BOOSTING UNDER-RESOURCED SPEECH RECOGNIZERS BY EXPLOITING OUT OF LANGUAGE DATA - CASE STUDY ON AFRIKAANS

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

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

ABSTRACT

Under-resourced speech recognizers may benefit from data in languages other than the target language. In this paper, we boost the performance of an Afrikaans speech recognizer by using already available data from other languages. To successfully exploit available multilingual resources, we use posterior features, estimated by multilayer perceptrons that are trained on similar languages. For two different acoustic modeling techniques, Tandem and Kullback-Leibler divergence based HMMs, the proposed multilingual system yields more than 10% relative improvement compared to the corresponding monolingual systems only trained on Afrikaans.

Index Terms— Multilingual speech recognition, posterior features, under-resourced languages, Afrikaans

1. INTRODUCTION

Previous studies have shown that Automatic Speech Recognition (ASR) may benefit from data in languages other than the target language only under certain conditions such as there being less than one hour of data for the training language [1, 2]. Usually, a language with large amounts of training data was used to simulate low amounts of training data [1, 2]. For instance Niesler [3] studied the sharing of resources on real under-resourced languages, including Afrikaans, inspired by multilingual acoustic modeling techniques proposed by Shultz and Waibel [4]. However, only marginal ASR performance gains were reported.

In this paper, we focus on Afrikaans and show how to significantly boost the performance of an Afrikaans speech recognizer that was trained on three hours of within language data, by using more than 100 hours of out of language data.

Standard ASR systems typically make use of phonemes as subword units to model human speech production. A phoneme is defined as the smallest sound unit of a language that discriminates between a minimal word pair [5, p. 78]. Although humans are able to produce a large variety of acoustic sounds, we assume that all those sounds, referred to as phones, across speakers and languages, share a common acoustic space $\mathcal{X}$. Of course, no single language makes use of all phones, and most languages only partially cover $\mathcal{X}$. However, we found in previous studies [1, 6] that the relation between phonemes of different languages can 1) be learned and 2) be exploited for cross-lingual acoustic model training or adaptation. Furthermore, we found that posterior features, estimated by multilayer perceptrons (MLPs), are particularly well suited for such tasks.

Inspired by this work, we use posterior features, estimated by MLPs that are trained on similar languages such as English, Dutch and Swiss German to successfully exploit available multilingual resources. According to the trees summarized by Blažek [7], Afrikaans and Dutch are Istveonic Germanic languages whereas British English and Swiss German are also Germanic languages, but located on different branches, namely Ingveonic and Erminonic Germanic, respectively. Intuitively, we would expect that Dutch data should provide most benefit. Previous studies [8] and a similarity analysis of Heeringa and de Wet [9] underpin this assumption.

Using two different acoustic modeling techniques for posterior features, namely Tandem [10] and Kullback-Leibler divergence based hidden Markov models (KL-HMM) [11], we investigate:

- *Crosslinguality:* We study how out of language data can be used to improve ASR performance of an under-resourced language and briefly discuss if there is a relation between similarity of the other language and performance gain on the target language.
- *Multilinguality:* We combine the resources of multiple languages in the form of posterior features by concatenating MLP outputs to boost ASR performance.
- *Context-dependency:* Since there are large amounts of out of language data, we enrich the exploited information by adding context dependency. More specifically, we train the MLPs on context-dependent targets.

This research was supported by the Swiss NSF through the project Interactive Cognitive Systems (ICS) under contract number 200021_132619/1 and the National Centre of Competence in Research (NCCR) in Interactive Multimodal Information Management (IM2) http://wwww.im2.ch

- *Context-dependent multilinguality:* Given the above approaches, we study the combination of outputs of multiple MLPs that were trained on context-dependent targets.

We first give a brief description of both applied acoustic modeling techniques in Section 2. In Section 3, we then present the databases that we used for the training of the MLPs as described in Section 4, and give an overview over the investigated systems in Section 5. Previous multi- and cross-lingual posterior feature studies that used more than one hour of target language data reported rather small or no improvements (up to 3.5% relative) [12, 13]. We will present experiments and results in Section 6 and show that the proposed systems yield more than 10% relative improvement compared to the monolingual recognizer for both acoustic modeling techniques.

2. ACOUSTIC MODELING

We study two different approaches to model posterior features: Tandem [10], illustrated in Figure 1 and a Kullback-Leibler divergence based HMM (KL-HMM) [11], illustrated in Figure 2. Both approaches involve the training/estimation of two different kind of distributions:

- *Phoneme posterior features:* The phoneme posterior features are phoneme posterior probabilities given the acoustics and estimated with an MLP that can be trained on any auxiliary dataset. Therefore we call it an *auxiliary MLP* and choose an out of language dataset with large amounts of available data with which to train. The language of the training data determines the number of output units K (number of phonemes) of the MLP. More details about the MLP training are given in Section 4.

  Once the MLP is trained, we consider a sequence of T acoustic feature vectors $X = \{x_t, \ldots, x_T\}$, namely Perceptual Linear Prediction (PLP) features, extracted from within language data. As seen in Figure 2, the phoneme posterior sequence $Z = \{z_1, \ldots, z_T\}$ is then estimated with the previously trained auxiliary MLP. To estimate $z_t = (z_t^1, \ldots, z_t^K)^\mathsf{T}$, we consider a nine frame temporal context $\{x_{t-4}, \ldots, x_{t+4}\}$. The described phoneme posterior estimation is identical for both acoustic modeling techniques.

- *HMM state distributions:* The HMM states $q^d : d \in \{1, \ldots, D\}$ are associated with the target language. Each phoneme of the target language is modeled with three states, thus the total number of states D is equal to three times the number of phonemes of the target language.

  The HMM state distributions consist of emission and transition probabilities. Based on anecdotal knowledge, we fix the transition probabilities a_{ij} for both acoustic modeling techniques (see Figures 1 and 2). The emission probabilities however are modeled differently for Tandem and KL-HMM. As we will describe later, Tandem (Section 2.1) uses Gaussian mixtures and KL-HMM (Section 2.2) uses a categorical distribution. The emission probabilities are trained from within language data only. Here, we assume that we have access to a limited amount of within language data.

In the remainder of this section, we briefly summarize both acoustic modeling techniques which will be compared to a state-of-the-art HMM/GMM system.

2.1. Tandem

The conventional Tandem approach models the emission probabilities of the HMM states q^d with mixtures of Gaussians. Figure 1 illustrates the HMM associated with a three-state-phoneme (q^1, q^2, q^3). To model the emission probabilities with Gaussians, the posterior features z_t need to be post-processed. More specifically, the log phoneme posteriors are decorrelated with a principal component analysis (PCA). The transformation matrix can be estimated on within language data. Usually, the resulting feature vector $w_t = (w_t^1, \ldots, w_t^L)^\mathsf{T}$, has a reduced dimensionality L.

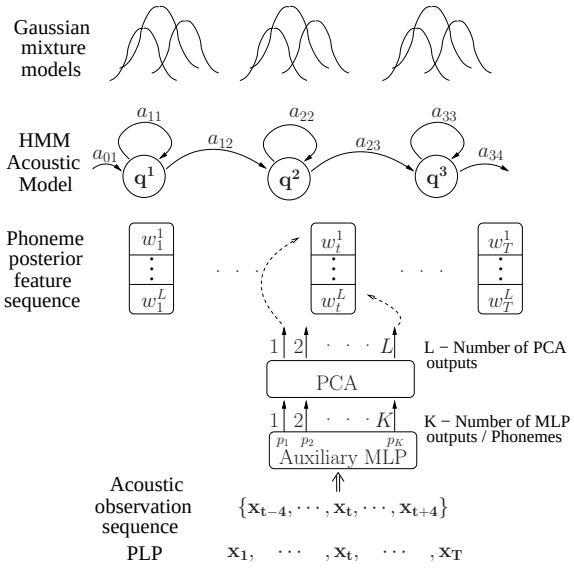

Fig. 1. Tandem: the emission probabilities of the HMM states are modeled with Gaussian mixtures and the MLP output is post-processed. For more details, see Section 5.1.1

2.2. Kullback-Leibler divergence based HMM

As illustrated in Figure 2, a KL-HMM is a particular form of HMM in which the emission probability of state q^d is parametrized by a categorical distribution $y_d =$

$(y_d^1, \ldots, y_d^K)^\mathsf{T}$, where K is the dimensionality of the features. A categorical distribution is a multinomial distribution where only one sample is drawn. In contrast to Tandem that uses Gaussian mixtures and therefore needs the post-processed features w_t, the categorical distributions can directly be trained from phoneme posterior probabilities z_t.

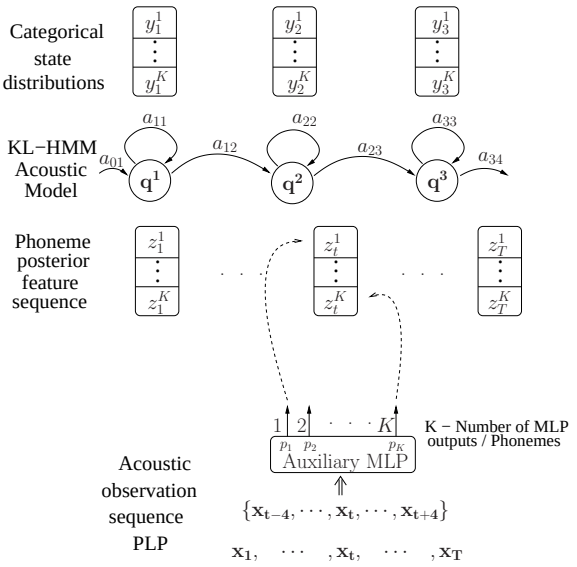

Fig. 2. KL-HMM: the emission probabilities are modeled with categorical distributions and the MLP output can directly be used. More details can be found in Section 5.1.2

For KL-HMM training and decoding, we use the divergence of Kullback and Leibler [14], which is nowadays often referred to as the symmetric variant of the KL divergence, as local score:

$$f_{SKL}(z_t, y_d) = \frac{1}{2} f_{KL}(z_t, y_d) + \frac{1}{2} f_{KL}(y_d, z_t) \qquad (1)$$

where

$$f_{KL}(x, y) = \sum_{k=1}^{K} x(k) \log \frac{x(k)}{y(k)} \qquad (2)$$

The cost function $\mathcal{F}_\mathcal{Q}(Z, Y)$ between the phoneme posterior sequence Z and the set of categorical distributions $Y = \{y_1, \ldots, y_D\}$ can be written as [1]:

$$\mathcal{F}_\mathcal{Q}(Z, Y) = \min_\mathcal{Q} \sum_{t=1}^{T} \left[f_{SKL}(z_t, y_{q_t}) - \log a_{q_{t-1} q_t} \right] \qquad (3)$$

where $\mathcal{Q} = \{q_1, \ldots, q_T\}$ stands for all allowed state paths and y_{q_t} is the categorical distribution associated with q_t, the state at time t. As already mentioned, the transition probabilities $a_{q_{t-1} q_t}$ are fixed. A more detailed description of training and decoding algorithms can be found in [1].

ID	Language	Number of phonemes	Amount of training data
AF	Afrikaans	38	3 h
CGN	Dutch	47	81 h
SZ	Swiss German	59	14 h
EN	British English	45	12.5 h

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

3. DATABASES

We used data from four different languages as summarized in Table 1. In this section, we describe the different databases.

3.1. LWAZI

We used the Afrikaans data that is available from the LWAZI corpus provided by the Meraka Institute, CSIR, South Africa[1] and described by Barnard et al. [15]. 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 consists of short words and phrases.

The Afrikaans database comes with a phoneme set that contains 38 phonemes (including silence) and a dictionary is also available [16]. The dictionary that we used contained 1585 different words. The HLT group at Meraka provided us with the training and test sets that will soon be released as official benchmarking sets. In total, about three hours of training data and 50 minutes of test data is available.

Since we did not have access to an appropriate language model, we trained a bi-gram phoneme model on the training set and only report phoneme accuracies in this study. The bi-gram phoneme model learned the phonotactic constraints of the Afrikaans language and has a phoneme perplexity of 15.

3.2. Corpus Gesproken Nederlands

Heeringa and de Wet [9] reported that standard dutch seems to be the best language from which to borrow acoustic data from, for the development of an Afrikaans ASR system. In this study, we used data of the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) [17] that contains standard Dutch pronounced by more than 4000 speakers from the Netherlands and Flanders. The database is divided into several subsets and we only used "*Corpus o*" that contains phonetically aligned read speech data. Corpus o uses 47 phonemes and contains 81 hours 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.

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

3.3. SpeechDat(II)

We also used data from the British English and the Swiss German part of SpeechDat(II)[2] data. The data is gender-balanced, dialect-balanced according to the dialect distribution in a language region, and age-balanced. The databases have been recorded over the telephone at 8 kHz and are subdivided into different corpora. We only used *Corpus S*, which contains ten read sentences from each of the 2000 speakers. To split the databases into training (1500 speakers), development (150 speakers) and testing (350 speakers) sets, we used the standard procedure that maintains the gender-, dialect- and age-distributions of the database, as described in [18]. Only the training and development portions were used for this study. The British English database uses 45 phonemes and contains 12.5 hours of data and the Swiss German database uses 59 phonemes and contains 14 hours of data.

4. MULTILAYER PERCEPTRONS

For each of the four languages (Afrikaans, Dutch, Swiss German, British English), we trained an MLP from 39 Mel-Frequency Perceptual Linear Prediction (MF-PLP) features (C0-C12+Δ+$\Delta\Delta$) in a nine frame temporal context (four preceding and following frames), extracted with the HTS variant[3] of the HTK toolkit. The number of parameters in each MLP was set to 10% of the number of available training frames, to avoid overfitting. We used Quicknet[4] software to train the MLPs.

- Afrikaans: We performed forced alignment to get the targets for the MLP training. 90% of the training set was used for training and 10% for cross-validation to stop training.
- Dutch: Corpus o of the CGN database is phonetically labeled, thus the targets for the MLP training were available. Again 90% of the data was used for training and 10% for cross-validation.
- Swiss German and British English: We performed forced alignment to get the targets for the MLP training. The standard training sets were used for training and the development sets for cross-validation.

5. SYSTEMS

In this section, we will describe the systems that we investigated to study crosslingual, multilingual and context-dependent aspects in the framework of under-resourced ASR. We will compare the performance of the Tandem approach with the performance of KL-HMM. Furthermore, we will also compare the proposed systems to an HMM/GMM baseline.

[2] http://www.speechdat.org/SpeechDat.html
[3] http://hts.sp.nitech.ac.jp/
[4] http://www.icsi.berkeley.edu/Speech/qn.html

5.1. Monolingual and crosslingual systems

We define a monolingual system as a system where we only use within language data for the training, i.e., the MLP for the feature extraction as well as the HMM for decoding are trained with Afrikaans data only.

A crosslingual system, on the other hand, is defined as a system that uses out of language data for the training of the auxiliary MLP. The HMM is trained from within language data and can either be a standard HMM as used for the conventional Tandem approach or a KL-HMM.

5.1.1. Tandem

As already discussed earlier (see Figure 1), each context-dependent phoneme is modeled with three states (q^i, q^j, q^k). As usually done, we first train context independent monophone models that serve as seed models for the context-dependent phoneme models. We use eight Gaussians per state to model the emission probabilities and use PCA for decorrelation. PCA can also be used to reduce the dimensionality to, for example, 30, as it is typically done [2, 13].

To balance the number of parameters with the amount of available training data, we apply conventional state tying with a decision tree that is based on the minimum description length principle [19]. Tandem training and decoding is performed with HTS.

5.1.2. KL-HMM

As for Tandem, for the KL-HMM system, we train context independent monophone models that serve as seed models for the three-state context-dependent phoneme models.

Since the KL-HMM system is handicapped by a low number of parameters [1] (only one K dimensional vector per state instead of eight Gaussians), we do not perform state tying, but build a decision tree to model unseen contexts during decoding. The decision tree is a modified version of the conventional approach proposed by Young et al. [20] but is based on minimum KL divergence instead of maximum likelihood [21].

5.2. Multilingual systems

As already proposed earlier [13], we can concatenate the output of several MLPs together. We refer to this kind of system to as multilingual. More specifically, we concatenate the output of multiple MLPs and renormalize the resulting vector to guarantee that the feature vectors can be interpreted as posterior distributions, as assumed by the KL-HMM. For the Tandem systems, we post-process the normalized vectors as already described Section 5.1.1.

5.3. Context-dependent MLP output systems

The concatenated MLP outputs of the multilingual systems have a higher dimensionality than the single MLP outputs of the mono- and cross-lingual systems and therefore can carry more information. To enrich the exploited information of the mono- and cross-lingual systems, we also explore MLPs that are trained on context-dependent targets instead of context independent targets [22]. Usually, the MLPs are trained on tied triphone states which are determined with the standard decision tree approach. However, we do not have acoustic models that are necessary to build such decision trees for all the languages. For the CGN database, for example, we have phoneme alignments only, but no acoustic models. Therefore, we limit ourselves to a simpler strategy in this study by setting an occupancy threshold. If a triphone does appear fewer times than a threshold, we simply back off to the monophone model. In our case, we adjusted the occupancy threshold such that the number of MLP outputs is equal to the dimensionality of the multilingual feature vector (which is 189).

6. EXPERIMENTS AND RESULTS

In this section, we analyze the performance of the different systems. For all the significance tests, we used the bootstrap estimation method [23] and a confidence interval of 95%.

6.1. HMM/GMM baseline

As shown in Table 2, we run a standard HMM/GMM system that was directly trained on the PLP features, as a baseline. Note that the results reported by van Heerden et al. [24], 63.1% phoneme accuracy, were the first set of results obtained for the data and the official train and test set were compiled after the official database release. Personal communication with HLT group at Meraka confirmed that the lower performance of our baseline can be attributed to the different data partitioning.

6.2. Monolingual and crosslingual systems

First, we analyze the performance of the mono- and cross-lingual systems. Despite the relatively low amount of Afrikaans training data (3 hours), but based on previous work [1], we expected the monolingual system to perform best. As discussed above, we expected the Dutch system to perform second.

As illustrated in Table 2, we trained the auxiliary MLP on one of the four languages, namely Afrikaans (AF), Dutch (CGN), Swiss German (SZ) and British English (EN), with context-independent targets (monophones). The HMM parameters of the Tandem systems as well as the KL-HMM parameters were always trained on the same Afrikaans data.

In all result tables, bold numbers are significantly better than the other results for a given modeling technique. For the KL-HMM systems, our hypothesis that the Afrikaans MLP performs best is confirmed. We believe that the Swiss German system (SZ-mono) is performing better than the British English system (EN-mono) in the case of KL-HMM because the dimensionality of the German posterior vectors is higher (59) than the English posterior vectors (45) and therefore system SZ-mono has more parameters.

System	Model	Feature dimension	Phoneme accuracy
Baseline	HMM/GMM	39	61.2 %
AF-mono	KL-HMM	38	**58.7** %
CGN-mono		47	**58.0** %
SZ-mono		59	55.3 %
EN-mono		45	52.2 %
AF-mono	Tandem	30	61.2 %
CGN-mono		30	**62.5** %
SZ-mono		30	60.4 %
EN-mono		30	60.4 %

Table 2. Afrikaans phoneme accuracy obtained from monolingual (AF) and cross-lingual (CGN, SZ, EN) systems. The system extension *-mono* stands for the context-independent MLP targets (monophones). The baseline uses PLP features, the KL-HMM systems raw posteriors and the Tandem systems processed posteriors.

For the Tandem technique, however, the Dutch system (CGN-mono) performs best. We believe that this is due to the PCA that transforms the phoneme posteriors and that the Dutch system performs best because it has the most training data available. We note that the Tandem performance correlates with the amount of available training data, whereas the KL-HMM performance tends to correlate with the language similarity.

In general, we observe that the Tandem systems outperform the KL-HMM systems by at least 2.5% absolute. This might be due to the fact that the employed Tandem system has received more attention than the recently proposed KL-HMM system. We observed earlier [1] that the KL-HMM system outperforms the Tandem system in scenarios with small amounts of data (less than one hour). The best Tandem system yields improvement compared to the HMM/GMM baseline, whereas the best KL-HMM system does not.

6.3. Multilingual systems

Although it has not been explicitly confirmed in the literature [3], we believed that the Afrikaans ASR performance could be boosted by properly combining acoustic information from multiple similar languages.

In this study, we investigate a normalized concatenation of MLP outputs. More specifically, we take different MLP outputs, concatenate them and then normalize the resulting feature vector. As shown in Table 2, during the mono- and

System	Model	Feature dimension	Phoneme accuracy
AF-mono	KL-HMM	38	58.7 %
AF-CGN		85	62.4 %
AF-CGN-SZ		144	64.0 %
AF-CGN-SZ-EN		189	**64.4** %
AF-mono	Tandem	30	61.2 %
AF-CGN		30	62.2 %
AF-CGN-SZ		30	62.0 %
AF-CGN-SZ-EN		30	62.1 %
AF-CGN-SZ-EN		189	**66.2** %

Table 3. Afrikaans phoneme accuracy obtained from multilingual systems. We concatenated different MLP outputs. The Tandem systems suffer from the dimensionality reduction that removes significant information.

cross-lingual experiments, for KL-HMM, system AF-mono performed best, then CGN-mono, SZ-mono and EN-mono. Therefore, we concatenate the MLPs one by one in that order.

As shown in Table 3, we get significant improvements for both acoustic modeling techniques, by concatenating MLP outputs.

However, we observe relatively small improvements for the Tandem systems. Grézl et al. [13] did not observe improvements at all if they concatenated MLP outputs. As we are going to discuss in more detail in Section 6.4, a dimensionality reduction during PCA removes some information. A reduction to 30 dimensions for example keeps only 91% of the data variance for system AF-CGN-SZ-EN as displayed in Figure 3. The monolingual system keeps 99% and the crosslingual systems between 97% and 98%. Tòth. et al. [12] reported significant decrease in performance if 95% of the variance is kept. Therefore, we trained a Tandem system without reducing the dimensionality (keeping the full decorrelated vector). Indeed, that system performs significantly better than all other Tandem systems.

6.4. Context-dependent MLP outputs

In Table 3 we showed that the multilingual systems outperformed the mono- and cross-lingual systems. However, it is not clear whether the improvement comes from the larger number of parameters or the multilingual information. Therefore, we increased the number of parameters of the mono- and cross-lingual systems by training MLPs on context-dependent outputs. For the sake of comparison, we used 189 triphone outputs for each MLP as described in Section 5.3. We expect the context-dependent MLP systems to perform better than the mono- and cross-lingual systems that were trained on monophone targets.

The results are reported in Table 4. For KL-HMM we observe an improvement of about 2% absolute for each system compared to Table 2. We observe that the multilingual

System	Model	Feat dim	PCA %-Var	Phoneme accuracy
AF-dep	KL-HMM	189	-	**59.9** %
CGN-dep		189	-	**60.2** %
SZ-dep		189	-	58.2 %
EN-dep		189	-	54.8 %
AF-dep	Tandem	30	98%	62.0 %
CGN-dep		30	93%	63.0 %
SZ-dep		30	96%	61.8 %
EN-dep		30	95%	61.7 %
AF-dep		189	100%	64.1 %
CGN-dep		189	100%	**66.5** %
SZ-dep		189	100%	65.8 %
EN-dep		189	100%	65.2 %

Table 4. Afrikaans phoneme accuracy obtained from monolingual and cross-lingual systems. The system extension *-dep* stands for the context-dependent MLP targets (triphones). For the Tandem systems, also the percentage of the variance that is kept by the PCA is given.

system *AF-CGN-SZ-EN* (Table 3) that has a similar number of parameters (feature vectors have the same dimensionality) performs significantly better than all mono- and cross-lingual systems presented in Table 4. Hence, we conclude that in the case of KL-HMM the multilingual information is exploited.

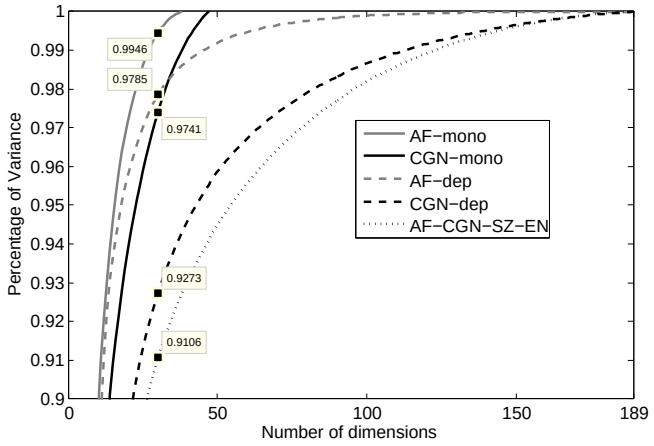

Fig. 3. PCA: Relation between the percentage of the variance that is kept and the dimensionality. The amount of training data and the number of different languages have influence.

For Tandem, we first reduced the dimensionality to 30 as we have done for the mono- and cross-lingual systems. Interestingly, we observe statistically similar performance for all languages except Dutch.

Table 4 also shows how much of the variance is kept with 30 dimensions for each language. Furthermore, Figure 3 plots the amount of variance that is kept for the Afrikaans and Dutch systems as a function of the number of retained di-

System (cont. dep.)	Model	Feat dim	PCA %-Var	Phon. acc.
AF-dep	KL-HMM	189	-	59.9 %
AF-CGN		378	-	64.9 %
AF-CGN-SZ		567	-	**65.4** %
AF-CGN-SZ-EN		756	-	**65.3** %
AF-dep	Tandem	189	100%	64.1 %
AF-CGN		189	99.6%	**67.5** %
AF-CGN-SZ		189	98.9%	**67.5** %
AF-CGN-SZ-EN		189	98.6%	**67.8** %
AF-CGN-SZ-EN		756	100%	63.7 %

Table 5. Afrikaans phoneme accuracy obtained from context dependent multilingual systems. We concatenated the outputs of different MLPs, trained on context-dependent targets.

mensions. We observe that the MLPs trained on monophones have the steepest curves. The more available training data, the higher variance and the more languages contained in the training data, the higher the variance, thus, the more dimensions we need to keep.

Using all the dimensions of the decorrelated feature vector yields significant improvement for all systems. The best performing system is CGN-dep, with the MLP that was trained on 81 hours of Dutch data. Hence, for Tandem, the cross-lingual systems that use MLPs trained on context-dependent targets, perform similar to the multilingual systems.

In this study, we fixed the number of context dependent output units to 189 (by using a simple occupancy threshold). The performance might further be increased if we use MLPs with more output units, which should be determined with an unsupervised data-driven technique.

6.5. Context-dependent multilingual systems

Since we observed improvement for the multilingual systems as well as for the context dependent systems, we also concatenate multiple outputs of the context-dependent MLPs. We expect the improvements to be cumulative.

As reported in Table 5, the KL-HMM based systems show the same behavior as in Table 3, but all the systems perform better. Hence the hypothesis that we can exploit cumulative performance gains is confirmed for KL-HMM systems.

Also for Tandem systems, the performance gains are cumulative. In contrast to the results in Table 3, using the full decorrelated vector for system AF-CGN-SZ-EN yields a lower performance for the context dependent MLP output concatenation. However, the 189 dimensions already contain 98.6% of the variance. Furthermore, we believe that the decrease in performance might be due to the high dimensionality (756) that the Tandem system is not able to handle anymore, i.e., 3 hours of Afrikaans data are not sufficient to train the high number of HMM parameters.

KL-HMM	Phoneme accuracy	rel. change
AF-mono	58.7%	-
+multilingual	64.4%	+9.7 %
+context	60.2%	+2.6 %
+multilingual+context	65.4%	+11.4 %
Tandem	**Phoneme accuracy**	**rel. change**
AF-mono	61.2%	-
+multilingual	66.2%	+8.2 %
+context	66.5%	+8.7 %
+multilingual+context	67.8%	+10.8 %

Table 6. Summary of the experimental results. KL-HMM gains more from multilinguality and Tandem from context-dependency. In both cases, the gains are additive.

Table 6 summarizes the gains that result from multilinguality, context dependency and both. We observe that multilingual information yields more improvement for the KL-HMM system. For the context dependency it is vice versa, there is more improvement for the Tandem systems. The relative performance gains of about 11% are higher then expected and we have shown that the ASR performance can be boosted with out of language data even if there are already three hours of data available in the target language.

7. CONCLUSION AND FUTURE WORK

In this study, we successfully exploited out of language data and boosted a monolingual speech recognizer that was trained on three hours of Afrikaans data.

First, for two investigated acoustic modeling techniques, the best multilingual system yields more than 10% relative improvement compared to the corresponding monolingual systems only trained on Afrikaans. To the best of our knowledge, such improvements have not been reported on Afrikaans before. Second, we also found that Tandem systems consistently outperform the KL-HMM systems. However, the performance of the KL-HMM is extremely satisfactory. Third, careful analysis of the experimental results revealed that the KL-HMM system mostly gains improvement by exploiting multilingual information whereas the Tandem system seems to benefit from both, contextual and multilingual information.

In the future, we will further investigate several aspects such as the training of MLPs with higher number of output units or weighted combinations of MLP outputs rather than normalized concatenations.

8. ACKNOWLEDGEMENT

The authors are grateful to the HLT group at Meraka, and especially Dr. Febe de Wet, for providing us with the training and test sets as well as the Afrikaans dictionary.

9. REFERENCES

[1] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *Proc. of ICASSP*, 2012, (to appear).

[2] Y. Qian, J. Xu, D. Povey, and J. Liu, "Strategies for using MLP based features with limited target-language training data," in *Proc. of ASRU*, 2011, pp. 354–358.

[3] T. Niesler, "Language-dependent state clustering for multilingual acoustic modelling," *Speech Communication*, vol. 49, pp. 453–463, 2007.

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

[5] Leonard Bloomfield, *Language*, New York: Holt, 1933.

[6] D. Imseng, H. Bourlard, J. Dines, P. N. Garner, and M. Magimai.-Doss, "Improving non-native ASR through stochastic multilingual phoneme space transformations," in *Proc. of Interspeech*, 2011, pp. 537–540.

[7] Vàclav Blažek, "On the internal classification of Indo-European languages: survey," 2005, `http://www.phil.muni.cz/linguistica/art/blazek/bla-003.pdf`.

[8] A. Constantinescu and G. Chollet, "On cross-language experiments and data-driven units for ALISP (automatic language independent speech processing)," in *Proc. of ASRU*, 1997, pp. 606–613.

[9] W. Heeringa and F. de Wet, "The origin of Afrikaans pronunciation: a comparison to west Germanic languages and Dutch dialects," in *Proc. of the Conf. of the Pattern Recognition Association of South Africa*, 2008, pp. 159–164.

[10] H. Hermansky, D.P.W. Ellis, and S Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of ICASSP*, 2000, pp. III–1635–1638.

[11] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proc. of Interspeech*, 2008, pp. 928–931.

[12] L. Tòth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian.," in *Proc. of Interspeech*, 2008, pp. 2695–2698.

[13] F. Grézl, M. Karafiàt, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. of ASRU*, 2011, pp. 359–364.

[14] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

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

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

[17] 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.

[18] D. Imseng, H. Bourlard, and M. Magimai.-Doss, "Towards mixed language speech recognition systems," in *Proc. of Interspeech*, 2010, pp. 278–281.

[19] Koichi Shinoda and Takao Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. of Eurospeech*, 1997, vol. I, pp. 99–102.

[20] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, 1994, pp. 307–312.

[21] David Imseng and John Dines, "Decision tree clustering for KL-HMM," Idiap-Communication Idiap-Com-01-2012, Idiap research institute, 2012.

[22] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. of Interspeech*, 2011, pp. 437–440.

[23] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. of ICASSP*, 2004, vol. 1, pp. I–409–412.

[24] C. van Heerden, E. Barnard, and M. Davel, "Basic speech recognition for spoken dialogues," in *Proc. of Interspeech*, 2009, pp. 3003–3006.