POOLING ASR DATA FOR CLOSELY RELATED LANGUAGES

Charl van Heerden, Neil Kleynhans, Etienne Barnard and Marelie Davel

cvheerden@csir.co.za, ntkleynhans@csir.co.za, ebarnard@csir.co.za, mdavel@csir.co.za

ABSTRACT

We describe several experiments that were conducted to assess the viability of data pooling as a means to improve speech-recognition performance for under-resourced languages. Two groups of closely related languages from the Southern Bantu language family were studied, and our tests involved phoneme recognition on telephone speech using standard tied-triphone Hidden Markov Models. Approximately 6 to 11 hours of speech from around 170 speakers was available for training in each language. We find that useful improvements in recognition accuracy can be achieved when pooling data from languages that are highly similar, with two hours of data from a closely related language being approximately equivalent to one hour of data from the target language in the best case. However, the benefit decreases rapidly as languages become slightly more distant, and is also expected to decrease when larger corpora are available. Our results suggest that similarities in triphone frequencies are the most accurate predictor of the performance of language pooling in the conditions studied here.

Index Terms— speech recognition, under-resourced languages, data pooling

1. INTRODUCTION

When developing automatic speech recognition (ASR) systems for under-resourced languages, the amount of training data available is an important limiting factor. Although a wide variety of approaches to this issue have been studied (see Section 2 below), it is safe to say that data scarcity remains the most significant obstacle to the development of high-quality ASR systems.

Many of the existing approaches to dealing with data scarcity utilize similarities between some or all of the phonemes in different languages in order to improve the accuracy of ASR. Typically, it is assumed that sufficient data from a well-resourced language is available, and that data is employed in various configurations to improve the performance of ASR in an under-resourced language. In the current contribution, we investigate a somewhat different approach, namely data sharing among groups of related languages that are all lacking in resources. In particular, we wish to investigate how similar languages need to be for straightforward pooling of ASR training data to be beneficial.

If this approach proves to be useful, it can be used widely, since the vast majority of the actively spoken languages occur in clusters of more or less closely related families. However, there is good reason to suspect that only very closely related languages will benefit from pooling in this way – and even then, only if the amount of training data in the target language is severely limited. The main evidence for these concerns comes from experience with dialects of well-resourced languages: it is well known, for example, that ASR systems trained on American English perform poorly when presented with British English, and that combining training data from these two dialects generally leads to a deterioration in performance.

We therefore experiment with two groups of languages that are strongly related, as discussed in Section 3. In order to assess the effect of data pooling without any confounding influences, we only pool phones with identical IPA symbols in different languages – details are provided in Section 4, which also describes the recognizers employed . In Section 5 we analyse the pooled data according to a number of distance measures and report results for a phoneme-recognition task, showing that the relatedness of languages is indeed crucial to the success of this approach. Our conclusions are summarized in Section 6, which also suggests further work.

2. RELATED WORK

Once ASR systems were being developed for resource-scarce languages, research related to the possibility of supplementing target language data with that of additional languages soon followed. The rationale is simple: since the statistical methods being employed during acoustic modelling require more data than is available for the target language in question, borrow additional matching data from "donor languages" where possible.

In this section we review the main approaches to data combination that have been investigated in the literature. We describe strategies for data combination, different approaches to model mapping, and prior studies dealing specifically with the languages relevant to this paper.

2.1. Data combination strategies

Several different data combination strategies have been investigated, with results strongly influenced by the amount of target language data available, the acoustic diversity of the available databases, as well as the acoustic and phonotactic similarity between the target and donor language(s). Approaches include:

- *Cross-language transfer*: using an existing recognizer without any adaptation to the target language. Predictably, this strategy typically provides poor results, and is only considered if the languages in question are closely related [1], or if no target language data is available. In the latter case, multilingual acoustic models (built from a number of different languages and simultaneously able to recognize any of these) have been shown to yield better results than monolingual donor models [2, 3].

- *Cross-language adaptation*: adapting an existing mono- or multilingual recognizer using limited training data and techniques such as Maximum Likelihood Linear Regression (MLLR) or Maximum A Posteriori (MAP) adaptation [4, 5, 6]. These techniques can produce better results than cross-language transfer, and if target language data is very limited, can also outperform bootstrapping (see below).

- *Data pooling*: combining data from different sources by pooling the data directly. Such multilingual models were first developed in the context of language identification [7] but are also used in speech recognition, especially as initial models from which to adapt or bootstrap [3] or, to a lesser extent, when bilingual speakers are being recognized [8].

- *Bootstrapping*: initially demonstrated in [9], acoustic models are initialized using models from a donor language (or languages) and then rebuilt using target language data only. In [10], bootstrapping from multilingual models was shown to outperform adaptation when both approaches were evaluated using approximately 15 minutes of (Portuguese) target speech. While useful gains were obtained using bootstrapping, accuracy only approached that of a monolingual target language system (developed using 16.5h of target language data) once improved alignments of 90 minutes of target speech were used. (These improved alignments were assumed to be available, but typically are not.)

The methods described above can also be combined. For example, data pooling can be used to create multilingual models as seed models for bootstrapping [10], or a donor language can be adapted to a target language prior to data pooling [11].

Whichever method is used, cross-language data sharing has only been shown to compensate for limited target language data, and improvements soon dwindle as more target language data becomes available.

2.2. Approaches to model mapping

Before applying any of the data combination approaches described above, some mapping is usually required between the acoustic models of the donor languages and those of the target language. These mappings can be based on linguistic knowledge, data analysis or a combination of the two approaches.

Linguistic knowledge is typically encoded in the phoneme inventory of each of the languages, and the phoneme sets mapped directly based on IPA or SAMPA equivalences [8, 3], or other prior phonetic knowledge. Data-driven mappings are based on some distance (or similarity) measure, various of which have been utilized [12, 13, 14]. Good results are obtained using "hierarchical clustering", employing linguistically motivated categories within which data-driven (within-category) clustering is performed [13, 15]. Note that while most of these experiments were applied to context-independent models, similar techniques are applicable to context-dependent models, as well as to sub-phonemic models.

Hierarchical clustering at the sub-phoneme level can be integrated with the standard decision tree building process typically used to cluster and combine context-dependent triphones during Hidden Markov Model-based (HMM-based) model building: data samples are tagged with their source language and this additional information made available during data-driven clustering, resulting in improved results [10].

2.3. Data sharing of Southern Bantu languages

None of the above studies dealt specifically with data from any of the Southern Bantu languages. In the only cross-lingual adaptation study that includes these languages (that we are aware of), monolingual systems in isiXhosa and isiZulu were compared with a multilingual system developed using IPA-based data pooling of the two languages, with language-specific questions added during tying of triphones [16]. The multilingual system outperformed the monolingual systems, but gains were small. (Optimal phoneme accuracies for both approaches ranged between for 60.5% and 61.3%.)

3. CORPUS AND LANGUAGES

Our experiments are based on the Lwazi ASR corpus which was developed as part of a project that aims to demonstrate the use of speech technology in information service delivery in South Africa [17]. The corpus contains data from each of the eleven official languages of South Africa – approximately 200 speakers per language (2,200 speakers in total), contributed

read and elicited speech, recorded over a telephone channel. Each speaker produced approximately 30 utterances; 16 of these were randomly selected from a phonetically balanced corpus and the remainder consist of short words and phrases.

For the purposes of the current research, we concentrate on two subsets of this corpus each containing a group of closely related languages. The three Sotho-Tswana languages (Sepedi, Setswana and Sesotho) form the first group, and three of the four Nguni languages (isiZulu, isiXhosa and isiNdebele) the second. (The fourth Nguni language in the Lwazi corpus, Siswati, was not included in the current study for reasons explained below.) These languages all belong to the Southern Bantu family of languages. Although they are used as first language by relatively large populations of speakers (all are considered as first language by several million speakers, with the exception of isiNdebele, which has only 700 000 first-language speakers), very few linguistic resources are available for these languages.

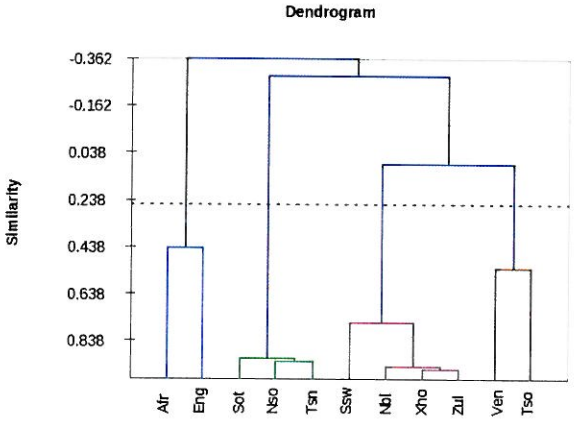

Fig. 1. Dendrogram calculated from confusion matrices for a multi-lingual text-based SVM classifier.

These languages all belong to the Southern Bantu family of languages [18]. We have previously studied their relationships using both orthographic and acoustic measures [19]. A typical dendrogram of the measured distances between the languages is shown in Fig. 1 (based on orthographic or text-based data); it can be seen that the two groups of languages selected here are indeed very closely related by these measures, and are therefore worthy candidates for the type of pooling considered here. Note also that Siswati is not as closely related to the other Nguni languages by these measures – it was therefore not included as a target language in the current study.

Language	# total min	# training min	# testing min
isiNdebele (Nbl)	609	517	92
isiZulu (Zul)	529	447	82
isiXhosa (Xho)	536	454	82
Sepedi (Nso)	548	465	83
Sesotho (Sot)	425	359	66
Setswana (Tsn)	443	379	64
Siswati (Ssw)	663	-	-

Table 1. Size of training and testing sets (in minutes) per language.

4. METHOD

4.1. ASR system overview

The ASR system developed to evaluate the effect of data pooling follows a standard Hidden Markov Model (HMM) design. Acoustic models consist of cross-word tied-state triphones modelled using a 3-state continuous density HMM. Each HMM state distribution is modelled by a 7-mixture multivariate Gaussian with diagonal covariance. The 39-dimensional feature vector consists of 12 static Mel-Frequency Cepstral Coefficients (MFCCs) with the 0'th cepstra, 13 delta and 13 delta-delta coefficients appended [20]. The final preprocessing step applies Cepstral Mean Normalization (CMN) which calculates a per utterance bias and removes it [21]. The different HMM state distributions were estimated by running multiple iterations of the Baum-Welch re-estimation algorithm. Once the triphone acoustic models were trained, a 40-class semi-tied transform [22] was estimated to further improve acoustic model robustness.

Our data pooling experiments are performed using the Lwazi ASR Corpus [17] and the Lwazi pronunciation dictionaries [23], as briefly described in Section 3. Table 1 indicates the amount of speech data in minutes for the different language-specific training and testing sets. Each language testing set was created by choosing 30 speakers at random, which were then excluded from the training data. In each case, we employed both the phonetically balanced sentences and the short phrases in our training and testing data.

4.2. Data combination approach

Our initial step in data pooling is to partition the languages into two groups: The *Nguni* group consists of isiZulu, isiNdebele and isiXhosa, while the *Sotho-Tswana* group includes Sepedi, Sesotho and Setswana. To increase our training data we systematically add speech data from languages in the same group to the target language.

Cross-language mapping is performed at the phoneme level based on the IPA-mapping described in the Lwazi phoneme set version 1.2., a phoneme set that is still undergoing further refinement [23].

Language combinations	# distinct phonemes
Sepedi	43
Sepedi + Setswana	46
Sepedi + Setswana + Sesotho	49
Sepedi + Setswana + Sesotho + isiZulu	65
Sesotho	41
Sesotho + Setswana	42
Sesotho + Setswana + Sepedi	49
Sesotho + Setswana + Sepedi + isiZulu	65
Setswana	34
Setswana + Sesotho	42
Setswana + Sesotho + Sepedi	49
Setswana + Sesotho + Sepedi + isiZulu	65

Table 2. The number of distinct phonemes for each Sotho-Tswana language cluster.

Language combinations	# distinct phonemes
isiNdebele	48
isiNdebele + isiZulu	54
isiNdebele + isiZulu + isiXhosa	61
isiNdebele + isiZulu + isiXhosa + Siswati	64
isiZulu	47
isiZulu + isiNdebele	54
isiZulu + isiNdebele + isiXhosa	61
isiZulu + isiNdebele + isiXhosa + Siswati	64
isiXhosa	53
isiXhosa + isiZulu	57
isiXhosa + isiZulu + isiNdebele	61
isiXhosa + isiZulu + isiNdebele + Siswati	64

Table 3. The number of distinct phonemes for each Nguni language cluster.

Table 2 shows the increase in the number of distinct phones when languages from the Sotho-Tswana group are added together (and also the count if isiZulu is added to these languages). Similarly, Table 3 shows the increasing count of distinct phones for the Nguni group. Column 1 in Tables 2 and 3, indicate the data pooling combinations which were used in the various ASR experiments.

5. COMBINATION RESULTS

In order to assess the performance of our combined ASR systems, phone recognition on the "base" languages is performed for all combined systems. We also measure several distances in order to quantify how far the languages are apart from one another.

5.1. Inter-phone comparisons

We investigate the "closeness" of languages by measuring several distances: the Bhattacharyya distances between overlapping multivariate normal distributions of monophone and triphone models, the Euclidean distance between overlapping phone durations and the cosine of the angle between phone-count vectors.

5.1.1. Comparison of acoustic similarities

The Bhattacharyya distance for multivariate Gaussian distributions,

$$D_B = \frac{1}{8}(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \ln \left(\frac{|\Sigma|}{\sqrt{|\Sigma_1 \Sigma_2|}} \right) \quad (1)$$

is used to calculate distances between corresponding states of corresponding phones in pairs of languages. In (1), μ_i denotes the mean vector of a particular multivariate distribution, Σ_i the corresponding covariance matrix and

$$\Sigma = \frac{\Sigma_1 + \Sigma_2}{2} \quad (2)$$

The Bhattacharyya distance is calculated for all monophones and triphones shared among languages. In order to obtain a single distance for both monophones and triphones, weighted sums are calculated, with each phone being weighted by the sum of its prior probabilities in the intersection of the languages being compared. The weighted sums, referred to as the acoustic distances, are displayed in tables 4 and 5.

5.1.2. Comparison of (tri)phone frequencies

Another way to assess the closeness of languages is to measure the similarities in the frequencies at which common monophones and triphones occur in those languages. We quantify these similarities in terms of the angle between the vectors containing the frequencies of all monophones / triphones in each of the languages:

$$\cos(\angle(\mathbf{x}, \mathbf{y})) = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}| \cdot |\mathbf{y}|}, \quad (3)$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors containing the phone / triphone frequencies in two different languages.

The higher this value is, the more overlap exists between both phones and phone counts in these languages. Tables 4 and 5 summarize these measurements for the two language groups studied here.

5.1.3. Comparison of phone durations

A third way to measure how close two languages are, is to consider the similarity between the durations of phones common to both languages. Phone durations are obtained by forced alignment, using tied-state triphone models together with the semi-tied transforms. The mean durations of common phones within each language are then compared, calculated as the sum of the squared differences between these mean durations, normalized by the sum of the mean durations of the phonemes in the pair of languages under consideration. (Normalization is performed to prevent differences between longer phone classes such as vowels to dominate the analysis.) These normalized distances are also presented in tables 4 and 5.

Distance	Nso-Sot	Nso-Tsn	Sot-Tsn
Acoustic distance	1.024	1.157	1.167
Similarity of frequencies	1.162e-02	1.148e-02	1.581e-02
Distance between durations	0.110	0.097	0.090

Table 4. Distance measures between the South African Sotho-Tswana languages, as described in Section 5.1

Distance	Nbl-Xho	Nbl-Zul	Xho-Zul
Acoustic distance	1.316	1.232	1.144
Similarity of frequencies	1.396e-02	1.649e-02	1.301e-02
Distance between durations	0.100	0.097	0.095

Table 5. Distance measures between the Nguni languages employed in this study, as described in Section 5.1

5.2. Recognition results

Figure 2 summarizes the phone-recognition accuracies that were obtained by pooling different sets of data. (In all cases, a flat language model was employed - that is, each phone was allowed to transition to any other phone. As a point of reference, our baseline recognizers were found to have word error rates ranging between 2% and 12% on a ten-word speaker-independent recognition task.) It can be seen that all languages seemed to benefit from the addition of data from closely related languages, except Sepedi. isiZulu in particular performed much better with the addition of isiNdebele and isiXhosa, with an improvement of approximately 2.6% absolute. To assess the magnitude of this improvement, one needs to keep in mind that asymptotic phone-recognition accuracy (with unlimited training data) using only bigram constraints is substantially less than 100%. In earlier work [17] we used parametric fits of accuracy against the amount of training data to estimate asymptotic phone-recognition accuracies for these languages. Based on those calculations, we estimate that the additional accuracy achieved by adding isiNdebele data to the isiZulu training data (our most beneficial pooling) is similar to the benefit that would be achieved by adding approximately another 250 minutes of isiZulu training data. Similarly, the the addition of the Sesotho data to the Setswana recognizer is found to be equivalent to the addition of approximately 110 minutes of Setswana data.

We also see that adding languages from other sub-families (such as isiZulu to the Sotho-Tswana languages) degrades performance significantly, and that the addition of Siswati data to the other Nguni languages is also detrimental in all cases.

Comparing these results with the distance measures shown in tables 4 and 5 suggests that similarity in triphone frequencies is the best predictor of how well data pooling will work. Sepedi, for example, is further away from Sesotho and Setswana than any of the other languages by this measure, and this correlates with the phone recognition results in figure 2, where Sepedi does not add any value to either Sesotho or Setswana. Sesotho and Setswana both improve when adding data from one to the other, as do the Nguni languages, with the angle between the isiZulu and isiNdeblele phone-count vectors being particularly small. The comparison of phone durations is somewhat aligned with the observed recognition accuracies (compare, for example, the relationship between Sepedi and Sesotho), but the measure of acoustic differences that we have employed does not seem to predict the behaviour of data pooling at all. This measure does not correlate with either the assessments of the other two measures (which are fairly comparable in ordering the six languages studied here) or the recognition results observed.

6. CONCLUSION

In this paper we investigated the effectiveness of pooling speech data to improve ASR system performance of resource-scarce languages. We have shown that for both the Nguni and Sotho-Tswana language groups, a non-negligible improvement in ASR correctness can be achieved by combining appropriate speech data sourced from closely related languages. In the best case, approximately 520 minutes of isiNdebele training data is found to improve accuracy to a similar extent as would be expected from approximately 250 minutes of isiZulu data. The next best improvement, to Setswana from approximately 420 minutes of Sesotho data, was seen to be equivalent to approximately 110 additional minutes of Setswana data. These provide rough guidelines for the benefit that can be achieved from pooling speech data from closely related languages families – namely, that two to four hours of cross-language data can give similar benefit to one hour of target-language data.

The factors that influence data combination, as described in Section 2, should of course be kept in mind. It would there-

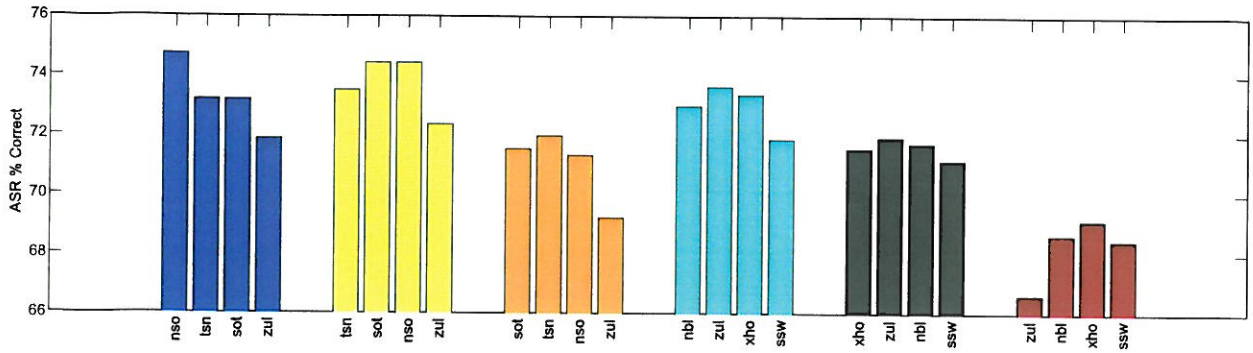

Fig. 2. ASR phone-recognition accuracies for Sepedi, Setswana, Sesotho, isiNdebele, isiXhosa and isiZulu. In each "cluster", the first bar indicates the baseline phone correctness for the particular language being recognized. Each subsequent bar is labelled with the language from which additional training data was added, in addition to all training data used for the previous bar. In this way, the 3^{rd} bar from the 2^{nd} (yellow) cluster, indicates the phone correctness obtained when recognizing Setswana, having used training data from Setswana, Sesotho and Sepedi.

fore be very interesting to repeat the comparisons performed here with different amounts of target and donor data, and also to investigate other language families with greater or lesser language similarities. It will also be interesting to see whether more elaborate data combination strategies can produce larger benefits from the combination of data from closely related languages.

Our results suggest that similarity in the frequencies of the various triphones is the best predictor of data-pooling performance amongst those measures studied here. This suggestion should be evaluated on data from other language families, and it may be fruitful to search for other measures that are even better predictors.

7. REFERENCES

[1] A. Constantinescu and G. Chollet, "On cross-language experiments and data-driven units for ALISP," in *Automatic Speech Recogntion and Understanding (ASRU)*, 1997, pp. 606–613.

[2] U. Bub, J Kohler, and B Imperl, "In-service adaptation of multilingual Hidden-Markov-Models," in *ICASSP*, Munich, Germany, 1997, pp. 1451–1454.

[3] T. Schultz and A. Waibel, "Multilingual and crosslingual speech recognition," in *Proceedings of the DARPA Workshop on Broadcast News Transcription and Understanding*, Landsdowne, VA, 1998, pp. 259–252.

[4] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusumy, "An evaluation of cross-language adaptation for rapid HMM development in a new language," in *ICASSP*, Adelaide, Australia, 1994, pp. 237–240.

[5] J. Kohler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," Seattle, WA, 1998.

[6] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *ICSLP*, Sydney, Australia, 1998, pp. 1819–1822.

[7] P. Dalsgaard and O. Andersen, "Identification of mono- and poly-phonemes using acoustic-phonetic features derived by a self-organising neural network," in *ICSLP*, Banff, Canada, 1992, pp. 547–550.

[8] U. Ackerman, B. Angelini, F. Brugnara, M. Federico, D. Giuliani, R. Gretter, G. Lazzari, and H. Niemann, "Speedata: Multilingual spoken data-entry," in *ICSLP*, Philadelphia, PA, 1996, pp. 2211–2214.

[9] L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna, "Testing generality in JANUS: A multilingual speech to speech translation system," in *ICASSP*, San Francisco, CA, 1992, vol. 1, pp. 209–212.

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

[11] C. Nieuwoudt and E. C. Botha, "Cross-language use of acoustic information for automatic speech recognition," *Speech Communication*, vol. 38, pp. 101–113, September 2002.

[12] O. Anderson, P. Dalsgaard, and W. Barry, "On the use of data-driven clustering techniques for language identification of poly- and mono-phonemes for four european languages," in *ICASSP*, Adelaide, Australia, 1994, pp. 121–124.

[13] J. Kohler, "Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks," Leusden, Netherlands, 1999, pp. 79–84.

[14] B. Imperl, Z. Kacic, B. Horvat, and A. Zgank, "Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones," *Speech Communication*, vol. 39, no. 3-4, pp. 353–366, 2003.

[15] T. Schultz and K. Kirchhoff, Eds., *Multilingual Speech Processing*, Elsevier, 2006.

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

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

[18] M. Paul Lewis, *Ethnologue: Languages of the World, Sixteenth edition*, SIL International, 2009.

[19] P.N. Zulu, G. Botha, and E. Barnard, "Orthographic measures of language distances between the official south african languages," *The Literator: Journal of literary criticism comparative linguistics and literary studies*, vol. 29, pp. 185–204, April 2008.

[20] Steve Young, "Large vocabulary continuous speech recognition: a review," in *of INCIS Project, Schedule 6 in (Small)*, 1996.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, 2009.

[22] M. J. F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, May 1999.

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