Motlicek *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:17
DOI 10.1186/s13636-015-0058-5

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Exploiting foreign resources for DNN-based ASR

Petr Motlicek*†, David Imseng†, Blaise Potard†, Philip N. Garner and Ivan Himawan

Abstract

Manual transcription of audio databases for the development of automatic speech recognition (ASR) systems is a costly and time-consuming process. In the context of deriving acoustic models adapted to a specific application, or in low-resource scenarios, it is therefore essential to explore alternatives capable of improving speech recognition results. In this paper, we investigate the relevance of foreign data characteristics, in particular domain and language, when using this data as an auxiliary data source for training ASR acoustic models based on deep neural networks (DNNs). The acoustic models are evaluated on a challenging bilingual database within the scope of the MediaParl project. Experimental results suggest that in-language (but out-of-domain) data is more beneficial than in-domain (but out-of-language) data when employed in either supervised or semi-supervised training of DNNs. The best performing ASR system, an HMM/GMM acoustic model that exploits DNN as a discriminatively trained feature extractor outperforms the best performing HMM/DNN hybrid by about 5 % relative (in terms of WER). An accumulated relative gain with respect to the MFCC-HMM/GMM baseline is about 30 % WER.

Keywords: Automatic speech recognition; Deep learning for speech; Acoustic model adaptation; Semi-supervised training

1 Introduction

Current automatic speech recognition (ASR) systems are based on statistical parametric methodologies and require large amounts of transcribed speech data during training. Therefore, there is a long-standing belief that "there is no data like more data" in the speech recognition community. In this spirit, a number of efforts have been undertaken to transcribe large amounts of speech data (i.e., the GALE project [1]) in order to improve performance. Unfortunately, transcribing speech is usually an expensive manual process. For that reason, several efforts towards the use of untranscribed data during training have been made in the past. However, the performance gains quickly saturate when continuously adding more data.

For many languages in the world, only very small amounts of transcribed data are available. Hence, many recent studies addressed the exploitation of foreign (i.e., out-of-domain or out-of-language) data for the training of ASR systems [2–4]. It was shown that foreign data usually helps in low-resourced scenarios. However, in general, there is little (or no) performance gain if a large amount of target data is available [4].

The main objective of this paper is to investigate whether foreign data can improve the performance of an ASR system already trained using matched (in-domain and in-language) data. We chose a recent bilingual speech database [5] for this work. The data was released in the context of the MediaParl project, which aims to recognize and annotate the proceedings of the cantonal parliament of Valais in Switzerland. The main characteristics of the parliamentary speech are that it is in two languages (French and German), accented, and reverberant. Initial results were presented in [5].

More specifically, this paper attempts to improve ASR acoustic models, by using state-of-the-art techniques based on deep neural nets (DNNs), towards the recognition of French MediaParl data. To this end, we set out with the following hypotheses:

1. More data should help. In general, most ASR techniques benefit from more data. However, preliminary experiments have shown that adding mismatched data may be detrimental [6]. At the outset, we do not know whether simply training on

*Correspondence: motlicek@idiap.ch
†Equal contributors
Idiap Research Institute, Rue Marconi 19, Martigny, Switzerland

more foreign (out-of-domain or language) data, or adapting to the in-domain data, is the better approach.
2. Out-of-language (but in-domain) data should help. Recent literature suggests that DNN approaches can benefit from out-of-language data. In contrast to other studies such as [7], this is pertinent in MediaParl in that the out-of-language German data is acoustically matched to the target French data.
3. Robustness to the characteristics of the channel should help. The MediaParl data is known to be quite reverberant; it follows that benefits should be gained from the utilization of techniques that are robust to reverberation.

We test these hypotheses in the general framework of deep learning methods through a hidden Markov model (HMM)/DNN architecture where the emission probabilities of the HMM states are estimated with DNNs. Nevertheless, HMM/GMM architectures that exploit an underlying shallow or flat generative model of context-dependent GMMs and HMMs [8, 9] are still popular and offer many highly effective techniques, developed over the last two decades, that can largely improve ASR performance. These techniques include discriminative training, unsupervised speaker adaptation, or noise robustness and are not directly applicable to DNN-based hybrid systems. To combine recent advances in acoustic modeling, namely DNNs, with state-of-the-art adaptation techniques, we employ the bottleneck approaches [10, 11], where the DNN acts as a feature extractor. Our extensive set of experiments compare DNN-based hybrid and bottleneck systems and test the previously mentioned hypotheses.

The remainder of the paper is organized as follows: Section 2 describes an application scenario. It gives more details about the environment, application, and used datasets. Section 3 briefly reviews previous works in applying neural networks for acoustic modeling and the conventional training procedures. Experimental setup, description of baseline systems, and evaluation procedures of individual hypotheses mentioned above are presented in Section 4, while Section 5 presents experimental results. Section 6 brings overall insight on achieved results, and Section 7 concludes the work.

2 Application scenario

2.1 Overview

Recently in the context of the MediaParl project, a bilingual spoken language database was introduced [5] to help to recognize and annotate the proceedings of the cantonal parliament of Valais in Switzerland. The main characteristics of the parliamentary speech are that it is reverberant and in two languages – French ("standard" French) and German (accented "standard" German).

From a speech processing point of view, this database is interesting in that it provides reverberant, multilingual, accented, and non-native speech. This study focuses on the French part of the database and evaluates state-of-the-art ASR techniques together with techniques that address some of the particularities of the database. The aim is to advance the baseline results already presented to state-of-the-art results. We draw on recent advances in ASR, notably in neural networks and the multilingual acoustic modeling (i.e., by leveraging data from different domains) that they enable. The databases used for this study are presented in the next section.

2.2 Databases

Three databases, summarized in Table 1, were used during this study: MediaParl containing both, matched data (French part) and out-of-language data (German part) recorded under the same acoustic conditions, and the out-of-domain databases ESTER [12] and BREF [13].

2.2.1 MediaParl

MediaParl provides 19 h of French data (MP-FR) as well as 18 h of German data (MP-GE). ASR evaluations will be performed using the French MediaParl test set (1.5 h). Since we evaluate our ASR system on French data, the German data is considered as out-of-language data, but contains speech of the same domain. Some of the speakers switch between the two languages. Therefore, the database may to a certain extent be used to study code-switched ASR. However, in contrast to for example [14], the code switches always occur on sentence boundaries.

The parliament debates always take place in the same closed room. Each speaker intervention can last from about 10 s up to 15 min. Speakers are sitting or standing when talking, and their voice is recorded through a single distant microphone.

The recordings took place in 2006 and 2009. The audio recordings of the year 2006 were compressed as "mp3", more specifically MPEG ADTS, layer III, v1, 128 kbps, 44.1 kHz, monaural with 16 bits per sample. The video recordings of the year 2009 were formatted as "avi" with uncompressed PCM (stereo, 48 kHz, 16 bits per sample) audio data. The recordings from 2009 that were processed

Table 1 Statistics of the datasets used for training: number of words in the dictionary and amount of speakers and data

Dataset	Dict.	No. of speakers	Data (h)
MP-FR (French)	11 k	106	19
MP-GE (German)	16 k	75	18
ESTER (French)	203 k	669	57
BREF (French)	13 k	218	114

at Idiap Research Institute are also available as video streams online (http://www.vs.ch/Navig/parlement.asp).

2.2.2 ESTER

ESTER is a database of standard French radio broadcast news [12], manually transcribed and annotated. It comprises a large number of speakers in various recording conditions. In this study, we retained a subset (57 h) of ESTER consisting of native speakers, in low noise conditions. The audio was provided as 16 bit, mono 16 kHz compressed in a lossless format (flac).

2.2.3 BREF

BREF is a large vocabulary, read-speech corpus of standard French. The audio is sampled at 16 kHz. The texts read were selected from 5 million words of the French newspaper "Le monde". The speakers read 11,000 distinct texts, chosen to maximize the number of distinct triphones. In total, it contains more than 114 hours of audio data. All the audio data was converted to 16 bit, mono 16 kHz RIFF prior to any processing.

Although there are some minor differences between standard French and Swiss French, we will consider them to be the same language. BREF and ESTER can both be considered as in-language data. From a domain point of view, the radio broadcasts from ESTER seem to be closer to our target MP-FR than the read speech of BREF.

3 Relation to previous work

The field of acoustic modeling for ASR has seen a lot of research on the usage of neural networks that are able to estimate phone posterior distributions. First, research works from the 1990s reveal that neural nets (NNs) can be used to directly estimate HMM observation probabilities (i.e., denoted as the *hybrid* approach [15, 16]). Later, NNs were also applied in ASR front ends for extracting discriminatively trained speech features (denoted as *tandem* [17] or *bottleneck* (BN) [10] approaches). In that case, the speech features were either the estimated phone posterior probabilities (usually decorrelated and with a reduced dimensionality (tandem)) or the activations of a narrow hidden layer (bottleneck). Tandem and bottleneck NN front ends are employed with conventional HMM/GMMs and can therefore benefit from many techniques that have been developed for the Gaussian mixture framework.

Although hybrid systems achieved good experimental results on a few large vocabulary tasks already more than two decades ago [18], they hardly outperformed HMM/GMMs. HMM/GMM systems performed better, due to the availability of speaker adaptation techniques (MAP [19] or MLLR [20]), HMM phone clustering [21], sequence-level discriminative training techniques (MMI or MPE [22]), or high-dimensional feature space transforms such as fMPE [23]. These techniques exploit particularities of the GMM framework and their training can easily be parallelized in a computer cluster setting, which historically gave such systems a significant advantage in scalability. Therefore, HMM/GMM-based ASR systems dominated over the last decades. Since that time, the processing capabilities of modern GPUs and the advent of more effective training algorithms for NNs revived their usage in ASR. Novel deep NN architectures (with many hidden layers) appeared, denoted as DNNs. However, the traditional error back-propagation algorithms may lead to a poor local minimum when the NN is trained from a set of randomly initialized parameters, and this effect gets more pronounced if the underlying NN structure is deep. Therefore, several forms of DNN pre-training algorithms were proposed for a better initialization of the parameters [24], for example by growing the neural network layer by layer without using the label information: treating each pair of layers in the network as a restricted Boltzmann machine (RBM), and each layer of the neural network can be trained using an objective criterion called contrastive divergence [25].

Early HMM/DNN architectures exploited five-layer DNN and monophone states as the modeling unit [26]. Later, the monophone phonetic representation of the DNN outputs were extended to context-dependent representations [27]. Such systems outperform HMM/GMMs on LVCSR tasks [28]. Current experimental results indicate that the decoding time of a five-layer HMM/DNN is almost the same as the one of a state-of-the-art HMM/GMM system.

Nevertheless, HMM/GMM systems still offer many efficient and relatively simple techniques, that can largely improve the recognition accuracies (e.g., speaker adaptation), which are not always directly transferable to HMM/DNN systems. In addition, these techniques can be combined with advantages of the DNN structures exploited in front ends (tandem, BN-ASR) to boost the performance of HMM/GMMs.

In context of multilingual NN training, and in line with our second hypothesis, several techniques were recently proposed. In [29, 30], features extracted from NN (trained in cross-lingual and multilingual manner) were applied in low-resource HMM/GMM acoustic modeling. Further, in [4], subspace GMM model was combined with cross-lingual BN features. The simultaneous NN training on many languages to extract multilingual BN features was proposed in [31]. Also, in the case of HMM/DNN hybrid, it has been shown that NN training of hidden layers on multiple languages can boost a cross-lingual transfer [32, 33], while the output layers are made language dependent. Besides multilingual adaptability properties of NN, a variety of extending algorithms were recently developed, such as (un)supervised speaker adaptation (e.g., [34, 35], or sequence discriminative training [36], although not yet

fully explored, especially in the combination with other types of NN training.

4 Experimental setup

We follow the state-of-the-art setup for training DNNs. To be able to easily reference a specific layer or set of parameters later in the paper, we briefly present the notation that we adopted.

4.1 Notation

The DNN is trained to classify the input acoustics, $\mathbf{o}_t$, into classes corresponding to the HMM states. After training, the DNN estimates the posterior probability $P(s|\mathbf{o}_t)$ of each state s given the acoustic observations $\mathbf{o}_t$ at time t. The neuron activations all calculated in such a way that all activations of the previous layer are multiplied by a weight vector, summed and passed through a non-linear activation function (e.g., sigmoid). We use the notation from [37]:

$$\mathbf{u}_l = \sigma(\mathbf{W}_l \mathbf{u}_{l-1} + \mathbf{b}_l), \text{for } 1 \leq l < L, \tag{1}$$

where $\mathbf{W}_l$ denotes the matrix of connection weights between $l-1$-th and l-th layers. $\mathbf{b}_l$ is the additive bias vector at the l-th layer, and $\sigma(x) = 1/(1+\exp(-x))$ is a sigmoid activation function.

For multi-class classification, where the classes correspond to the HMM states, the posterior probability $P(s|\mathbf{o}_t)$ can be estimated using the softmax:

$$P(s|\mathbf{o}_t) = \frac{\exp\{a_L(s)\}}{\sum_{s'} \exp\{a_L(s')\}}, \tag{2}$$

where $\mathbf{a}_L = \mathbf{W}_L \mathbf{u}_{L-1} + \mathbf{b}_L$ is the activation at the output layer L.

The DNN is trained using the standard error back-propagation procedure and the optimization is done through stochastic gradient descent (SGD) by minimizing a negative log posterior probability cost function over the set of training examples $O = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \; -\sum_{t=1}^{T} \log P\big(s_t|\mathbf{o}_t\big), \tag{3}$$

where $\theta = \{\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L\}$ is the set of parameters of the network. The ground truth, i.e., the most likely state s at time t, is obtained by performing forced alignment of the acoustic vectors with the transcripts.

4.2 Dictionaries

For the French dictionary, we used BDLex [38] that comprises 38 phonemes including the silence token "sil". For German, we employed PhonoLex [39] using 55 phonemes including "sil". The phonemes of the dictionaries are represented using the speech assessment methods phonetic alphabet (SAMPA) [40] that supports multiple languages including French and German.

To compensate for many unseen words such as abbreviations and names in both languages, we trained a grapheme-to-phoneme tool (Phonetisaurus [41]) from existing dictionaries to derive finite state transducer - based mappings of sequences of letters (graphemes) to their acoustic representation (phonemes).

4.3 Acoustic modeling techniques

We employ three different classes of acoustic models: (1) traditional HMM/GMM, where the emission probabilities of the HMM states are modeled using GMMs; (2) hybrid HMM/DNN, where the emission probabilities of the HMM states are directly estimated using the DNN; and (3) BN-HMM/GMM, where a front-end (trained DNN) estimates BN features that are subsequently modeled by a HMM/GMM.

In the baseline systems, the acoustic signal is parameterized with standard 13-dimensional MFCC features, combined with their first- and second-order derivatives, and means are normalized per-speaker.

More details about the acoustic modeling techniques are given in the subsequent sections:

HMM/GMM — We employ a conventional three-state left-to-right context-dependent HMM/GMM trained on acoustic parameters (MFCCs in the case of the baseline systems). The tree-based clustering returns about 4000 tied states and about 50K Gaussians are used. A forced alignment at the HMM-state level is performed to generate the targets for subsequent DNN training.

DNNs — The hybrid and the bottleneck system both depend on the underlying DNN. In this paragraph, we briefly review DNN techniques that are applied in this paper:

- *Unsupervised generative pre-training* – DNNs have several more layers than conventional multilayer perceptrons and therefore many more parameters.

We use a generative pre-training approach [42] to initialize the DNN parameters, subsequently trained using SGD. The network is pre-trained incrementally layer-by-layer, and each pair of layers is treated as a RBM [43]. The first RBM uses Gaussian-Bernoulli units, and the following RBMs have Bernoulli-Bernoulli units. This pre-training approach is completely unsupervised and does not require transcriptions. It was already shown that pre-training is language-independent [44, 45]. However, it seems to be still unclear what makes some data suitable for unsupervised pre-training [44].

Since the effect of pre-training on the databases used in this paper has already been investigated [6], we focused on other aspects in the present study. In all our DNN-based experiments, we performed pre-training on the MP-FR dataset.

- *DNN adaptation* – For all the experiments that exploit foreign databases, DNN model adaptation using a condition-specific layer is performed at the end of training. This procedure should avoid overfitting the DNN to the out-of-domain data and generalizes to the target data. The idea is similar to multilingual DNN approaches in which hidden layers are shared, while the output layers are made language-specific (e.g., [31, 46]). The adaptation procedure is graphically visualized in Fig. 1. Starting with the DNN models trained using foreign data, the output layer is replaced by a new layer in which we randomly initialise the W_L, which is the matrix of connection weights between the layer $L-1$ and the output layer L. The network is then retrained using the MP-FR data, which most closely matches the evaluation set.
- *Supervised and semi-supervised training* – We distinguish between supervised and semi-supervised training. During supervised training, the data is manually transcribed, i.e., the original transcription is exploited during training.

During semi-supervised training, the auxiliary data is un-transcribed, i.e., the data is exploited in a semi-supervised fashion. A two-pass system can be applied: (1) a DNN is trained using manually transcribed MP-FR data and (2) used to generate posterior probabilities for each frame of the foreign un-transcribed data. We assume that the class with the highest posterior probability is the correct one and use these automatically generated labels during training on the whole dataset. In one of our earlier studies [6], we also investigated the employment of different confidence measures, but only marginal improvements were achieved. Therefore, in this work, we simply use all the foreign data without confidence-based data selection.

HMM/DNN (hybrid) — Nine consecutive speech feature frames serve as input for the DNN. The DNN comprises five layers (three-hidden layers) with the following number of nodes: 351 (or 702, cf. Section 5.4), 2000, 2000, 2000, K, where K is given by the number of tied states in the HMM/GMM baseline. After RBM pre-training, a fully connected language-specific DNN is trained using SGD and the cross-entropy criterion. To prevent over-fitting, 10 % of the training set is used for cross-validation. As hitherto mentioned, all outputs of the nodes in the last layer are transformed using the softmax function, whereas the sigmoid activation function is applied in all other layers.

BN-HMM/GMM — Similarly to HMM/DNN, the phone classes of BN-HMM/GMMs are context-dependent triphones, and the input is given by nine consecutive speech feature frames. A six-layer bottleneck (BN) DNN is trained with the following number of nodes in each layer: 351 (or 702), 2000, 2000, 30, 2000, K, where K is the number of tied-states in the HMM/GMM baseline. Similarly to HMM/DNN, the softmax function is applied at the output; the sigmoid transfer function is applied in hidden layers, except for the BN layer which is purely linear. The BN features are generated using forward pass on the BN layer. We append (first and second order) derivatives and perform per-speaker mean normalization before using them for HMM/GMM training. HMM/GMM model is then trained using in-domain (MP-FR) data. The number of model parameters remains similar to the case of conventional HMM/GMM.

4.4 Evaluation

System evaluation is performed on the test set of MP-FR, consisting of 1.5 h of French speech. A conventional trigram ARPA language model was trained from three different sources: transcripts from the training set of MP-FR, French text from the Swissparl corpus containing Swiss Parliament proceedings, and French text from Europarl—a multilingual corpus of European Parliament proceedings [47]; Europarl contains about 50 million words for each language and is used to overcome data sparsity of the MediaParl and Swissparl texts.

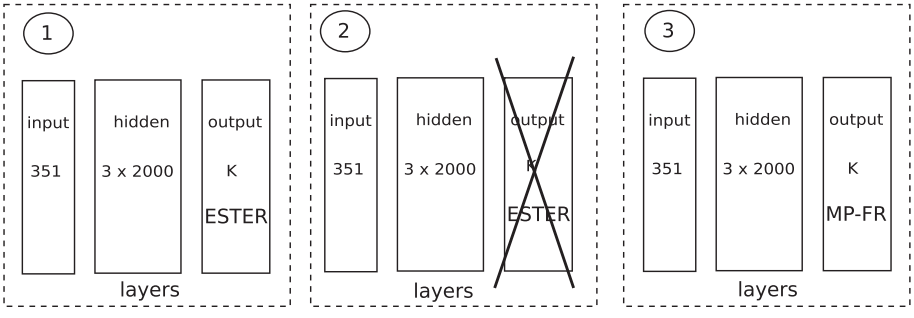

Fig. 1 Network configurations and DNN adaptation procedure in case of supervised training: (*1*) DNN is trained with foreign data (ESTER), (*2*) the output layer is replaced by a new layer with randomly initialised W_L connections, and (*3*) retraining of the network on in-domain (MP-FR) data

5 Results

5.1 Baseline

As baseline acoustic models for developing French MediaParl ASR, we trained our three classes of models (HMM/GMM, HMM/DNN, BN-HMM/GMM) on MP-FR, using MFCC coefficients. The baseline ASR systems are thus trained using 19 h of French MediaParl (matched domain and language) data, as given in Table 1. As expected, results in terms of word error rates (WERs) on the evaluation set of MP-FR (the first line of Table 2) indicate significantly better performance of the DNN-based systems compared to HMM/GMM.

5.2 More data helps

In line with hypothesis one, we explore whether commonly used (though out-of-domain) French corpora of transcribed speech can improve the performance of the MP-FR ASR system. To do so, we used manually transcribed recordings from the ESTER and BREF databases (described in Section 2.2) for training all three different acoustic models.

Table 2 presents different training scenarios, combining in-domain with out-of-domain French data. Experimental results suggest that a blind combination of both BREF and ESTER (the largest database) does not lead to the lowest WERs.

Furthermore, we observe that out-of-domain data significantly decreases the ASR performance for the HMM/GMM system. For the DNN-based ASR systems, however, a combination of MediaParl with either the ESTER or BREF datasets improves the ASR performance. This effect is more pronounced for the BN-HMM/HMM system.

The best performance is achieved by extending MP-FR with ESTER data only. We suppose ESTER is more suitable compared to BREF because the speaking style (broadcast speech vs. read speech) is closer to the MediaParl database (parliamentary debates). Since adding the BREF data does not yield further improvement, we do not consider adding the BREF data for the subsequent experiments.

5.3 Out-of-language data

For many languages in the world, or for specific accents or dialects as is often the case in our scenario, there are only very small amounts of transcribed data available. Therefore, we investigate whether foreign (out-of-domain, or out-of-language) data can improve the performance of an ASR system that has already been trained on a significant amount of matched data (in our case, 19 h of MP-FR). More specifically, this section investigates how important the domain and language are by using three foreign datasets: ESTER-57, ESTER-20 (i.e., a 20 h randomly chosen subset of ESTER-57), and MP-GE (18 h). For each dataset, we compare supervised and semi-supervised training as previously mentioned.

Experimental results are given in Table 3. Note that by comparing the results of the system based on supervised training of this table to the results presented in Table 2, the effect of the DNN adaptation can be seen. Results indicate that ESTER data, employed as out-of-domain but in-language data resource, bring the largest improvement: the results obtained with the larger ESTER-57 are the best in all categories, and we can observe a slight advantage in all cases for the smaller in-language (ESTER-20) data compared to the system trained on a roughly equivalent amount (18 h) of in-domain but out-of-language data (MP-GE).

5.4 Reverberation

MediaParl (in-domain) data was recorded with a single distant microphone in a reverberant environment (i.e., a large chamber where political debates take place). Such an acoustic condition, in addition to bilingualism, is another challenge in ASR, often causing serious performance degradation. Due to the application of a single distant microphone scenario, the typical reverberation reduction

Table 2 ASR performance in WERs of conventional MFCC features modeled by HMM/GMM-, hybrid (HMM/DNN)-, and BN-HMM/GMM-based systems on MP-FR evaluation data while taking into account different in-language speech resources for training

System	HMM/GMM (%)	HMM/DNN (%)	BN-HMM/GMM (%)
MP-FR (19 h) (baseline)	17.1	14.9	14.5
+ESTER-57 (+57 h)	17.6	14.0	13.7
+BREF-114 (+114 h)	19.5	14.3	14.0
+ESTER-57 + BREF-114 (171 h)	20.3	14.2	13.7

Table 3 ASR performance in WERs of conventional MFCC features modeled by hybrid (HMM/DNN)- and BN-HMM/GMM-based systems on MP-FR evaluation data while taking into account different foreign data for training

System (training)	HMM/DNN		BN-HMM/GMM	
	Supervised (%)	Semi-supervised (%)	Supervised (%)	Semi-supervised (%)
MP-FR (19 h)	14.9	–	14.5	–
+ESTER-20 (20 h)	14.3	14.8	13.9	14.0
+ESTER-57 (57 h)	14.0	14.5	13.7	13.9
+MP-GE (18 h)	14.6	14.7	14.2	14.1

techniques relying on multiple microphones cannot be investigated.

Two basic approaches are often used in this case: (1) application of some filtering or pre-processing (e.g., modulation filtering [48]) or (2) employment of a robust front end more resistant to reverberation.

The latter approach is preferred in ASR with techniques such as TRAP [49], RASTA post-processing [50], or frequency domain linear prediction (FDLP) [51]. They usually attempt to model temporal characteristics in critically warped frequency sub-bands over relatively long windows. In our experimental setup, FDLP feature extraction is applied, which has already been shown to improve recognition of reverberated speech [52]. FDLP features are to some extent complementary to MFCCs [53] and the combination of MFCCs and FDLPs yields further improvement.

Similarly to previous acoustic models built on MFCCs, 13-dimensional FDLP features augmented with first- and second-order derivatives are exploited, in addition to MFCCs. The size of the HMM/GMM models remains the same. For the DNNs, the input layer size is extended to 702 nodes due to the doubled size of the speech feature frames and the nine-frame temporal context used. Other layers are unmodified. Results are shown in Table 4 and confirm that the combination of MFCC and FDLP features yields significant improvements to all the ASR systems.

5.5 Combined approaches

Table 5 gives an overview over both DNN-based acoustic modeling techniques, for all three hitherto explored areas, (1) more data helps, (2) out-of-language data, and (3) reverberation. Similar to Table 3, both supervised and semi-supervised training algorithms are applied.

In addition, Table 5 also compares both DNN-based acoustic modeling approaches when speaker adaptation (and conventional discriminative training in the case of BN-HMM/GMM) is exploited during training and decoding (performance denoted as *final*). More specifically, the speaker normalization using feature-space maximum likelihood linear regression (fMLLR), also known as constrained MLLR (CMLLR [54]), is applied. The fMLLR in both types of ASR systems has 78×79 parameters and is estimated using the HMM/GMM-based system applying speaker adaptive training (SAT) [55]. The BN-HMM/GMM system exploits discriminative training using first feature-space boosted MMI (fBMMI) and then model-space boosted MMI. Note that the fBMMI is similar to the form of fMPE described in [56], but uses the objective function of boosted MMI (BMMI) [57] instead of that of MPE. The systems called *final* represent current state-of-the-art in acoustic modeling, either based on a traditional HMM/GMM framework, or a hybrid (HMM/DNN) approach. We are aware of other recently proposed DNN training schemes to further compensate for unseen speakers, mismatched acoustic backgrounds, or replacing traditional cross-entropy training by new algorithms, as discussed in Section 3. These techniques may further improve performance, though their combination with already applied training methods were not yet fully studied.

Table 4 ASR performance in WERs of MFCCs and their combination with FDLPs modeled by HMM/GMM-, hybrid (HMM/DNN)-, and BN-HMM/GMM-based systems on MP-FR evaluation data

System	HMM/GMM (%)	HMM/DNN (%)	BN-HMM/GMM (%)
MFCC	17.1	14.9	14.5
MFCC+FDLP	16.5	14.4	13.9

The acoustic models are trained on MP-FR (19 h) data

Table 5 ASR performance in WERs of MFCC + FDLP features modeled by hybrid (HMM/DNN)- and BN-HMM/GMM-based systems on MP-FR evaluation data while taking into account different foreign data for training

System	HMM/DNN		BN-HMM/GMM	
(training)	Supervised (%)	Semi-supervised (%)	Supervised (%)	Semi-supervised (%)
MP-FR (19 h)	14.4	–	13.9	–
– final	12.9		12.5	
+ESTER-57 (57 h)	13.6	14.2	12.9	13.5
– final	12.7	13.0	12.1	12.6
+MP-GE (18 h)	14.0	14.1	13.2	13.2
– final	12.7	12.8	12.4	12.6

Systems denoted as "final" exploit enhanced algorithms, as described in Section 5.5

6 Discussion

Deep learning methods have been shown to currently offer the most efficient acoustic modeling framework for ASR. Even if this study does not present novel training or decoding techniques for DNNs, the rigorous evaluation of two powerful implementations of DNN architectures on this particular database yield conclusions that can easily be generalized. The hypotheses under investigation are closely aligned with the development of a production ASR system (in this case, the MediaParl application [5]). We therefore believe that the hypotheses and experimental results are useful for other researchers and for building modern ASR systems using foreign data.

The main goal of this study was to improve the performance of DNN-based French ASR through the exploitation of several foreign resources for the MediaParl task. Within the scope of this paper, three main hypotheses were set out:

- First, we hypothesized that more data should help to improve the acoustic models. This was tested on several out-of-domain data resources matching the target language (French). As can be seen from results in Table 2, the hypothesis is confirmed for DNN-based acoustic models. Both HMM/DNN and BN-HMM/GMM acoustic models yield relative improvement on MP-FR evaluation data up to about 6 % WERs, when another source of in-language data was used during training. Results also reveal that ESTER-57 data are more appropriate for training than BREF-114. Furthermore, the combination of both foreign resources did not bring any further improvement compared to the training on ESTER-57 data only. This indicates that careful consideration should be given to the foreign data used, and in particular its similarity—in terms of, e.g., speaking style or noise level—to the target; these experiments demonstrate that for foreign databases, larger does not necessarily mean better. Note also that HMM/GMMs exploiting conventional spectral-based MFCC features did not benefit from additional resources.
- The second hypothesis tested whether out-of-language (but in-domain) data can be helpful. Table 3 introduces MP-GE (18 h of German MediaParl in-domain) data and compares it to roughly the same amount (20 h) of ESTER data (and to the above used ESTER-57 data). Results demonstrate that a slight improvement of about 2 % in terms of relative WERs can be achieved for HMM/DNN and BN-HMM/GMM models using MP-GE. However, these results are notably worse than those obtained with roughly the same amount of out-of-domain, but in-language data. We also tested a scenario when foreign data was used in a semi-supervised way; the performance gains obtained by both kinds of foreign data were then noticeably lower than in the supervised case, but followed the same trend.
- The last hypothesis was related to specific particularities of MediaParl data. We tested whether the use of techniques robust to reverberation can improve the accuracy of French MediaParl ASR. In this study, FDLP features in addition to conventional MFCCs were employed. According to Table 4, this resulted in about 3 % and 4 % relative WER improvement for HMM/DNN and BN-HMM/GMM models, respectively. These gains are complementary to the previous improvements, as can be seen from Table 5 (e.g., relative gain of about 5 and 7 % in the case of ESTER-57 applied as foreign data).

Finally, Table 5 shows results for the case when advanced techniques (fMLLR transformed features for HMM/DNN and SAT plus discriminative training for BN-HMM/GMM) were applied (denoted as *final*). With respect to the MP-FR baselines, 15 and 17 % relative gains were achieved for the best HMM/DNN and BN-HMM/GMM setups (exploiting ESTER-57 as foreign data). The systems based on semi-supervised training (i.e., no transcripts for the foreign resources) perform similarly to the systems trained on MP-FR data only.

7 Conclusions

We have shown that foreign data can improve DNN-based ASR systems even if it has already been trained on reasonable amounts of target data. This effect is more pronounced if transcripts for the foreign data are available (supervised training). In the case of semi-supervised training, when no transcripts are available for the foreign data, experimental results reveal that speaker-independent systems still benefit significantly from foreign data. However, if speaker adaptation techniques are employed, un-transcribed foreign data was not able to improve the performance of the baseline systems trained on target data only.

DNN-based ASR systems significantly outperform conventional HMM/GMMs. The BN-based systems, that use DNNs as discriminative feature extractors followed by a GMM-based back-end that exploits advanced adaptation techniques, perform consistently better than state-of-the-art hybrid systems. The best performance of 12.1 % WER was achieved by BN-HMM/GMM exploiting in-language (but out-of-domain) data during training, which yields about 5 % relative WER improvement compared to the best performing HMM/DNN hybrid (12.7 % WER). An accumulated relative gain with respect to the simple MFCC HMM/GMM system trained on in-domain data only is about 30 % WER.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The work was supported by Samsung Electronics Co. Ltd., South Korea, under the project "Multi-Lingual and Cross-Lingual Adaptation for Automatic Speech Recognition". The work was also partially supported by Eurostars Programme powered by Eureka and the European Community under the project "D-Box: A generic dialog box for multi-lingual conversational applications", and by EC FP7 "Speaker Identification integrated project" under grant agreement no. 607784. The authors would like to express their gratitude to the MediaParl project, funded by the Parliament Service of the State of Valais, Switzerland for their financial support and for providing access to the audio-video recordings.

Received: 15 December 2014 Accepted: 13 May 2015
Published online: 26 June 2015

References

1. J Cohen, in *Automatic Speech Recognition Understanding Workshop (ASRU)*. The gale project: a description and an update, (2007), pp. 237–237. doi:10.1109/ASRU.2007.4430115
2. NT Vu, F Kraus, T Schultz, in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*. Multilingual a-stabil: a new confidence score for multilingual unsupervised training, (2010), pp. 183–188
3. S Thomas, ML Seltzer, K Church, H Hermansky, in *Proc. of ICASSP*. Deep neural network features and semi-supervised training for low resource speech recognition, (2013), pp. 6704–6708
4. D Imseng, P Motlicek, H Bourlard, PN Garner, in *Speech Communication*. Using out-of-language data to improve an under-resourced speech recognizer, (2014), pp. 142–151
5. D Imseng, H Bourlard, H Caesar, PN Garner, G Lecorvé, A Nanchen, in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*. MediaParl: Bilingual mixed language accented speech database, (2012), pp. 263–268
6. D Imseng, B Potard, P Motlicek, A Nanchen, H Bourlard, in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*. Exploiting un-transcribed foreign data for speech recognition in well-resourced languages, (2014), pp. 2322–2326
7. Y Huang, D Yu, Y Gong, C Liu, in *Proc. of Interspeech*. Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration, (2013), pp. 2360–2364
8. LR Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE. **77**(2), 257–286 (1989)
9. LR Rabiner, BH Juang, in *IEEE ASSP Magazine*. An introduction to hidden Markov models, (1986), pp. 5 –16
10. F Grezl, M Karafiat, L Burget, in *Proc. of Interspeech*. Investigation into bottle-neck features for meeting speech recognition, (2009), pp. 2947–2950
11. F Grézl, M Karafiát, M Janda, in *Proc. of ASRU*. Study of probabilistic and bottle-neck features in multilingual environment, (2011), pp. 359–364
12. S Galliano, E Geoffrois, G Gravier, J-F Bonastre, D Mostefa, K Choukri, in *Proc. of the International Conference on Language Resources and Evaluation*. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news, (2006)
13. LF Lamel, J-L Gauvain, M Eskenazi, M Eskenazi, in *Proceedings of Eurospeech*. BREF, a large vocabulary spoken corpus for french, (1991), pp. 505–508
14. L Dau-Cheng, *et al*, in *Proc. of Interspeech*. SEAME: a Mandarin-english code-switching speech corpus in South-East Asia, (2010), pp. 1986–1989
15. H Bourlard, N Morgan, *Continuous speech recognition: a hybrid approach*. (Kluwer Academic Publishers, 1994)
16. S Renals, H Bourlard, M Cohen, H Franco, Connectionist probability estimators in HMM speech recognition. IEEE Trans. Audio Speech Lang. Process. **2**(1), 161–174 (1994)
17. H Hermansky, DPW Ellis, S Sharma, in *Proc. of ICASSP*. Tandem connectionist feature extraction for conventional HMM systems, (2000), pp. 1635–1638
18. MM Hochberg, S Renals, A Robinson, DG Cook, in *Proc. of ICASSP*. Recent improvements to the ABBOT large vocabulary CSR system, (1995), pp. 69–72
19. J-L Gauvain, C-H Lee, in *Proc. of ICASSP*. Speaker adaptation based on MAP estimation of HMM parameters, vol. 2, (1993), pp. 558–561
20. MJF Gales, Maximum likelihood linear transformation for HMM-based speech recognition. Report CUED/F-INFENG/TR291, Cambridge University Engineering Department (1997)
21. SJ Young, JJ Odell, PC Woodland, in *Proceedings of the Workshop on Human Language Technology*. Tree-based state tying for high accuracy acoustic modelling, (1994), pp. 307–312
22. D Povey, PC Woodland, in *Proc. of ICASSP*. Minimum phone error and I-smoothing for improved discriminative training, (2002), pp. 105–108
23. D Povey, B Kingsbury, L Mangu, G Saon, H Soltau, G Zweig, in *Proc. of ICASSP*. FMPE: discriminatively trained features for speech recognition, (2005), pp. 961–964
24. D Erhan, PA Manzagol, Y Bengio, S Bengio, P Vincent, in *Proc. 12th. Int. Conference on Artificial Int. Statist. (AISTATS)*. The difficulty of training deep architectures and the effect of unsupervised pre-training, (2009), pp. 153–160
25. G Hinton, in *Tech. Rep. UTML TR 2010-003*. A practical guide to training restricted Boltzmann machines (University of Toronto, 2010)
26. A-r Mohamed, GE Dahl, G Hinton, Deep belief networks for phone recognition. NIPS workshop on deep learning for speech recognition (2009)
27. D Yu, L Deng, GE Dahl, in *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition, (2010). http://research.microsoft.com/apps/pubs/default.aspx?id=143619
28. GE Dahl, D Yu, L Deng, A Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. **20**(1), 30–42 (2012)
29. S Thomas, H Hermansky, in *Proc. of Interspeech*. Cross-lingual and multistream posterior features for low resource LVCSR systems, (2010), pp. 877–880
30. S Thomas, S Ganapathy, H Hermansky, in *Proc. of ICASSP*. Multilingual MLP features for low-resource LVCSR systems, (2012), pp. 4269–4272
31. K Vesely, M Karafiát, F Grezl, M Janda, E Egorova, in *IEEE Spoken Language Technology Workshop (SLT)*. The language-independent bottleneck features, (2012), pp. 336–341
32. J-T Huang, J Li, D Yu, L Deng, Y Gong, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, (2013), pp. 7304–7308
33. A Ghoshal, P Swietojanski, S Renals, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multilingual training of deep neural networks, (2013), pp. 7319–7323
34. P Swietojanski, S Renals, in *Proc. IEEE Workshop on Spoken Language Technology*. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models (Lake Tahoe, USA, 2014)
35. T Ochiai, S Matsuda, X Lu, C Hori, S Katagiri, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speaker adaptive training using deep neural networks, (2014), pp. 6349–6353
36. K Vesely, A Ghoshal, L Burget, D Povey, in *INTERSPEECH*. Sequence-discriminative training of deep neural networks, (2013), pp. 2345–2349
37. P Swietojanski, A Ghoshal, S Renals, in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition, (2013). doi:10.1109/ASRU.2013.6707744
38. G Perennou, in *Proc. of ICASSP*. B.D.L.E.X: a data and cognition base of spoken French, vol. 11, (1986), pp. 325–328
39. F Schiel, Aussprache-Lexikon PHONOLEX (2013). http://www.phonetik.uni-muenchen.de/forschung/Bas/BasPHONOLEXeng.html
40. JC Wells, SAMPA computer readable phonetic alphabet (2013). http://www.phon.ucl.ac.uk/home/sampa/
41. J Novak, N Minematsu, K Hirose, C Hori, H Kashioka, P Dixon, in *Proc. of Interspeech*. Improving WFST-based G2P conversion with alignment constraints and RNNLM N-best rescoring, (2012), pp. 2526–2529
42. D Erhan, Y Bengio, A Courville, P-A Manzagol, P Vincent, S Bengio, Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. **11**, 625–660 (2010)
43. G Hinton, S Osindero, Y Teh, A fast algorithm for deep belief nets. Neural Nets. **18**, 1527–1554 (2006)
44. P Swietojanski, A Ghoshal, S Renals, in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR, (2012), pp. 246–251
45. K Vesely, M Hannemann, L Burget, in *Proc. of ASRU*. Semi-supervised training of deep neural networks, (2013), pp. 267–272
46. D Imseng, P Motlicek, P Garner, H Bourlard, in *Proc. of ASRU*. Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition, (2013), pp. 332–337
47. P Koehn, in *Proceedings of the 10th Machine Translation Summit*. Europarl: a parallel corpus for statistical machine translation, (2005), pp. 79–86
48. A Kusumoto, T Arai, K Kinoshita, N Hodoshima, N Vaughan, Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. Speech Commun. **45**, 101–113 (2005)
49. H Hermansky, in *Proc. of ASRU*. Trap-tandem: data-driven extraction of temporal features from speech, (2003), pp. 255–260
50. B Kingsbury, N Morgan, in *Proc. of ICASSP*, vol. 2, (1997), pp. 1259–1262
51. M Athineos, D Ellis, in *Proc. of ASRU*. Frequency-domain linear prediction for temporal features, (2003), pp. 261–266

52. S Ganapathy, S Thomas, P Motlicek, H Hermansky, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Applications of signal analysis using autoregressive models for amplitude modulation, (2009), pp. 341–344
53. S Ganapathy, S Thomas, H Hermansky, Modulation frequency features for phoneme recognition in noisy speech. J. Acoust. Soc. Am. **125**(1), EL8–EL12 (2009)
54. M Gales, Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)
55. S Matsoukas, R Schwartz, H Jin, L Nguyen, in *DARPA Speech Recognition Workshop*. Practical implementations of speaker-adaptive training, (1997)
56. D Povey, in *Proceedings of Interspeech*. Improvements to fMPE for discriminative training of features, (2005), pp. 2977–2980
57. D Povey, D Kanevsky, *et al*, in *Proceedings of IEEE ICASSP*. Boosted MMI for model and feature-space discriminative training, (2008), pp. 4057–4060