UNSUPERVISED CROSS-LINGUAL KNOWLEDGE TRANSFER IN DNN-BASED LVCSR

Pawel Swietojanski, Arnab Ghoshal and Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB
p.swietojanski@sms.ed.ac.uk, {a.ghoshal,s.renals}@ed.ac.uk

ABSTRACT

We investigate the use of cross-lingual acoustic data to initialise deep neural network (DNN) acoustic models by means of unsupervised restricted Boltzmann machine (RBM) pretraining. DNNs for German are pretrained using one or all of German, Portuguese, Spanish and Swedish. The DNNs are used in a tandem configuration, where the network outputs are used as features for a hidden Markov model (HMM) whose emission densities are modeled by Gaussian mixture models (GMMs), as well as in a hybrid configuration, where the network outputs are used as the HMM state likelihoods. The experiments show that unsupervised pretraining is more crucial for the hybrid setups, particularly with limited amounts of transcribed training data. More importantly, unsupervised pretraining is shown to be language-independent.

Index Terms— Cross-lingual ASR, Deep Neural Networks, RBM pretraining, GlobalPhone

1. INTRODUCTION

In cross-lingual speech recognition, knowledge from one or more languages is used to improve speech recognition for a target language that is typically low-resourced. A number of techniques for cross-lingual acoustic modelling have been published including the use of: global phone sets [1, 2]; multilingual posterior features in tandem [3, 4] and Kullback-Liebler hidden Markov model (KL-HMM) systems [5]; subspace Gaussian mixture models (SGMMs) with a shared multilingual phonetic subspace [6, 7]; and cross-lingual bootstrapping with unsupervised training of the target language [8]. These approaches rely on transcribed audio data for building automatic speech recognition (ASR) systems in some source languages that may or may not be linguistically related to the target language. These approaches assume that only a small volume of transcribed target language audio is available, and in some cases the target language audio is assumed to be entirely untranscribed [2, 8].

Here we are concerned with building acoustic models with limited amounts of transcribed audio. We also assume that we have untranscribed audio in the chosen language, as well as in other languages. The key question that we address is how to usefully employ these untranscribed acoustic data for speech recognition of the target language. We consider this in the context of deep neural network (DNN) acoustic models for the target language, which can take advantage of the untranscribed audio using unsupervised pretraining techniques. We use layer-wise restricted Boltzmann machine (RBM) initialisation of a DNN [9], an unsupervised procedure, in which a deep generative model of the acoustic data is estimated and used to initialise the weights of the DNN, which are then refined using supervised training on transcribed acoustic data in the target language. The generative model may be of acoustics in the same language (in-domain) or a different language (out-of-domain). Through these experiments we aim to develop a better understanding of cross-lingual knowledge transfer, as well as unsupervised pretraining.

We use the DNNs in both tandem and hybrid configurations. In tandem systems, DNNs are used to generate discriminative features, based on a linear transformation of either the network outputs—posterior features [10]—or the outputs of a narrow hidden layer—-bottleneck features [11]. These features are then usually concatenated with some standard acoustic features, for example, mel-frequency cepstral coefficients (MFCC) or perceptual linear prediction (PLP) coefficients, and used as the feature vector in an HMM-GMM system. In a hybrid system, the trained DNN is used to provide scaled likelihood estimates for the states of an HMM [12]. Due to computational constraints, earlier uses of hybrid systems were limited to estimating scaled likelihoods for monophone states using multi-layer perceptrons (MLPs) with two layers [13] and recurrent networks [14]. More recently DNNs with up to 9 layers have been used with outputs corresponding to both monophone states [15] and context-dependent tied states [16]. The principal modelling and algorithmic difference to previous systems is the use of RBM pretraining [9].

Previous uses of neural networks in cross-lingual acoustic modelling have mainly focussed on tandem approaches that require transcribed data in source languages. Examples include: the direct use of posterior features obtained from a source language network [17]; the use of cross-lingual bottleneck features [3]; training/initialising a neural network using transcribed source language acoustics, then retraining the

This research was supported by EPSRC Programme Grant grant, no. EP/I031022/1 (Natural Speech Technology)

network with transcribed target language acoustics, using a phoneset mapping where necessary [18, 4]; and posterior features derived from networks trained to estimate articulatory features [19]. To the best of our knowledge, this is the first work where unlabeled acoustic data from a different language is successfully used to improve speech recognition accuracy.

2. DNNS FOR ASR

DNNs are L-layer MLPs with a softmax output layer, which we train to classify the input acoustics into classes corresponding to HMM states. After training, the output of the DNN is an estimate of the posterior probability $P(y|\mathbf{o}_t)$ of each state y given the acoustic observations $\mathbf{o}_t$ at time t. The computation performed by the network may be written as:

$$\mathbf{u}_l = \sigma(\mathbf{W}_l \mathbf{u}_{l-1} + \mathbf{b}_l), \qquad \text{for } 1 \le l < L$$
$$P(y|\mathbf{o}_t) = \frac{\exp(\mathbf{W}_L \mathbf{u}_{L-1} + \mathbf{b}_L)}{\sum_{\tilde{y}} \exp(\mathbf{W}_L \mathbf{u}_{L-1} + \mathbf{b}_L)},$$

where $\mathbf{u}_l$ is the input to the $l+1$-th layer, with $\mathbf{u}_0 = \mathbf{o}_t$; $\mathbf{W}_l$ is 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 non-linearity, also known as the activation function.

We use stochastic gradient descent to train DNNs, minimising a negative log posterior probability cost function over the set of training examples $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$:

$$\theta^* = \arg\min_\theta E(\mathcal{O}) \approx \arg\min_\theta - \sum_{t=1}^{T} \log P(y_t|\mathbf{o}_t),$$

where $\theta = \{\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L\}$ is the set of parameters of the network, and y_t is the most likely state at time t obtained by a forced-alignment of the acoustics with the transcript. While training deep networks directly results in a difficult optimization problem, an unsupervised pretraining phase using greedy layer-wise training of RBMs [9] or stacked autoencoders [20] have been shown to give good results. More recently, supervised layer-wise training with early stopping was shown to achieve comparable or better results than unsupervised pretraining on a relatively large speech recognition task [21]. For our investigation of unsupervised cross-lingual pretraining, RBMs were a natural first choice due to their previous successful application in speech recognition [15, 16].

RBMs are bipartite undirected graphical models, with a set of nodes corresponding to observed random variables (also called visible units) and a set of nodes corresponding latent random variables (or hidden units), that only allow interactions between the two sets of variables (that is, between the visible and hidden units) but not within each set of nodes. The joint probability of the visible units $\mathbf{v}$ and hidden units $\mathbf{h}$ is defined as:

$$P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z_{h,v}} e^{-E(\mathbf{v},\mathbf{h})},$$

where $Z_{h,v}$ is the normalising partition function. Visible units are real-valued for speech observations and binary-valued otherwise; hidden units are always binary-valued.

In the case of binary visible units, we have a Bernoulli-Bernoulli RBM whose energy function is:

$$E_{\text{B-B}}(\mathbf{v}, \mathbf{h}) = -\mathbf{v}^T \mathbf{W} \mathbf{h} - \mathbf{b}^T \mathbf{v} - \mathbf{a}^T \mathbf{h},$$

and for real-valued visible units we use a diagonal covariance Gaussian-Bernoulli RBM whose energy function is given by:

$$E_{\text{G-B}}(\mathbf{v}, \mathbf{h}) = -\mathbf{v}^T \mathbf{W} \mathbf{h} - \frac{1}{2}(\mathbf{v} - \mathbf{b})^T (\mathbf{v} - \mathbf{b}) - \mathbf{a}^T \mathbf{h}.$$

$\mathbf{W}$ is a symmetric weight matrix defining interactions between vectors $\mathbf{v}$ and $\mathbf{h}$ while $\mathbf{b}$ and $\mathbf{a}$ are additive bias terms. RBM pretraining maximises the likelihood of the training samples using the contrastive divergence algorithm [9]. When multiple layers have to be initialised the parameters of the given layer are frozen and its output is used as the input to the higher layer which is optimised as a new RBM. This procedure is repeated until the desired number of layers is reached.

When used in tandem configuration [10], the DNN outputs correspond to posterior probabilities of the context-independent phones in the language (in our case, 44 for German). The outputs are Gaussianized by taking logarithms, decorrelated using principal components analysis (PCA), and concatenated with MFCCs. The PCA step also reduces the dimensionality from 44 to 25 (this guaranteed keeping at least 95% of variance—on average it was 98%), producing a combined 64-dimensional feature for the HMM-GMM acoustic model. In the hybrid setup, the outputs correspond to tied triphone states. Depending on the amount of training data used, the number of tied triphone states may vary from a few hundred to a few thousand (roughly 550 to 2500 in our case). To obtain scaled likelihoods, the posterior probability estimates produced by the network were divided by the prior probabilities [12].

3. EXPERIMENTS

For testing cross-lingual knowledge transfer in DNNs, we use the GlobalPhone corpus [22]. The corpus consists of recordings of speakers reading newspapers in their native language. There are 19 languages from a variety of geographical locations: Asia (Chinese, Japanese, Korean), Middle East (Arabic, Turkish), Africa (Hausa), Europe (French, German, Polish), and Americas (Costa Rican Spanish, Brazilian Portuguese). There are about 100 speakers per language and about 20 hours of audio material. Recordings are made under relatively quiet conditions using close-talking microphones; however acoustic conditions may vary within a language and between languages.

Our setup is similar to that reported in [7]. We use German as our in-domain language and we simulate different degrees of available resources by selecting random 1 and 5 hour

Table 1. Word error rates (%) for HMM-GMM systems on GlobalPhone German development set.

Training	Features	Amount of training data		
		15hr	5hr	1hr
ML	MFCC	16.17	18.40	23.11
ML	LDA+MLLT	15.53	18.41	22.31
fBMMI+BMMI	LDA+MLLT	15.19	18.19	21.53

subsets of the total 15 hours of labeled training speech data. When using the 1 and 5 hour subsets, the entire 15 hours of audio from the training set were used for the RBM-based unsupervised pretraining. We contrast this with RBM pretraining using unlabeled acoustic data from three other languages: Portuguese (26 hours), Spanish (22 hours) and Swedish (22 hours), as well as with pretraining using all the languages (85 hours).

3.1. Baseline results

Before discussing the results on GlobalPhone, it is important to note that the results reported in various sources (for example, [1, 3, 7, 23]) are not directly comparable. This primarily because of the differences between LMs, which are much more significant than other differences, such as the use of MFCC vs PLP features. Following previous work [7], we use LMs that were included in an earlier release of the corpus, but are not available in later releases. The differences between the results reported here and those in [7] are due to the fact that we found it beneficial to interpolate the provided LM with one trained on the training transcripts.

We build standard maximum-likelihood (ML) trained HMM-GMM systems, using 39-dimensional MFCC features with delta and acceleration coefficients, on the full 15-hour training set for GlobalPhone German, as well as the 5-hour and 1-hour subsets, using the Kaldi speech recognition toolkit [24]. The number of context-dependent triphone states for the three systems are 2564, 1322 and 551, respectively, with an average of 16, 8 and 4 Gaussians, respectively, per state. The word error rates (WER) of the different baselines are presented in Table 1.

Since the tandem systems use phone posteriors obtained using a window of 9 frames, we compare them with a baseline system where 9 frames (4 on each side of the current frame) of 13-dimensional MFCCs are spliced together and projected down to 40 dimensions using linear discriminant analysis (LDA). We also use a single maximum likelihood linear transform [25] on the features thus obtained using LDA. The combined system is referred to as LDA+MLLT. We compare the hybrid setup to a HMM-GMM system that uses both model and feature-space discriminative training using boosted maximum mutual information (BMMI) estimation [26], referred to as fBMMI+BMMI in Table 1.

3.2. DNN configuration and results

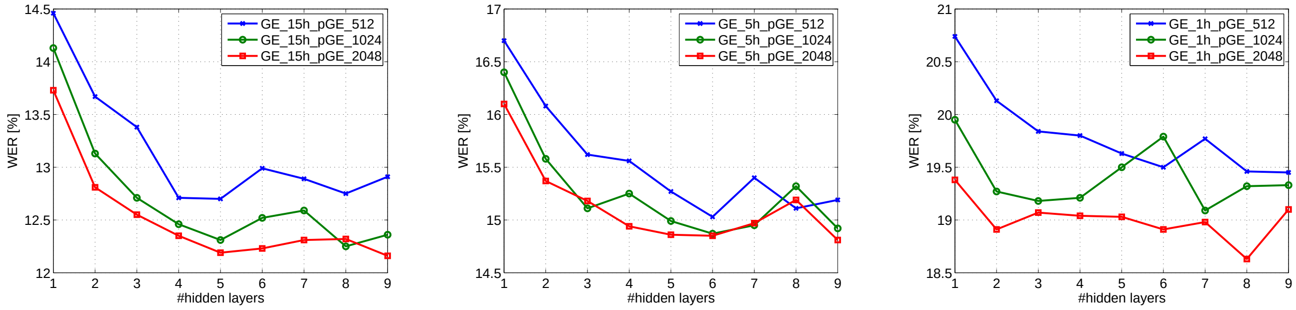
For training DNNs, our tools utilise the Theano library [27], which supports transparent computation using both CPUs and GPUs. We use 12 PLP coefficients and the energy term appended with the delta and acceleration coefficients for a 39-dimensional acoustic feature vector. The features are globally normalised to zero mean and unit variance, and 9 frames (4 on each side of the current frame) are used as the input to the networks. The choice of PLP features was initially motivated by the desire to have information that is complementary to MFCCs for the tandem configuration.

The initial network weights (both for RBM pretraining, and when no pretraining was done) were chosen uniformly at random: $w \sim U[-r, r]$, where $r = 4\sqrt{6/(n_j + n_{j+1})}$ and n_j is the number of units in layer j. We choose the pretraining hyper-parameters as follows: learning rate for Bernoulli-Bernoulli RBM is 0.08, and for Gaussian-Bernoulli RBM in the input layer it is 0.005. Mini-batch size is 100. Fine-tuning is done using stochastic gradient descent on 256-frame mini-batches and an exponentially decaying schedule, learning at a fixed rate (0.08) until improvement in accuracy on cross-validation set between two successive epochs falls below 0.5%. The learning rate is then halved at each epoch until the overall accuracy fails to increase by 0.5% or more, at which point the algorithm terminates. While learning, both RBM and DNN gradients were smoothed with a first-order low-pass momentum (0.5).

In the tandem setup, the networks are up to five layers deep since the tandem systems were not found to improve in terms of WER with deeper networks (Fig. 2). The networks have 1024 hidden units per layer, which was found to outperform 512 hidden units and to have similar WER to 2048 hidden units. In contrast, the hybrid system benefits from deeper architectures (Fig 3), as well as wider hidden layers with 2048 units, even when fine-tuning using just 1 hour of transcribed speech (Fig. 1).

We find that the hybrid systems provide lower WER than the corresponding tandem systems. Additionally, and perhaps most importantly, unsupervised RBM pretraining is found to be language-independent. Pretraining is found to be more effective for hybrid systems than for tandem systems, and the effect is most pronounced when the hybrid systems are fine-tuned using limited amounts of transcribed data. In fact, with 1 hour of transcribed speech the hybrid system only outperformed the baseline HMM-GMM system when pretraining was done. However, for both the tandem and hybrid configurations, we see no correlation between the amount of data used for pretraining (which varied between 15 and 85 hours) and the WER obtained by the fine-tuned system.

For the different DNN configurations shown in figures 2 and 3, we pick the ones with the lowest WER on the development set and use them to decode the evaluation set. The results are shown in in tables 2 and 3. For the different amounts



(a) 15h hybrid system

(b) 5h hybrid system

(c) 1h hybrid system

Fig. 1. German development set WERs for hybrid systems with different sizes of hidden layers (512, 1024 and 2048 hidden units) for the three training sets.

Table 2. Tandem system WER results on German eval set

System description	Amount of training data		
	15hr	5hr	1hr
ML using LDA+MLLT	24.53	27.56	34.08
DNN random initialised	22.05	25.10	31.84
DNN pretrained on GE	21.39	24.60	30.91
DNN pretrained on PO	21.21	24.43	31.29
DNN pretrained on SP	21.48	24.23	30.74
DNN pretrained on SW	21.62	24.44	30.52
DNN pretrained on All	21.48	24.49	30.98

Table 3. Hybrid system WER results on German eval set

System description	Amount of training data		
	15hr	5hr	1hr
fBMMI+BMMI using LDA+MLLT	24.13	27.08	33.11
DNN random initialised	21.52	25.03	33.54
DNN pretrained on GE	20.09	22.78	28.70
DNN pretrained on PO	20.00	22.44	28.79
DNN pretrained on SP	20.03	22.64	28.40
DNN pretrained on SW	20.20	22.89	28.92
DNN pretrained on All	20.14	22.70	28.72

of training data, the best HMM-GMM, tandem, and hybrid results are summarised in figure 4.

4. DISCUSSION

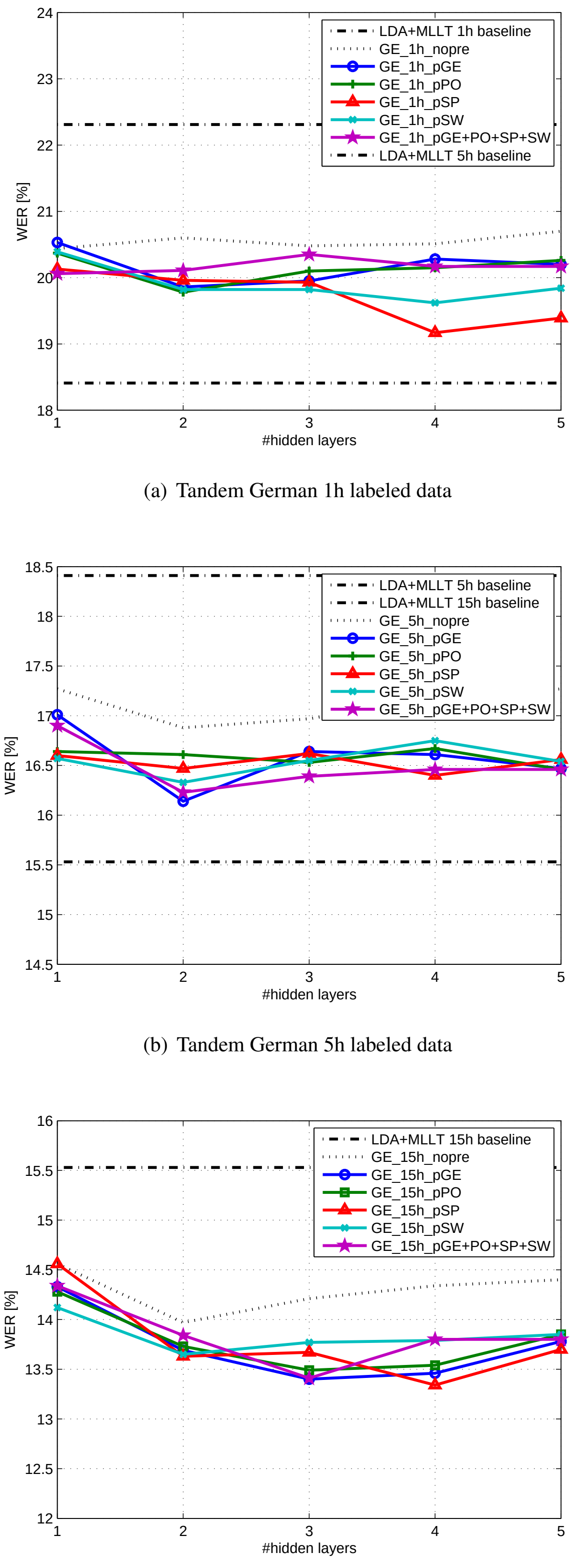
In this work we examined the usability of unlabeled data from one or more languages to improve recognition accuracy of a different, possibly low-resourced, language in a fully unsupervised fashion. These experiments suggest that unsupervised RBM-based initialisation of DNNs is language-independent, allowing hybrid setups to be built from as little as 1 hour of labelled fine-tuning data. This simple approach reduces the cost of building an ASR system in a new language by not only requiring less transcribed data, but less amount of data to be collected in the first place.

One may think of cross-lingual speech recognition as an exercise in judicious application of prior knowledge, whether in the linguistic sense of mapping between phonesets, or in the statistical sense of sharing model parameters between languages. Unsupervised pretraining of DNNs fits in this framework. In fact, Erhan et al. [28] explain unsupervised pretraining as "an unusual form of regularization" that restrict the subsequent supervised (and discriminative) learning to points of the parameter space corresponding to a better generative model of the data. Our results strongly suggest that RBM-based unsupervised pretraining is able to learn characteristics of human speech that are largely language-independent. It is possible, even likely, that this characteristic will be demonstrated by other unsupervised pretraining techniques as well, for example, pretraining using stacked autoencoders [20]

While pretraining is seen to be language-independent, no clear pattern emerges when going from 15 to 85 hours of data for pretraining. This raises two questions that have not been sufficiently addressed in literature: what makes some data suitable for unsupervised pretraining, and what are sufficient amounts of suitable pretraining data. It is possible that cross-corpus variability offset gains from pretraining on a mixture of languages; it is also possible that more data is simply not necessary. Better embeddings of the data may be obtained by imparting domain knowledge: for example, pretraining and fine-tuning in a speaker-adaptive fashion may be helpful in a cross-lingual setting. Finally, our approach is complimentary to other cross-lingual ASR approaches, and it is easy to imagine combining cross-lingual DNNs and SGMMs using the tandem approach.

5. REFERENCES

[1] T Schultz and A Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

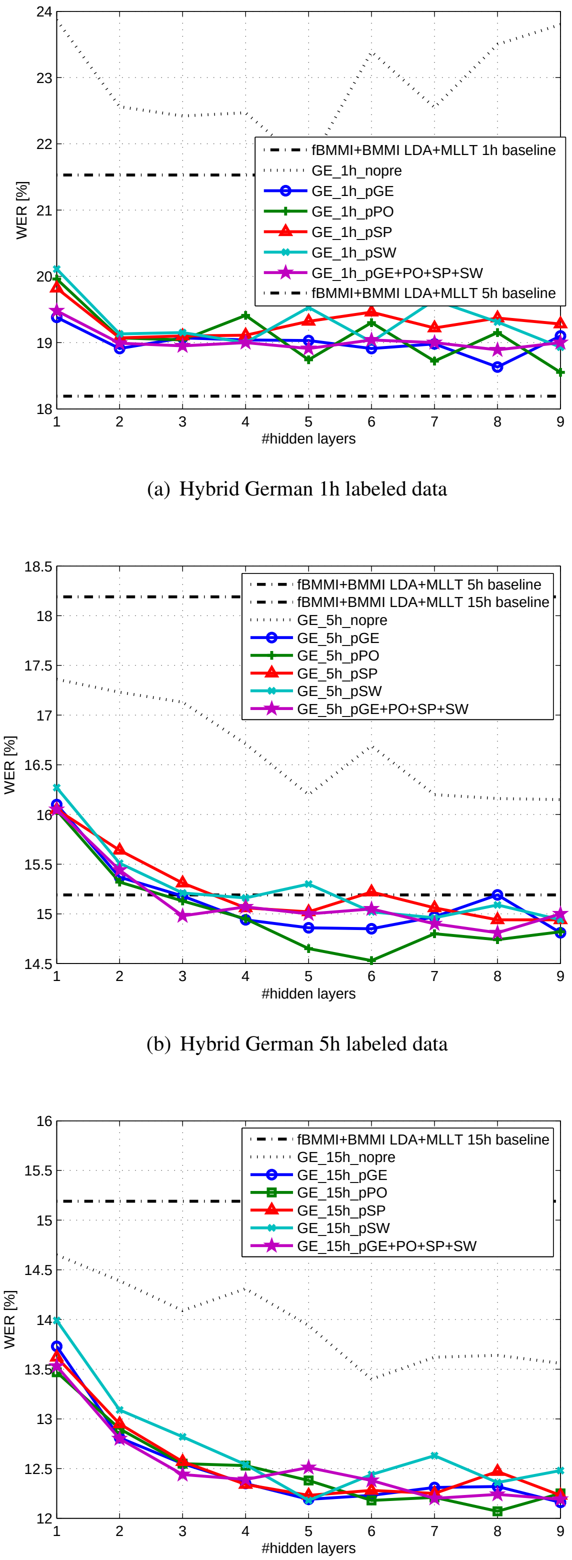


(a) Tandem German 1h labeled data

(b) Tandem German 5h labeled data

(c) Tandem German 15h labeled data

Fig. 2. Tandem HMM-GMM setup. Results on devset.



(a) Hybrid German 1h labeled data

(b) Hybrid German 5h labeled data

(c) Hybrid German 15h labeled data

Fig. 3. Hybrid setup. Results on devset.

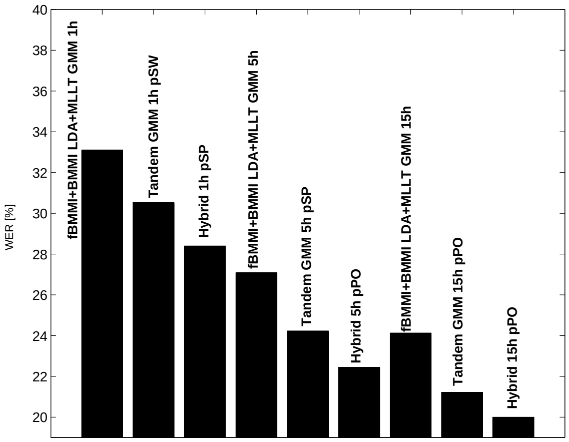


Fig. 4. A summary of results on the evaluation set.

[2] T Schultz and A Waibel, "Experiments on cross-language acoustic modeling," in *Proc. Eurospeech*, 2001.

[3] F Grézl, M Karafiát, and M Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. IEEE ASRU*, 2011.

[4] S Thomas, S Ganapathy, and H Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. IEEE ICASSP*, 2012.

[5] D Imseng, H Bourlard, and PN Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *Proc. IEEE ICASSP*, 2012.

[6] L Burget, P Schwarz, M Agarwal, P Akyazı, K Feng, A Ghoshal, O Glembek, N Goel, M Karafiát, D Povey, A Rastrow, RC Rose, and S Thomas, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proc. IEEE ICASSP*, 2010.

[7] L Lu, A Ghoshal, and S Renals, "Regularized subspace Gaussian mixture models for cross-lingual speech recognition," in *Proc. IEEE ASRU*, 2011.

[8] NT Vu, F Kraus, and T Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in *Proc. IEEE ICASSP*, 2011.

[9] G Hinton, S Osindero, and Y Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[10] H Hermansky, DPW Ellis, and S Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. IEEE ICASSP*, 2000.

[11] F Grézl, M Karafiát, S Kontár, and J Černocký, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. IEEE ICASSP*, 2007.

[12] H Bourlard and N Morgan, *Connectionist Speech Recognition—A Hybrid Approach*, Kluwer Academic, 1994.

[13] S Renals, N Morgan, H Bourlard, M Cohen, and H Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 161–174, 1994.

[14] AJ Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 298–305, 1994.

[15] A Mohamed, GE Dahl, and G Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[16] GE Dahl, D Yu, L Deng, and A Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[17] A Stolcke, F Grézl, M-Y Hwang, X Lei, N Morgan, and D Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. IEEE ICASSP*, 2006.

[18] S Thomas and H Hermansky, "Cross-lingual and multistream posterior features for low resource LVCSR systems," in *Proc. Interspeech*, 2010.

[19] O Çetin, M Magimai-Doss, K Livescu, A Kantor, S King, C Bartels, and J Frankel, "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs," in *Proc IEEE ASRU*, 2007.

[20] Y Bengio, P Lamblin, D Popovici, and H Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pp. 153–160. MIT Press, 2007.

[21] F Seide, G Li, X Chen, and D Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE ASRU*, 2011.

[22] T Schultz, "GlobalPhone: a multilingual speech and text database developed at Karlsruhe University," in *Proc. ICLSP*, 2002.

[23] P Lal, *Cross-Lingual Automatic Speech Recognition using Tandem Features*, Ph.D. thesis, The University of Edinburgh, 2011.

[24] D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlíček, Y Qian, P Schwarz, J Silovský, G Stemmer, and K Veselý, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, December 2011.

[25] R Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. IEEE ICASSP*, May 1998, vol. 2, pp. 661–664.

[26] D Povey, D Kanevsky, B Kingsbury, B Ramabhadran, G Saon, and K Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. IEEE ICASSP*, 2008, pp. 4057–4060.

[27] J Bergstra, O Breuleux, F Bastien, P Lamblin, R Pascanu, G Desjardins, J Turian, D Warde-Farley, and Y Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. SciPy*, 2010.

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