Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 41 (2017) 180–194

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Acoustic adaptation to dynamic background conditions with asynchronous transformations

Oscar Saz *, Thomas Hain

Speech and Hearing Group, University of Sheffield, 211 Portobello St., Sheffield S1 4DP, UK

Received 11 January 2016; received in revised form 23 May 2016; accepted 25 June 2016
Available online 4 July 2016

Abstract

This paper proposes a framework for performing adaptation to complex and non-stationary background conditions in Automatic Speech Recognition (ASR) by means of asynchronous Constrained Maximum Likelihood Linear Regression (aCMLLR) transforms and asynchronous Noise Adaptive Training (aNAT). The proposed method aims to apply the feature transform that best compensates the background for every input frame. The implementation is done with a new Hidden Markov Model (HMM) topology that expands the usual left-to-right HMM into parallel branches adapted to different background conditions and permits transitions among them. Using this, the proposed adaptation does not require ground truth or previous knowledge about the background in each frame as it aims to maximise the overall log-likelihood of the decoded utterance. The proposed aCMLLR transforms can be further improved by retraining models in an aNAT fashion and by using speaker-based MLLR transforms in cascade for an efficient modelling of background effects and speaker. An initial evaluation in a modified version of the WSJCAM0 corpus incorporating 7 different background conditions provides a benchmark in which to evaluate the use of aCMLLR transforms. A relative reduction of 40.5% in Word Error Rate (WER) was achieved by the combined use of aCMLLR and MLLR in cascade. Finally, this selection of techniques was applied in the transcription of multi-genre media broadcasts, where the use of aNAT training, aCMLLR transforms and MLLR transforms provided a relative improvement of 2–3%.

Keywords: Speech recognition; Acoustic adaptation; Factorisation; Dynamic background; Media transcription

1. Introduction

Complex and dynamic acoustic backgrounds usually cause significant loss of performance on Large Vocabulary Continuous Speech Recognition (LVCSR) systems in many scenarios. Research has focused mostly on situations where the background is stationary or, at least, synchronous with the speech, following the assumption that the characteristics of the background noise remain unchanged through each utterance to decode. Multiple techniques, designed for ASR systems based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), have been

* Corresponding author at: Speech and Hearing Group, University of Sheffield, 211 Portobello St., Sheffield S1 4DP, UK. Fax: +44 (0) 114 222 1810.
E-mail addresses: o.saztorralba@sheffield.ac.uk (O. Saz).

http://dx.doi.org/10.1016/j.csl.2016.06.008

reported to provide solid improvements in ASR tasks (Li et al., 2014). These techniques can be categorised depending on whether they operate in the acoustic space, the feature space or in the model space.

Acoustic-based techniques aim to remove the background noise in the audio via some speech enhancement techniques like Wiener filtering or Minimum Mean-Square Error (MMSE) (Ephraim and Malah, 1984, 1985). Several works have reported significant improvement in recognition rates on benchmark tasks using such techniques (Astudillo et al., 2009; Paliwal et al., 2010). In a similar approach are techniques based on missing features that aim to reconstruct the clean speech signal from the input noisy signal, also with successful results (Cooke et al., 2001). More recently, techniques based on exemplars and Non-negative Matrix Factorisation (NMF) have provided also substantial gains in several tasks (Raj et al., 2010; Schuller et al., 2010).

Techniques in the feature space aim to enhance or transform the input features in order to reduce the mismatch with the GMM–HMM model used for decoding. These include Stereo-based Piecewise Linear Compensation for Environment (SPLICE) (Droppo et al., 2001) or Multi-Environment Model-based Linear Normalization (MEMLIN) (Buera et al., 2007), which have been successfully employed in conventional benchmarks for robust ASR. Another well-known technique is Constrained Maximum Likelihood Linear Regression (CMLLR) (Gales, 1998), which has been widely used to reduce variability caused by multiple sources, like speaker or background.

Model space techniques aim to re-estimate and adapt the parameters of the GMM–HMM model used for recognition. Methods like Parallel Model Combination (PMC) (Gales and Young, 1996) or Vector Taylor Series (VTS) (Moreno et al., 1996) have been especially targeted to speech recognition in noisy environments, while adaptation techniques like Maximum a Posteriori (Gauvain and Lee, 1994) and Maximum Likelihood Linear Regression (MLLR) (Gales and Woodland, 1996) have been used for adaptation to different speakers or different background conditions. Other types of model-based methods are adaptive training regimes, where the parameters of the GMM–HMM are re-estimated jointly with the parameters of some of the previously mentioned techniques. In Speaker Adaptive Training (SAT), for instance, MLLR transforms trained from a set of target speakers are used to update the model parameters (Anastasakos et al., 1996). Extending this, other types of adaptive training regimes have been used for adaptation to the background effects (Kalinli et al., 2010; Liao and Gales, 2007).

The assumption of stationarity and synchrony of the background noise is true for corpora such as NOISEX (Varga and Steeneken, 1993) or Aurora (Hirsch and Pearce, 2000), traditional benchmarks for noise adaptation and compensation techniques. These corpora were generated by adding noise to clean speech signals. This process guaranteed that a single type of noise was added to each utterance. However, in naturally occurring audio, the assumption of stationarity is often not valid. Non-stationary background effects, such as music or overlapping speech, can be common in many tasks, including the multimedia domain and meeting recognition. Furthermore, acoustic background conditions can, by nature, be independent, and hence asynchronous to the target speaker. Typical examples of asynchronous acoustic events can be applause, laughter or door slamming. The common feature of both non-stationary and asynchronous events is that their acoustic properties are not tied to the beginning and end of a speaker utterance; hence, modelling them as a single static environment does not have to be optimal.

The work in this paper aims to deal with asynchrony and non-stationarity of the background in ASR tasks, performing a thorough evaluation of the benefits that an explicit modelling of non-stationary backgrounds can provide. The initial technique will be asynchronous CMLLR (aCMLLR) transforms, which will provide adaptation to dynamic acoustic backgrounds in the feature space. This work will be then expanded with two further techniques: an asynchronous Noise Adaptive Training (aNAT) regime will be defined to provide asynchronous adaptation in the model space; and factorisation using cascading aCMLLR and MLLR transforms following work in Seltzer and Acero (2011, 2012). The proposed techniques will be evaluated in state-of-the-art GMM–HMM systems, with acoustic front-ends like Perceptual Linear Predictive (PLP) features (Hermansky, 1990), and Deep Neural Network (DNN)-front-ends like bottlenecks features (Grezl and Fousek, 2008; Liu et al., 2014). This paper expands and describes a common framework for the techniques briefly introduced in Saz and Hain (2013) and Saz et al. (2015).

This paper is organised as follows: Section 2 will introduce a novel technique to perform asynchronous background adaptation with feature transforms. Section 3 will describe the two extensions to this method for adaptive training and factorisation. Section 4 will evaluate the proposed techniques with asynchronous transforms in a controlled scenario with WSJCAM0, and Section 5 will provide the results on the automatic transcription of Multi-Genre Broadcasts (MGB). Finally, Section 6 will present the conclusions to this work.

2. Asynchronous background adaptation with feature transforms

Constrained Maximum Likelihood Linear Regression (CMLLR) (Gales, 1998) is an adaptation technique initially defined for adapting GMM-based HMMs to a specific speaker, where the same linear transform is applied to both the means and the covariances of the GMM. Due to this property, CMLLR can be equally interpreted as a linear transform applied directly to the input feature vector, which is very useful in many practical situations. The linear transform is given by a transform matrix (A) and a bias vector (b), which are estimated from the data of the desired speaker. Modelling then uses the transformed feature vectors y, as given by Equation 1, to perform decoding.

$$y = Ax + b \tag{1}$$

For a speaker *spk*, the pair of transform matrix and bias vector $W_{spk} = \{A_{spk}, b_{spk}\}$ is further referred to as the CMLLR transform for that speaker. Such a transform can also be trained on several utterances from different speakers in a given acoustic background *bck*, for the purpose of background adaptation: $W_{bck} = \{A_{bck}, b_{bck}\}$. As in all the family of MLLR-based adaptation techniques, CMLLR can be used for supervised adaptation, with manually transcribed data, or for unsupervised adaptation, using the output of an initial recognition stage. Since adaptation data are usually sparse, CMLLR and MLLR techniques use regression classes in order to cluster model parameters together. All phonemes or acoustic units within a regression class will share the same transform, and the number of regression classes can be optimised based on the amount of data available.

For the purpose of background adaptation, it is usually required to have a priori knowledge of the acoustic background that is present in every given utterance. If this information is not available an assumption has to be made as to the background of a given utterance, either by known context or by an initial system for background detection and classification. Furthermore, CMLLR applies the same linear transform throughout the whole utterance, implicitly assuming that the background has stationary properties during the utterance. The use of CMLLR on non-stationary backgrounds or backgrounds whose conditions change asynchronously with the target speech will result in suboptimal modelling and a loss in the potential improvement provided by the adaptation.

This paper proposes the use of asynchronous CMLLR (aCMLLR) transforms. In the aCMLLR framework, the transform applied to the input feature vector $x(t)$ for each frame is different, as in Equation 2, with the objective of producing a frame-by-frame adaptation to the acoustic background present across the utterance.

$$y(t) = A_{bck}(t)\,x(t) + b_{bck}(t) \tag{2}$$

The implementation of Equation 2 is complex, as it would require a continuous space of background transforms that would optimise the search for each frame t, so a set of constraints and assumptions has to be made to develop an implementation of this technique. The specifics of performing speech recognition with aCMLLR transforms as well as the training regime of the transforms are described next.

2.1. Decoding with aCMLLR transforms

In order to develop an implementation of Equation 2, the first assumption will be regarding the set of transforms available. This work will assume that there is a finite number, N, of previously trained CMLLR transforms to apply ($W_{bck}^1, \ldots, W_{bck}^n, \ldots, W_{bck}^N$). Under this constraint, Equation 2 is simplified into Equation 3:

$$y(t) = A_{bck}^{c(t)}\,x(t) + b_{bck}^{c(t)} \tag{3}$$

where $c(t)$ is the index of the pre-existing transform W_{bck}^n, which better compensates the background for the acoustic frame $x(t)$. The problem in Equation 2 has been simplified to finding the optimal sequence c of backgrounds for each frame from the pool of N existing backgrounds.

One approach to perform this search could be to use a system that would classify each frame $x(t)$ as being contaminated by one of the N valid background conditions. In order to train such classification system, it is required the existence of sufficient data where the acoustic background has been fully annotated at a frame level. Semi-supervised and unsupervised training techniques could overcome the lack of such data. Even a classification system

with good classification performance might not produce significant improvements in recognition rates, mainly because the objective function of classification, namely maximise frame classification rate, does not match the objective function of the decoder, which is to maximise the overall likelihood.

An alternative solution will be used in this paper to overcome these limitations. This approach uses Viterbi decoding as an on-line framewise classifier, setting as only objective in the whole process the overall maximisation of the likelihood during the Viterbi search. A new HMM topology is proposed for this as shown in Fig. 1. This figure presents model structure changes with two possible background transforms, W^1_{bck} and W^2_{bck}, but it easily generalises to any number, N, of backgrounds. The usual three-state left-to-right HMM topology with an entry and an exit state is modified by creating two sequences of states from entry to exit. The HMM states in the upper and lower branches share the same GMMs (state 1 with state 4, state 2 with state 5, etc.), but these are modified by two different CMLLR transforms. While states 1, 2 and 3 are transformed by W^1_{bck}; states 4, 5 and 6 are transformed by W^2_{bck}. Additional transitions between the upper and lower paths are included in order to allow the ability to change from one background transform to the other with each new input frame.

With the topology in Fig. 1 decoding can be done following the Maximum Likelihood (ML) criterion and the Viterbi decoding algorithm. On the optimal path, each frame will be transformed by the background transform that provides the highest increase in the overall likelihood. Fig. 1 can be slightly modified by removing the transitions between the upper and lower branches. In this case, the same CMLLR transform will be applied to the whole phonetic unit, and changes from one background to another will only be allowed when the phonetic unit changes. This possibility will be referred to as *phone synchronous*, while the original one in Fig. 1 will be referred to as *fully asynchronous*. Given that this decoding topology involves applying different CMLLR transform across time, the Jacobian of each transform must be included in the likelihood calculation, as it is the case when using multiple regression classes in standard CMLLR adaptation. Given the expansion in the model structure shown in Fig. 1, an increase in computation can be expected, which without further optimisation can limit the use of such model in tasks where speed is required, such as online decoding.

This frame-by-frame asynchronous decoding is similar to the proposal for online Vocal Tract Length Normalization (VTLN) in Miguel et al. (2008). In that work the model space was augmented to consider different VTLN warping values, and the decoder automatically chose the path that maximised the total likelihood through the augmented space. This approach was shown to improve over traditional static VTLN (Lee and Rose, 1998). Furthermore, the presented topology can be seen similar to approaches for decomposing speech and noise in HMMs (Varga and Moore, 1990) or to Subspace Gaussian Mixture Models (SGMMs) (Povey et al., 2011a) where different subspaces can be trained to model different background conditions. The proposal in this paper is more flexible than these two cases, which required to retrain the GMMs in order to cover new backgrounds, while the use of CMLLR transforms requires less acoustic data for adaptation and is more modular to include new background conditions.

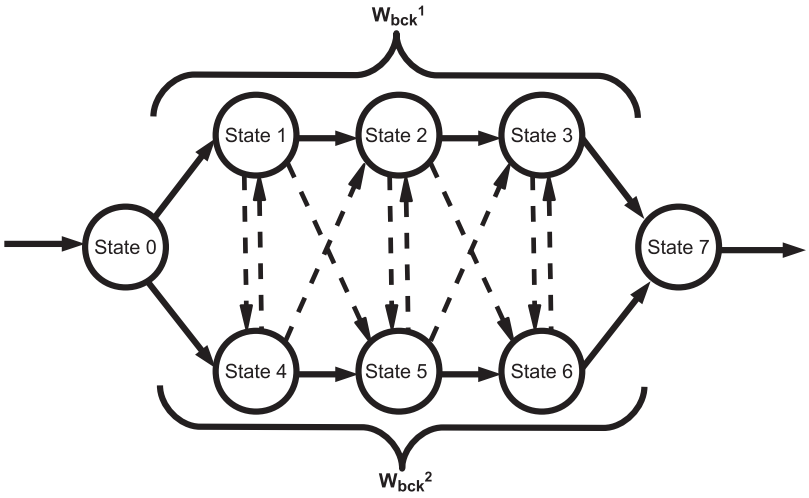


Fig. 1. Asynchronous decoding with two background transforms (W^1_{bck} and W^2_{bck}). State self transitions have been removed for clarity.

2.2. Training of aCMLLR transforms

So far the discussion has concentrated on recognition in an ML framework. However this does not immediately allow to infer appropriate training regimes. Usually, transforms are trained following the same assumption that each utterance presents a stationary acoustic background. This way, all the frames from the utterances that share the same background will be used in the estimation of the transform W_{bck}. This is not the optimal way for estimation of the transforms in the case of non-stationary noise, as not at all the frames in an utterance will share the same background.

The same topology presented in Fig. 1 can be used when learning transforms from adaptation data. Supposing an initial set of background CMLLR transforms exists, trained in a synchronous manner, this topology allows alignment of the input speech to the best sequence of states, as in regular CMLLR training, but also provides the optimal sequence of backgrounds for each frame. With this alignment information, a new set of transforms can be updated from the training data statistics.

The complete procedure for training of aCMLLR transforms can then be summarised as follows: All the training utterances will be separated into N classes according to their acoustic background assuming that this is invariant for the utterance. Then, N background CMLLR transforms will be trained on the adaptation data. The data are then pooled and the CMLLR transforms are used to produce an asynchronous alignment to these data with the topology in Fig. 1. Finally, the aCMLLR transforms are then re-estimated with all the frames from the complete set of utterances which have been aligned to each of the asynchronous transforms. This re-estimation process can be iterated in order to produce better alignments and obtain a better estimation of the transforms. While this approach requires a certain knowledge of the backgrounds in the training data to initialise the transforms, it will be seen how loose and unreliable information can be sufficient to perform this procedure.

3. Extensions to the aCMLLR background adaptation

3.1. Asynchronous Noise Adaptive Training

As in other adaptation techniques using (C)MLLR transforms, an adaptive training setup can be implemented using aCMLLR transforms. This asynchronous Noise Adaptive Training (aNAT) is performed following the same procedure as described in the literature (Anastasakos et al., 1996; Kalinli et al., 2010; Liao and Gales, 2007). After the full procedure for aCMLLR training is performed, alignment of the train data can be done using the asynchronous topology in Fig. 1. Again, statistics for each state are collected and a full retraining of the GMM–HMM parameters is performed. At this stage, the Gaussians belonging to the states in the parallel paths in Fig. 1 are not shared anymore, as they will have been retrained on new statistics. Following this, new aCMLLR transforms are trained on top of the aNAT model and decoding is done as described previously.

3.2. Factorisation of asynchronous background and speaker

Earlier in GMM–HMM-based systems, factorisation of the different sources of variability has been proposed as a way to further improve the performance of ASR systems in varying conditions (Gales, 2001). Typically the sources of variability considered for factorisation approaches are the speaker variability against the environmental factors, including channel and background conditions. Techniques based on joint factorisation of sources of variability, for example Joint Factor Analysis (JFA) (Yin et al., 2007), as used in speaker verification tasks, are now being considered for use in ASR tasks. SGMMs aim to incorporate the factorisation directly in the HMM topology (Povey et al., 2011a) via the use of subspace models. Other approaches are based on jointly combining transforms for the speakers and the environments. This has been done by combining Vector Taylor Series (VTS) and MLLR transforms in Wang and Gales (2011), CMLLR transforms in Seltzer and Acero (2011), and CMLLR and MLLR transforms in Seltzer and Acero (2012). Factorisation techniques based in eigenspace MLLR adaptation have also been studied recently (Saz and Hain, 2014).

Seltzer and Acero (2011) propose factorisation by means of CMLLR transforms applied in cascade. Here background transforms W_{bck} are trained for every possible background and across all speakers. Further speaker transforms W_{spk} are trained on top of the background transforms for each speaker across all backgrounds. In decoding stage, for

every input utterance x spoken by speaker *spk* in background *bck* the feature vectors are transformed to y as in Equation 4.

$$y = A_{spk}\left(A_{bck}x + b_{bck}\right) + b_{spk} \tag{4}$$

Using both transforms in cascade was shown to improve results over conventional CMLLR adaptation on environment and speaker. Also, the speaker transforms, which had been estimated on feature frames already adapted to the environment, were shown to perform well when used across previously unseen backgrounds. It is straightforward to generalise this proposal to deal with non-stationary and asynchronous backgrounds following Equation 5:

$$y(t) = A_{spk}\left(A_{bck}^{c(t)}x(t) + b_{bck}^{c(t)}\right) + b_{spk} \tag{5}$$

Similar factorisation can be achieved following the work in Seltzer and Acero (2012) using CMLLR and MLLR transforms. In this case, a background-based CMLLR transform is used to transform the input features, while a factorised speaker-based MLLR transform is used to transform the models. Using asynchronous adaptation on the CMLLR transforms, it is also straightforward to provide factorisation with cascading aCMLLR/MLLR transforms. Further work in factorised adaptation for ASR has argued the need for the components of the factorisation to be orthogonal (Seo et al., 2014) in order to avoid correlation between the different factorised elements. Although the use of feature-space and model-space MLLR transforms does not involve orthogonality, Seltzer and Acero (2012) argue that it achieves a better factorisation between the different variability factors in speech than, for instance, using cascading feature-space transforms.

4. Benchmark results: WSJCAM0

An initial evaluation of the proposed techniques was performed on a modified version of the WSJCAM0 corpus. WSJCAM0 is a re-recording of the original WSJ sentences uttered by a collection of British speakers (Robinson et al., 1995). WSJ is a very common benchmark for the evaluation of acoustic modelling techniques and speaker adaptation tasks in ASR (Paul and Baker, 1992) and it was the base for the creation of Aurora4 (Parihar et al., 2004), which is also commonly used as a benchmark in background adaptation for robust ASR. For the purpose of our experiments, 7387 utterances from 86 speakers were used for training, and 1315 utterances from 18 speakers were used for evaluation. Besides this *Clean* data, new train and test sets were generated including highly diverse background conditions, which will be referred to as *Diverse* data. In the *Diverse* data sets, 7 possible background conditions appear with the distribution of segments seen in Table 1 for Train and Test data. The Signal-to-Noise Ratio (SNR) of the new utterances in the *Diverse* data was uniformly distributed from 5 db to 15 dB. Furthermore, in the *Diverse* sets, segments were drawn randomly for the recordings using the close-talking microphone and the table-top microphone, with a 50% distribution of a given sample being from either source. The *Clean* sets corresponded only to the close-talking microphone recordings.

The ASR experiments were performed on two of the original WSJ tasks: the 5000-word closed vocabulary task with a 2-gram Language Model (LM); and the 20,000-word open vocabulary task with a 3-gram LM. The baseline ASR system used was built using a Hidden Markov Model Toolkit (HTK) (Young et al., 2006) setup. Crossword triphone models with 3 states per model and 16 Gaussians per state were used. A total of 1816 physical states were trained using the Maximum Likelihood (ML) criterion. Thirty-nine-dimension feature vectors were used with 13 PLP features (Hermansky, 1990) and their first and second derivatives. Cepstral Mean Normalization (CMN) was applied to the static features for each utterance. The standard WSJ lexicons and language models were used in decoding.

Table 1
Distribution of segments in the *Diverse* data set.

	Clean	Music		Noisy			
		Orchestral	Popular	Traffic	Restaurant	Applause	Outdoors
Train	2,504	1,238	1,249	598	582	630	586
Test	433	226	215	120	101	118	102

Table 2
Baseline recognition results (WER) for modified WSJCAM0.

Train	Test	5K set	20K set	Total
Clean	*Clean*	5.7%	13.0%	9.4%
Clean	*Diverse*	27.0%	39.1%	33.1%
Diverse	*Diverse*	14.9%	25.5%	20.3%

The baseline results in Word Error Rate (WER) achieved with this setup are presented in Table 2. The results in the *Diverse* testset showed considerable decrease in performance compared to the *Clean* testset, 33.1% against 9.5%. This indicated that the background conditions included in the data had a negative influence in ASR performance. The use of models trained on *Diverse* data showed an improvement, reaching 20.3% WER, but performance was still far from the result on *Clean* test data.

4.1. Adaptation experiments

The next set of experiments studied different types of adaptation, speaker and background adaptation, in these data sets. For background adaptation, single-regression-class block-diagonal CMLLR transforms were used, while for speaker adaptation 5-regression-class block-diagonal MLLR transforms were used. Background adaptation was performed in supervised and unsupervised fashion, depending on whether the adaptation was made on data from the training set, with ground truth transcriptions, or from the test set, with errorful transcriptions from ASR. Speaker adaptation was only studied in unsupervised mode, as none of the speakers in the test set appeared in the training set.

The results for these experiments are shown in Tables 3 and 4. Speaker adaptation achieved a 15% relative improvement on the *Clean* test set over *Clean* models, and outperformed background adaptation on the *Diverse* test set for both models. The gains were more pronounced in the mismatched condition with *Clean* models, with background adaptation only providing 3% improvement over *Diverse* models. For the *Clean* models, the use of supervised CMLLR adaptation to the 7 backgrounds using the training set improved the use of unsupervised adaptation.

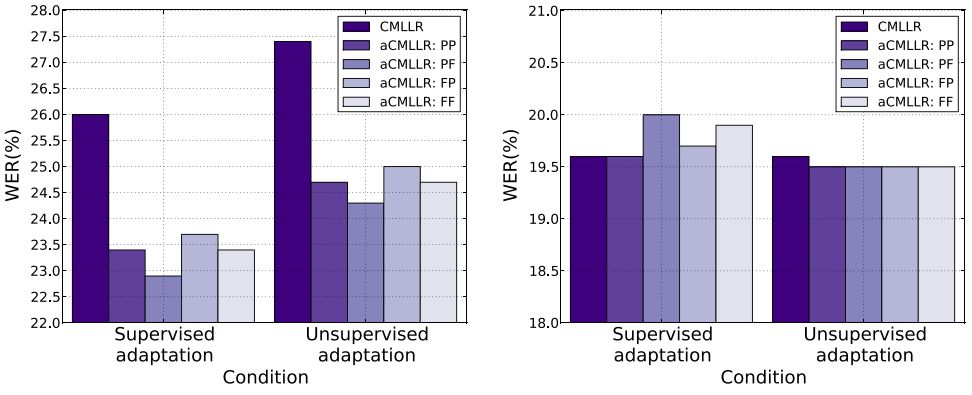
Asynchronous background adaptation was then performed on this setup. The topology used for asynchronous adaptation was based on 7 CMLLR transforms, previously trained in a synchronous fashion for the 7 types of background included in the *Diverse* data. Fig. 2 shows the results for four possible cases depending whether the GMM–HMM models used were trained on *Clean* on *Diverse* data and whether the adaptation was made unsupervised or supervised. Four possible conditions were also defined based on the situations where the topology in training and testing was *Phone synchronous* or *Fully asynchronous* according to the definition explained in Section 2.1. From these results, it was observed that using a *Phone synchronous* topology in training and *Fully asynchronous* in decoding provided the largest improvements. The best result in the condition with *Clean* models was 22.9%, which was an improve-

Table 3
Speaker-based MLLR adaptation results for modified WSJCAM0.

Train	Test	5K set	20K set	Total	Rel. impr.
Clean	*Clean*	4.7%	11.4%	8.1%	14.7%
Clean	*Diverse*	18.9%	30.1%	24.6%	25.7%
Diverse	*Diverse*	13.1%	22.1%	17.6%	13.3%

Table 4
Background-based CMLLR adaptation results for modified WSJCAM0.

Train	Test	Condition	5K set	20K set	Total	Rel. impr.
Clean	*Diverse*	Supervised	20.2%	31.7%	26.0%	21.5%
		Unsupervised	21.3%	33.4%	27.4%	17.2%
Diverse	*Diverse*	Supervised	14.2%	24.9%	19.6%	3.4%
		Unsupervised	14.2%	24.8%	19.6%	3.4%



(a) Clean *train* (b) Diverse *train*

Fig. 2. Results for different topologies in training and decoding of aCMLLR transforms (first letter indicates training topology, second letter indicates testing topology, P stands for *Phone synchronous*, and F stands for *Fully asynchronous* as defined in Section 2.1).

ment of 30.8% over the baseline, 9% better than the synchronous background adaptation. With the use of *Diverse* models, the best result achieved was 19.5%, only 4.0% below the baseline and barely 0.6% better than the case with synchronous background adaptation.

The results in Fig. 2 showed a divergence in the performance of the aCMLLR framework in the *Clean* and *Diverse* train conditions. While large gains were achieved with *Clean* models, no significant changes were seen with *Diverse* models. Also, the initial gain for using background-based CMLLR transformations over the baseline was smaller (3.4% relative) in the *Diverse* train conditions. This effect in what is a classical multicondition training scenario might arise from the way the modified WSJCAM0 corpus was built. In this case, the same 7 types of noises were used in train and test, and the baseline GMM–HMM models will have implicitly learnt the characteristics of each background, thus giving little room for improvement for any type of background-based adaptation. This highlights the needs for realistic scenarios in robust ASR, where a variety of backgrounds and noises appear differently in the train and test data.

4.2. Factorisation experiments

Factorisation techniques were then studied, based on the use of CMLLR/MLLR cascading transforms. Both synchronous and asynchronous CMLLR background transforms were used, together with subsequent MLLR speaker transforms. The results in Table 5 show that the WER in the mismatched condition was reduced to 19.7% with the use of asynchronous background transforms (40.5% relative improvement), while synchronous background adaptation only achieved 34.7% relative improvement. However, in the *Diverse* models, asynchronous adaptation produced an increase of 0.3% in WER compared to the synchronous case. These results were consistent in an early exploratory work carried out with a more limited number of background conditions (Saz and Hain, 2013).

A final analysis of these results was done regarding the ability of the proposed method to adapt the *Clean* models to the different types of backgrounds existing in the *Diverse* set. Table 6 shows the results for the baseline and the best result (aCMLLR/MLLR cascade) for the 7 background conditions. The proposed method improved significantly even in the Clean background and in the Outdoors noise background, which had the lowest initial WER. Meanwhile,

Table 5
Factorisation results for modified WSJCAM0.

Train	Test	Adaptation	5K set	20K set	Total	Rel. impr.
Clean	*Diverse*	CMLLR/MLLR	16.9%	26.2%	21.6%	34.7%
		aCMLLR/MLLR	14.6%	24.7%	19.7%	40.5%
Diverse	*Diverse*	CMLLR/MLLR	12.5%	21.5%	17.1%	15.8%
		aCMLLR/MLLR	12.8%	21.9%	17.4%	14.3%

Table 6
Recognition results per background condition in *Clean* trained models for baseline model and aCMLLR/MLLR factorised adaptation.

	Clean	Music		Noisy			
		Orchestra	Popular	Traffic	Outdoors	Restaurant	Applause
Baseline	14.1%	36.9%	48.0%	40.4%	28.9%	37.3%	58.2%
aCMLLR/MLLR	10.6%	20.9%	27.8%	22.0%	17.7%	22.0%	32.1%
Rel. Impr.	33.0%	43.4%	42.1%	45.5%	38.8%	41.0%	44.8%

Table 7
RTFs of the decoding for different model architectures.

	HMM	Phone synchronous HMM	Fully asynchronous HMM
5K task	5.5	22.4	65.2
20K task	12.8	43.9	102.8

the backgrounds with the highest degradation, like popular music, traffic noise or applause, achieved relative improvements of up to 45%.

4.3. Computational complexity

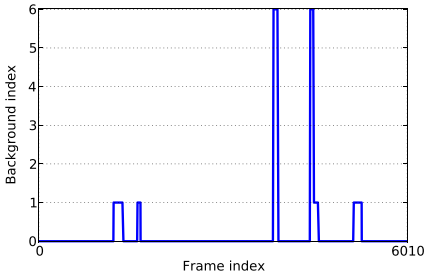
When discussing the new HMM structure shown in Fig. 1, it was expected that computational complexity would increase as the extra paths in the HMM are added to model different backgrounds. To evaluate the extent of such increase, the Real Time Factors (RTF) of the decoding of both WSJCAM0 tasks were studied in 3 cases: Standard HMM structure, phone synchronous HMM structure and fully asynchronous HMM structure. The results are presented in Table 7 and show increases of 4× and 3× in RTF for the phone synchronous structure in the 5k and 20k tasks, respectively, and of 11× and 7× for the fully asynchronous structure. All results were achieved in similar conditions, where the decoding of the test set was submitted as an array job to a computing grid using the OpenSGE grid scheduler. The actual physical machines in the grid where the processes ran had 32 hyper-threaded Intel Xeon cores at 2.6 GHz each.

This indicated that when time is a constraint in the decoding stage, the phone synchronous structure in decoding should be preferred. Although a limitation for some tasks, such as online decoding, the increase in computation time does not impair the proposed method, especially when performance is the main goal. A way of reducing this increase in time can be reducing the number of modelled backgrounds, as this will reduce the number of extra paths added to the HMM structure. Also, according to the measurements in Table 7, the extra increase in decoding time becomes less relevant as the vocabulary size and language model complexity increase, possibly due to the acoustic decoding accounting for less of the actual computation time.

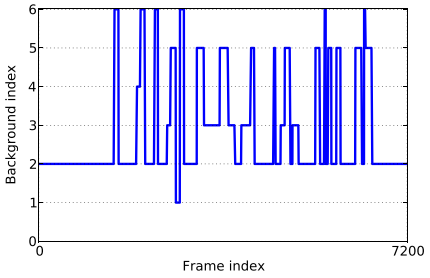
The increase in computation time is linked to the ability of the new HMM structure to switch backgrounds. Fig. 3 presents the actual background paths for two files in the 5K task. Fig. 3(a) and 3(c) shows this for a clean signal using phone synchronous and fully asynchronous HMMs, while Fig. 3(b) and 3(d) shows them for a signal contaminated with music. Using a *Phone synchronous* HMM structure yields a much lower rate of background switching, with only 0.18% and 0.57% of frames being decoded with a different background than their precedent frame. In the *Fully asynchronous* structure, changes occur more frequently, up to 1.16% and 4.28% of the total frames change the background compared to their precedent frame. As the figures show, a more non-stationary background like music produced an increase in background changes across frames.

5. Transcription of multi-genre broadcasts

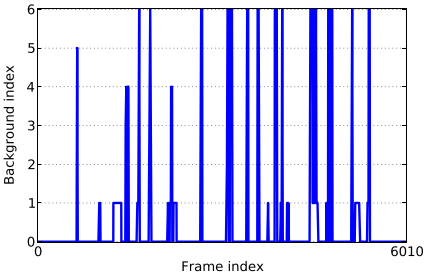
Full evaluation of the proposed techniques was conducted on the data available for Task 1 of the Multi-Genre Broadcast (MGB) challenge, which is speech-to-text transcription of broadcast television (Bell et al., 2015). The MGB challenge aimed to evaluate and improve several speech technology tasks in the area of media broadcasts, extending the



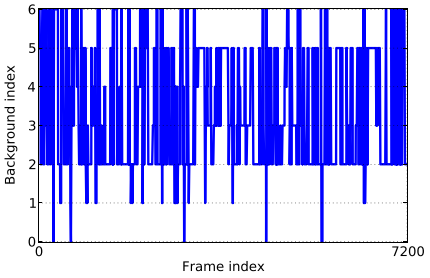
(a) Clean background, phone synchronous HMM



(b) Music background, phone synchronous HMM



(c) Clean background, fully asynchronous HMM



(d) Music background, fully asynchronous HMM

Fig. 3. Frame-wise backgrounds selected in decoding for two 5K task files with *Phone synchronous* and *Fully asynchronous* HMMs with 7 possible backgrounds (indexes 0 to 6).

Table 8
Distribution of train and test data for experimentation with media broadcasts.

Genre	Training			Testing		
	Shows	Audio	Speech	Shows	Audio	Speech
Advice [ADV]	264	193.1 h.	155.2 h.	4	3.0 h.	2.5 h.
Children's [CHI]	415	168.6 h.	112.5 h.	8	3.0 h.	2.0 h.
Comedy [COM]	148	73.9 h.	50.8 h.	6	3.2 h.	2.2 h.
Competition [COP]	270	186.3 h.	142.8 h.	6	3.3 h.	2.8 h.
Documentary [DOC]	285	214.2 h.	149.9 h.	9	6.8 h.	4.6 h.
Drama [DRA]	145	107.9 h.	68.8 h.	4	2.7 h.	1.4 h.
Events [EVE]	179	282.0 h.	206.9 h.	5	4.4 h.	2.2 h.
News [NEW]	487	354.4 h.	309.9 h.	5	2.0 h.	1.8 h.
Total	2193	1580.4 h.	1196.7 h.	47	28.4 h.	19.6 h.

work of other evaluations like Hub4 (Pallett et al., 1996), TDT (Cieri et al., 1999), Ester (Galliano et al., 2006), Albayzin (Zelenak et al., 2012) and MediaEval (Larson et al., 2013). All experiments are based on the official MGB challenge training set and evaluated on the official development set. Table 8 presents the statistics of these data in terms of number of shows, and amount of audio and speech in the training and testing sets. One of the goals of the task was to study recognition performance across diverse broadcast genres, for that reason both training and test data were labelled according to 8 possible genres: advice, children's, comedy, competition, documentary, drama, events and news.

The baseline system configuration used for this task was a DNN–GMM–HMM system with the following setup. A DNN was used as a front-end for extracting a set of 26 bottleneck features. Such DNN took as input 15 contiguous log-filterbank frames and consisted of 4 hidden layers of 1745 neurons plus the 26-neuron bottleneck layer, and an output layer of 8000 triphone state targets. The state-level Minimum Bayes Risk (sMBR) criterion was used as the optimisation criterion and Stochastic Gradient Descent (SGD) was used for parameter updating. DNN training used

the TNet (Vesely et al., 2010) and Kaldi (Povey et al., 2011b) toolkits. The input feature vectors for training the GMM–HMM system were 65-dimensional, including the 26 dimensional bottleneck features, as well as 13 dimensional PLP features together with their first and second derivatives. GMM–HMM models were trained using 16 Gaussian components per state and around 8000 distinct triphone states.

The manual segmentation as provided for the development set was used; no automatic speech segmentation was required for the experiments. However, for speaker adaptation ground truth was not available, so automatic speaker clustering had to be used. The clustering system used was based on the Bayesian Information Criterion (BIC) (Chen and Gopalakrishnan, 1998) and was similar to the one used by the University of Sheffield in the system submitted for the longitudinal diarisation of broadcast television task of the MGB challenge, with a speaker error rate of 41.7% (Milner et al., 2015). The diarisation task proves especially challenging in broadcast data, as shown by the general results achieved by the participating groups in the MGB challenge (Bell et al., 2015). For these experiments single-regression-class block-diagonal CMLLR and 5-regression-class block-diagonal MLLR transforms were used.

The only transcription available for the 1200 hours of training speech was the original BBC subtitles, aligned to the audio data using a lightly supervised approach (Long et al., 2013). Given that this can produce unreliable transcripts for some segments, the training data were filtered and only 700 hours of speech were used for training. For filtering, a segment-level confidence measure was calculated based on posterior estimates obtained with a DNN as in Zhang et al. (2014). Only segments with higher values of confidence were maintained in the training set, with the threshold set to obtain 700 hours of speech.

Decoding was carried out in two stages; in a first stage, lattices were generated using a 2-gram LM; this was followed by rescoring of these lattices using a 4-gram LM and obtaining the 1-best output. The outputs were scored with the official MGB scoring package (Bell et al., 2015), which was based on NIST scoring tools (Fiscus, 2007), and the evaluation done in terms of global WER and genre-specific WERs. Both language models were built using the SRI LM toolkit (Stolcke, 2002) from more than 700 million of words from subtitles provided as material for the challenge. The vocabulary size used in decoding was a 50,000 word list, constructed from the most frequent words in the subtitles provided for language model training. Pronunciations were obtained using the Combilex pronunciation dictionary (Richmond et al., 2010), which was provided to the challenge participants. When a certain word was not contained in the lexicon, automatically generated pronunciations were obtained using the Phonetisaurus toolkit (Novak et al., 2012). These pronunciations were expanded to incorporate pronunciation probabilities, learnt from the alignment of the acoustic training data (Hain, 2005).

5.1. Results

Several experiments were carried out in this setup. All results are presented in Table 9. The initial baseline performance obtained with the unadapted system was 31.0% WER. A large variability in results could be observed across genres, ranging from 16.3% WER for news shows, to 44.7% for comedy shows. Synchronous adaptation was studied first by means of show-specific CMLLR transforms, targeted to capture variability due to the background condition

Table 9
WER and relative improvement for transcription of multi-genre broadcasts with ML-trained models.

	ADV	CHI	COM	COP	DOC	DRA	EVE	NEW	Total
Baseline									
Baseline	25.2%	30.8%	44.7%	27.3%	28.9%	42.1%	34.9%	16.6%	31.0%
Synchronous adaptation									
Show CMLLR	25.2%	30.3%	44.5%	27.2%	28.7%	42.0%	34.3%	16.3%	30.8%
Asynchronous adaptation									
aCMLLR	25.1%	30.6%	44.3%	27.0%	28.8%	42.0%	34.5%	16.3%	30.7%
+ Show aCMLLR	24.8%	30.1%	44.1%	26.9%	28.4%	41.3%	34.3%	16.2%	30.5%
Asynchronous adaptive training									
aNAT	25.0%	30.5%	44.2%	27.3%	28.6%	42.0%	34.6%	16.2%	30.7%
+ Show aCMLLR	24.6%	29.7%	43.7%	26.7%	28.2%	41.4%	34.1%	15.9%	30.3%
+ Speaker MLLR	24.5%	29.4%	43.5%	26.5%	28.2%	41.1%	33.7%	15.9%	30.1%
Rel. impr.	2.8%	4.5%	2.7%	2.9%	2.4%	2.4%	3.4%	4.2%	2.9%

Table 10
WER and relative improvement for transcription of multi-genre broadcasts with MPE-trained models.

	ADV	CHI	COM	COP	DOC	DRA	EVE	NEW	Total
Baseline	23.6%	28.5%	42.3%	25.5%	27.5%	39.7%	32.8%	15.5%	29.2%
aNAT+CMLLR+MLLR	23.1%	28.1%	41.4%	25.2%	27.0%	39.1%	32.0%	15.1%	28.6%
Rel. impr.	2.1%	1.4%	2.1%	1.2%	1.8%	1.5%	2.4%	2.6%	2.0%

in each show. This gave a slight improvement of 0.2%, which showed the difficulty of trying to model the types of backgrounds present in broadcast shows by using traditional approaches.

An initial set of aCMLLR transforms were trained on the training data using the following procedure. First, 8 genre-based CMLLR transforms were trained from the training set. Assuming that each genre will present certain distinct background conditions, these transforms were used as initialisation in the training of a global aCMLLR transform with 8 possible backgrounds. The use of this transform achieved 0.3% absolute improvement over the baseline, also improving the use of the show-based CMLLR transforms. Finally, this global aCMLLR transform was used as initialisation for training show-based aCMLLR transforms on the test shows, adding an extra 0.2% absolute reduction of the WER.

The final set of experiments involved an adaptive retraining of the GMM–HMM parameters following the aNAT procedure. This new model only provided an improvement of 0.3%, similar to using the aCMLLR transforms on the baseline GMM–HMM model. However, training show-based aCMLLR transforms on top of the adaptively trained model boosted the improvement to 0.8% absolute. This showed how adaptive training provided a better flexibility of the model to adapt to specific background conditions existing in each show. Finally, the factorisation approach using MLLR speaker transforms on top of the aNAT model and show-based aCMLLR transforms was tested. This only increased the improvement to 0.9% absolute (2.9% relative), which reflects the difficulty of performing accurate speaker clustering in this task and how this actually hampers speaker adaptation.

Finally, Minimum Phone Error (MPE) training (Povey and Woodland, 2002) was performed in this dataset, since it is well-known to provide significant improvements in GMM–HMM systems with bottleneck features (Grezl et al., 2009). The results of the baseline models and the MPE–aNAT retrained models with aCMLLR and MLLR transformations are presented in Table 10. The use of the MPE training criterion provides a 1.8% absolute improvement over the ML baseline. This gain may seem lower than those usually reported using MPE over ML, but in the MGB tasks the transcriptions of the training set are errorful, which has been reported to produce a decrease in MPE performance (Long et al., 2013). Finally, the use of MPE–aNAT with asynchronous CMLLR and MLLR reduces the WER to 28.6%. Although the relative gain over the MPE baseline is reduced to 2.0% (from 2.8% using ML models), this was found to still be significant.

6. Conclusions

As a summary, this paper has presented a complete set of tools for improving the performance of ASR systems in complex background conditions. It has been shown how asynchronous CMLLR transforms can be successfully trained and used in the decoding of large vocabulary speech. Furthermore, two extensions to this asynchronous adaptation have also been proposed. In the first one, it has been shown how it is possible to generate adaptively trained models using asynchronous transforms, leading to a more flexible modelling of the backgrounds. Also, the possibility of stacking transforms to deal with multiple sources of variability, like background and speaker, can be enhanced by the use of asynchronous modelling of the background.

An evaluation of the proposed techniques in a benchmark task, a modified version of WSJCAM0, has shown large improvements, up to 40% in overall when using *Clean* trained models on a very *Diverse* test set. This improvement occurs across a very varied range of acoustic conditions, with significant improvement being achieved also for clean test data. The evaluation in WSJCAM0 provided a good insight into the strengths of the techniques and into how to achieve the best performance from them.

Finally, an evaluation of the techniques in a very complex scenario has shown that they can achieve gain in a real environment. The transcription of multi-media broadcasts, going beyond the transcription of broadcast news, is a difficult task, where the existence of multiple and dynamic background conditions (noise, music, overlapping speech, etc.) is a major cause of performance degradation. This is better exemplified by the challenging performance achieved

in some genres like comedy or drama, over 40% WER. The proposed techniques have been shown to reduce absolute WER from 31.0% to 30.1% using ML models and from 29.2% to 28.6% using MPE models, which represents a significant improvement for this task. The techniques have also been successfully used as part of a more complex system, submitted to the MGB challenge (Saz et al., 2015), and that provided one of the top performances in the overall evaluation (Bell et al., 2015). This indicates that these techniques can perform well in large tasks and in complex environments.

In further analysis, the proposed aCMLLR background adaptation technique has been shown to work in situations where some knowledge about the background conditions exists, as in WSJCAM0 experiments, and where no knowledge of the background conditions is present, as in broadcast shows. This shows how the proposed training regime can work even with a very loose knowledge of the background situation for initialisation, which makes this technique especially suitable for the most complex situations where the acoustics are unknown.

Finally, other work has shown that asynchronous background modelling can be used for other tasks beyond ASR, as for instance identifying the genre of broadcast shows (Saz et al., 2014). This opens possibilities for future applications of asynchronous background adaptation, which could prove very useful in areas where modelling and compensation of dynamic and challenging background conditions are especially important.

7. Data access management

The original WSJCAM0 corpus is available via the Linguistic Data Consortium with catalogue number LDC95S24. The modified version of WSJCAM0 used in the article is available with DOI 10.15131/shef.data.3363466. All the data related to the MGB challenge, including audio files, subtitle text and scoring scripts, are available via special licence with the BBC on http://www.mgb-challenge.org/. All recognition outputs and scoring results are available with DOI 10.15131/shef.data.3248584.

Acknowledgement

This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

References

Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., 1996. A compact model for speaker-adaptive training. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP). Philadelphia, PA, pp. 1137–1140.

Astudillo, R.F., Hoffmann, E., Manderlatz, P., Orglmeister, R., 2009. Speech enhancement for automatic speech recognition using complex Gaussian mixture priors for noise and speech. In: Proceedings of the 2009 Non-Linear Speech Processing (NOLISP) Workshop. Vic, Spain, pp. 60–67.

Bell, P., Gales, M., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., et al., 2015. The MGB challenge: evaluating multi-genre broadcast media recognition. In: Proceedings of the 2015 IEEE Automatic Speech Recognition and Understanding Workshop. Scottsdale, AZ, pp. 687–693.

Buera, L., Lleida, E., Miguel, A., Ortega, A., Saz, O., 2007. Cepstral vector normalization based on stereo data for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 15 (3), 1098–1113.

Chen, S.S., Gopalakrishnan, P.S., 1998. Clustering via the Bayesian information criterion with applications in speech recognition. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seattle, WA, pp. 645–648.

Cieri, C., Graff, D., Liberman, M., Martey, N., Strassel, S., 1999. The TDT–2 text and speech corpus. In: Proceedings of the 1999 DARPA Broadcast News Workshop. Herndon, VA.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34 (3), 267–285.

Droppo, J., Deng, L., Acero, A., 2001. Evaluation of the SPLICE algorithm on the AURORA2 database. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech). Aalborg, Denmark, pp. 217–220.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time amplitude estimator. IEEE Trans. Acoust. 32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. 33 (2), 443–445.

Fiscus, J., 2007. Speech Recognition Scoring Toolkit (SCTK) version 2.4.0. <http://www.itl.nist.gov/iad/mig/tools/>.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12 (2), 75–98.

Gales, M.J.F., 2001. Acoustic factorisation. In: Proceedings of the 2001 IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop. Madonna di Campiglio, Italy, pp. 77–80.

Gales, M.J.F., Woodland, P.C., 1996. Mean and variance adaptation within the MLLR framework. Comput. Speech Lang. 10 (4), 249–264.

Gales, M.J.F., Young, S.J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4 (5), 352–359.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.F., Mostefa, D., Choukri, K., 2006. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy, pp. 139–142.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.

Grezl, F., Fousek, P., 2008. Optimizing bottleneck features for LVCSR. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas, NV, pp. 4729–4732.

Grezl, F., Karafiat, M., Burget, L., 2009. Investigation into bottle-neck features for meeting speech recognition. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech). Brighton, UK, pp. 2947–2950.

Hain, T., 2005. Implicit modelling of pronunciation variation in automatic speech recognition. Speech Commun. 46, 171–188.

Hermansky, H., 1990. Perceptual Linear Predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87 (4), 1738–1752.

Hirsch, H.G., Pearce, P., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP). Beijing, China, pp. 29–32.

Kalinli, O., Seltzer, M.L., Droppo, J., Acero, A., 2010. Noise adaptive training for robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. 18 (8), 1889–1901.

Larson, M., Anguera, X., Reuter, T., Jones, G., Ionescu, B., Schedl, M., et al. (Eds.), 2013. Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop. Barcelona, Spain.

Lee, L., Rose, R., 1998. A frequency warping approach to speaker normalization. IEEE Trans. Speech Audio Process. 6 (1), 49–60.

Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE Trans. Audio Speech Lang. Process. 22 (4), 745–777.

Liao, H., Gales, M.J.F., 2007. Adaptive training with joint uncertainty decoding for robust recognition of noisy data. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Honolulu, HI, pp. 389–392.

Liu, Y., Zhang, P., Hain, T., 2014. Using neural network front-ends on far field multiple microphones based speech recognition. In: Proceedings of the 2014 International Conference on Acoustic, Speech and Signal Processing (ICASSP). Florence, Italy, pp. 5579–5583.

Long, Y., Gales, M.J.F., Lanchantin, P., Liu, X., Seigel, M.S., Woodland, P.C., 2013. Improving lightly supervised training for broadcast transcriptions. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech). Lyon, France, pp. 2187–2191.

Miguel, A., Lleida, E., Rose, R., Buera, L., Saz, O., Ortega, A., 2008. Capturing local variability for speaker normalization in speech recognition. IEEE Trans. Audio Speech Lang. Process. 16 (3), 578–593.

Milner, R., Saz, O., Deena, S., Doulaty, M., Ng, R., Hain, T., 2015. The 2015 Sheffield system for longitudinal diarisation of broadcast media. In: Proceedings of the 2015 IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop. Scottsdale, AZ, pp. 632–638.

Moreno, P., Raj, B., Stern, R.M., 1996. A vector Taylor series approach for environment-independent speech recognition. In: Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Atlanta, GA, pp. 733–736.

Novak, J.R., Minematsu, N., Hirose, K., 2012. WSFT-based grapheme-to-phoneme conversion: open source tools for alignment, model-building and decoding. In: Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing. San Sebastián, Spain.

Paliwal, K.-K., Lyons, J.-G., So, S., Stark, A.-P., Wojcicki, K.-K., 2010. Comparative evaluation of speech enhancement methods for robust automatic speech recognition. In: Proceedings of the 4th International Conference on Signal Processing and Communication Systems (ICSPCS). Gold Coast, Australia, pp. 1–5.

Pallett, D., Fiscus, J., Garofalo, J., Przybocki, M., 1996. 1995 Hub–4 dry run broadcast materials benchmark test. In: Proceedings of 1996 DARPA Speech Recognition Workshop. Harriman, NY.

Parihar, N., Picone, J., Pearce, D., Hirsch, H.G., 2004. Performance analysis of the AURORA large vocabulary baseline system. In: Proceedings of the 12th European Signal Processing Conference (EUSIPCO). Vienna, Austria, pp. 553–556.

Paul, D.B., Baker, J.M., 1992. The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the 5th DARPA Speech and Natural Language Workshop. Harriman, NY, pp. 357–362.

Povey, D., Woodland, P.C., 2002. Minimum phone error and I–smoothing for improved discriminative training. In: Proceedings of the 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Orlando, FL, pp. 105–108.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., et al., 2011a. The subspace Gaussian mixture model – a structured model for speech recognition. Comput. Speech Lang. 25 (2), 404–439.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Ondrej, G., Nagendra, G., et al., 2011b. The Kaldi speech recognition toolkit. In: Proceedings of the 2011 IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop. Big Island, HA.

Raj, B., Virtanen, T., Chaudhuri, S., Singh, R., 2010. Non-negative matrix factorization based compensation of music for automatic speech recognition. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech). Makuhari, Japan, pp. 717–720.

Richmond, K., Clark, R., Fitt, S., 2010. On generating combilex pronunciations via morphological analysis. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech). Makuhari, Japan, pp. 1974–1977.

Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S., 1995. WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition. In: Proceedings of the 1995 International Conference on Acoustic, Speech and Signal Processing (ICASSP). Detroit, MI, pp. 81–84.

Saz, O., Hain, T., 2013. Asynchronous factorisation of speaker and background with feature transforms in speech recognition. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech). Lyon, France, pp. 1238–1242.
Saz, O., Hain, T., 2014. Using contextual information in joint factor eigenspace MLLR for speech recognition in diverse scenarios. In: Proceedings of the 2014 International Conference on Acoustic, Speech and Signal Processing (ICASSP). Florence, Italy, pp. 6314–6318.
Saz, O., Doulaty, M., Hain, T., 2014. Background-tracking acoustic features for genre identification of broadcast shows. In: Proceedings of the 2014 IEEE Spoken Language Technology (SLT) Workshop. South Lake Tahoe, CA, pp. 118–123.
Saz, O., Doulaty, M., Deena, S., Milner, R., Ng, R., Hasan, M., et al., 2015. The 2015 Sheffield system for transcription of multi-genre broadcast media. In: Proceedings of the 2015 IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop. Scottsdale, AZ, pp. 624–631.
Schuller, B., Weninger, F., Wollmer, M., Sung, Y., Rigoll, G., 2010. Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, TX, pp. 4562–4565.
Seltzer, M.L., Acero, A., 2011. Separating speaker and environmental variability using factored transforms. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech). Florence, Italy, pp. 1097–1100.
Seltzer, M.L., Acero, A., 2012. Factored adaptation using a combination of feature-space and model-space transforms. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech). Portland, OR, pp. 1792–1795.
Seo, H., Kang, H.-G., Seltzer, M.L., 2014. Factored adaptation of speaker and environment using orthogonal subspace transforms. In: Proceedings of the 2014 International Conference on Acoustic, Speech and Signal Processing (ICASSP). Florence, Italy, pp. 3275–3279.
Stolcke, A., 2002. SRILM – an extensible language modeling toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP). Denver, CO, pp. 901–904.
Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.
Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proceedings of the 1990 International Conference on Acoustic, Speech and Signal Processing (ICASSP). Albuquerque, NM, pp. 845–848.
Vesely, K., Butget, L., Grezl, F., 2010. Parallel training of neural networks for speech recognition. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech). Makuhari, Japan, pp. 2934–2937.
Wang, Y., Gales, M.J.F., 2011. Speaker and noise factorisation on the AURORA4 task. In: Proceedings of the 2011 International Conference on Acoustic, Speech and Signal Processing (ICASSP). Prague, Czech Republic, pp. 4584–4587.
Yin, S.C., Rose, R.C., Kenny, P., 2007. A joint factor analysis approach to progressive model adaptation in text-independent speaker verification. IEEE Trans. Audio Speech Lang. Process. 15 (7), 1999–2010.
Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., et al., 2006. The HTK Book version 3.4. Cambridge, UK.
Zelenak, M., Schulz, H., Hernando, J., 2012. Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign. EURASIP J. Audio Speech Music Process. 19, 1–9.
Zhang, P., Liu, Y., Hain, T., 2014. Semi-supervised DNN training in meeting recognition. In: Proceedings of the 2014 IEEE Spoken Language Technology (SLT) Workshop. South Lake Tahoe, CA, pp. 141–146.