Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 31 (2015) 65–86

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Environmentally robust ASR front-end for deep neural network acoustic models☆

T. Yoshioka [a,b,*], M.J.F. Gales [a]

[a] *Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, UK*
[b] *NTT Communication Science Laboratories, Hikari-Dai, Seika-Cho, Kyoto 619-0237, Japan*

Received 29 May 2014; received in revised form 25 September 2014; accepted 25 November 2014
Available online 4 December 2014

Abstract

This paper examines the individual and combined impacts of various front-end approaches on the performance of deep neural network (DNN) based speech recognition systems in distant talking situations, where acoustic environmental distortion degrades the recognition performance. Training of a DNN-based acoustic model consists of generation of state alignments followed by learning the network parameters. This paper first shows that the network parameters are more sensitive to the speech quality than the alignments and thus this stage requires improvement. Then, various front-end robustness approaches to addressing this problem are categorised based on functionality. The degree to which each class of approaches impacts the performance of DNN-based acoustic models is examined experimentally. Based on the results, a front-end processing pipeline is proposed for efficiently combining different classes of approaches. Using this front-end, the combined effects of different classes of approaches are further evaluated in a single distant microphone-based meeting transcription task with both speaker independent (SI) and speaker adaptive training (SAT) set-ups. By combining multiple speech enhancement results, multiple types of features, and feature transformation, the front-end shows relative performance gains of 7.24% and 9.83% in the SI and SAT scenarios, respectively, over competitive DNN-based systems using log mel-filter bank features.

Keywords: Environmental robustness; Deep neural network; Front-end; Meeting transcription

1. Introduction

Overcoming performance degradation caused by background noise and reverberation has been one of the main challenges facing automatic speech recognition (ASR) for the last two decades. With the recent increased adoption of the ASR technology by industry, it is becoming a pressing problem to make ASR systems robust against environmental distortion. A range of approaches has been proposed to tackle this problem, including moment normalisation (de la Torre et al., 2005; Hilger and Ney, 2006), speech enhancement (Macho et al., 2002), feature enhancement (Stouten,

☆ This paper has been recommended for acceptance by H. Van Hamme.

* Corresponding author at: NTT Communication Science Laboratories, Hikari-Dai, Seika-Cho, Kyoto 619-0237, Japan. Tel.: +81 774 93 5326; fax: +81 774 93 5158.
E-mail addresses: yoshioka.takuya@lab.ntt.co.jp (T. Yoshioka), mjfg@eng.cam.ac.uk (M.J.F. Gales).

http://dx.doi.org/10.1016/j.csl.2014.11.008

2006; Yoshioka and Nakatani, 2013), feature transformation (Droppo et al., 2001), and acoustic model adaptation (Kalini et al., 2010; Wang and Gales, 2012; Lu et al., 2013). Most of the existing robustness techniques have been evaluated using conventional acoustic models consisting of Gaussian mixture models (GMMs) and hidden Markov models (HMMs).

In the last couple of years, significant progress has been made on acoustic modelling using context-dependent deep neural networks (DNNs), which has significantly changed the nature of acoustic models. A DNN is a multi-layer perceptron (MLP) with many hidden layers. The multiple layers of nonlinear processing allow the acoustic model to learn complex decision boundaries between HMM states. Following great initial success in several different tasks (Dahl et al., 2012; Mohamed et al., 2012a), the DNN-based acoustic model is now becoming an essential component of today's ASR systems. It has been shown that this novel acoustic modelling approach is less sensitive to variations in speech features than the conventional approach based on GMMs and that some of the classical techniques developed for GMM-based systems, such as heteroscedastic linear discriminant analysis (HLDA), have a limited impact on the recognition performance of DNN-based systems under certain conditions (Seide et al., 2011). Therefore, it is important to examine the usefulness of existing environmental robustness techniques in DNN-based systems and to analyse how the effects of the different approaches interact.

The DNN-based acoustic models have two widely used configurations (Hinton et al., 2012). In the first configuration, which is called a *DNN–HMM hybrid* or simply a *hybrid*, a DNN is utilised to compute the posterior probabilities of context-dependent HMM states based on observed feature vectors (Morgan and Bourlard, 1995; Renals et al., 1994; Dahl et al., 2012). A Viterbi decoding is then performed with these posteriors. The second configuration, which is called an *MLP tandem* or a *tandem* for short, uses the DNN to perform a nonlinear discriminative feature transformation, which yields MLP features (Hermansky et al., 2000; Grezl et al., 2007). These features are merged with a standard set of features, such as MFCCs or PLP coefficients, to form a new set of features, which are collectively called TANDEM[1] features, and then input into a GMM-HMM acoustic model. This type of acoustic models has often been used with a speaker adaptive training (SAT) set-up since various adaptation techniques are available for GMM-HMM acoustic models, including cluster adaptive training (Gales, 2000) and CMLLR (Gales, 1998). More recent, but less established, architectures, such as stacked hybrids (Knill et al., 2013) and convolutional neural networks[2] (CNNs) (Abdel-Hamid et al., 2012), are not considered in this paper.

While DNN acoustic models have been successfully applied to noisy speech recognition tasks with both configurations, little has been revealed about which element of the DNN acoustic model is susceptible to environmental distortion and which robustness techniques still matter in the DNN-based systems. The DNN–HMM hybrid approach was first applied to a noisy ASR task by Seltzer et al. (2013). They achieved the best published result on the Aurora 4 data set using a multi-condition hybrid acoustic model trained with the drop-out technique. They also showed that performing speech enhancement on both training and test sets with the method described in Yu et al. (2008) degraded the recognition performance. Geiger et al. (2014) showed that a non-negative matrix factorisation-based enhancement method improved the recognition performance of a heterogeneous acoustic model consisting of GMM-HMMs and a long short-term memory network in the CHiME2 medium vocabulary task. Li and Sim (2013) attempted to exploit a vector Taylor series (VTS) adaptation to improve the noisy digit recognition performance of a hybrid acoustic model. A few papers were also presented at ICASSP 2014 that proposed front-end processing schemes using clean/noisy stereo corpus (Narayanan and Wang, 2014; Li and Sim, 2014; Weninger et al., 2014). On the other hand, a large body of work applied MLP tandem acoustic models, with both shallow and deep configurations, to tasks related to environmental robustness such as meeting and lecture transcriptions based on distant microphones (Hain et al., 2012; Stolcke, 2011; Chang et al., 2013). These previous efforts showed that, while DNN acoustic models greatly outperform conventional GMM-based models in acoustically adverse environments, they still suffer from acoustic degradation. However, since the previous work made little use of existing robustness techniques developed for GMM-based models, the performance gain that can be obtained from the existing techniques when DNNs are being used is unknown. Furthermore,

[1] In this paper, we capitalise the term 'tandem' to refer to a feature vector. When we refer to an acoustic model configuration, we use lower-case letters.

[2] A conference paper discussing the use of CNNs in speaker independent meeting transcription was published just before submission of this paper (Renals and Swietojanski, 2014).

most previous work used small to medium vocabulary tasks with simulated data, such as Aurora 2 and Aurora 4. and considered only additive noise.

In this paper, we investigate the individual and combined impact of various front-end approaches for environmental robustness on the performance of DNN-based systems. To this end, after describing the data sets used in this work and our baseline DNN-based acoustic models in Section 2, we examine the environmental robustness of DNN acoustic models trained on corrupted data in Section 3. Our investigation uses the AMI meeting corpus (Carletta et al., 2006) with a single distant microphone set-up, which enables us to evaluate the practical relevance of the techniques being investigated. In Section 4, we classify various front-end processing approaches based on functionality and evaluate each class of approaches individually to reveal its fundamental usefulness in DNN-based systems. Based on this investigation, we propose a front-end processing pipeline in Section 5 to allow different classes of approaches to be incorporated efficiently in a single system. Using this front-end, we evaluate the combined effects of different classes of approaches with both hybrid SI and tandem SAT configurations. Finally, we conclude the paper in Section 6. Although many of the techniques considered in this paper have been described previously, there has been no work that has investigated their individual and combined effects on the DNN acoustic models in practically relevant tasks with both SI and SAT set-ups.

Note that stereo corpus-based feature transformation and noise insensitive parameterisation are not covered in this paper. The former learns a mapping from corrupted features to their underlying clean features using a set of clean and noisy feature pairs, where the mapping may be modelled using piece-wise linear functions (Droppo et al., 2001) or neural networks (Narayanan and Wang, 2014; Li and Sim, 2014; Weninger et al., 2014). These methods have been successful for artificially simulated data. However, they assume a stereo corpus to be collectable from target or acoustically similar environments while the way to do this has yet to be established for practical applications.[3] The latter approach seeks acoustic features that are inherently robust against noise. Such features include power normalised cepstral coefficients (Kim and Sterm, 2012) and frequency domain linear prediction (Thomas et al., 2008). To limit the scope of this paper, we focus on conventional features, such as MFCCs and log mel-filter bank outputs.

2. Experimental framework

2.1. Tasks and data sets

This work used the AMI meeting corpus. The corpus consists of recordings of meetings conducted in English at three different sites. Each meeting has four participants. The meetings are either scenario-based role playing discussions or natural unconstrained conversations. Many of the meeting participants are non-native speakers and hence they may have very different accents and speaking styles. Each meeting was recorded with an eight-channel circular microphone array placed on a table in a meeting room, providing eight-channel synchronised acoustic signals. In this paper, we focus on a single distant microphone (SDM) task in which only the first channel is used.[4] The large distances between the microphone and the speakers mean that the speech signals are distorted by reverberation and background noise. The background noise is almost stationary and the SNR does not vary significantly over different meetings. As regards reverberation, the three meeting rooms seem to have similar reverberation times. However, since the relative positions of individual speakers in relation to the microphone vary significantly and some speakers appear to be moving within an utterance, different utterances may have different reverberation characteristics. See the corpus website (https://www.idiap.ch/dataset/ami) for further details.

We divided the corpus into training, development test, and evaluation test sets in the same way as Breslin et al. (2011). Specifically, we selected a set of eight meetings (ES2009a–d and IS2009a–d) as a development test set and another set of eight meetings (ES2008a–d and IS2008a–d) as an evaluation test set. The remaining portion of the corpus was used for training. The development and evaluation sets each included eight speakers while the training set consisted of utterances from 175 speakers. We excluded overlapping speech segments from both the training and test sets, which left 59 h of speech for training, 2.7 h for the development test, and 2.6 h for the evaluation. Note that these

[3] Du et al. (2014) describes an attempt to overcome this limitation by generating pseudo-clean features.

[4] Our work on a multiple distant microphone task is described in Yoshioka et al. (2014).

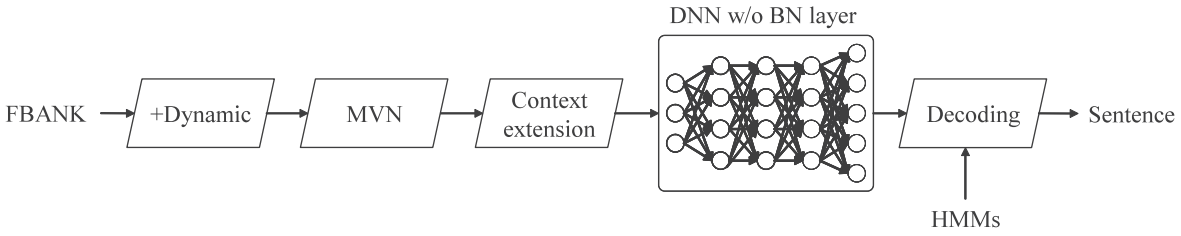

Fig. 1. Processing flow diagram of baseline hybrid SI system.

quantities refer to the quantities of audio after segmentation. Speech segments and speaker identities were obtained from manual annotations.[5]

In our experiments, we considered both SI and SAT set-ups. With the SI set-up, adaptation was not performed in either the training or test stages. The speaker information was used only for mean and variance normalisation (MVN). Specifically, we performed normalisation processing on a per speaker and per meeting basis. We did this so that the SI and SAT systems used the same input features, which enabled us to accurately evaluate the gains from adaptation techniques on top of speaker and session-level MVN. With the SAT set-up, the acoustic models were trained with a speaker adaptive approach. The test data were swept over multiple times, which means that decoding was performed with off-line processing.

2.2. *Baseline systems*

In this section, we describe the ways in which we built our baseline SI and SAT systems and show the word error rates (WERs) of these systems. We employed a DNN-HMM hybrid configuration for the SI set-up because it achieved lower WERs than a tandem configuration in our preliminary tests (see Yoshioka et al., 2014 for a comparison of the hybrid and tandem SI performances). On the other hand, we adopted the MLP tandem configuration to build our SAT systems, which allowed us to exploit conventional adaptation techniques including CMLLR and MLLR. The tandem configuration enables SAT to be performed with any input feature type, which matters in the experiments using expanded feature sets described later. These two baseline systems are described in Sections 2.2.1 and 2.2.2, respectively.

2.2.1. *DNN–HMM hybrid SI system*

Fig. 1 shows the processing flow of the baseline hybrid SI system. The input features consisted of 24-channel log mel-filter bank outputs, which are called FBANK features. We employed these features because they provided lower WERs than MFCCs, which is consistent with the findings of other studies (Mohamed et al., 2012b; Deng et al., 2013). These input features were concatenated with their delta parameters up to the third order, resulting in a stream of 96-dimensional feature vectors. Then MVN was performed on a speaker-by-speaker, meeting-by-meeting basis. The resultant normalised feature vectors were spliced with neighbouring feature vectors within a context window. This context-extended feature set was fed into a DNN to compute the posterior probabilities of individual context-dependent HMM states. Decoding was performed based on these posteriors with the Viterbi algorithm. In our default settings, the context window consisted of nine consecutive frames (four frames on each side) and the DNN consisted of five hidden layers, each with 1500 units, followed by a softmax output layer. Our previous experiments showed that changing the context window size and the network topology had little impact on the recognition performance for the same data set (Yoshioka et al., 2014).

The hybrid SI system was trained according to the standard recipe (Dahl et al., 2012), consisting of two stages. In the first stage, an underlying GMM-HMM system was built using 13 MFCCs (including C0). The initial step of the GMM-HMM system construction was to augment the training MFCC set with their delta coefficients up to the third order, which yielded 52-dimensional feature vectors. This was followed by speaker and meeting-level MVN. These

[5] Although accurate speaker diarisation is a challenging task from a single microphone input, our SAT experiments allow us to understand the impact of adaptation on the recognition performance in adverse acoustic environments and the insights obtained from the experiments will be useful for other tasks.

Table 1
WERs (%) of baseline SI systems.

System	WER (%)		
	Dev	Eval	Avg
MPE GMM-HMM	54.7	55.6	55.2
DNN–HMM hybrid	43.5	42.6	43.1

normalised feature vectors were projected onto a 39-dimensional feature space by HLDA. Then, a maximum likelihood GMM-HMM acoustic model was trained to model the HLDA features. The acoustic model consisted of approximately 4000 context-dependent states and 16 Gaussians per state. The model was further refined by MPE training, which yielded the baseline GMM-HMM SI system.

In the second stage, a DNN was trained to predict the context-dependent HMM states from context-extended FBANK feature vectors. The second stage began with the forced alignment of the entire training data set to produce frame-level state labels. Then, the DNN was trained to predict these labels from the nine-frame context-extended feature vectors. This DNN training was achieved by layerwise discriminative pre-training (Seide et al., 2011), followed by fine-tuning based on a cross entropy criterion. The DNN trained in the second stage and the HMMs obtained in the first stage constitute the baseline hybrid SI acoustic model. At the test stage, decoding was performed by bigram lattice generation with a 40K-word language model, followed by trigram lattice rescoring and confusion network rescoring. The language model was built from a variety of sources including transcriptions of AMI, ICSI, NIST, and ISL meetings, Callhome, Switchboard, Gigaword, and extra web data (Breslin et al., 2011).

Table 1 shows the performance of the baseline hybrid SI system and that of the MPE-trained GMM-HMM system. The DNN acoustic model improved the average WER from 55.2% to 43.1%, providing a relative improvement of 21.9%.[6] This gain is consistent with other work and indicates the usefulness of the DNN acoustic models in adverse acoustic conditions. Section 4 investigates various approaches to further improve this competitive baseline system.

2.2.2. *DNN tandem SAT system*

Fig. 2 shows the processing flow of the baseline tandem SAT system, which is based on the bottleneck configuration (Grezl et al., 2007). As in the SI system, input features consisting of 24 FBANK coefficients were appended with their first, second, and third-order delta coefficients and then mean and variance-normalised on a speaker-by-speaker, meeting-by-meeting basis. Then, each of the resultant 96-dimensional feature vectors was extended using a nine-frame context window, yielding a sequence of 864-dimensional feature vectors. The tandem DNNs used in this work consisted of four hidden layers, each with 1500 units, followed by one hidden bottleneck (BN) layer and a softmax output layer, where the BN layer had only 26 units. The linear outputs from the BN layer were further converted by a global semi-tied covariance (STC) transform (Gales, 1999) and concatenated with the corresponding MFCC-derived HLDA features to form 65-dimensional TANDEM feature vectors. The TANDEM features were recognised with a GMM-HMM system adapted with global CMLLR and MLLR mean transforms (Gales, 1998).

Our tandem system construction followed the rapid training pipeline described in Park et al. (2011). A DNN with a BN layer was trained in the same way as the DNN of a hybrid SI system. After the DNN was trained, each training feature vector was forwarded through the network and the linear outputs from the BN layer were computed. A global STC transform was optimised for and applied to these BN features. Then each BN feature vector was concatenated with the corresponding HLDA feature vector to generate a 65-dimensional TANDEM feature vector. A GMM-HMM acoustic model was trained with an ML criterion using these TANDEM features. With this ML model, global CMLLR transforms were obtained for each speaker to eliminate speaker-specific variations. Finally, based on these speaker-normalised TANDEM features, MPE training was performed to obtain a discriminative SAT acoustic model.

At the test stage, adaptation was performed with supervisions generated by the baseline hybrid SI system. Global CMLLR and MLLR mean transforms were optimised for each speaker at each meeting and used to adapt the test

[6] The WER of 43.1% is comparable to performance figures reported by other sites on the same corpus with similar acoustic model configurations (Swietojanski et al., 2013), although the partitions of the corpus are slightly different.

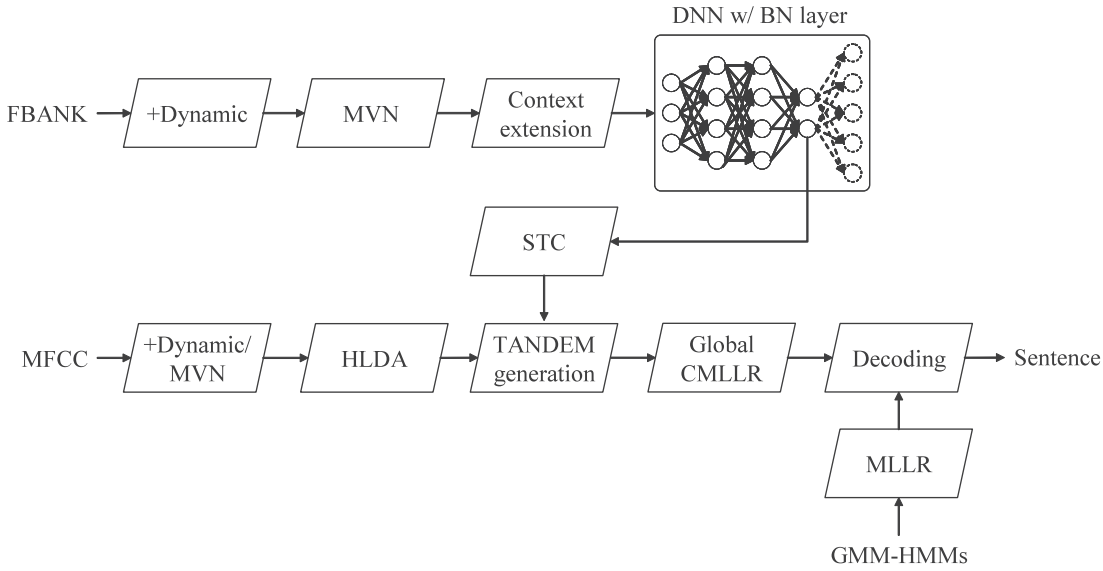

Fig. 2. Processing flow diagram of baseline tandem SAT system.

Table 2
WERs (%) of baseline SAT systems.

System	WER (%)		
	Dev	Eval	Avg
MPE-SAT GMM-HMM	48.8	50.2	49.5
DNN tandem	40.7	40.9	40.8

TANDEM features and the TANDEM-feature GMMs, respectively. Decoding was performed by bigram lattice generation, followed by trigram lattice rescoring and confusion network rescoring as with the SI set-up.

Table 2 shows the WERs of the baseline tandem SAT system and the MPE-SAT GMM-HMM system. The results clearly show the advantage of the DNN tandem acoustic model over the GMM-HMM acoustic model with conventional non-MLP features. The DNN tandem SAT system outperformed the MPE-SAT GMM-HMM system by 8.7% absolute, or 17.6% relative. Comparing Tables 1 and 2, we can see that the tandem SAT system achieved a lower WER than the DNN–HMM hybrid SI system, indicating the benefit of SAT. Note that the GMM-HMM SAT system used adaptation supervisions generated by the MPE-trained GMM-HMM SI system. When we used the supervisions produced by the hybrid SI system, the WERs were improved to 47.3% and 48.7% for the development and evaluation test sets, respectively, although these numbers were still far greater than those of the tandem SAT system.

In the following sections, results should not be compared across tables because different tables may use different configurations.

3. Assessment of environmental robustness of deep neural networks

First, we conducted a set of experiments to identify the causes of performance limitation. While acoustic features extracted from a distant microphone are certainly degraded by background noise and reverberation, it is unclear how much of the recognition error can be attributed to the acoustic degradation. In addition, while training of a DNN-based acoustic model (in particular, a hybrid model) consists of two stages, i.e., state alignment generation and DNN parameter learning, it is an open question as to which stage is more susceptible to acoustic distortion. This section describes our experimental results related to these questions.

To examine the impact of environmental distortion on recognition performance, we used meeting audio recorded with individual headset microphones (IHMs) in addition to the SDM recordings. The AMI corpus contains both distant

Table 3
SDM vs. IHM comparison with hybrid SI set-up. The SDM numbers slightly differ from those in Table 1 due to different hidden layer sizes (1000 units in this table vs. 1500 units in Table 1).

Data set	WER (%)		
	Dev	Eval	Avg
SDM	43.8	43.0	43.4
IHM	28.2	24.6	26.4

Table 4
WERs with cross-set training, hybrid SI set-up. The last two rows can also be found in Table 3.

System			WER (%)		
Alignment, HMM	Input	Topology	Dev	Eval	Avg
IHM	SDM	1000×5	41.8	40.8	41.3
		2000×5	41.7	40.6	41.2
SDM	IHM	1000×5	30.6	27.0	28.8
SDM	SDM	1000×5	43.8	43.0	43.4
IHM	IHM	1000×5	28.2	24.6	26.4

microphone recordings and those recorded with IHMs that were synchronised with distant microphones. Since the IHM data are little affected by environmental distortion, the contrast between the performance of the SDM and IHM data sets illuminates the performance loss caused by environmental distortion. To enable an accurate evaluation of the performance difference between the SDM and IHM set-ups, the IHM training and test sets did not contain overlapping speech segments.

Table 3 contrasts the system trained and tested on the IHM data set with the system trained and tested on the SDM data set, both using the hybrid SI configuration. The IHM system was built exactly in the same way as the SDM system. Note that, in this experiment, we used DNNs with 1000×5 hidden units. A significant performance gap can be seen between these two systems, which means that the acoustic degradation caused by the background noise and reverberation constituted a major cause of the recognition errors of the SDM system. Specifically, 39.2% of the word errors made by the SDM system can be attributed to environmental distortion. This clearly indicates that, while acoustic modelling based on DNNs provides significant performance gains, such acoustic models are still liable to be harmed by environmental distortion even when they are trained on corrupted data.

A further experiment was performed to investigate whether state alignments or DNN parameters are more susceptible to environmental distortion and thus need improvement. To this end, we developed two hybrid SI systems. One was based on a DNN trained on the SDM data set using the state alignments generated by performing forced alignment on the IHM data set with the IHM MPE system. Thus, this system shared the alignments and the HMMs with the IHM system described above. The other system was trained on the IHM data set while the state alignments and the HMM set were taken from the SDM MPE system.

The WERs obtained with these cross-set training set-ups are listed in Table 4. When we used the SDM data as inputs, the average WER was as high as 41.3% even though the alignments and the HMM set were taken from the IHM system. Increasing the number of hidden units to 2000×5 did not change the performance at all. On the other hand, when we used the IHM data as the inputs, the average WER was significantly improved, achieving 28.8%, in spite of the use of the SDM-based state alignments and the SDM-derived HMM set. These results clearly indicate that the DNN parameters are prone to acoustic distortion and needs improvement for environmental robustness.

4. Approaches to environmental robustness

In Section 3, we showed that making DNN parameters insensitive to distortion would result in a higher degree of environmental robustness of a recognition system. In this section, we review various environmental robustness techniques, classify them into distinguishable categories, and evaluate the efficacy of individual classes of approaches in the SDM meeting transcription task with the hybrid SI set-up. Our aim is to examine whether each class of approach

can fundamentally improve the performance of DNN-based acoustic models. Since the prerequisite for such environmentally robust ASR is insensitivity to the irrelevant variations that exist in observed speech, some of the approaches considered in this section can also be applied to general ASR tasks.

Following the classification of Yu et al. (2013), we can distinguish three categories of robustness approaches based on where the operation takes place in the processing flow of a system.

- Front-end processing – this enhances, transforms, or expands a set of features input into a DNN to make them more robust against environmental distortion.
- Network adaptation – this adjusts DNN parameters to the characteristics of each environment or speaker to make the DNN less sensitive to variations resulting from differences in speakers and environments.
- Output transformation – this transforms the outputs from the DNN (i.e., context-dependent HMM posteriors or TANDEM features) for each environment or speaker to eliminate irrelevant variations from these DNN outputs.

This paper exclusively considers the front-end processing approaches and only explains the other two approaches briefly below.

The effectiveness of network adaptation, which directly modifies DNN parameters, has been established mainly in supervised speaker adaptation tasks. A representative network adaptation approach is regularised retraining, which modifies the DNN parameters to improve the classification accuracy for a given adaptation data set while keeping the transformed model from deviating too much from the original model. Such regularisation can be achieved by using an L2 penalty (Liao, 2013) or a Kullback–Leibler divergence penalty (Yu et al., 2013). While this approach has been successful in supervised speaker adaptation tasks, there has been limited evidence of its usefulness in large vocabulary unsupervised adaptation tasks, particularly those with high error rates.

For output transformation, the form of the transform depends on the acoustic model configuration. When the acoustic model is based on a tandem configuration, existing feature transformation and GMM adaptation techniques can be employed since the outputs from the DNN are recognised with a conventional GMM-HMM acoustic model. For example, front-end CMLLR (Gales and Flego, 2012) and global CMLLR may be used to perform environment- or speaker-specific TANDEM feature transformation while cluster adaptive training (Gales, 2000), regression-class CMLLR and MLLR may be utilised for GMM adaptation. The effect of SAT using a tandem configuration was confirmed in the previous section. With a hybrid configuration, the state posteriors generated by the DNN may be further converted with a linear transform. Such a transform can be estimated with an output discriminative linear transform (Yao et al., 2012) or similar techniques.

4.1. Characteristics of front-end processing approaches

The objective of front-end processing is to convert an observed speech signal to a set of DNN input features that are insensitive to environmental distortion while simultaneously containing a sufficient amount of discriminant information. This can be approached in three different ways.

- Speech enhancement – this attempts to recover clean features by exploiting models of speech, noise, and the way in which speech and noise interact to form a noisy speech signal. In the classification adopted here, speech enhancement refers to a process exploiting an interaction model of speech and noise and therefore acts before MVN as MVN processing makes it impossible to mathematically formulate the relationship between clean speech and noise. Example methods are spectral subtraction, Wiener filtering, and front-end VTS (FE-VTS).
- Feature transformation – this class of approaches, namely where the feature transforms are learned from data unlike the speech enhancement approaches, aims at reducing irrelevant variations from the MVN-processed features and possibly enhancing the discriminant information inherent in the features. FMPE (Povey et al., 2005), global CMLLR, and linear input network (LIN) (Neto et al., 1996) can be placed in this category. Note that we classify stereo corpus-based speech enhancement methods, such as stereo piecewise linear compensation for environment (SPLICE) (Droppo et al., 2001), as a feature transformation approach.

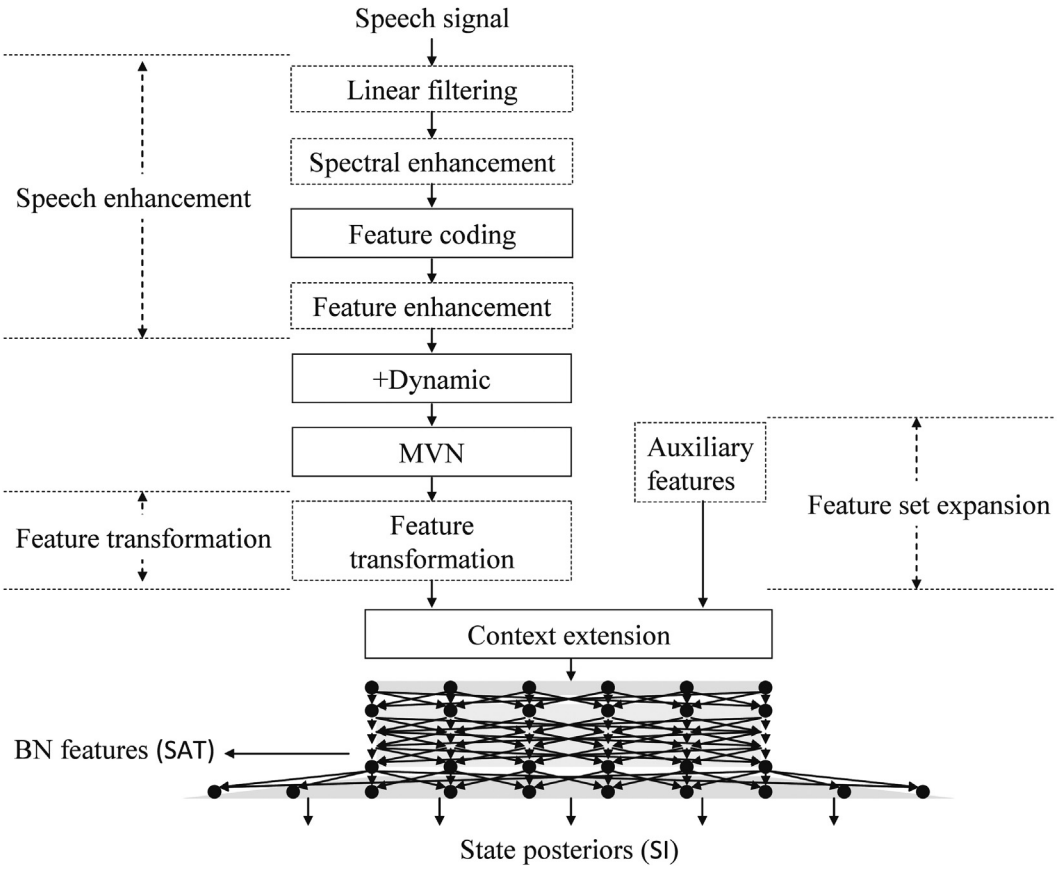

Fig. 3. Stages of feature extraction interleaved with possible front-end robustness approaches (robustness approaches shown in dashed boxes).

- Feature set expansion – this expands a feature set by including auxiliary features that are obtained in different ways from the primary feature extraction path. The objective is to provide the DNN with complementary information to improve the classification accuracy.

Fig. 3 shows the processing flow of robust feature extraction, which highlights where the operations of each category of approaches take place. In the rest of this section, we review the approaches of each category and discuss their characteristics. Then, we describe the results of a series of experiments conducted to evaluate the impact of individual classes of approaches. Based on these experimental results, in the next section a front-end processing pipeline is proposed to further investigate the combined effects of different approaches.

4.1.1. Speech enhancement

Features extracted from clean speech signals contain much more discriminant information than those of corrupted speech. Therefore, if clean features can be recovered to some extent, better DNN parameters will be obtained. The aim of speech enhancement is to achieve this by removing background noise and reverberation from observed speech.

A characteristic common to all speech enhancement approaches is that they exploit explicit models of the relationship between clean speech, noise, and noisy speech rather than statistically learning such relationships from data. This allows these approaches to be applied on a relatively short time scale basis, for example, on a frame-by-frame, block-by-block, or utterance-by-utterance basis. It is also important to note that most speech enhancement methods efficiently exploit a much longer acoustic context than the DNNs for noise and reverberation estimation. The DNNs cannot necessarily learn environmental characteristics by extending the temporal coverage of a context window as demonstrated in Yoshioka et al. (2014) probably because such a long context window may make training difficult.

As shown in the diagram in Fig. 3, the speech enhancement approaches can be further classified according to the quantities that they modify. Below, we discuss the general characteristics of these approaches to clarify the commonalities and differences.

4.1.1.1. Linear filtering. The first approach, linear filtering, processes time-domain signals, or almost equivalently, complex-valued short time Fourier transform (STFT) coefficients with a linear time-invariant filter or a filter that slowly changes with time. The enhanced time-domain or STFT-domain speech signals are transformed into FBANK features. This approach requires an array of synchronised microphones to reduce additive noise (Tashev, 2009). By contrast, for reverberation reduction, or dereverberation, there are established families of linear filtering algorithms that can be applied to single microphone signals (Yoshioka et al., 2012). In one such method called weighted prediction error (WPE) minimisation (Yoshioka and Nakatani, 2012), enhanced STFT coefficient y_t is computed as

$$y_t = x_t - \sum_{k=T_\perp}^{T_\top} g_k^* x_{t-k}, \tag{1}$$

where x_t and $(g_{T_\perp}, \ldots, g_{T_\top})$ denote an observed STFT coefficient and a set of filter coefficients, respectively, with t being a frame index. Note that the frequency bin index is omitted for conciseness. Thanks to the linear convolutional nature of the room impulse responses, which constitute reverberation, there is a set of filter coefficients that can cancel the effect of reverberation (Abed-Meraim et al., 1997). In practice, the filter coefficients need to be estimated based on the observed speech. See Yoshioka et al. (2012) and the references therein for filter estimation methods and a detailed discussion.

A distinctive characteristic of linear filtering as opposed to the other speech enhancement approaches is that this approach is unlikely to produce unnatural artefacts or irregular transitions between frames. This is because this approach is based on a time-invariant filter, whereas spectral and feature enhancement approaches process signals on a frame-by-frame basis. Thanks to this property, the features obtained with linear filtering can be directly fed into a standard speech recognition pipeline unlike those generated by the other enhancement approaches as discussed later.

4.1.1.2. Spectral enhancement. The second approach, spectral enhancement, operates after obtaining the magnitudes or squared magnitudes of the observed STFT coefficients. The principle is to estimate the noise components of each short time frame of the observed speech and then remove them from the observations. The removal operation can be performed, for example, by assuming the power spectra of speech and noise to be additive, i.e.,

$$|x_t|^2 = |s_t|^2 + |n_t|^2, \tag{2}$$

where s_t and n_t denote speech and noise STFT coefficients, respectively. There are a large number of spectral enhancement methods for background noise reduction, including spectral subtraction (Boll, 1979), log-spectral amplitude estimation (Ephraim, 1985), and the two-stage Wiener filtering used in the ETSI advanced front-end (ETSI ES 202 050 Ver. 1.1.5, 2005). There are several spectral enhancement methods that aim at reducing reverberation (Lebart et al., 2001; Kameoka et al., 2009).

We found that, when an acoustic model was trained on corrupted data, it was good to use the enhanced speech only for computing static features. The enhanced static features were combined with dynamic features obtained from the original un-enhanced speech. Similar results were reported for GHM-HMMs in Droppo and Acero (2008). Although no clear explanation has been given, we suspect that this is because the frame-by-frame operation performed by spectral enhancement causes unnatural transitions between neighbouring static features. The dynamic features computed from the enhanced static features could thus be less reliable than those computed from the original static features.

4.1.1.3. Feature enhancement. As an alternative to spectral enhancement, enhancement may be achieved after coding observed signals into features. Typically, the feature enhancement approach also assumes the additivity of speech and noise power spectra (although there are several methods that account for the modelling errors resulting from this additivity assumption (Deng et al., 2004)), which may be formulated in the FBANK domain as

$$\boldsymbol{x}_t = \boldsymbol{s}_t + \boldsymbol{h} + \log\left(1 + \exp(\boldsymbol{n}_t - \boldsymbol{x}_t - \boldsymbol{h})\right), \tag{3}$$

where $\boldsymbol{x}_t$, $\boldsymbol{s}_t$, and $\boldsymbol{n}_t$ denote FBANK feature vectors of corrupted speech, clean speech, and additive noise, respectively, at time t, and $\boldsymbol{h}$ denotes a static convolutional noise vector. Unlike the speech enhancement approach, most feature enhancement methods exploit a statistical model of clean features, such as GMMs and ergodic HMMs, to effectively compensate for environmental distortion. A variety of methods have been proposed to compensate for the degradation caused by background noise, including a front-end vector Taylor series (VTS) (Moreno et al., 1996; Stouten, 2006) and an unscented transform (Shinohara and Akamine, 2009). There are also a few methods that compensate for the effect of reverberation in the feature domain (Krueger and Haeb-Umbach, 2010; Yoshioka and Nakatani, 2013).

The feature enhancement approach is similar to spectral enhancement in the sense that both approaches modify observed speech on a frame-by-frame basis. Therefore, in our experiments, enhancement was performed only on static features, and delta coefficients were computed from original un-enhanced speech. It is sometimes argued that using a clean feature model allows the feature-domain approach to generate features that are less distorted than those obtained with spectral enhancement as demonstrated in Yoshioka and Nakatani (2013). Although this seems to be true for clean training tasks, it has yet to be clarified whether the same argument applies to multi-condition or matched training tasks when a DNN-based acoustic model is used.

4.1.2. Feature transformation

As shown in the diagram in Fig. 3, the (original or enhanced) static FBANK features are combined with dynamic features and then mean and variance normalised to reduce the speaker-specific variations. The features generated by MVN may be further transformed to reduce the remaining irrelevant variations while enhancing the discriminant information inherent in the features. Feature transformation is usually achieved with a data-driven approach since a simple relationship, such as Eqs. (1) and (2), cannot be formulated between the normalised clean and noisy features. In the following, three forms for transformation are identified and discussed.

The first form modifies feature vectors on a frame-by-frame basis, thus applying different modifications to different feature frames. If stereo data consisting of clean and noisy feature pairs can be used, SPLICE (Droppo et al., 2001) may be used, in which the transform takes the following form:

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \sum_{c=1}^{C} \gamma_{t,c} \boldsymbol{b}_c \tag{4}$$

with $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ being original and transformed feature vectors, respectively. Each $\boldsymbol{b}_c$ is called a correction vector and the set of correction vectors, $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_C)$, defines the transform. Which correction vector to use is determined by $\gamma_{t,c}$ and computed as the posterior probability of a C-component GMM of the original feature vectors. A stereo training data set is used to train the correction vector set. When stereo data are unavailable, FMPE may be used as an alternative way of estimating the correction vector set without utilising such a stereo corpus. With FMPE, the correction vector set is trained by interleaved updates of the correction vector set and an acoustic model parameter set, where the respective updates are achieved with MPE and ML criteria (Povey et al., 2005). The mathematical link between SPLICE and FMPE is discussed in Deng et al. (2005).

Apart from the frame-by-frame transformation approach, when speaker (or environment) labels are available, global CMLLR feature transformation may be performed to de-emphasise the differences between speakers (or environments).[7] With global CMLLR, each feature vector of speaker s is transformed as

$$\boldsymbol{y}_t = \boldsymbol{A}^{(s)} \boldsymbol{x}_t + \boldsymbol{b}^{(s)}, \tag{5}$$

where $\boldsymbol{A}^{(s)}$ and $\boldsymbol{b}^{(s)}$ define a transform for this speaker. The speaker transform is estimated for each speaker contained in the training and test data sets by using a GMM-HMM acoustic model based on FBANK features. Such an FBANK model can be efficiently constructed from a baseline MFCC-based GMM-HMM acoustic model with single pass retraining (SPR) (Young et al., 2009). Since FBANK features are closely correlated with each other, it is essential to use full covariance matrices or semi-tied covariance (STC) transforms (Gales, 1999). In our experiments, we used a

[7] Although we use the term "speaker adaptation", adaptation was actually performed at the meeting and speaker level in our experiments, i.e., the fact that some of the speakers were present at multiple meetings was not exploited.

global STC transform. Thus, SPR was configured to produce both FBANK-space GMM parameters and a global STC transform. Therefore, the speaker transforms actually used in our experiments had the following form:

$$\boldsymbol{y}_t = \boldsymbol{A}^{(s)}\boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}^{(s)}, \tag{6}$$

where $\boldsymbol{U}$ denotes the global STC transform. We used CMLLR rather than DNN-based discriminative methods, such as LIN. According to Li and Sim (2010), since CMLLR is based on generative principles, it is more effective than LIN for unsupervised adaptation in high error rate tasks.

Yet another approach is utterance-based transformation, the concept of which may be implemented by quantised CMLLR (Q-CMLLR) as explained below. The fundamental idea is to cluster the utterances in the training set into disjoint classes in some appropriate way and train a global CMLLR transform for each utterance class. At the test stage, each test utterance is clustered into one of the classes. Then the transform associated with the selected class is applied to the utterance. Unlike the speaker transformation approach described above, utterance-based transformation can be applied to SI scenarios. Although frame-based transformation is also applicable to SI tasks, there is a fundamental difference between the utterance-based and frame-based approaches. While the frame-based approach selects the transform to apply based on a relatively short acoustic context (i.e., the centre frame plus several adjacent frames), the utterance-based approach uses all the feature vectors within an utterance to select the transform, thus utilising a much longer acoustic context than the DNNs. This may be viewed as performing very rapid adaptation with one recognition pass (Yao et al., 2011).

As our preliminary test showed little performance gain from using FMPE,[8] our experiments investigate the other two approaches, namely the utterance-based and speaker-based approaches for SI and SAT scenarios, respectively.

4.1.3. Feature set expansion

With the goal of providing complementary information to help a DNN discriminate between different HMM states, feature set expansion expands the default feature set, consisting of log mel-filter bank outputs and their dynamic parameters, by adding auxiliary features. It is desirable for the auxiliary features to be extracted with little additional computational cost. A simple example is to add static MFCCs to the feature set, which indeed provides meaningful performance gains (Section 4.2.3).

It is useful to note that the feature set expansion approach becomes effective only with DNN-based acoustic models. This is because the DNNs are much more insensitive to the increase in input dimensionality than GMMs when a sufficient quantity of training data is available. Furthermore, as the DNNs can deal with correlations between different features, it is possible to include multiple features that have strong correlations. This research direction was initiated by the work described in Plahl et al. (2011), in which the authors combined three different coding systems, each extracting MFCCs, PLP coefficients, or gammatone features, using an MLP with a BN layer. The feature combination improved the performance of their Spanish recognition system. Noise aware training, whereby each observed feature vector is concatenated with a noise feature estimate, may be regarded as expanding the feature set (Seltzer et al., 2013).

Feature set expansion is particularly interesting for noisy speech recognition because it offers an alternative way of exploiting speech enhancement and feature transformation outputs. Conventionally, enhanced or transformed features are substituted for the original un-enhanced features as explained earlier. Instead, it is possible to use both the original and enhanced features as inputs into a DNN (Weninger et al., 2014). This allows us to utilise multiple speech enhancement and feature transformation systems that cannot be arranged in tandem. Another motivation for expanding the feature set is to complement features masked by environmental noise. Since some speech features are lost in acoustically adverse environments due to the masking property of noise, using extra features would lead to performance improvement when they are extracted in a way that emphasises parts of the observed speech that are likely to be disregarded in the primary feature extraction path.

4.2. Experimental results

Now, we show and discuss the results of our experiments conducted to evaluate the effectiveness of each class of robustness approaches. The SDM data sets were used for these experiments with the SI set-up.

[8] This is also shown by the experimental results presented at http://wissap.iiit.ac.in/proceedings/TSai_L7.pdf.

Table 5
Effect of dereverberation with linear filtering on hybrid SI system. ✗ symbol means 'disabled'.

Dereverberation	WER (%)		
	Dev	Eval	Avg
✗	43.5	42.6	43.1
WPE	**42.0**	**41.1**	**41.6**
AWPE	42.2	41.4	41.8

Table 6
Effects of additive noise compensation techniques on the hybrid SI system with dereverberation enabled. Spectral and feature enhancements performed with IMCRA and FE-VTS, respectively. ✗ and √ symbols mean 'disabled' and 'enabled', respectively.

Enhancement target		WER (%)			Enhanced features used to
Spectrum	Feature	Dev	Eval	Avg	
✗	✗	42.0	41.1	41.6	Replace original features
√	✗	**41.3**	40.9	41.1	Replace original features
✗	√	41.4	40.5	41.0	Replace original features
√	√	42.0	41.0	41.5	Replace original features
√	√	41.4	**40.4**	**40.9**	Expand feature set

4.2.1. Speech enhancement

We started our investigation with a linear filtering experiment. In this experiment, we employed a single microphone dereverberation method called WPE (Yoshioka et al., 2012) to eliminate reverberant noise from both the training and test data sets. WPE is based on STFT-domain linear filtering as shown by Eq. (1), where the filter coefficient set, $(g_{T_\perp}, \ldots, g_{T_\top})$, is adaptively estimated from observed STFT coefficients within a sliding time block. We used a time block of two seconds with a shift of 2 s. Both filter estimation and filtering operations were performed on unsegmented data. We also evaluated adaptive WPE (AWPE), which is based on adaptive linear filtering (Yoshioka et al., 2009).

In this experiment, a hybrid SI system was rebuilt from scratch, i.e., we reconstructed an underlying GMM-HMM acoustic model by using dereverberated MFCCs and used it to produce state alignments. With these state alignments, we trained a DNN that predicts the HMM states from dereverberated FBANK features.

Table 5 contrasts the WER of the baseline hybrid SI system with that of the hybrid SI systems with dereverberation. We can see that both dereverberation methods improved the recognition performance for both the development and evaluation data sets. The impact of dereverberation on the performance of the DNN acoustic models was also examined in our conference paper (Yoshioka et al., 2014), where we explored different model configurations and DNN topologies with both SI and SAT set-ups. The results in Table 5 and those described in our earlier work show linear filtering approaches to be effective with DNN-based acoustic models. To realise a quick experimental turn-around, all the following experiments used the state alignments generated with WPE and retrained only the DNN.

Next, let us examine the performance gains that spectral and feature enhancement methods can provide. In this experiment, we performed either or both spectral and feature enhancement to suppress the background noise from the dereverberated speech data generated in the experiment described above. As a spectral enhancement method, we combined an optimally modified log spectral amplitude (OMLSA) estimator (Cohen, 2002) and an improved minima controlled recursive averaging (IMCRA) estimator (Cohen, 2003) for speech and noise spectral estimation, respectively. Feature enhancement was performed using FE-VTS with a first order approximation (Moreno et al., 1996). Our FE-VTS system employed a speech GMM trained on a 10-h random subset of the IHM training data set. A noise Gaussian was estimated for each utterance with an expectation-maximisation algorithm with the initial noise Gaussian computed from the IMCRA-based noise estimates (Stouten, 2006). Note that spectral enhancement was applied to an unsegment speech stream whereas feature enhancement was performed on each utterance independently.

Table 6 lists the WERs obtained with various configurations. When we performed enhancement only once, either at the spectrum level or at the feature level, small but consistent performance gains were obtained (see the second and third rows of Table 6). On the other hand, the use of both the spectral and feature enhancement methods did not

Table 7
Performance contrast of hybrid SI systems with and without Q-CMLLR in addition to dereverberation processing. The third-order delta coefficients were not used. ✗ and √ symbols indicate 'disabled' and 'enabled', respectively.

Q-CMLLR	WER (%)		
	dev	eval	avg
✗	41.9	40.9	41.4
√	**41.4**	**40.1**	**40.8**

provide any performance gains (see the fourth row of the table). This is probably because FE-VTS is based on the assumption that each utterance is contaminated by stationary additive and convolutional noise,[9] which is violated by spectral enhancement processing. These results indicate that both spectral and feature enhancements can provide less distorted DNN input features when they are used alone and that they should not be performed simultaneously.

It is worth mentioning that no such negative effect was observed when we augmented the original feature set by adding these two types of enhanced static features (one obtained with IMCRA and the other with FE-VTS) into the original unenhanced feature set (see the bottom row of Table 6). This confirms our argument in Section 4.1.3 that feature set expansion enables multiple enhancement results to be incorporated. Based on this result, we employ this configuration in Section 5, where we examine the combined effects of different approaches.

Our conclusion from the results described above is that speech enhancement techniques substantially improve the performance of DNN-HMM hybrid acoustic models. Such performance gains have not been reported in the previous work (Seltzer et al., 2013). Our results show that speech enhancement techniques are still useful for DNN-based acoustic models at least in tasks that are acoustically similar to AMI.

4.2.2. Feature transformation

The next class of approaches we investigate is feature transformation with Q-CMLLR, which is applied after MVN. The idea is to cluster the utterances into disjoint classes and train and apply global CMLLR transforms associated with each class.

The feature transforms and the associated acoustic model were trained and evaluated in the following way.

1. Cluster the utterances in the training set using MFCC features. Clustering was performed using a 256-component GMM generated by merging the GMMs in the ML GMM-HMM acoustic model. For each utterance, the frame-level log likelihoods of each GMM component were computed without using prior probabilities and averaged over all the frames within the utterance. The index of the GMM component giving the largest average log probability was used as the clustering result.
2. Train a global CMLLR transform for each cluster of the training utterances.
3. Train a DNN on the transformed data set.
4. At test time, each test utterance is assigned to the closest cluster, which the transform associated with was applied to the utterance. The transformed feature vectors were used to perform decoding.

The transformed features were again normalised before being fed into the DNN. The additional computational cost required at run time was very small. In this and the following experiments, the third-order delta coefficients were not used.

Table 7 compares the WERs of the hybrid SI systems with and without Q-CMLLR. We can see that transforming the feature vectors with Q-CMLLR before feeding them into the DNN resulted in certain degrees of improvement on both test sets. This result indicates that utterance-based feature transformation using Q-CMLLR achieves the further elimination of irrelevant feature variations. Other improvements have been obtained by performing clustering with i-vectors (Glembek et al., 2011) in a similar way to Yao et al. (2011). These are beyond the scope of this paper and will be reported separately.

[9] In principle, FE-VTS can deal with varying noise if such noise can be automatically tracked with sufficient accuracy. However, in most cases, this is a difficult task and no solution has yet been established.

4.2.3. Feature set expansion

Finally, we look at feature set expansion, which aims at improving frame-level classification accuracy by using auxiliary features. In this experiment, we considered the following types of features that can be computed with negligible or minor additional cost.

- MFCCs.
- Intra-frame delta coefficients.
- PLP coefficients.
- Gammatone cepstral coefficients (GTCCs).

The first two types of features can be computed with simple linear operations. The MFCCs can be computed simply by multiplying FBANK features with a discrete cosine transform matrix and discarding higher-order cepstral coefficients. In our experiment, we used the 0-12th MFCCs. It should be noted that the FBANK-to-MFCC transformation cannot be subsumed by the first layer of a DNN because of the speaker and meeting-level MVN. Therefore, these two types of features, while being strongly correlated, deliver slightly different information to the DNN and provide complementary information for improving the recognition performance as we will see later. This suggests that we may easily create a different feature vector simply by applying MVN to any linear transformations of the FBANK feature vector. One interesting linear transform would be a filter that emphasises spectral peaks and dips. Such a filter may be created by applying the delta operation along the frequency axis instead of the time axis, resulting in a linear transform with the following structure when using a regression window of five channels:

$$\begin{bmatrix} & & \ddots & & & & & \\ \cdots & c_{-2} & c_{-1} & 0 & c_1 & c_2 & \cdots & \\ & \cdots & c_{-2} & c_{-1} & 0 & c_1 & c_2 & \cdots \\ & & & & & \ddots & & \end{bmatrix}, \tag{7}$$

where $c_i = i/(2\sum_{\theta=1}^{2}\theta^2)$. The first and last FBANK features are replicated to fill the regression window at both ends of a vector. Combined with the speaker and meeting-level MVN, the outputs of this linear transform may provide features in which spectral peak and dip patterns are emphasised in such a way that makes the variations resulting from speaker differences less pronounced. We call these features intra-frame delta coefficients and represent them with the ∇ symbol to emphasise that the delta operation acts within each frame. It is possible to apply the transform twice, resulting in intra-frame double delta coefficients, which we denote by ∇^2.

Instead of using linear transforms, auxiliary information may be generated by using a different coding system such as PLP or a gammatone filter bank. Since these coding systems employ a different frequency warping from the mel-scale warping, the PLP coefficients and GTCCs emphasise spectral contents in different ways from the FBANK and MFCC features. Our system computes the GTCCs as described in Schlüter et al. (2007) by using the gammatone filter bank implementation by Slaney, which is available at http://www.slaney.org/malcolm/pubs.html. Finally, it is possible to further include the temporal delta coefficients of the above-described auxiliary features in the feature set although they provided little performance gain in our experiments.

Table 8 lists the WERs obtained with various feature sets. The use of the linear transformation-based features, i.e., MFCCs or intra-frame delta coefficients, resulted in an absolute improvement of 1.0% in spite of its simplicity. The use of different coding systems also resulted in similar performance improvements; adding PLP coefficients or GTCCs to the feature set provided absolute WER reductions of 0.9% or 1.0%, respectively. Combining more than two different types of features provided little further gain. Because less effort is required when using linear transformation-based features than when using different coding systems, we used the former features in the subsequent experiments.

5. Combined effects of environmental robustness techniques

Having evaluated the individual effects of different robustness approaches, we now describe a front-end processing pipeline that combines different classes of approaches to create a single set of features for DNN training. This front-end

Table 8
WERs of hybrid SI systems with expanded feature sets. The third-order delta coefficients were not used. Dereverberated features were used as inputs into DNNs.

Feature set (#features per frame)	WER (%)		
	Dev	Eval	Avg
FBANK+Δ/Δ^2(72)	41.9	40.9	41.4
+ ∇/∇^2 (120)	40.9	39.8	40.4
+ MFCC (85)	41.1	**39.7**	40.4
+ PLP (85)	40.7	40.3	40.5
+ GTCC (88)	40.8	40.0	40.4
+ MFCC + $\Delta_{\text{MFCC}}/\Delta^2_{\text{MFCC}}$ (111)	40.6	40.2	40.4
+ MFCC + ∇/∇^2 (133)	**40.4**	39.8	**40.1**
+ GTC + PLP (101)	41.1	39.9	40.5

is used to evaluate the combined effects of various approaches with both SI and SAT set-ups. We refer to the resultant feature set as a final feature set and the FBANK-based original feature set as a base feature set.

Fig. 4 shows the pipeline used here, where the components inside the shaded plate constitute the baseline front-end. The proposed front-end functions as follows.

- An observed speech signal is first processed with a linear filter. In the following experiment, a single-microphone dereverberation filter based on WPE was employed to mitigate the effect of reverberation.
- Other speech enhancement algorithms that act in the power spectrum or FBANK domains are applied to extract less distorted static features. The resultant static features are added to the final feature set after being processed with MVN. This approach allows multiple enhancement results to be exploited as explained in Table 6 and the associated text. In our experiments, we performed spectral enhancement and FBANK enhancement based on IMCRA and FE-VTS, respectively.
- Other kinds of features are also added to the feature set. Static MFCCs and intra-frame delta coefficients were used in the following experiment.
- Feature transformation is performed on the MVN-processed FBANK features, where Q-CMLLR and speaker-level global CMLLR were used with the SI and SAT set-ups, respectively.

The front-end processing pipeline shown in Fig. 4 is generic in the sense that various algorithms can be employed to realise the function that each class of approaches is supposed to provide.

5.1. Evaluation with SI set-up

The combined effects of various front-end approaches to environmental robustness were evaluated in an SDM meeting transcription task with the hybrid SI set-up. The same configurations were used as those for the previous experiments except that DNN's hidden layers were enlarged to 2000 units to accommodate multiple types of auxiliary features.

Table 9 shows the recognition performance of various front-end processing stages. We can see that, except for Q-CMLLR, all the approaches resulted in certain degrees of improvement. Linear filtering for dereverberation improved the WER by 1.5% absolute. Adding the MFCCs and intra-frame delta coefficients reduced the WER by 1.0% absolute. By merging this extended feature set with the enhanced static features, a further improvement of 0.6% was obtained. With the combination of all these approaches, the proposed front-end achieved a WER of 39.7%, which outperforms our competitive baseline by 7.24% relative. It is important to note that the performance gains shown in Table 9 are very close to those shown in Tables 5, 6, and 8, where the individual approaches were examined separately. This means that the performance gains from dereverberation, feature set expansion, and noise reduction are mostly additive when these processing steps are arranged as shown in Fig. 4. In other words, this suggests that the front-end of Fig. 4 allows different robustness approaches to be exploited in DNN-based SI systems in such a way that the individual approaches contribute to the improvement.

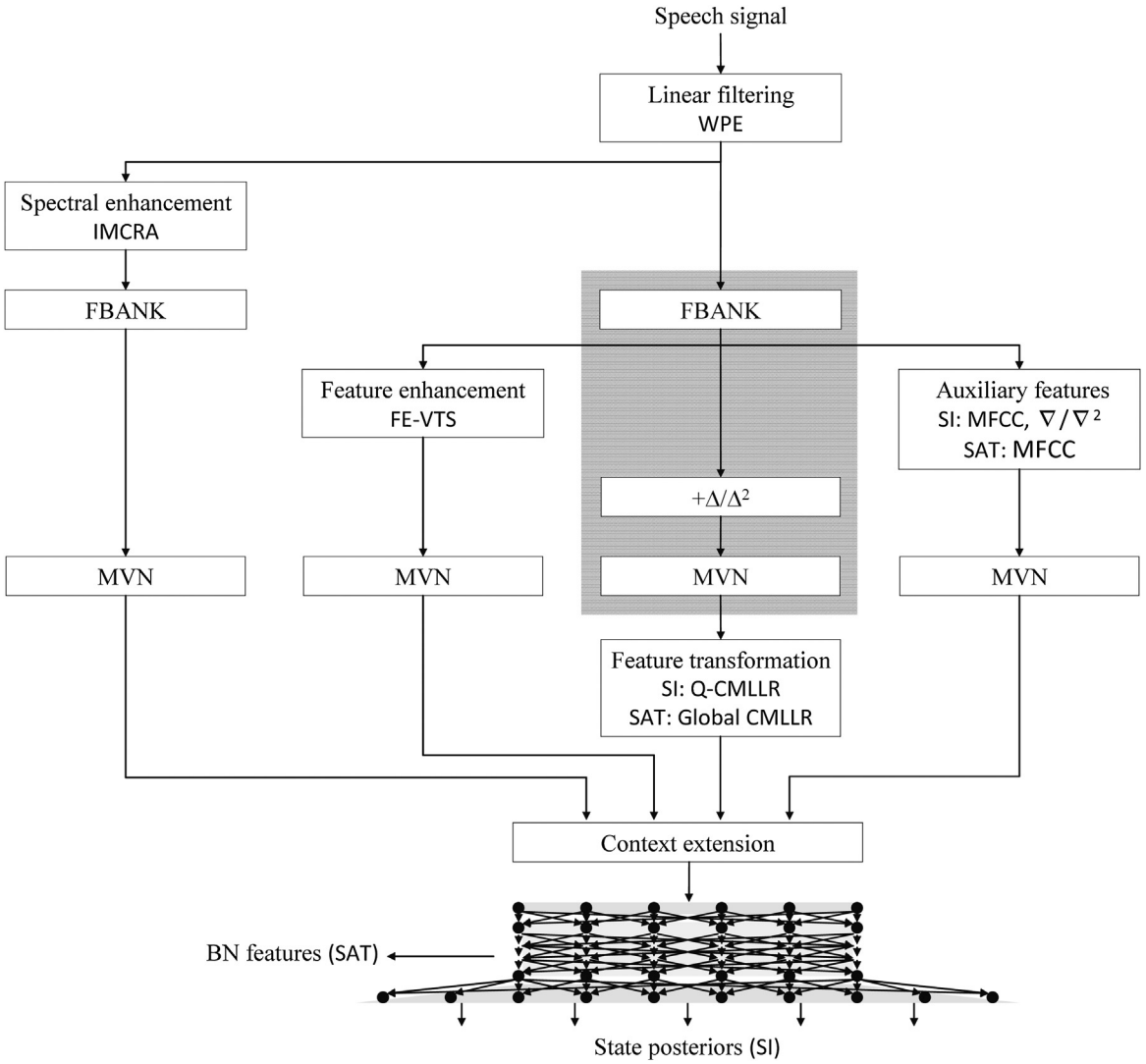

Fig. 4. Front-end processing pipeline incorporating different classes of approaches. Algorithms employed in our experiments are shown in parentheses. The shaded area shows the baseline front-end processing.

Table 9
WERs of various front-end processing stages with hybrid SI acoustic models. 2000 hidden units were used per layer. Third-order delta coefficients were not used.

Front-end	WER (%)		
	Dev	Eval	Avg
FBANK baseline	43.1	42.4	42.8
+ dereverberation	41.8	40.7	41.3
+ MFCC + ∇/∇^2	40.5	40.1	40.3
+ denoising	**40.0**	**39.3**	**39.7**
+ Q-CMLLR	40.9	39.5	40.2

Table 10
WERs of various front-end processing stages with tandem SAT acoustic models. 2000 hidden units were used per layer. Third-order delta coefficients were not used.

Front-end	WER (%)		
	Dev	Eval	Avg
FBANK baseline	40.1	41.3	40.7
+ dereverberation	38.9	39.3	39.1
+ MFCC	38.5	38.5	38.5
+ denoising	38.4	38.7	38.6
+ global CMLLR	**36.6**	**36.7**	**36.7**
+ global CMLLR	36.9	37.0	37.0
+ global CMLLR	38.4	38.6	38.5

With the framework shown in Fig. 4, feature transformation undertaken with Q-CMLLR did not improve the performance when it was combined with other approaches. This could be because normalisation is applied to the base and auxiliary feature sets at different levels. On one hand, Q-CMLLR, applied to the base feature set, groups utterances across speakers and transforms them. On the other hand, the auxiliary features are MVN-processed at the meeting and speaker level. This mismatch in the normalisation level could have caused inconsistency between the base and auxiliary features and thus made the feature vector distribution more confusing. If feature transformation could be consistently performed on the expanded feature set, further performance gains would be expected.

5.2. *Evaluation with SAT set-up*

An integration test was also conducted with the SAT set-up. Again, we used a DNN with five hidden layers, each consisting of 2000 units except for the BN layer. Adaptation supervisions were generated by the baseline SI hybrid system.

In Table 10, the first group of rows (i.e., rows 1–5) shows the WERs of various front-end processing stages. As in the hybrid SI experiments, linear filtering for dereverberation improved the WER by 1.6% absolute. Adding MFCC features to the feature set further reduced the WER by 0.6%. However, unlike in the SI experiments, the use of enhanced features resulted in no improvement, indicating that SAT using BN features subsumed the spectral and feature enhancements. Note that, even with conventional GMM-HMMs, the effect of spectral and feature enhancement processing is very limited in large tasks when SAT is performed (Rennie et al., 2011). On the other hand, feature transformation using speaker-based CMLLR yielded a substantial performance gain of 1.8%. The joint use of linear filtering, feature set expansion, and feature transformation achieved a WER of 36.7%, surpassing our competitive tandem SAT baseline by 9.83% relative.

Further experiments were conducted to see whether there is an overlapping effect between CMLLR feature transformation and other robustness approaches. The results are shown in the second row group in Table 10. The first row in this group (i.e., the sixth row in Table 10) shows the performance of a system using dereverberated and speaker-adapted FBANK features. The second row (i.e., the seventh row in Table 10) shows the recognition performance when only CMLLR feature transformation was used. Comparing the performance figures of these two rows, we can see that linear filtering and speaker-level CMLLR transformation provide gains that are mostly additive, indicating that these two approaches have distinct effects. On the other hand, a comparison of the fifth and sixth rows shows that the use of auxiliary features (MFCCs in this experiment) provided only a minor performance gain. However, it should be noted that this does not necessarily mean that speaker-level CMLLR transformation absorbs the effects of any kinds of auxiliary features. Indeed, very recently, Saon et al. (2013) showed substantial improvements by combining speaker-specific i-vectors and PLP features adapted by global CMLLR and feeding them into a DNN. Thus, it would be possible to improve the performance of tandem SAT systems by using auxiliary features specific to speakers and environments.

We can see that the performance improvement was largely gained from linear filtering (i.e., dereverberation) and global CMLLR. This could imply that static transformations are more effective for DNN-HMM acoustic models than modifying features on a per frame basis.

Finally, we can also see that, even with the best performing configuration, the performance gap was noticeable between the IHM and SDM conditions. This gap will be narrowed to some extent by further advances in acoustic modelling and speech enhancement. For example, use of CNNs was shown to improve the recognition performance in a meeting transcription task (Renals and Swietojanski, 2014) and our internal noisy speech recognition task. More fundamentally, features that are typically used might be insufficient to support speech information in adverse acoustic environments. Peters et al. showed that the human speech recognition performance dropped when a speech signal underwent typical feature extraction steps that filter out phases and reduce frequency resolutions (Peters et al., 1999). Since DNNs are insensitive to the increase in input dimensionality, seeking better acoustic features and associated modelling schemes would be an important future research direction (Sainath et al., 2013).

6. Conclusion

In this paper, we described a front-end processing pipeline for improving the environmental robustness of DNN-based acoustic models, which have recently been greatly changing the nature of acoustic models. First, we showed that it is essential to improve the frame-level classification accuracy if we are to achieve a higher degree of environmental robustness. Then, we classified various robust front-end processing approaches based on their functionality and characteristics. After examining the impact of each class of approaches individually, we described our proposed front-end processing pipeline, which efficiently combines different classes of approaches. By using this front-end, the combined effects of various robustness approaches were investigated with both SI and SAT set-ups in an SDM meeting transcription task.

The conclusions drawn from the series of experiments are as follows.

- All the approaches examined in this paper, i.e., linear filtering, spectral and feature enhancement, feature transformation, and feature set expansion can improve the recognition performance of the DNN acoustic models when they are used alone.
- With the SI set-up, the effects of linear filtering, spectral and feature enhancement, and feature set expansion are almost additive.
- With the SAT set-up, the effects of linear filtering and feature transformation (i.e., FBANK transformation with speaker-level global CMLLR) are additive while the other approaches can yield only minor or no performance gains.

Our experimental results show that the front-end techniques can substantially improve the recognition performance of the DNN acoustic models in adverse acoustic environments and suggest the importance of combining multiple approaches. Some of the front-end techniques examined in this paper yielded significant performance gains over the baseline DNN–HMM hybrid system submitted to the REVERB challenge (http://reverb2014.dereverberation.com/) from the first author's research group. This indicates that our conclusions can be carried over to other tasks that are acoustically similar to AMI. We hope that the front-end pipeline described in the paper would serve as a framework for investigating the interaction of different front-end processing and acoustic modelling approaches, such as CNNs and rectified linear units (Dahl et al., 2013), in the future.

Acknowledgements

Anton Ragni and Xie Chen gave invaluable help with the tools utilised in this work.

References

Abdel-Hamid, O., Mohamed, A., Jiang, H., Penn, G., 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 4277–4280.

Abed-Meraim, K., Moulines, E., Loubaton, P., 1997. Prediction error method for second-order blind identification. IEEE Trans. Signal Process. 45, 694–705.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Breslin, C., Chen, K., Gales, M.J.F., Knill, K., 2011. Integrated online speaker clustering and adaptation. In: Proc. Interspeech, pp. 1085–1088.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Reidsma, D., Wellner P, W.P., 2006. The AMI meeting corpus: a pre-announcement. In: Proceedings of International Workshop on Machine Learning for Multimodal Interaction, pp. 28–39.

Chang, S.Y., Meyer, B.T., Morgan, N., 2013. Spectro-temporal features for noise-robust speech recognition using power-law nonlinearity and power-bias subtraction. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 7063–7067.

Cohen, I., 2002. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process. Lett. 9, 113–116.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. 11, 466–475.

Dahl, G.E., Sainath, T.N., Hinton, G.E., 2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 8609–8613.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 30–42.

de la Torre, A., Peinado, A.M., Segura, J.C., P’erez-Córdoba, J.L., Benítez, M.C., Rubio, A.J., 2005. Histogram equalization of speech representation for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 13, 355–366.

Deng, L., Droppo, J., Acero, A., 2004. Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. IEEE Trans. Speech Audio Process. 12, 133–143.

Deng, L., Li, J., Huang, J.T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., Acero, A., 2013. Recent advances in deep learning for speech research at Microsoft. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 8604–8608.

Deng, L., Wu, J., Droppo, J., Acero, A., 2005. Analysis and comparison of two speech feature extraction/compensation algorithms. IEEE Signal Process. Lett. 12, 477–480.

Droppo, J., Acero, A., 2008. Environmental robustness. In: Benesty, J., Sondhi, M.M., Huang, Y. (Eds.), Springer Handbook of Speech Processing. Springer, pp. 653–679.

Droppo, J., Acero, A., Deng, L., 2001. Evaluation of the SPLICE algorithm on the Aurora2 database. In: Proc. Eurospeech, pp. 217–220.

Du, J., Dai, L.R., Huo, Q., 2014. Synthesized stereo mapping via deep neural networks for noisy speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 1764–1768.

Ephraim, Y., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33, 443–445.

ETSI ES 202 050 Ver. 1.1.5, 2005. Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12, 75–98.

Gales, M.J.F., 1999. Semi-tied covariance matrices for hidden Markov models. IEEE Trans. Speech Audio Process. 7, 272–281.

Gales, M.J.F., 2000. Cluster adaptive training of hidden Markov models. IEEE Trans. Speech Audio Process. 8, 417–428.

Gales, M.J.F., Flego, F., 2012. Model-based approaches for degraded channel modelling in robust ASR. In: Proc. Interspeech.

Geiger, J.T., Weninger, F., Gemmeke, J.F., Wöllmer, M., Schller, B., Rigoll, G., 2014. Memory-enhanced neural networks and NMF for robust ASR. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 1037–1046.

Glembek, O., Burget, L., Matějka, P., Karafiát, M., Kenny, P., 2011. Simplification and optimization of i-vector extraction. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 4516–4519.

Grezl, F., Karafiat, M., Kontar, S., Cernocky, J., 2007. Probabilistic and bottle-neck features for LVCSR of meetings. In: Proc. Int. Conf. Acoust., Speech, Signal Process., IV-757–IV-760.

Hain, T., Burget, L., Dines, J., Garner, P.N., Grezl, F., El Hannani, A., Huijbregts, M., Karafiát, M., Lincoln, M., Wan, V., 2012. Transcribing meetings with the AMIDA systems. IEEE Trans. Audio Speech Lang. Process. 20, 486–498.

Hermansky, H., Ellis, D., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMM systems. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 1635–1638.

Hilger, F., Ney, H., 2006. Quantile based histogram equalization for noise robust large vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 14, 845–854.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kinsgbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 29, 82–97.

Kalini, O., Seltzer, M.L., Droppo, J., Acero, A., 2010. Noise adaptive training for robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. 18, 1889–1901.

Kameoka, H., Nakatani, T., Yoshioka, T., 2009. Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 45–48.

Kim, C., Sterm, R.M., 2012. Power-normalized cepstral coefficients (pncc) for robust speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 4101–4104.

Knill, K.M., Gales, M.J.F., Rath, S.P., Woodland, P.C., Zhang, C., Zhang, S.X., 2013. Investigation of multilingual deep neural networks for spoken term detection. In: Proc. Workshop on Automatic Speech Recognition and Understanding, pp. 138–143.

Krueger, A., Haeb-Umbach, R., 2010. Model-based feature enhancement for reverberant speech recognition. IEEE Trans. Audio Speech Lang. Process. 18, 1692–1707.

Lebart, K., Boucher, J.M., Denbigh, P.N., 2001. A new method based on spectral subtraction for speech dereverberation. Acta Acust. Unit. Acust. 87, 359–366.

Li, B., Sim, K.C., 2010. Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. In: Proc. Interspeech, pp. 526–529.

Li, B., Sim, K.C., 2013. Noise adaptive front-end normalization based on vector Taylor series for deep neural networks in robust speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 7408–7412.
Li, B., Sim, K.C., 2014. An ideal hidden-activation mask for deep neural networks based noise-robust speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 200–204.
Liao, H., 2013. Speaker adaptation of context dependent deep neural networks. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 7947–7951.
Lu, L., Chin, K.K., Ghoshal, A., Renals, S., 2013. Joint uncertainty decoding for noise robust subspace Gaussian mixture models. IEEE Trans. Audio Speech Lang. Process. 21, 1791–1804.
Macho, D., Mauuary, L., Noé, B., Cheng, Y.M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., Saadoun, F., 2002. Evaluation of a noise-robust DSR front-end on AURORA databases. In: Proc. Int. Conf. Spoken Language Process, pp. 17–20.
Mohamed, A., Dahl, G.E., Hinton, G., 2012a. Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. 20, 14–22.
Mohamed, A., Hinton, G., Penn, G., 2012b. Understanding how deep belief networks perform acoustic modelling. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 4273–4276.
Moreno, P.J., Raj, B., Stern, R.M., 1996. A vector Taylor series approach for environmental-independent speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 733–736.
Morgan, N., Bourlard, H., 1995. Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach. IEEE Signal Process. Mag. 12, 24–42.
Narayanan, A., Wang, D., 2014. Joint noise adaptive training for robust automatic speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 2523–2527.
Neto, J.P., Martins, C., Almeida, L.B., 1996. Speaker-adaptation in a hybrid HMM–MLP recognizer. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 3382–3385.
Park, J., Diehl, F., Gales, M.J.F., Tomalin, M., Woodland, P.C., 2011. The efficient incorporation of MLP features into automatic speech recognition systems. Comput. Speech Lang. 25, 519–534.
Peters, S.D., Stubley, P., Valin, J.M., 1999. On the limits of speech recognition in noise. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 365–368.
Plahl, C., Schlüter, R., Ney, H., 2011. Improved acoustic feature combination for LVCSR by neural networks. In: Proc. Interspeech, pp. 1237–1240.
Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H., Zweig, G., 2005. FMPE: Discriminatively trained features for speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 961–964.
Renals, S., Morgan, N., Bourlard, H., Cohen, M., Franco, H., 1994. Connectionist probability estimators in HMM speech recognition. IEEE Trans. Speech Audio Process. 2, 161–174.
Renals, S., Swietojanski, P., 2014. Neural networks for distant speech recognition. In: Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays, pp. 172–176.
Rennie, S., Dognin, P., Fousek, P., 2011. Robust speech recognition using dynamic noise adaptation. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 4592–4595.
Sainath, T.N., Kingsbury, B.r., Mohamed, A., Ramabhadran, B., 2013. Learning filter banks within a deep neural network framework. In: Proc. Workshop on Automatic Speech Recognition and Understanding, pp. 297–302.
Saon, G., Soltau, H., Nahamoo, D., Picheny, M., 2013. Speaker adaptation of neural network acoustic models using i-vectors. In: Proc. Workshop on Automatic Speech Recognition and Understanding, pp. 55–59.
Schlüter, R., Bezrukov, I., Wagner, H., Ney, H., 2007. Gammatone features and feature combination for large vocabulary speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process., IV-649–IV-652.
Seide, F., Li, G., Chen, X., Yu, D., 2011. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. Workshop on Automatic Speech Recognition and Understanding, pp. 24–29.
Seltzer, M.L., Yu, D., Wang, Y., 2013. An investigation of deep neural networks for noise robust speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 7398–7402.
Shinohara, Y., Akamine, M., 2009. Bayesian feature enhancement using a mixture of unscented transformations for uncertainty decoding of noisy speech. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 4569–4572.
Stolcke, A., 2011. Making the most from multiple microphones in meeting recordings. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 4992–4995.
Stouten, V., 2006. Robust automatic speech recognition in time-varying environments. Katholieke University Leuven, Ph.D. Thesis.
Swietojanski, P., Ghoshal, A., Renals, S., 2013. Hybrid acoustic models for distant multichannel large vocabulary speech recognition. In: Proc. Workshop on Automatic Speech Recognition and Understanding, pp. 285–290.
Tashev, I., 2009. Sound Capture and Processing: Practical Approaches. Wiley.
Thomas, S.S., Ganapathy, S., Hermansky, H., 2008. Recognition of reverberant speech using frequency domain linear prediction. IEEE Signal Process. Lett., 681–684.
Wang, Y., Gales, M.J.F., 2012. Speaker and noise factorization for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 2149–2158.
Weninger, F., Watanabe, S., Tachioka, Y., Schuller, B., 2014. Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 4656–4660.
Yao, K., Gong, Y., Liu, C., 2011. A feature space transformation method for personalization using generalized i-vector clustering. In: Proc. Interspeech.
Yao, K., Yu, D., Seide, F., Su, H., Deng, L., Gong, Y., 2012. Adaptation of context-dependent deep neural networks for automatic speech recognition. In: Proc. IEEE Workshop on Spoken Language Technology, pp. 366–369.

Yoshioka, T., Chen, X., Gales, M.J.F., 2014. Impact of single-microphone dereverberation on DNN-based meeting transcription systems. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 5527–5531.

Yoshioka, T., Nakatani, T., 2012. Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. IEEE Trans. Audio Speech Lang. Process. 20, 2707–2720.

Yoshioka, T., Nakatani, T., 2013. Noise model transfer: novel approach to robustness against nonstationary noise. IEEE Trans. Audio Speech Lang. Process. 21, 2182–2192.

Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., Kellermann, W., 2012. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process. Mag. 29, 114–126.

Yoshioka, T., Tachibana, H., Nakatani, T., Miyoshi, M., 2009. Adaptive dereverberation of speech signals with speaker-position change detection. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 3733–3736.

Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C., 2009. The HTK Book version 3.4.1. Cambridge University Engineering Department, Cambridge, UK.

Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., Acero, A., 2008. Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor. IEEE Trans. Audio Speech Lang. Process. 16, 1061–1070.

Yu, D., Yao, K., Su, H., Li, G., Seide, F., 2013. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In: Proc. Int. Conf. Acoust., Speech, Signal Process, pp. 7893–7897.