, require high computational load, each element is equally important
	temporal	(Qian et al. 2017; Rethage et al. 2017)	retain the complete information	large data size, require heavy computational load
front (masking-based)	IBM	(Grais et al. 2016; Wang et al. 2014; Wang and Wang 2013)	little interference	much magnitude distortion
	IRM	(Grais et al. 2016; Huang et al. 2015; Narayanan and Wang 2013; Wang et al. 2014; Weninger et al. 2014b)	little magnitude distortion	much interference
	PSM	(Erdogan et al. 2015)	less phase distortion and little interference	do not restore the complete phase information
	cIRM	(Williamson and Wang 2017a)	learn a complete relationship between noisy and clean speech for both magnitude and phase	relative complex to compute
	MA	(Narayanan and Wang 2013; Wang and Wang 2013; Weninger et al. 2014b)	most straightforward way	no enhanced phase
	SA	(Erdogan et al. 2015; Grais et al. 2016; Huang et al. 2015)	directly optimise the objective	no enhanced phase
	PSA	(Weninger et al. 2015)	considered phase information when predicting mask	
back	tandem double-stream	(Sharma et al. 2000; Wöllmer et al. 2009) (Geiger et al. 2014c; Weninger et al. 2014c)	explicitly make use of discriminative features	HMM model dependent
	hybrid	(Geiger et al. 2014d)		
	multi-condit. train.	(Seltzer et al. 2013; Wang and Wang 2013)	straightforward and very efficient	require many data in different noisy scenarios
	model adapt.	(Mirsamadi and Hansen 2015)	flexible to different noisy environments	require a relatively large amount of adaptation data, otherwise easy overfitted
	NAT	(Karanasou et al. 2014; Seltzer et al. 2013; Yu et al. 2015)	easy to be implemented	require another disassociated model to estimate noise and cannot be optimised jointly
	dynamic NAT	(Xu et al. 2014a)	more efficient to deal with non-stationary noise	more efforts to estimate noise
	multi-task train.	(Chen et al. 2015; Giri et al. 2015)	exploit the clean-sensitive speech	could lose some discriminative features
joint	re-training	(Weninger et al. 2013, 2014a)	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	(Gao et al. 2015; Lee et al. 2016, 2017; Mimura et al. 2016; Ravanelli et al. 2017; Wang and Wang 2016)	exploit the complementary of enhancement networks and speech recognition networks	tricky to combine the two networks
	end-to-end	(Amodei et al. 2016; Qian et al. 2016)	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

Those methods are summarised at different ASR processing stages (*front*-end, *back*-end, and *joint* front- and back-end).

performance of the latter. In particular, when setting λ 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 tandem (Wöllmer et al. 2009), double-stream (Geiger et al. 2014c), or hybrid structures (Geiger et al. 2014d).

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 (Amodei et al. 2016). Therefore, HMM-independent approaches are more than necessary than ever before. The widely used approach comes to *multi-condition* training (Seltzer et al. 2013). 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 (Mirsamadi and Hansen 2015). Alternatively, a part of neural network parameters can be modified. For example, the authors of the work (Mirsamadi and Hansen 2015) 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 (Giri et al. 2015)) when training, which is often termed as *Noise-Aware Training* (NAT) (Seltzer et al. 2013). 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 to automatically learn the relationship between noisy speech and noise, which is beneficial to predict phonetic targets (Seltzer et al. 2013). 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* (Dehak et al. 2011), which were originally developed for speaker recognition. The i-vector can be calculated either from the hand-crafted features such as commonly used MFCCs (Karanasou et al. 2014) or from the automatically extracted bottleneck representations by DNN (Yu et al. 2015); and either from the raw noisy features or from the enhanced features, e.g., Vector Taylor Series (VTS) (Yu et al. 2015).

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 Xu et al. (2014a), where the authors used masking-based approaches (cf. Section 3.2) to estimate the noise that varies along time. This approach performs more efficient especially for non-stationary noise, whereas it 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 Giri et al. (2015) and Chen et al. (2015), 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

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

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.

5 JOINT FRONT- AND BACK-END TRAINING TECHNIQUES

Most research efforts on flighting 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 to employ the enhanced speech obtained in the front-end to *re-train* the pre-trained AM in the back-end (Weninger et al. 2013). This simply remains everything unchanged but a further re-training process on the AM.

A more sophisticated *joint* DNN structure was proposed in Lee et al. (2016, 2017), 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 (Lee et al. 2016). Then, join 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 Narayanan and Wang (2014) 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 all of the 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 Gao et al. (2015), Mimura et al. (2016), and Wang and Wang (2016).

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 Ravanelli et al. (2017). 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 shown great promise in the latest ASR systems (Amodei et al. 2016; Sainath et al. 2015). Its central idea is to jointly optimise the parameters of the networks at the front-end that automatically learn the inherent representations from low-level features/signals for the task at hand and the networks at the back-end that provide final predictions. For noisy speech recognition, a quite recent and well-developed framework has been reported in Qian et al. (2016), 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 (Qian et al. 2016): (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.

6 MULTI-CHANNEL TECHNIQUES

Microphone arrays and *multi-channel* processing techniques have recently played an increasingly significant role in the development of robust ASR (Barker et al. 2015; Kinoshita et al. 2016). 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* (Marro et al. 1998; McCowan and Bourlard 2003). 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 evolute 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 Van Veen and Buckley (1988)) 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 channels, as mentioned. The reviewed techniques are summarised and compared in Table 3.

6.1 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 (Barker et al. 2015; Vincent et al. 2016).

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 $Y(n) = S(n)\mathbf{d} + A(n)$, where the vector $Y(n) = (Y_1(n), \ldots, Y_M(n))^T$ 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 $A(n)$ 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 $Y(n)$ as $\mathbf{x}(n) = \mathbf{w}^H Y(n)$, 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 $\mathbf{R}_{VV} = (1/N) \sum_n \mathbf{Y}(n)\mathbf{Y}^H(n)$ (Mestre and Lagunas 2003). 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 Cox et al. (1987)

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

The MVDR beamformer is not robust against an inaccurately estimated steering vector $\mathbf{d}$ (Khabbazibasmenj et al. 2012). In contrast, GEV beamformer requires no DOA estimate and is based on maximising the output signal-to-noise ratio (Warsitz and Haeb-Umbach 2007). The beamformer filter coefficients for a given frequency bin are found as the principal eigenvector of a generalised eigenvalue problem as required by Warsitz and Haeb-Umbach (2007)

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

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 (Ochiai et al. 2017): (i) beamformers with a *mask estimation* network (Heymann et al. 2015, 2016a, 2016b; Menne et al. 2016) and (ii) beamformers with a *filter estimation* network (Li et al. 2016; Meng et al. 2017; Xiao et al. 2016a). Both approaches aim to obtain an enhanced signal based on the formalisation of the conventional filter-and-sum beamformer 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 3.2), which are then applied to calculate the spatial covariance matrix further followed by a calculation of filter coefficients by Equations (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 first investigated in Heymann et al. (2015), 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 Heymann et al. (2016a), 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 Heymann et al. (2016a) were competitive with the ones in previous work (Heymann et al. 2015). 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., $\mathbf{d} = \mathcal{P}(\mathbf{R}_{VV})$.

Table 3. A Summary of Representative *Multi-channel* Approaches Based on Deep Learning for Environmentally Robust Speech Recognition

stage	approaches	typical publications	advantages	disadvantages
front	mask estimation	(Erdogan et al. 2016; Heymann et al. 2015, 2016a; Menne et al. 2016)	avoid relying on a DOA estimation	require large-scale training data
	filter coefficients estimation	(Li et al. 2016; Meng et al. 2017; Xiao et al. 2016a)	easily to be integrated with DNN AM as a joint network	based on the simulated data in all possible scenarios
	post-filter estimation	(Pertilä and Nikunen 2014)	do not require explicit estimates of the signal and noise statistics	require large-scale simulated data
joint	channel concatenation	(Liu et al. 2014; Swietojanski et al. 2013)	require no knowledge of microphone array geometry and signal information	unclear on a severe mismatch among multiple channels
	cross-channel max-pooling	(Swietojanski et al. 2014)	able to pick the most informative channel	unable to make use of the spatial information found in multi-channel signals
	factoring spatial & spectral filtering	(Sainath et al. 2017)	robust to varying target speaker direction of arrival	additional computational cost
	end-to-end	(Hoshen et al. 2015; Ochiai et al. 2017)	automatically extracted the underlying and salient representations over multiple channels	heavy parameters tuning and computational load

Those nethods are summarised at different ASR processing stages (*front*-end and *joint* front- and back-end).

The effectiveness of all these mentioned approaches has been demonstrated in the fourth CHiME Challenge (Heymann et al. 2016b; Menne et al. 2016).

A typical filter coefficients-based approach was evaluated in Xiao et al. (2016a), 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 (Xiao et al. 2016a). A similar investigation was also shown in Li et al. (2016) and Meng et al. (2017).

As for post-filtering, very few recent articles 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 (Pertilä and Nikunen 2014).

6.2 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 (Hoshen et al. 2015; Liu et al. 2014; Swietojanski et al. 2014). In Swietojanski et al. (2013), 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 (Swietojanski et al. 2013) and was further verified in Liu et al. (2014).

A more sophisticated approach was proposed in Swietojanski et al. (2014). 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

Table 4. Benchmarks for Four Selected Standard Corpora (i.e., Aurora-4, CHiME-2, CHiME-4, and AMI)

publications	single channel front-end	single channel back-end	single channel joint	multiple channels front	multiple channels joint	model	WER (or SDR)
Aurora-4							
(Narayanan and Wang 2013)	IBM/IRM (MA; Mel)	MCT				DNN	16.50%
(Seltzer et al. 2013)		NAT; MCT; re-training				DNN	12.40%
(Kundu et al. 2016)		NAT, MTL					8.80%
(Qian et al. 2016)			end-to-end			VDCNN	8.81%
CHiME-2							
(Weninger et al. 2013)	mapping: MFCC	Re-training				BLSTM	26.73%
(Geiger et al. 2014b)		multi-stream,re-training				BLSTM	41.42%
(Weninger et al. 2014b)	masking: IRM (MA, SA; Mel)					LSTM	17.68 (SDR)
(Geiger et al. 2014d)		hybrid				BLSTM	22.20%
(Weninger et al. 2014a)	mapping: log Mel	re-training				BLSTM	22.16%
(Erdogan et al. 2015)	masking: PSM (SA; log Mel)					BLSTM	14.76 (SDR)
(Han et al. 2015)	mapping: log mag	hybrid				DNN	≈25%
(Chen et al. 2015)	masking: IRM (SA; log Mel)	MTL; hybrid				BLSTM	16.04%
(Narayanan and Wang 2015)	masking: IRM (MA; Mel)		joint			DNN	15.40%
(Weninger et al. 2015)	masking: IRM (PSA; Mel)					BLSTM	13.76%
(Wang and Wang 2016)	masking: IRM (power spec.)	multi-stream; model adapt.; MCT	joint			DNN	10.63%
CHiME-4[a]							
(Menne et al. 2016)				mask est.		BLSTM	//; 4.0%,5.2%
(Heymann et al. 2016b)				mask est.		WRN & BLSTM	1.7%,9.9%; 3.1%,3.9%
(Xiao et al. 2016b)				filter coeff. est.; mask est.		LSTM	21.4%,20.9%; 5.0%,6.4%
(Erdogan et al. 2016)		re-training; hybrid		mask est.		BLSTM	//; 3.4%,4.4%
(Qian and Tan 2016)		NAT			end-to-end	VDCNN & LSTM	12.9%,13.9%; 6.3%,6.4%
AMI[b]							
(Swietojanski et al. 2013)		MCT			channel concatenation	DNN-HMM	57.30%
(Swietojanski et al. 2014)					cross-channel max-pooling	CNN	49.40%
(Liu et al. 2014)					channel concatenation	DNN	44.80%

(Continued)

Table 4. Continued

| publications | single channel | | | multiple channels | | | WER |
	front-end	back-end	joint	front	joint	model	(or SDR)
(Qian et al. 2016)			end-to-end			VDCNN	46.90%
(Xiao et al. 2016a)				filter coeff. est.	end-to-end	DNN	42.20%
(Ochiai et al. 2017)				mask est.	end-to-end		39.00%

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 (Zagoruyko and Komodakis 2016); VDCNN: Very Deep CNN.
[a] results are provided for the one and six channel signals (separated by ';') on the simmulated and real subsets of the evaluation dataset (separated by ',');
[b] only the results for MDM subset are provided.

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 (Swietojanski et al. 2014).

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 (Hoshen et al. 2015). 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 (Hoshen et al. 2015). A follow-up work was reported in Sainath et al. (2017), 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. 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 (Sainath et al. 2017).

7 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 1, 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 4.

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 that can optimise all processing stages simultaneously; and (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 flighting 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 accomplished 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 (Barker et al. 2015; Kinoshita et al. 2016; Vincent et al. 2016). 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.

REFERENCES

Alex Acero. 2012. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Vol. 201. Springer Science & Business Media, Berlin.

Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, and others. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of the International Conference on Machine Learning (ICML'16)*. New York, NY. 173–182.

Yekutiel Avargel and Israel Cohen. 2007. System identification in the short-time fourier transform domain with crossband filtering. *IEEE Trans. Audio Speech Lang. Process.* 15, 4 (Mar. 2007), 1305–1319.

Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. 2015. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'15)*. 504–511.

Jon Barker, Emmanuel Vincent, Ning Ma, Heidi Christensen, and Phil Green. 2013. The PASCAL CHiME speech separation and recognition challenge. *Comput. Speech Lang*. 27, 3 (May 2013), 621–633.

Steven Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Sign. Process.* 27, 2 (Apr. 1979), 113–120.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, and others. 2005. The AMI meeting corpus: A pre-announcement. In *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*. 28–39.

Zhuo Chen, Shinji Watanabe, Hakan Erdoğan, and John R. Hershey. 2015. Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'15)*. Dresden, Germany, 1–5.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of the Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST'14)*. 103–111.

Henry Cox, Robert M. Zeskind, and Mark M. Owen. 1987. Robust adaptive beamforming. *IEEE Trans. Acoust. Speech Sign. Process.* 35, 10 (Oct. 1987), 1365–1376.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. 2017. Generative adversarial networks: An overview (submitted for publication).

George E. Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Process.* 20, 1 (Jan. 2012), 30–42.

Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* 19, 4 (May 2011), 788–798.

Li Deng. 2011. Front-end, back-end, and hybrid techniques for noise-robust speech recognition. In *Robust Speech Recognition of Uncertain or Missing Data*. Springer, Berlin, 67–99.

Yariv Ephraim and David Malah. 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Sign. Process.* 32, 6 (Dec. 1984), 1109–1121.

Yariv Ephraim and David Malah. 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Sign. Process.* 23, 2 (Apr. 1985), 443–445.

Hakan Erdogan, Tomoki Hayashi, John R. Hershey, Takaaki Hori, Chiori Hori, Wei-Ning Hsu, Suyoun Kim, Jonathan Le Roux, Zhong Meng, and Shinji Watanabe. 2016. Multi-channel speech recognition: LSTMs all the way through. In *Proceedings of the CHiME-4 Workshop*.

Hakan Erdogan, John R. Hershey, Shinji Watanabe, and Jonathan Le Roux. 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 708–712.

Hakan Erdogan, John R. Hershey, Shinji Watanabe, Michael I. Mandel, and Jonathan Le Roux. 2016. Improved MVDR beamforming using single-channel mask prediction networks. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'16)*. 1981–1985.

Xue Feng, Yaodong Zhang, and James Glass. 2014. Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 1759–1763.

Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2015. Joint training of front-end and back-end deep neural networks for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 4375–4379.

J.-L. Gauvain and Chin-Hui Lee. 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech Aud. Process.* 2, 2 (Apr. 1994), 291–298.

Jürgen Geiger, Jort F. Gemmeke, Björn Schuller, and Gerhard Rigoll. 2014a. Investigating NMF speech enhancement for neural network based acoustic models. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'14)*. 2405–2409.

Jürgen Geiger, Erik Marchi, Felix Weninger, Björn Schuller, and Gerhard Rigoll. 2014b. The TUM system for the REVERB challenge: Recognition of reverberated speech using multi-channel correlation shaping dereverberation and BLSTM recurrent neural networks. In *Proceedings of the REVERB Workshop, Held in Conjunction with ICASSP 2014 and HSCMA 2014*. 1–8.

Jürgen Geiger, Felix Weninger, Jort F. Gemmeke, Martin Wöllmer, Björn Schuller, and Gerhard Rigoll. 2014c. Memory-enhanced neural networks and NMF for robust ASR. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 6 (June 2014), 1037–1046.

Jürgen Geiger, Zixing Zhang, Felix Weninger, Björn Schuller, and Gerhard Rigoll. 2014d. Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'14)*. 631–635.

Ritwik Giri, Michael L. Seltzer, Jasha Droppo, and Dong Yu. 2015. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 5014–5018.

Yifan Gong. 1995. Speech recognition in noisy environments: A survey. *Speech Commun.* 16, 3 (Apr. 1995), 261–291.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS'14)*. 2672–2680.

E. M. G. Grais, Gerard Roma, Andrew J. R. Simpson, and Mark D. Plumbley. 2016. Combining mask estimates for single channel audio source separation using deep neural networks. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'16)*. 3339–3343.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv:1308.0850* (Aug. 2013).

Kun Han, Yuxuan Wang, DeLiang Wang, William S. Woods, Ivo Merks, and Tao Zhang. 2015. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* 23, 6 (Apr. 2015), 982–992.

John H. L. Hansen and Mark A. Clements. 1991. Constrained iterative speech enhancement with application to speech recognition. *IEEE Trans. Sign. Process.* 39, 4 (Apr. 1991), 795–805.

John H. L. Hansen and Bryan L. Pellom. 1998. An effective quality evaluation protocol for speech enhancement algorithms. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'98)*. 2819–2822.

Jahn Heymann, Lukas Drude, Aleksej Chinaev, and Reinhold Haeb-Umbach. 2015. BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'15)*. 444–451.

Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach. 2016a. Neural network based spectral mask estimation for acoustic beamforming. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*. 196–200.

Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach. 2016b. Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition. In *Proceedings of the 4th International Workshop on Speech Processing in Everyday Environments (CHiME'16)*. 12–17.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Sign. Process. Mag.* 29, 6 (Nov. 2012), 82–97.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (July 2006), 504–507.

Hans Günter Hirsch and Harald Finster. 2005. The simulation of realistic acoustic input scenarios for speech recognition systems. In *Proceedings of the Conference of the International Speech Communications Association (INTERSPEECH)*. Lisbon, Portugal, 2697–2700.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neur. Comput.* 9, 8 (Nov. 1997), 1735–1780.

Yedid Hoshen, Ron J. Weiss, and Kevin W. Wilson. 2015. Speech acoustic modeling from raw multichannel waveforms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 4624–4628.

Yi Hu and Philipos C. Loizou. 2008. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio Speech Lang. Process.* 16, 1 (Jan. 2008), 229–238.

Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. 2014. Deep learning for monaural speech separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 1562–1566.

Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. 2015. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 23, 12 (Dec. 2015), 2136–2147.

Takaaki Ishii, Hiroki Komiyama, Takahiro Shinozaki, Yasuo Horiuchi, and Shingo Kuroiwa. 2013. Reverberant speech recognition based on denoising autoencoder. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'13)*. 3512–3516.

Penny Karanasou, Yongqiang Wang, Mark J. F. Gales, and Philip C. Woodland. 2014. Adaptation of deep neural network acoustic models using factorised i-vectors. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'14)*. 2180–2184.

Arash Khabbazibasmenj, Sergiy A. Vorobyov, and Aboulnasr Hassanien. 2012. Robust adaptive beamforming based on steering vector estimation with as little as possible prior information. *IEEE Trans. Sign. Process.* 60, 6 (June 2012), 2974–2987.

Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A. P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, and others. 2016. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP J. Adv. Sign. Process.* 2016, 1 (Dec. 2016), 1–19.

Souvik Kundu, Gautam Mantena, Yanmin Qian, Tian Tan, Marc Delcroix, and Khe Chai Sim. 2016. Joint acoustic factor learning for robust deep neural network based automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*. 5025–5029.

Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neur. Comput.* 1, 4 (1989), 541–551.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (Oct. 1999), 788–791.

Kang Hyun Lee, Shin Jae Kang, Woo Hyun Kang, and Nam Soo Kim. 2016. Two-stage noise aware training using asymmetric deep denoising autoencoder. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*. 5765–5769.

Kang Hyun Lee, Woo Hyun Kang, Tae Gyoon Kang, and Nam Soo Kim. 2017. Integrated DNN-based model adaptation technique for noise-robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. 5245–5249.

Christopher J. Leggetter and Philip C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Comput. Speech Lang.* 9, 2 (Apr. 1995), 171–185.

Bo Li, Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, and Michiel Bacchiani. 2016. Neural network adaptive beamforming for robust multichannel speech recognition. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'16)*. 1976–1980.

Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech. Lang. Proces.* 22, 4 (Apr. 2014), 745–777.

Yan Liu, Yang Liu, Shenghua Zhong, and Songtao Wu. 2017. Implicit visual learning: Image recognition via dissipative learning model. *ACM Trans. Intell. Syst. Technol.* 8, 2 (Jan. 2017), 31:1–31:24.

Yulan Liu, Pengyuan Zhang, and Thomas Hain. 2014. Using neural network front-ends on far field multiple microphones based speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 5542–5546.

Philipos C. Loizou. 2013. *Speech Enhancement: Theory and Practice*. Taylor Francis, Abingdon, UK.

Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. 2013. Speech enhancement based on deep denoising autoencoder. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'13)*. 436–440.

Andrew L. Maas, Quoc V. Le, Tyler M. OÑeil, Oriol Vinyals, Patrick Nguyen, and Andrew Y. Ng. 2012. Recurrent neural networks for noise reduction in robust ASR. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'12)*. 22–25.

Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS'16)*. 2802–2810.

Claude Marro, Yannick Mahieux, and Klaus Uwe Simmer. 1998. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Trans. Speech Audio Process.* 6, 3 (May 1998), 240–259.

Iain McCowan and Herv'e Bourlard. 2003. Microphone array post-filter based on noise field coherence. *IEEE Trans. Speech Audio Process.* 11, 6 (Nov. 2003), 709–716.

Zhong Meng, Shinji Watanabe, John R. Hershey, and Hakan Erdogan. 2017. Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. 271–275.

Tobias Menne, Jahn Heymann, Anastasios Alexandridis, Kazuki Irie, Albert Zeyer, Markus Kitza, Pavel Golik, Kulikov Ilia, Lukas Durde, Ralf Schlater, Hermann Ney, Reinhold Haeb-Umbach, and Athanasios Mouchtaris. 2016. The RWTH /UPB/FORTH system combination for the 4th CHiME challenge evaluation. In *Proceedings of the 4th International Workshop on Speech Processing in Everyday Environments (CHiME'16)*. 49–51.

Xavier Mestre and Miguel Angel Lagunas. 2003. On diagonal loading for minimum variance beamformers. In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology*. 459–462.

Daniel Michelsanti and Zheng-Hua Tan. 2017. Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'17)*. 2008–2012.

Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2016. Joint optimization of denoising autoencoder and DNN acoustic model based on multi-target learning for noisy speech recognition. In *Proceedings of theConference of the International Speech Communication Association (INTERSPEECH'16)*. 3803–3807.

Seyedmahdad Mirsamadi and John H. L. Hansen. 2015. A study on deep neural network acoustic model adaptation for robust far-field speech recognition. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'15)*. 2430–2434.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, and others. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (Feb. 2015), 529–533.

Asunción Moreno, Børge Lindberg, Christoph Draxler, Gaël Richard, Khalid Choukri, Stephan Euler, and Jeffrey Allen. 2000. SPEECHDAT-CAR. A large speech database for automotive environments. In *Proceedings of the the 2nd International Conference on Language Resources and Evaluation (LREC'00)*.

Arun Narayanan and DeLiang Wang. 2013. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13)*. 7092–7096.

Arun Narayanan and DeLiang Wang. 2014. Joint noise adaptive training for robust automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 2504–2508.

Arun Narayanan and DeLiang Wang. 2015. Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training. *IEEE/ACM Trans. Audio Speech Lang. Process.* 23, 1 (Jan. 2015), 92–101.

Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, and John R. Hershey. 2017. Multichannel end-to-end speech recognition. In *Proceedings of the the 34th International Conference on Machine Learning (ICML'17)*. 2632–2641.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv:1609.03499* (Sep. 2016).

ITU-T Recommendation P.862. 2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. 2011. The importance of phase in speech enhancement. *Speech Commun.* 53, 4 (Apr. 2011), 465–494.

Se Rim Park and Jinwon Lee. 2016. A fully convolutional neural network for speech enhancement. *arXiv:1609.07132* (Sep. 2016).

Santiago Pascual, Antonio Bonafonte, and Joan Serrà. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv:1703.09452* (Mar. 2017).

David Pearce and Hans-Günter Hirsch. 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'00)*. 29–32.

David Pearce and J. Picone. 2002. *Aurora Working Group: DSR Front End LVCSR Evaluation AU/384/02*. Institute for Signal & Information Processing, Mississippi State University, Tech. Rep (2002).

Pasi Pertilä and Joonas Nikunen. 2014. Microphone array post-filtering using supervised machine learning for speech enhancement. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'14)*. 2675–2679.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson. 2017. Speech enhancement using Bayesian WaveNet. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'17)*. 2013–2017.

Yanmin Qian, Mengxiao Bi, Tian Tan, and Kai Yu. 2016. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Trans. Audio Speech. Lang. Process.* 24, 12 (Dec. 2016), 2263–2276.

Yanmin Qian and Tian Tan. 2016. The SJTU CHiME-4 system: Acoustic noise robustness for real single or multiple microphone scenarios. In *Proceedings of the CHiME-4 Workshop*.

Schuyler R. Quackenbush, Thomas Pinkney Barnwell, and Mark A. Clements. 1988. *Objective Measures of Speech Quality*. Prentice-Hall, Upper Saddle River, NJ.

Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio. 2017. A network of deep neural networks for distant speech recognition. In *Proceedings of the IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP'17)*. 4880–4884.

Dario Rethage, Jordi Pons, and Xavier Serra. 2017. A Wavenet for speech denoising. *arXiv:1706.07162* (June 2017).

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 3 (Dec. 2015), 211–252.

Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 4580–4584.

Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew W. Senior, Kean K. Chin, Ananya Misra, and Chanwoo Kim. 2017. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 5 (May 2017), 965–979.

George Saon, Tom Sercu, Steven Rennie, and Hong-Kwang J. Kuo. 2016. The IBM 2016 english conversational telephone speech recognition system. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'16)*. 7–11.

Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. "Your word is my command": Google search by voice: A case study. In *Advances in Speech Recognition*. Springer, 61–90.

Markus Schedl, Yi-Hsuan Yang, and Perfecto Herrera-Boyer. 2016. Introduction to intelligent music systems and applications. *ACM Trans. Intell. Syst. Technol.* 8, 2 (Oct. 2016), 17:1–17:8.

Björn Schuller, Felix Weninger, Martin Wöllmer, Yang Sun, and Gerhard Rigoll. 2010. Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'10)*. 4562–4565.

Michael L. Seltzer, Dong Yu, and Yongqiang Wang. 2013. An investigation of deep neural networks for noise robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13)*. 7398–7402.

Sangita Sharma, Dan Ellis, Sachin S. Kajarekar, Pratibha Jain, and Hynek Hermansky. 2000. Feature extraction using nonlinear transformation for robust speech recognition on the aurora database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)*. 1117–1120.

Soundararajan Srinivasan, Nicoleta Roman, and DeLiang Wang. 2006. Binary and ratio time-frequency masks for robust speech recognition. *Speech Commun.* 48, 11 (Nov. 2006), 1486–1501.

Pawel Swietojanski, Arnab Ghoshal, and Steve Renals. 2013. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU'13)*. 285–290.

Pawel Swietojanski, Arnab Ghoshal, and Steve Renals. 2014. Convolutional neural networks for distant speech recognition. *IEEE Sign. Process. Lett.* 21, 9 (Sep. 2014), 1120–1124.

George Trigeorgis, Fabien Ringeval, Raymond Bruckner, Erik Marchi, Mihalis Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*. Shanghai, China, 5200–5204.

Barry D. Van Veen and Kevin M. Buckley. 1988. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Mag.* 5, 2 (Apr. 1988), 4–24.

Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni. 2013. The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13)*. 126–130.

Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. 2006. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Process.* 14, 4 (July 2006), 1462–1469.

Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. 2016. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. (submitted for publication).

Tuomas Virtanen, Rita Singh, and Bhiksha Raj. 2012. *Techniques for Noise Robustness in Automatic Speech Recognition*. John Wiley & Sons, Hoboken, NJ.

DeLiang Wang. 2005. *On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis*. Springer US, Boston, MA, 181–197.

Yuxuan Wang, Arun Narayanan, and DeLiang Wang. 2014. On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 12 (Dec. 2014), 1849–1858.

Yuxuan Wang and DeLiang Wang. 2013. Towards scaling up classification-based speech separation. *IEEE Trans. Audio Speech Lang. Process.* 21, 7 (July 2013), 1381–1390.

Yuxuan Wang and DeLiang Wang. 2015. A deep neural network for time-domain signal reconstruction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'15)*. 4390–4394.

Zhong-Qiu Wang and DeLiang Wang. 2016. A joint training framework for robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 24, 4 (Apr. 2016), 796–806.

Ernst Warsitz and Reinhold Haeb-Umbach. 2007. Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Trans. Audio Speech Lang. Process.* 15, 5 (July 2007), 1529–1539.

Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R. Hershey, and Björn Schuller. 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*. 91–99.

Felix Weninger, Florian Eyben, and Björn Schuller. 2014. Single-channel speech separation with memory-enhanced recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14)*. 3709–3713.

Felix Weninger, Jordi Feliu, and Björn Schuller. 2012. Supervised and semi-supervised suppression of background music in monaural speech recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'12)*. 61–64.

Felix Weninger, Jürgen Geiger, Martin Wöllmer, Björn Schuller, and Gerhard Rigoll. 2013. The munich feature enhancement approach to the 2nd CHiME challenge using BLSTM recurrent neural networks. In *Proceedings of the 2nd CHiME Workshop on Machine Listening in Multisource Environments*. 86–90.

Felix Weninger, Jürgen Geiger, Martin Wöllmer, Björn Schuller, and Gerhard Rigoll. 2014a. Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. *Comput. Speech Lang.* 28, 4 (July 2014), 888–902.

Felix Weninger, John R. Hershey, Jonathan Le Roux, and Björn W. Schuller. 2014b. Discriminatively trained recurrent neural networks for single-channel speech separation. In *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP'14)*. 577–581.

Felix Weninger, Shinji Watanabe, Jonathan Le Roux, J. Hershey, Yuuki Tachioka, Jürgen Geiger, Björn Schuller, and Gerhard Rigoll. 2014c. The MERL/MELCO/TUM system for the REVERB challenge using deep recurrent neural network feature enhancement. In *Proceedings of the REVERB Workshop, Held in Conjunction with ICASSP 2014 and HSCMA 2014*. 1–8.

Donald S. Williamson and DeLiang Wang. 2017a. Speech dereverberation and denoising using complex ratio masks. In *Proceedings of the IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP'17)*. 5590–5594.

Donald S. Williamson and DeLiang Wang. 2017b. Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* 25, 7 (July 2017), 1492–1501.

Martin Wöllmer, Florian Eyben, Alex Graves, Björn Schuller, and Gerhard Rigoll. 2010a. Improving keyword spotting with a tandem BLSTM-DBN architecture. In *Proceedings of the Advances in Non-Linear Speech Processing: International Conference on Nonlinear Speech Processing (NOLISP'10)*. 68–75.

Martin Wöllmer, Florian Eyben, Björn W. Schuller, Yang Sun, Tobias Moosmayr, and Nhu Nguyen-Thien. 2009. Robust in-car spelling recognition—A tandem BLSTM-HMM approach. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'09)*. 2507–2510.

Martin Wöllmer, Björn Schuller, Florian Eyben, and Gerhard Rigoll. 2010b. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE J. Select. Top. Sign. Process.* 4, 5 (Oct. 2010), 867–881.

Martin Wöllmer, Zixing Zhang, Felix Weninger, Björn Schuller, and Gerhard Rigoll. 2013. Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'13).* 6822–6826.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144* (Oct. 2016).

Bingyin Xia and Changchun Bao. 2013. Speech enhancement with weighted denoising auto-encoder. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'13).* 3444–3448.

Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L. Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu. 2016a. Deep beamforming networks for multi-channel speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16).* 5745–5749.

Xiong Xiao, Chenglin Xu, Zhaofeng Zhang, Shengkui Zhao, Sining Sun, and Shinji Watanabe. 2016b. A study of learning based beamforming methods for speech recognition. In *Proceedings of the CHiME Workshop.* 26–31.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. *Achieving Human Parity in Conversational Speech Recognition.* Technical Report MSR-TR-2016-71. Microsoft Research.

Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2014a. Dynamic noise aware training for speech enhancement based on deep neural networks. In *Proceedings of the Conference of the International Speech Communication Association (INTER-SPEECH'14).* 2670–2674.

Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2014b. An experimental study on speech enhancement based on deep neural networks. *IEEE Sign. Process. Lett.* 21, 1 (Jan. 2014), 65–68.

Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2014c. NMF-based target source separation using deep neural network. *IEEE Sign. Process. Lett.* 21, 1 (Jan. 2014), 65–68.

Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2015. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* 23, 1 (Jan. 2015), 7–19.

Yi-Hsuan Yang and Homer H. Chen. 2012. Machine recognition of music emotion: A review. *ACM Trans. Intell. Syst. Technol.* 3, 3 (May 2012), 40:1–40:30.

Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita, Roland Maas, Tomohiro Nakatani, and Walter Kellermann. 2012. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *IEEE Sign. Process. Mag.* 29, 6 (Nov. 2012), 114–126.

Chengzhu Yu, Atsunori Ogawa, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, and John H. L. Hansen. 2015. Robust i-vector extraction for neural network adaptation in noisy environment. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'15).* 2854–2857.

Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv:1605.07146* (May 2016).

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV'14).* 818–833.

Zixing Zhang, Nicholas Cummins, and Björn Schuller. 2017. Advanced data exploitation for speech analysis—An overview. *IEEE Sign. Process. Mag.* 34 (July 2017). 24 pages.

Zixing Zhang, Joel Pinto, Christian Plahl, Björn Schuller, and Daniel Willett. 2014. Channel mapping using bidirectional long short-term memory for dereverberation in hand-free voice controlled devices. *IEEE Trans. Cons. Electron.* 60, 3 (Aug. 2014), 525–533.

Zixing Zhang, Fabien Ringeval, Jing Han, Jun Deng, Erik Marchi, and Björn Schuller. 2016. Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH'16).*