IDIAP RESEARCH REPORT

CROSSLINGUAL TANDEM-SGMM: EXPLOITING OUT-OF-LANGUAGE DATA FOR ACOUSTIC MODEL AND FEATURE LEVEL ADAPTATION

Petr Motlicek David Imseng Philip N. Garner

Idiap-RR-39-2013

NOVEMBER 2013

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Crosslingual Tandem-SGMM: Exploiting Out-Of-Language Data for Acoustic Model and Feature Level Adaptation

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

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland
{Petr.Motlicek,David.Imseng,Phil.Garner}@idiap.ch

Abstract

Recent studies have shown that speech recognizers may benefit from data in languages other than the target language through efficient acoustic model- or feature-level adaptation. Crosslingual Tandem-Subspace Gaussian Mixture Models (SGMM) are successfully able to combine acoustic model- and feature-level adaptation techniques. More specifically, we focus on under-resourced languages (Afrikaans in our case) and perform feature-level adaptation through the estimation of phone class posterior features with a Multilayer Perceptron that was trained on data from a similar language with large amounts of available speech data (Dutch in our case). The same Dutch data can also be exploited on an acoustic model-level by training globally-shared SGMM parameters in a crosslingual way. The two adaptation techniques are indeed complementary and result in a crosslingual Tandem-SGMM system that yields relative improvement of about 22% compared to a standard speech recognizer on an Afrikaans phoneme recognition task. Interestingly, eventual score-level combination of the individual SGMM systems yields additional 3% relative improvement.

Index Terms: Automatic speech recognition, Acoustic model adaptation, Under-resourced languages

1. Introduction

Developing a state-of-the-art speech recognizer from scratch for a given language is expensive. The main reason for this is the need of large amounts of training data to be used for the development. Data collection involves a lot of manual work, not only in time for the speakers to be recorded, but also for annotation of the subsequent recordings. Therefore, the need for training data is one of the main barriers in porting current systems to many languages. On the other hand, large databases already exist for many (especially economically viable) languages.

Previous studies have shown that automatic speech recognition (ASR) can benefit from data in languages other than the target language [1, 2, 3, 4]. Two different approaches to exploit out-of-language data have been investigated:

- acoustic model-level: for instance Niesler [1] studied the sharing of resources inspired by multilingual acoustic modeling techniques proposed by Schultz [5]. However, only marginal ASR performance gains were reported.
- feature-level: previous studies [6, 2, 3] found that the relation between phonemes of different languages can be learned and exploited for cross-lingual acoustic model training or adaptation. Posterior-based features, estimated by Multilayer Perceptrons (MLPs), are particularly well suited for such tasks. We successfully used posterior-based features to boost the performance of an Afrikaans speech recognizer [4].

In this paper, we also focus on Afrikaans (i.e., within-language) and exploit 80 hours of Dutch (i.e, out-of-language) data for crosslingual adaptation on the feature-level. Therefore, an MLP is initially trained on the Dutch data and subsequently applied to estimate posterior-based features. These features are first exploited in a conventional HMM/GMM system (Tandem system). Results reveal that they significantly outperform conventional acoustic features (MFCCs). Recent study [7] has shown that subspace Gaussian mixture models (SGMMs) usually outperform standard HMM/GMMs. In line of these findings, we build an SGMM system exploiting posterior-based features, called a Tandem-SGMM system, and show that it yields improvement compared to the standard Tandem system. Furthermore, SGMMs allow exploitation of crosslingual information through the training of globally-shared model parameters on out-of-language data. In that context, we build a crosslingual Tandem-SGMM system that exploits out-of-language data on the feature-level as well as on the acoustic model-level which yields the best performance on Afrikaans. Eventually, we perform score-level combination of the individual SGMM systems to demonstrate their complementary recognition outputs.

In the remainder of this paper, we briefly review the concept of posterior-based features as well as SGMMs (Section 2). In Section 3, we introduce the experimental datasets before we present experiments with results in Section 4. Section 5 presents score-level combination performance and Section 6 concludes the work.

2. Related work

In this section, we first review the concept of posterior-based features and then briefly summarize the SGMM acoustic modeling technique.

2.1. Posterior-based features

The posterior-based features are phone class posterior probabilities given the acoustics and estimated with an MLP that can be trained on any auxiliary dataset (out-of-language data). The language of the training data determines the number of output units K (number of phone classes) of the MLP. The phone classes can for example be context-independent monophones or context-dependent triphones.

Once the MLP is trained, we consider a sequence of T acoustic feature vectors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ extracted from within-language data. The phone class posterior sequence $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$ is then estimated with the previously trained auxiliary MLP.

The conventional Tandem approach [8] models the emission probabilities of HMM states with mixtures of Gaussians. To model the emission probabilities with Gaussians, the posterior-based features z_t need to be post-processed. More specifically, the log-phone class posteriors are decorrelated, usually with a principal component analysis (PCA). The transformation matrix can be estimated on within-language data. Usually, the resulting feature vector $\boldsymbol{r}_t = (r_t^1, \ldots, r_t^L)^\mathsf{T}$ has a reduced dimensionality L.

Some posterior-based feature studies exploited in multilingual framework reported rather small improvements (up to 3.5% relative) [2, 3]. We successfully used posterior-based features to boost the performance of an Afrikaans speech recognizer (10% relative) [4]. In the earlier study [4], we trained a Dutch MLP with 189 context-dependent outputs. In this paper, we train the Dutch MLP on about ten times more context-dependent targets and expect even higher relative improvement.

2.2. SGMM

The subspace Gaussian mixture model (SGMM) [9] is a way of compactly representing a large collection of mixture-of-Gaussian models. In the case of a conventional Gaussian mixture model (GMM), the likelihood is given as

$$p(\mathbf{x} \mid j) = \sum_{i=1}^{M_j} w_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji}), \quad (1)$$

where j is the state and the parameters of the model are w_{ji}, $\boldsymbol{\mu}_{ji}$ and $\boldsymbol{\Sigma}_{ji}$. The SGMM in the basic case is given as

$$p(\mathbf{x} \mid j) = \sum_{i=1}^{I} w_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_i) \quad (2)$$

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i \mathbf{v}_j \quad (3)$$

$$w_{ji} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_j}{\sum_{l=1}^{I} \exp \mathbf{w}_l^T \mathbf{v}_j}, \quad (4)$$

where $\mathbf{v}_j$ are state-specific vectors (with dimension similar to the dimension of speech features) and I is the number of Gaussians in the shared GMM structure. The globally-shared model parameters $\mathbf{w}_i$, $\mathbf{M}_i$, and $\boldsymbol{\Sigma}_i$ (full covariance) embody most of the free parameters in the system and can be trained using out-of-domain data. This idea has been explored in the application of SGMMs to multilingual speech recognition [10], where the global parameters were estimated by tying across multiple languages. Since, for smaller systems, most of the parameters are globally-shared, this can lead to more robust parameter estimates. Extension of SGMMs with sub-states (replacing $\mathbf{v}_j$ with mixtures $\mathbf{v}_{jm}$ and introducing sub-state weights c_{jm}) towards large-scale models is described in [11].

Although SGMMs have shown their capabilities to outperform conventional GMMs in monolingual as well as multilingual scenarios, acoustic models were always applied on relatively simple acoustic features (e.g., MFCCs) [7]. In crosslingual or multilingual scenarios, the SGMM adapted models were never compared or combined with feature-level adaptation techniques. The goal of this paper is, to first apply both concepts, posterior-based features and SGMMs independently on Afrikaans - an under-resourced language. Then we will show that both concepts (feature- and model-level out-of-language adaptations) are complementary and can be successfully combined, leading to an improved ASR on within-language data (Afrikaans).

ID	Language	Number of phonemes	Amount of training data
AF	Afrikaans	38	3 h
CGN	Dutch	47	81 h

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

3. Datasets

To evaluate posterior-based features, SGMMs and both concepts in combination, we used data from Afrikaans (referred to as the within-language data). In [12], it was reported that standard Dutch seems to be the best language from which to borrow acoustic data for the development of an Afrikaans ASR system. Our recent studies confirmed that hypothesis [4]. We therefore used Dutch as out-of-language data. The data for both languages are summarized in Table 1.

The Afrikaans data is available from the LWAZI corpus [13]. 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 [14] that defines the phoneme set containing 38 phonemes (including silence). The dictionary that we used comprises 1585 different words. The HLT group at Meraka provided us with the training and test sets. In total, about three hours of training data and 50 minutes of test data is available (after voice activity detection). Since we did not have access to an appropriate language model (LM) (i.e., we did not find any reasonable source (URL) of text data in Afrikaans on the Web which could be collected by a Web crawler and subsequently used to build our LM), we trained a bi-gram phoneme model on the training set and only report phoneme error-rates (PERs) in this study. The bi-gram phoneme model learned the phonotactic constraints of the Afrikaans language. Phoneme perplexity measured on the test set is about 15.

We used data of the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) [15] 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.

4. Experiments and results

Throughout all experiments, the bi-gram phoneme model described in Section 3 was applied, hence PERs are reported. In addition to three hours of Afrikaans training data, we also performed experiments with much smaller training subsets (i.e., 5 minutes and 1 hour) to further simulate under-resourced scenarios and to investigate the performance of HMM/GMMs and SGMMs. The model sizes for HMM/GMMs and SGMMs trained on Afrikaans are always reported for the 3 hours scenario.

In the remainder of this section, we first describe the conventional HMM/GMM and SGMM systems employing standard acoustic features. Then, we present how to exploit crosslingual information on the feature-level by using different kinds of posterior-based features. At the end of this sec-

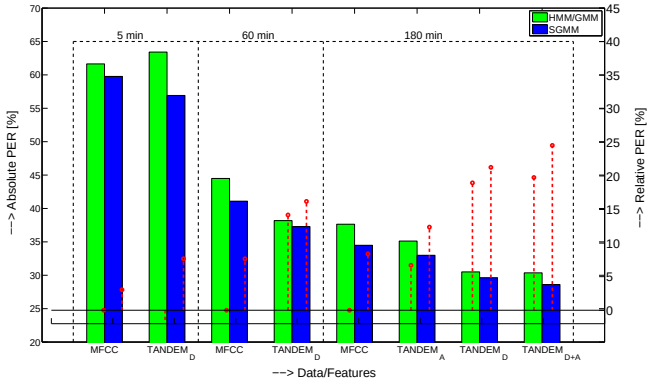

Figure 1: PER results: HMM/GMM and SGMM systems trained using MFCC and posterior-based (Tandem) features evaluated on Afrikaans. Relative improvements of each system (red dashed lines measured on the right vertical axis) are given with respect to HMM/GMMs trained on MFCCs (for given training scenario).

tion, we exploit crosslingual information on both the acoustic model- and the feature-level using the crosslingual Tandem-SGMM system.

4.1. Standard features

The standard acoustic features in our study were conventional Mel-frequency cepstral coefficient (MFCC) based speech features extended with Δ+$\Delta\Delta$ (total dimension D=39). First, we trained an HMM/GMM as well as an SGMM system with these features, extracted on Afrikaans data, and hypothesized that the SGMM system outperforms the HMM/GMM system.

Our baseline system was a conventional crossword context-dependent HMM/GMM system that used 12K Gaussians in approximately 1800 HMM-states, determined with state tying. The SGMM system[1] was also trained from the MFCC features and used tree clustering to obtain a total number of sub-states (i.e., vectors $\mathbf{v}_{jm}$) similar to the number of Gaussians (12K) of the HMM/GMM system. The gender-independent universal background model (UBM) had 500 Gaussians ($I = 500$) and the phonetic subspace dimension (S) was 40.

Figure 1 graphically presents the results. It can be clearly seen that SGMMs trained on MFCC features outperform HMM/GMMs for all training scenarios (5 min, 1 hour and 3 hours). Hence our hypothesis is confirmed.

4.2. Posterior-based features

In this section, we will present systems that used different kinds of posterior-based features: Tandem$_A$ - features estimated by an MLP trained on Afrikaans data; Tandem$_D$ - features estimated by an MLP trained on Dutch data; and Tandem$_{D+A}$ - the simple concatenation of Tandem$_D$ and Tandem$_A$.

4.2.1. Tandem$_A$

These features were estimated with an MLP uniquely trained on Afrikaans data. We trained the MLP using 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 [16] of the HTK toolkit. The number of parameters in the MLP was set to 10% of the number of available training frames, to avoid overfitting. Quicknet [17] software was used to train the MLP.

To obtain triphone targets, we used the standard HMM/GMM system presented in Section 4.1 with a modified decision tree. The HMM/GMM system described in Section 4.1 uses 1800 HMM-states. State tying was based on the minimum description length (MDL) criterion that automatically determined the number of tied triphones [18]. As described by [18], the MDL criterion has a hyper-parameter, c, which controls the weight of the term that penalizes models with large amounts of triphones. In a recent study [4], we achieved good results with 189 context-dependent targets. Therefore, we used $c = 16$ to obtain 187 triphone targets. During PCA, we reduced the dimensionality of the posterior-based features to keep 99% of the variance which gave us 48-dimensional posterior features.

The purpose of the Tandem$_A$ features was to investigate whether posterior-based features yield better performance than conventional MFCC features for the HMM/GMM as well as for the SGMM systems. We did not train an MLP for the 5 min and 1 hour data scenarios. Therefore, Figure 1 only displays Tandem$_A$ results for the 3 hours scenario. The HMM/GMM system had 12K Gaussians and 1800 context-dependent states and the SGMM system had 8K sub-states and also 1800 context-dependent states (S=50, D=48). Indeed, the Tandem$_A$ features yield better performance than the standard MFCC features.

4.2.2. Tandem$_D$

We hypothesized that the posterior-based features can successfully exploit out-of-language data. Therefore, we trained an MLP on 80 hours of Dutch data similar to the MLP previously trained on Afrikaans data. To obtain the triphone targets, we trained an HMM/GMM system on all the Dutch data and used $c = 16$ during state tying (see Section 4.2.1) resulting in about 1800 triphone targets. After PCA, we kept 286 dimensions.

For the 3 hours scenario, the HMM/GMM system with 12K Gaussians and 1800 context-dependent states was then trained using the Tandem$_D$ features (Afrikaans data passed through the MLP trained on Dutch). In case of SGMMs, it is usually assumed that the subspace dimension S is about the size of the feature dimension D. However, to preserve roughly the same amount of model parameters while having reasonable UBM size, we used the setting $S = 200$, $I = 100$.

Figure 1 confirms our hypothesis. For the 3 hours scenario, the Tandem$_D$ systems clearly outperform the MFCC as well as the Tandem$_A$ systems.

Furthermore, we also evaluated the Tandem$_D$ features for the 5 min and the 1 hour scenario. Therefore, we adapted the number of parameters in the HMM/GMM as well as the SGMM system accordingly. Figure 1 shows that SGMMs are able to exploit the out-of-language information through feature-level adaptation for both scenarios, whereas the HMM/GMM system with the Tandem$_D$ features performs worse than the MFCC based system for the 5 min scenario. It seems that only the SGMM system, which is able to share global parameters, is capable of handling large-dimensional features with very small amounts of data.

[1] Note: although we aim to minimize PER, SGMMs are built to have approximately the same number of parameters as the HMM/GMM system trained on the same data.

features	HMM/GMM	SGMM	crosslingual SGMM
MFCC	37.6	34.5	32.9
TANDEM$_A$	35.1	33.0	n/a
TANDEM$_D$	30.5	29.6	**29.1**
TANDEM$_{D+A}$	30.4	**28.6**	n/a

Table 2: Afrikaans PER [%] results: HMM/GMM and SGMM system results are the same as in Figure 1. The crosslingual SGMM and Tandem-SGMM systems can be trained and evaluated only for MFCC and Tandem$_D$ features.

4.2.3. Tandem$_{D+A}$

Eventually, we hypothesized that the two MLPs, trained on different languages, generate complementary features. Therefore, we also evaluated Tandem$_{D+A}$, the feature concatenation of Tandem$_D$ and Tandem$_A$. Although HMM/GMM system yields only marginal improvement (about 0.1% PER absolute), SGMMs improved by about 1% PER absolute with respect to Tandem$_D$ features, as shown in Figure 1 and Table 2.

4.3. Crosslingual SGMM

In the previous section, we have already seen that SGMMs trained using posterior-based features estimated using the Dutch (out-of-language) MLP outperforms conventional Tandem features estimated with an MLP trained on within-language data. In other words, out-of-language data was exploited on the feature-level.

In this section, we go one step further and hypothesize that the Tandem-SGMM system can exploit out-of-language data on both levels (i.e., also model-level) simultaneously. As proposed in [10], the globally-shared model parameters $\mathbf{w}_i$, $\mathbf{M}_i$, and $\mathbf{\Sigma}_i$ together with the UBM can be trained in cross-lingual ways (i.e., parameters can be estimated on well-resourced data). Similar to the Tandem approach, Dutch training data can be employed in SGMMs to perform model-level adaptation. The state-specific SGMM parameters are then re-estimated in a maximum likelihood fashion on within-language data (3 hours of Afrikaans). We refer to this system as the crosslingual SGMM system. Note that we evaluated the crosslingual SGMM system with standard MFCC features as well as Tandem features. The latter is referred to as the crosslingual Tandem-SGMM.

Table 2 shows the decoding results. Two trends can be observed: (1) from top-to-bottom: the results improve when the features change. The Afrikaans posteriors outperform standard MFCC features and the Dutch posteriors (additionally concatenated with the Afrikaans posteriors) are significantly better than the Afrikaans posteriors; (2) from left-to-right: the acoustic modeling technique changes and the results also improve. SGMMs outperform standard HMM/GMMs and the crosslingual SGMM system that exploits Dutch and Afrikaans data during training performs best.

A crosslingual SGMM system that use the Afrikaans posteriors (either through Tandem$_A$ or Tandem$_{D+A}$) was excluded since the globally-shared SGMM model parameters cannot be trained using the Afrikaans posteriors.

5. Score-level combination

Although Tandem-SGMM largely improves recognition performance for all types of posterior features, crosslingual SGMM yields significant improvement only when built using conventional MFCCs, as shown in Table 2. Nevertheless, to highlight complementary properties of the crosslingual Tandem-SGMM (with respect to previous Tandem-SGMMs), we perform score-level combination of the SGMM systems.

system	PER[%]
without crosslingual Tandem-SGMM	28.3
with crosslingual Tandem-SGMM	**28.0**

Table 3: Afrikaans PER [%] results: Score-level combination of the SGMM systems trained on MFCC, TANDEM$_A$, TANDEM$_D$, and TANDEM$_{D+A}$ features, while also exploiting crosslingual Tandem-SGMM system.

More specifically, we employ ROVER - a standard technique allowing to combine "symbol" sequences taken as outputs of different recognition systems [19]. ROVER can be seen as a simple approach measuring complementarity of recognition systems based on counting simultaneous and dependent errors. It assumes that significant recognition gain can be achieved if the combined systems exhibit different (heterogeneous) recognition errors.

Results on score-level combination for the SGMM systems are given in Table 3. First, recognition outputs of four individual SGMM systems (trained using MFCC, TANDEM$_A$, TANDEM$_D$, and TANDEM$_{D+A}$ features) are combined. Then, crosslingual Tandem-SGMM is also exploited in ROVER combination. Achieved improvement (about 0.3% PER absolute) suggests that crosslingual SGMM yields complementary performance to previous SGMM systems.

6. Conclusions

In this paper, we successfully trained an SGMM-based speech recognizer on posterior features (Tandem-SGMM) and subsequently exploited out-of-language information on a feature-level by training the MLP used for the feature generation on out-of-language data. Furthermore, we used the same out-of-language data for acoustic model adaptation by training globally-shared SGMM parameters in a crosslingual way (crosslingual SGMM) and showed that the two adaptation techniques are complementary.

Using Afrikaans data as an example of under-resourced languages, and Dutch as the similar well-resourced language, we showed that crosslingual posterior features are superior to standard acoustic features and that the crosslingual SGMM outperforms the standard SGMM. The crosslingual Tandem-SGMM system that combines feature-level and acoustic model-level adaptation largely improved phoneme recognition error-rates by about 22% relative (compared to an un-adapted MFCC based HMM/GMM system trained purely on within-language data). Subsequent score-level combination brings additional 3% relative PER improvement.

7. Acknowledgements

This 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 multi-lingual conversational applications". D. Imseng was supported by the Swiss NSF through the project Interactive Cognitive Systems (ICS) under contract number 200021_132619/1.

8. References

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

[2] 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.

[3] 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.

[4] 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.

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

[6] 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, pp. 4869–4872.

[7] N. T. Vu, T. Schultz, and D. Povey, "Modeling gender dependency in the subspace GMM framework," in *Proc. of ICASSP*, 2012, pp. 4345–4348.

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

[9] D. Povey *et al.*, "Subspace gaussian mixture models for speech recognition," in *Proc. of ICASSP*, 2010, pp. 4330–4333.

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

[11] D. Povey, M. Karafiat, A. Ghoshal, and P. Schwarz, "A symmetrization of the subspace gaussian mixture model," in *Proc. of ICASSP*, 2011, pp. 4504–4507.

[12] 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, www.let.rug.nl/heeringa/dialectology/papers/prasa08.pdf.

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

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

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

[16] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, *The HMM-based speech synthesis system version 2.0*, 2007, http://hts.sp.nitech.ac.jp/.

[17] D. Johnson, *Quicknet*, 2005, http://www.icsi.berkeley.edu/Speech/qn.html.

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

[19] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. of ASRU*, 1997, pp. 347–354.