IMPACT OF DEEP MLP ARCHITECTURE ON DIFFERENT ACOUSTIC MODELING TECHNIQUES FOR UNDER-RESOURCED SPEECH RECOGNITION

David Imseng[1], *Petr Motlicek*[1], *Philip N. Garner*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{dimseng,motlicek,pgarner,bourlard}@idiap.ch

ABSTRACT

Posterior based acoustic modeling techniques such as Kullback–Leibler divergence based HMM (KL-HMM) and Tandem are able to exploit out-of-language data through posterior features, estimated by a Multi-Layer Perceptron (MLP). In this paper, we investigate the performance of posterior based approaches in the context of under-resourced speech recognition when a standard three-layer MLP is replaced by a deeper five-layer MLP. The deeper MLP architecture yields similar gains of about 15% (relative) for Tandem, KL-HMM as well as for a hybrid HMM/MLP system that directly uses the posterior estimates as emission probabilities. The best performing system, a bilingual KL-HMM based on a deep MLP, jointly trained on Afrikaans and Dutch data, performs 13% better than a hybrid system using the same bilingual MLP and 26% better than a subspace Gaussian mixture system only trained on Afrikaans data.

Index Terms— KL-HMM, Tandem, hybrid system, deep MLPs, under-resourced speech recognition

1. INTRODUCTION

Under-resourced speech recognition is a very challenging task. The main reason for this is the large amount of data that is usually required to train current recognizers. Therefore, acoustic modeling techniques that are able to exploit out-of-language data such as Kullback–Leibler divergence based HMM (KL-HMM) [1], Tandem [2] or Subspace Gaussian mixture models (SGMMs) [3] have been developed and extensively studied. KL-HMM and Tandem both exploit out-of-language data through posterior features, estimated by a Multi-Layer Perceptron (MLP) that was trained on out-of-language data. SGMMs on the other hand exploit out-of-language data through parameter sharing.

Recently, it has been shown that deep MLP architectures can greatly improve the performance of automatic speech recognition (ASR) systems [4]. Most deep MLP based ASR studies use hybrid HMM/MLP systems, where the MLP output is directly used to model the emission probability of the HMM states. However, if the MLP output is used as a feature [5, 6], conclusions tend to be more ambiguous, i.e. it is not clear if deeper MLP architectures are beneficial.

In this study, we build on our previous results [1] and investigate how deep MLP architectures affect the performance of posterior based acoustic modeling techniques that are particularly well suited for under-resourced ASR. As an additional reference point, we also evaluate SGMMs that do not rely on posterior features.

Taking Afrikaans as a representative of an under-resourced language (target language), we use large amounts of out-of-language data to improve an Afrikaans speech recognizer. Since Afrikaans is similar to Dutch, we intuitively expect that Dutch data provides most benefit for an Afrikaans speech recognizer [7]. Indeed, we already compared how English, Dutch and Swiss German data influence the performance of an Afrikaans speech recognizer and found that Dutch data yielded most improvement [8]. Hence, in this paper, we will use Dutch as a representative of the well-resourced language. In this context, we already compared phoneme accuracies of KL-HMM, Tandem, SGMM, conventional maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation systems [1]. Here, we compare word error rates (WERs) of KL-HMM, Tandem, SGMM and hybrid HMM/MLP systems. For KL-HMM, Tandem and HMM/MLP, we also investigate the impact of a deep MLP compared to the standard MLP.

The remainder of this paper is structured as follows: Section 2 described the databases that are used in this work. Section 3 then introduces all the investigated systems and Section 4 presents the experimental results.

The work was supported by Samsung Electronics Co. Ltd, South Korea, under the project "Multi-Lingual and Cross-Lingual Adaptation for Automatic Speech Recognition", and by Eurostars Programme powered by Eureka and the European Community under the project "D-Box: A generic dialog box for multilingual conversational applications".

ID	Language	number of	Amount of	
		phonemes	trn data	test data
AF	Afrikaans	38	3 h	50 min
CGN	Dutch	47	81 h	-

Table 1. Summary of the different languages with number of phonemes and amount of available data.

2. DATABASES

We used data from Afrikaans and Dutch as summarized in Table 1. In this section, we describe the two databases.

2.1. LWAZI

The Afrikaans data is available from the LWAZI corpus provided by the Meraka Institute, CSIR, South Africa[1] and described by [9]. The database consists of 200 speakers, recorded over a telephone channel at 8 kHz. Each speaker produced approximately 30 utterances, where 16 were randomly selected from a phonetically balanced corpus and the remainder consisted of short words and phrases.

The Afrikaans database comes with a dictionary [10] that defines the phoneme set containing 38 phonemes (including silence). The dictionary that we used contained 1,585 different words. The HLT group at Meraka provided us with the training and test sets. In total, about 3 h of training data and 50 min of test data is available (after voice activity detection).

The bi-gram language model, built on the training sentences, has 1.1% out-of-vocabulary words and a perplexity of about 19 on the test set.

2.2. Corpus Gesproken Nederlands

We used data of the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) [11] that contains standard Dutch pronounced by more than 4,000 speakers from the Netherlands and Flanders. The database is divided into several subsets and we only used *Corpus o* because it contains phonetically aligned *read* speech data pronounced by 324 speakers from the Netherlands and 150 speakers from Flanders. *Corpus o* uses 47 phonemes and contains 81 h of data after the deletion of silence segments that are longer than one second. It was recorded at 16 kHz, but since we use the data to perform ASR on Afrikaans, we downsampled it to 8 kHz prior to feature extraction.

3. SYSTEMS

In this section, we describe the systems under investigation. The systems can be divided into three different categories: (a) monolingual systems, using only Afrikaans data; (b) crosslingual systems, using only Dutch data during MLP training; and (c) bilingual systems, using Afrikaans and Dutch data during MLP training. This, coupled with the various different architectures, leads to quite a lot of systems. For a summary, see Table 5.

Afrikaans	HL	HU	OU	TRN	DEV
Standard	1	1,366	1,447	35.0%	30.8%
Deep	3	6,636	1,447	41.8%	35.0%

Table 2. Summary of the Afrikaans MLP training. The number of hidden layers (HL), the total number of hidden units (HU) and the number of output units (OU) are given. Frame accuracies on the training (TRN) and cross-validation set (DEV) are shown as well. Note that we fixed the number of hidden units to be the same than for the Dutch MLPs presented in Section 3.2.

3.1. Monolingual systems

The monolingual systems serve as reference systems only. In this paper, we evaluate a conventional HMM/GMM system, an SGMM system and two hybrid HMM/MLP systems, one based on a three-layer MLP (standard hybrid system) and one based on a five-layer MLP (deep hybrid system).

3.1.1. HMM/GMM

The HMM/GMM system is a standard cross-word context-dependent speech recognizer that models each triphone with three states and is based on 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) features (C_0–$C_{12} + \Delta + \Delta\Delta$), extracted with the HTK toolkit [12]. As usually done, we first trained context-independent monophone models that were then used as seed models for the context-dependent triphone models. We used eight Gaussians per state to model the emission probabilities. To balance the number of parameters with the amount of available training data, we applied conventional state tying with a decision tree that is based on the minimum description length principle [13], resulting in 1,447 tied states.

3.1.2. Monolingual SGMM

The SGMM acoustic modeling technique allows compact representation of large collection of mixture-of-Gaussian models and has shown its capability to outperform conventional HMM/GMMs in monolingual as well as cross- or multi-lingual scenarios [3, 14]. For the monolingual SGMM system, we trained all the parameters from Mel-Frequency cepstrum coefficients (MFCCs), using Afrikaans data only. In total we used 500 Gaussians and the substate phone-specific vectors had 50 dimensions.

[1] http://www.meraka.org.za/hlt

Dutch	HL	HU	OU	TRN	DEV
Standard	1	1,366	1,789	59.0%	56.5%
Deep	3	6,636	1,789	64.2%	60.3%

Table 3. Summary of the Dutch MLP training. The number of hidden layers (HL), the total number of hidden units (HU) and the number of output units (OU) are given. Frame accuracies on the training (TRN) and cross-validation set (DEV) are shown as well.

3.1.3. Monolingual HMM/MLP

The monolingual HMM/MLP systems used the same 1,447 tied states as the HMM/GMM system presented in Section 3.1.1. For the standard hybrid system, we trained a three-layer MLP and for the deep hybrid system, we trained a five-layer MLP (each hidden layer had similar number of hidden units) using Quicknet software [15]. We randomly split the three hours of Afrikaans training data into an *MLP training set* (90%) and an *MLP cross-validation set* (10%). We trained the MLPs from the 39-dimensional MF-PLP features in a nine frame temporal context (four preceding and following frames). More details about the MLP training are given in Table 2. For this study, the only difference between the three-layer and the five-layer network was in the number of parameters (and in the number of hidden layers). We did not employ more elaborated training procedures such as pre-training or dropout. The resulting posterior probabilities were divided by the priors and then directly used as emission probabilities.

3.2. Crosslingual systems

The crosslingual systems exploit Dutch data during MLP training. More specifically, we trained a standard and a deep MLP with all the available Dutch data. As we already did in earlier studies [1], we developed a standard HMM/GMM system with all the Dutch training data to obtain 1,987 tied states targets. We set the number of parameters for the standard MLP to 10% of the available number of training frames, resulting in a hidden layer with 1,366 units. As suggested by studies on deep MLPs [16], we targeted about 2000 hidden units per layer in the deeper MLP and therefore set the number of parameters to 50% of the available number of training frames, leading to a total of 6,636 hidden units distributed to three hidden layers. We used 90% of the training set for MLP training and 10% for cross-validation to stop training. More details about the MLP training are given in Table 3.

In this paper, we investigated two approaches that benefit from exploiting out-of-language data through posterior features: Tandem and KL-HMM. In both approaches, Afrikaans data is passed through the MLP trained on Dutch and the resulting posterior features are then used to train the HMM parameters (see Sections 3.2.1 and 3.2.2). Since the hybrid HMM/MLP approach is bound to the tied states target used during the MLP training, we did not evaluate a crosslingual HMM/MLP system. However, as an additional reference point, we also evaluated a crosslingual SGMM system that did not use the Dutch posterior features, but used the Dutch data for global parameter training as described in [1].

3.2.1. Crosslingual Tandem

Similar to the conventional HMM/GMM system, for the Tandem system, we trained context-independent monophone models that served as seed models for the three-state context-dependent triphone models. Because of the ambiguous results from earlier studies [5, 6], we evaluate a *standalone Tandem* system (similar to the system in [6]) as well as an *augmented Tandem* system, where *augmented* refers to our concatenating of MF-PLP features with the posterior features (similar to the system in [5]). We used eight Gaussians per state to model the emission probabilities. As in our previous study [1], we used PCA for dimensionality reduction and fixed the dimensionality such that 99% of the variance was preserved. This procedure resulted in 286-dimensional features (we used the same feature dimensionality for the posteriors of the standard and the deep MLP). To have comparable Tandem systems, we run PCA again after concatenating MF-PLP features with posterior features and reduced the dimensionality to 286.

3.2.2. Crosslingual KL-HMM

The KL-HMM acoustic modeling technique can directly model raw posterior features. Therefore no post-processing is necessary. In the KL-HMM acoustic modeling approach, the HMM states are parametrized with reference posterior distributions (categorical distributions) that can be trained by minimizing the Kullback–Leibler divergence between the categorical distributions and the posterior features. More details about training and decoding in the KL-HMM framework can be found in, for instance, [1]. Similar to HMM/GMM and Tandem, the KL-HMM system was trained based on the context-independent monophone models that served as seed models for the three-state context-dependent triphone models. For KL-HMM, we applied a decision tree clustering reformulated as dictated by the KL criterion [17]. We found in our previous study that the best KL-HMM performance is achieved with a fully developed tree (about 15,000 tied states), therefore we did the same for this study.

3.2.3. Crosslingual SGMM

SGMMs can be naturally exploited in under-resourced scenarios, since most of the model parameters can be estimated on well-resourced datasets. Therefore, we use the crosslingual SGMM system as an additional reference point in this study. To exploit out-of-language data, the SGMM model parameters can be divided into HMM-state specific and shared

AF & Dutch	HL	HU	OU	TRN	DEV
Standard	1	1,366	1,447	48.3%	38.3%
Deep	3	6,636	1,447	53.1%	42.1%

Table 4. Summary of the MLP trained on Dutch first and the re-trained on Afrikaans. The number of hidden layers (HL), the total number of hidden units (HU) and the number of output units (OU) are given. Frame accuracies on the training (TRN) and cross-validation set (DEV) are shown as well.

parameters. The crosslingual SGMM used Dutch data during training of the globally-shared (language-independent) parameters and Afrikaans data for the training of the HMM-state specific parameters [3]. Similar to the monolingual SGMM system, we used 500 Gaussians and the substate phone-specific vectors had 50 dimensions.

3.3. Bilingual systems

Inspired by a recent study [6], the bilingual systems that we present are based on MLPs that were trained on Afrikaans and Dutch data. More specifically, we took the two Dutch MLPs (standard and deep) trained in Section 3.2 and removed the output layer. Then, we appended a new randomly initialized output layer and trained the MLP (all layers) to estimate posterior probabilities for the 1,447 Afrikaans tied states by using Afrikaans data. More details about the MLP training are given in Table 4. In this study, we investigated three acoustic modeling techniques that are able to exploit the posterior probabilities estimated with the bilingually trained MLP: hybrid HMM/MLP, Tandem and KL-HMM. Again, SGMM serves as a reference not using posterior features.

3.3.1. Bilingual HMM/MLP

The bilingual HMM/MLP systems are essentially the same systems as the monolingual HMM/MLP ones presented in Section 3.1.3. The monolingual HMM/MLP systems used the posterior probabilities estimated with the MLP only trained on Afrikaans data, and the bilingual HMM/MLP systems employed the posterior probabilities estimated with the MLP first trained on Dutch data and then re-trained on Afrikaans data.

3.3.2. Bilingual Tandem

Similar to the crosslingual Tandem systems, presented in Section 3.2.1, we trained a standalone and an augmented Tandem system based on three-state context-dependent triphone models. We used eight Gaussians per state to model the emission probabilities and used PCA for decorrelation. To preserve 99% of the variance we reduced the feature dimensionality to 146.

3.3.3. Bilingual KL-HMM

The bilingual KL-HMM system resembles the crosslingual KL-HMM system, presented in Section 3.2.2. The 1,789 dimensional Dutch posterior features were replaced by 1,447 dimensional feature vectors, trained on Dutch and on Afirkaans data.

3.3.4. Bilingual SGMM

The bilingual SGMM system used Dutch and Afrikaans data during training of the globally-shared parameters and Afrikaans data only for the training of the HMM-state specific parameters. We used 500 Gaussians and the substate phone-specific vectors had 50 dimensions.

4. EXPERIMENTS

In this section, we first discuss the hypotheses under investigation, then present the experimental results.

4.1. Prior expectations

Given the systems described in Section 3, we hypothesize the following:

1. Based on the success of deep architectures in recent studies [4], we hypothesize that the deep MLP architectures yield improvement for all systems.
2. Recent literature [5] suggests that adding hidden layers does not improve the performance of a augmented Tandem system. We therefore assume that MLP output post-processing reduces the performance gain resulting from deeper MLP architectures and hypothesize that:
   (a) hybrid systems gain most from a deeper MLP architecture because they directly use the estimated posteriors probabilities as emission probabilities.
   (b) KL-HMM gains more than Tandem because the posterior features are directly modeled without post-processing.
3. Multilingual data was successfully used to generate deep neural network features for low resource speech recognition [6]. Therefore, we hypothesize that the gains from the deep MLP architecture and the out-of-language data exploitation are complementary.

4.2. Results

The experimental results are summarized in Table 5. All the systems based on deep MLPs outperform the equivalent system based on the standard MLP, hence hypothesis 1 is demonstrated.

	System	Std.	Deep	Rel. Gain
Monoling.	HMM/GMM	11.4%	-	-
	SGMM	9.5%	-	-
	HMM/MLP	12.3%	9.9%	20%
Crossling.	Tandem	10.5%	9.4%	10%
	+MF-PLP	9.7%	9.5%	2%
	KL-HMM	9.6%	9.0%	6%
	SGMM	8.5%	-	-
Biling.	HMM/MLP	9.3%	8.0%	14%
	Tandem	9.9%	8.4%	15%
	+MF-PLP	9.7%	8.9%	8%
	KL-HMM	8.0%	7.0%	13%
	SGMM	8.5%	-	-

Table 5. Achieved word error rates (WERs) of the monolingual, crosslingual and bilingual systems described in Section 3. *Std.* stands for the standard (three-layer) MLP and *deep* for the deep (five-layer) MLP. The relative gain by using the deeper MLP is also given.

For the bilingual scenario, HMM/MLP, KL-HMM and standalone Tandem yield very similar improvement if the standard and deep MLP performance are compared. Therefore we must reject hypothesis 2. We evaluated a standalone and an augmented Tandem system. Our results are in line with earlier studies [5, 6] where it was found that deep MLPs yield improvement for standalone systems [6], but only to a limited extend for augmented Tandem systems [5]. It seems reasonable to conclude that the concatenation of the MLP output with MF-PLP features diminishes the advantage of the deep MLP architecture.

Although the experimental results suggest that the relative gain decreases in cross- and bi-lingual scenarios compared to the monolingual HMM/MLP system, it seems that the gains from out-of-language data exploitation and a deep MLP architecture are still complementary. Thus, hypothesis 4 is demonstrated.

The bilingual KL-HMM systems yields the best performance (13% relative improvement compared to the hybrid HMM/MLP system). We attribute the advantage of the KL-HMM system to the fact that the hybrid system is bound to the tied state targets used during the MLP training. Hence the hybrid system uses about 1,500 tied states. The KL-HMM system on the other hand is more flexible and allows more tied states to be used, even in under-resourced scenarios. The parsimonious use of parameters of the KL-HMM system (categorical distributions) allows training of an HMM with 15,000 tied states, only using three hours of Afrikaans data.

Furthermore, Table 5 also reveals that the crosslingual and the bilingual SGMM perform similarly. The crosslingual environment is particularly well suited for the SGMM system because the shared parameters can be trained on Dutch data and the language specific parameters on Afrikaans data. In the bilingual case however, the 3 h of Afrikaans data are dominated by the 80 h of Dutch data during the shared parameter training. The MLP based systems yield more improvement from the bilingual setup because the MLPs estimate Afrikaans tied states posteriors instead of Dutch tied states posteriors in the crosslingual case.

5. CONCLUSION

We investigated under-resourced speech recognition in the context of an Afrikaans speech recognizer that benefits from Dutch data, and compared how the performance of posterior based approaches changes if a standard three-layer MLP is replaced by a deeper five-layer MLP. We have shown that the deeper MLP structure equally improved a hybrid HMM/MLP and a standalone Tandem system as well as a KL-HMM system. Further, experiments revealed that gains from the deeper MLP architecture and out-of-language data exploitation are complementary. The best performing bilingual system, KL-HMM based on the MLP that was jointly trained on Afrikaans and Dutch data, performs 13% better than a hybrid system using the same bilingual MLP and yields 26% relative improvement if compared to a monolingual SGMM system only trained on Afrikaans data.

We therefore conclude that deep MLP architectures are suitable for under-resourced speech recognition, with the KL-HMM being the most promising.

6. REFERENCES

[1] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, “Using out-of-language data to improve an under-resourced speech recognizer,” *Speech Communication*, 2013.

[2] A. Stolcke, F. Grezl, 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. of ICASSP*, 2006, vol. 1, pp. 321–324.

[3] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas, “Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models,” in *Proc. of ICASSP*, 2010, pp. 4334–4337.

[4] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[5] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2012, pp. 246–251.

[6] S Thomas, M. L. Seltzer, Church K., and Hermansky H., "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. of ICASSP*, 2013, pp. 6704–6708.

[7] W. Heeringa and F. de Wet, "The origin of Afrikaans pronunciation: a comparison to west Germanic languages and Dutch dialects," in *Proceedings of the Conference of the Pattern Recognition Association of South Africa*, 2008, pp. 159–164, `www.let.rug.nl/heeringa/dialectology/papers/prasa08.pdf`.

[8] D. Imseng, H. Bourlard, and P. N. Garner, "Boosting under-resourced speech recognizers by exploiting out of language data - case study on Afrikaans," in *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages*, 2012, pp. 60–67.

[9] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," in *Proc. of Interspeech*, 2009, pp. 2847–2850.

[10] M. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proc. of Interspeech*, 2009, pp. 2851–2854.

[11] N. Oostdijk, "The spoken Dutch corpus. Overview and first evaluation.," in *In Proceedings of the Second International Conference on Language Resources and Evaluation*, 2000, vol. II, pp. 887–894.

[12] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[13] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. of Eurospeech*, 1997, pp. 99–102.

[14] P. Motlicek, D. Povey, and M. Karafiat, "Feature and score level combination of subspace Gaussians in LVCSR task," in *Proc. of ICASSP*, 2013, pp. 7604–7608.

[15] D. Johnson, "ICSI quicknet software package," `http://www.icsi.berkeley.edu/Speech/qn.html`, 2004.

[16] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[17] D. Imseng, J. Dines, P. Motlicek, P. N. Garner, and H. Bourlard, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proc. of Interspeech*, 2012.