Capitalising on North American speech resources for the development of a South African English large vocabulary speech recognition system

Herman Kamper[a], Febe de Wet[a,b], Thomas Hain[c], Thomas Niesler[a,*]

[a] *Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa*
[b] *Human Language Technology Competency Area, CSIR Meraka Institute, Pretoria, South Africa*
[c] *Department of Computer Science, University of Sheffield, UK*

Abstract

South African English is currently considered an under-resourced variety of English. Extensive speech resources are, however, available for North American (US) English. In this paper we consider the use of these US resources in the development of a South African large vocabulary speech recognition system. Specifically we consider two research questions. Firstly, we determine the performance penalties that are incurred when using US instead of South African language models, pronunciation dictionaries and acoustic models. Secondly, we determine whether US acoustic and language modelling data can be used in addition to the much more limited South African resources to improve speech recognition performance. In the first case we find that using a US pronunciation dictionary or a US language model in a South African system results in fairly small penalties. However, a substantial penalty is incurred when using a US acoustic model. In the second investigation we find that small but consistent improvements over a baseline South African system can be obtained by the additional use of US acoustic data. Larger improvements are obtained when complementing the South African language modelling data with US and/or UK material. We conclude that, when developing resources for an under-resourced variety of English, the compilation

[*] Corresponding author. Tel.: +27 21 808 4118.

Email addresses: kamperh@sun.ac.za (Herman Kamper), fdw@sun.ac.za (Febe de Wet), t.hain@dcs.shef.ac.uk (Thomas Hain), trn@sun.ac.za (Thomas Niesler)

of acoustic data should be prioritised, language modelling data has a weaker effect on performance and the pronunciation dictionary the smallest.

Keywords: under-resourced languages, accented speech, South African English, varieties of English

1. Introduction

We will describe the development of large vocabulary speech recognition systems for South African English (SAE), which is considered an under-resourced variety of English because exceedingly little annotated speech data is currently available (Davel et al., 2011; Kamper et al., 2012a). However, although SAE may be considered under-resourced, other varieties of English, notably North American (US) English, have abundant resources for the development of speech technology. The primary aim of the research presented in this paper is to determine how best to capitalise on these existing and extensive language, pronunciation and acoustic modelling resources in the development of our South African (SA) speech transcription system. We consider the following two research scenarios and have structured the paper accordingly.

Firstly, we investigate the performance penalty incurred when a specific SA system component is absent and its US counterpart is used instead. To achieve this, we perform a balanced comparison in which SA and US systems are developed under equivalent model training conditions using speech corpora of similar size and character. We highlight language, pronunciation and acoustic differences through cross-domain experiments in which SA language models, pronunciation dictionaries and acoustic models are replaced by their US counterparts and vice-versa. By balancing the training conditions of the SA and US systems we try to minimise performance differences due to mismatches in training corpus nature and size. The results of this investigation identify the components upon which future SA resource collection and system development efforts should focus and the components which can feasibly be replaced by their US counterparts.

Secondly, we determine whether the extensive US acoustic and language modelling resources could be used to improve on the performance of an SA system. Here the US dataset is not limited artificially in size. Rather, the complete and much larger US dataset is considered and experiments are performed in order to determine the extent to which the US resources can be useful in the development of the SA system. These experiments reflect a typical under-resourced setting in which the in-domain data is limited but can be supplemented from extensive out-of-domain sources.

2. Background

Several studies have considered modelling approaches for different varieties of the same language. For example, Chengalvarayan (2001) dealt with the recognition of American, Australian and British varieties of English and showed that a single acoustic model obtained by pooling data outperformed a system employing separate models for each variety in parallel. Other authors have considered adaptation approaches in which a model trained on one variety is adapted using data from another variety. For example, Kirchhoff and Vergyri (2005) adapted Modern Standard Arabic acoustic models for improved recognition of Egyptian Conversational Arabic. Similarly, Despres et al. (2009) found that an accent-independent model which has been adapted with accented data outperformed both accent-specific and accent-independent models for Northern and Southern varieties of Dutch. Recently, selective data sharing across language varieties through the use of appropriate decision-tree state clustering algorithms has received some attention (Caballero et al., 2009; Kamper et al., 2012b). In these studies, the multilingual modelling approaches first proposed by Schultz and Waibel (2001) were extended to apply to multiple varieties of the same language.

There has also been increased recent interest in the development of systems using limited speech resources. Several authors have considered schemes in which models trained on one set of languages are used to obtain models for a new target language, either through a bootstrapping approach (Le and Besacier, 2009) or through multilingual unsupervised training (Vu et al., 2011). Currently, new adaptation and modelling approaches such as the use of MLP-based features (Qian et al., 2011; Vu et al., 2012), deep neural networks (Swietojanski et al., 2012), and subspace Gaussian mixture modelling (Zhang et al., 2012; Imseng et al., 2012) are being extended in order to deal with the challenges presented by limited resources.

Large vocabulary speech recognition of SAE has received very limited attention in the literature. Davel et al. (2011) presented a segmentation technique for harvesting SAE speech from the internet and showed that this is a viable option for creating corpora from publicly-available sources. Kamper et al. (2012b) considered speech recognition of the different accents of English spoken in South Africa using a fairly constrained dialogue-oriented telephone speech corpus. A larger wideband corpus of prompted SAE is currently being compiled as part of the NCHLT project (De Vries et al., 2011).

In contrast to SAE, there has been a long-standing focus on the continuous improvement of large vocabulary speech recognition systems for US English, notably in the broadcast news domain (Woodland et al., 1997; Gales et al., 2006). This task has been extended to other languages and language varieties including British (UK) English (Abberley et al., 1998), Italian (Cettolo, 2000), French (Gauvain et al., 2005), Turkish (Arısoy et al., 2007), dialects of German (Hecht et al., 2002), and varieties of Dutch (Van Leeuwen et al., 2009; Despres et al., 2009). All of these can be considered well-resourced languages because in each case substantial speech resources are available.

The research we present here is an experimental investigation and is based on established methodologies. Our contribution is a systematic and contrastive study which shows how data from the well-resourced US domain can be used to replace or supplement corresponding SA resources and thereby support the development of large vocabulary speech recognition in this under-resourced variety of English. Furthermore, in contrast to many previous studies, we consider not only acoustic modelling, but also pronunciation and language modelling in our contrastive experiments. Our results can be used to determine where resource development efforts should be focused in an under-resourced domain and in which ways more extensive resources from a well-resourced variety can be expected to be useful.

3. Speech resources

3.1. South African acoustic data

The work presented here is based on a recently-compiled corpus of SA broadcast news (Kamper et al., 2012a). The broadcast news domain is attractive for the development of large vocabulary speech recognition systems in under-resourced environments because it provides both a ready source of audio data as well as a variety of speech styles and quality. The SA corpus consists of approximately 20 hours of audio recordings from one of the country's main radio news channels, *SAFM*. News bulletins were broadcast between 1996 and 2006 and are a mix of newsreader speech, interviews, and crossings to reporters. These varying channel conditions were manually annotated for each sentence-level segment in the corpus. In addition to channel condition, speaker identity and accent were noted for each segment. The majority of the speakers in the corpus (contributing approximately 80% of the data) can be considered native speakers of English. Audio was sampled at 16 kHz and stored with 16-bit precision.

The corpus was divided into training (SA_ACtrain) and test (SA_test) partitions as indicated in Table 1. The first chronological 17.10 hours of data (extending up to March 2005) was used for training and the last 2.65 hours (April 2005 to March 2006) for testing. Some speaker overlap between the training and test sets exists because data from the same newsreaders is present in both. In particular, of the 535 speakers in the training set, 34 are also present in the test set.

Table 1: Composition of the South African acoustic training (SA_ACtrain) and test (SA_test) sets.

	SA_ACtrain	SA_test
Segments	9147	1412
Speakers	535	107
Speech (h)	17.10	2.65

3.2. North American acoustic data

Extensive resources are available for North American (US) broadcast news through the Linguistic Data Consortium (LDC). Unfortunately there is no fully transcribed corpus covering the same epoch as the SA data described in the previous section. A US corpus of comparable structure and size was therefore derived from the HUB-4 1996/1997 data (Graff et al., 1997; Fiscus et al., 1998). Since the SA data was collected from a single radio channel, we used US data from only two shows: *CNN Prime Time News* and *CNN The World Today*. The data was further divided into training (US_ACtrain) and test (US_test) sets as indicated in Table 2. US_ACtrain stretches from 14 April 1996 to 5 October 1996, with US_test continuing to the end of December 1997. The representation of the two *CNN* shows are approximately equal in both sets. As for the SA corpus, audio was sampled at 16 kHz and stored with 16-bit precision.

In addition to US_ACtrain, which we used for the balanced system comparison presented in Section 5, we extracted a more extensive US training set for use in the data augmentation and adaptation experiments described in Section 6. Using all the HUB-4 1996/1997 data but excluding the data in US_test, we compiled US_AC130h consisting of approximately 130 hours of US acoustic data. This additional training set is representative of the more extensive US resources and is also indicated in Table 2.

Table 2: Composition of the North American acoustic training (US_AC130h, US_ACtrain) and test (US_test) sets. US_ACtrain was developed to be comparable to SA_ACtrain while US_AC130h is representative of the extensive availability of US acoustic resources.

	US_AC130h	US_ACtrain	US_test
Segments	36 669	5799	770
Speakers	5555	1115	202
Speech (h)	129.31	17.27	2.70

3.3. Text sources

A corpus of newspaper text was collected from a number of major South African newspapers, including *The Financial Mail*, *Business Day*, *The Sunday Times*, *The Times*, *Sunday World*, *The Sowetan*, *The Herald*, *The Algoa Sun* and *The Daily Dispatch*. From this text an SA language model training set (SA_LMtrain) consisting of approximately 109 million words and including material from January 2000 to March 2005 was compiled.

A similarly sized corpus of US language model training material was released by the LDC (MacIntyre, 1998). This set (US_LMtrain) consists of approximately 130 million words and was collected from transcribed news broadcasts aired between January 1992 and June 1996. In addition to the SA and US data we have also considered the use of UK language model training material. For this purpose we have used a 30 million word corpus (UK_LMtrain) of transcribed BBC broadcasts stretching from early 1997 to the end of 1999.

Finally, the transcripts of the SA and US broadcast news training sets were also both available for language modelling purposes

3.4. Pronunciation dictionaries

An SA training pronunciation dictionary (used during alignment) for the 14 622 unique words in SA_ACtrain was developed by a phonetic expert (Loots and Niesler, 2010). Subsequently, pronunciations for the most frequent words in the SA language modelling data (Section 3.3) were determined by the same phonetic expert to obtain a recognition pronunciation dictionary (SA_dict60k) with 60 698 words and on average 1.25 pronunciations per word.

Similar US pronunciation dictionaries were derived from the background dictionary described in (Wan et al., 2008). A training dictionary was generated for the 13 148 unique words in US_ACtrain. To these, pronunciations for the most frequent words in the US language modelling data (Section 3.3) were added to obtain a recognition dictionary (US_dict60k) with 59 642 words and on average 1.02 pronunciations per word. A separate training dictionary was also generated for the unique words in US_AC130h. All pronunciation dictionaries (both SA and US) are based on the same set of 45 phones. This phone set was derived from ARPABET, as described by Rabiner and Juang (1993), after mapping the rare phones /ɨ/, /ʍ/ and /ɾ̺/ to /ə/, /w/ and /t/ respectively, and deleting the glottal stop /ʔ/.

A comparison between SA, UK and US pronunciation dictionaries has revealed that, in general, SA pronunciations are closer to their UK than to their US counterparts, with the bulk of the differences being accounted for by vowels (Loots and Niesler, 2010). In relation to other varieties, the phonetics of South African English is often characterised by the vowels in the words "kit" and "bath" (Bowerman, 2004). The former is referred to as the "kit split", and describes the process by which the close front vowel , as used in US English, becomes the centralised /ə/. The vowel used in the SA pronunciation of "bath" is the back /a/, whereas the common US pronunciation would employ the fronted /æ/. In contrast to US English, SA English is non-rhotic, meaning that the liquid /r/ is not pronounced in words such as "start" and "star".

Our initial intention when embarking upon this research was to compare the language, pronunciation and acoustic differences between SA, US and UK varieties of English. However, despite some effort, we were unable to acquire either an appropriate UK acoustic corpus or a suitable UK pronunciation dictionary. Nevertheless, UK language modelling experiments are presented in Section 6.

4. General experimental procedure

All acoustic models (AMs) were developed following the same procedure, which is similar to that proposed in (Hain et al., 2010). Audio data was converted into a stream of 39-dimensional mel-frequency perceptual linear prediction (MF-PLP) feature vectors (Woodland et al., 1997). Cepstral means were subtracted on a per-utterance basis and subsequently cepstral variance normalisation was performed on a per-bulletin basis. Using the HTK tools (Young et al., 2009), state-clustered phonetic decision-tree tied triphone HMMs were trained using a three-state left-to-right model topology and 16 Gaussian mixtures per state. Language models (LMs) were trained using the SRILM toolkit (Stolcke, 2002). Trigram language models were used throughout, with Kneser-Ney smoothing and Katz backoff (Chen and Goodman, 1999). The first-best output from the HTK HDecode tool (Young et al., 2009) was used in all recognition experiments. All word error rates (WERs) were computed using the NIST Scoring Toolkit SCTK (NIST, 2009).

5. Substituting SA with US resources

In this section we investigate the effect of replacing SA language models, pronunciation dictionaries and acoustic models with their US counterparts and vice-versa. The aim is to assess the performance penalties involved when incorporating US system components into our SA system. The results will enable us to differentiate between speech resources that can feasibly be inherited from the US domain and speech resources which are best developed separately for the SA domain. For the experiments in this section, SA and US models were trained on similar amounts of data from comparable sources in order to ensure that results are not skewed by differences in training corpus character or size.

5.1. Models and experimental setup

Using the procedure described in Section 4, a baseline SA acoustic model (SA_AM1) with 2624 states was trained on SA_ACtrain. A comparable baseline US acoustic model (US_AM1) with 2697 states was trained on US_ACtrain.

In order to arbitrarily exchange language models and pronunciation dictionaries between the SA and US systems, a consistent vocabulary is required. In the following experiments we therefore restricted the vocabulary of the SA and US language models and pronunciation dictionaries to the 39 423 words that are common to SA_dict60k and US_dict60k. Note that this restriction does not affect the size of the acoustic training set. For the SA system, language models were trained separately on SA_LMtrain and SA_ACtrain and then linearly interpolated to yield the baseline SA_LM40k model. Similarly, language models were trained separately on US_LMtrain and US_ACtrain and linearly interpolated to yield US_LM40k. A development set consisting of the transcriptions of the chronologically most recent 5% of SA_ACtrain (Table 1) was used to optimise the language interpolation weights for SA_LM40k. An analogous strategy was followed for US_LM40k, in this case using the most recent 5% of US_ACtrain (Table 2). In both cases language model performance was observed to be insensitive to the precise values of the interpolation weights, which exhibited wide, flat perplexity minima.

The perplexities and out-of-vocabulary (OOV) rates measured on SA_test and US_test using the two baseline language models are given in Table 3. Note that, since the two language models are based on the same 40k vocabulary, their OOV rates correspond. Under mismatched conditions substantial perplexity increases of more than 80% are observed. Recognition pronunciation dictionaries containing SA and US pronunciations for the same 39 423-word vocabulary were also compiled and are referred to as SA_dict40k and US_dict40k, respectively.

Table 3: The 40k trigram language model perplexities and OOV rates measured on SA_test and US_test.

Language model	SA_test		US_test	
	Perplexity	OOVs	Perplexity	OOVs
SA_LM40k	129.6	3.78%	343.4	1.53%
US_LM40k	238.6	3.78%	188.1	1.53%

5.2. Language model swaps

Table 4 shows recognition performance when exchanging language models between the SA and US systems. Configurations 1 and 4 are the SA and US 40k baselines, respectively. By comparing the cross-domain results of these two configurations it is evident that there is a large mismatch between the SA baseline system and the US test set and vice-versa, with WERs more than doubling in both cases (from 28.1% to 58.7% on SA_test and from 30.9% to 62.6% on US_test). By comparing the performance of configurations 1 and 2 on SA_test and configurations 4 and 3 on US_test, an absolute increase of approximately 5% in WER is observed in both cases when using the language model from the other domain.

Table 4: WERs (%) measured on SA_test and US_test when exchanging SA and US language models (LMs) between systems.

Configuration			SA_test	US_test
AM	LM	Dictionary		
1. SA_AM1	SA_LM40k	SA_dict40k	28.1	62.6
2. SA_AM1	US_LM40k	SA_dict40k	32.8	57.2
3. US_AM1	SA_LM40k	US_dict40k	49.5	36.0
4. US_AM1	US_LM40k	US_dict40k	58.7	30.9

5.3. Pronunciation dictionary swaps

Table 5 presents recognition results for systems in which pronunciation dictionaries have been exchanged. Configurations 1 and 4 are the 40k baseline systems first shown in Table 4. When the US dictionary is used in the SA system (configuration 5), a degradation of 8% in WER is observed on SA_test. This degradation is far less than that observed on US_test when using an SA dictionary in a US system (configuration 6) leading to an increase of almost 17% in WER.

Although it seems that mismatched dictionaries (8% to 17% penalty) have a more severe impact on WER than mismatched language models (∼5% penalty, Section 5.2), we show in the following that this is due to the mismatch between the dictionaries used during training and recognition. Table 5 shows that a configuration in which the domain of the acoustic model is inconsistent with the dictionary exhibits worse performance than a setup in which only the test set is mismatched. For example, the combination of the US acoustic model and US dictionary (configuration 4) yields better performance on SA_test (58.7%) than the combination of the US acoustic model and SA dictionary (configuration 6) on the same test data (62.3%). This highlights the interdependence of the acoustic model and the pronunciation dictionary. In order to verify this interpretation, we have trained an SA acoustic model using an alignment obtained using US pronunciations. Pronunciations for 9249 words in SA_ACtrain were available in our US background dictionary, covering only 55% of the SA_ACtrain training set. In a two-model re-estimation procedure, SA_AM3 was trained on this subset of SA_ACtrain. To provide a fair baseline, a matching model set (SA_AM2) was trained on the same subset of SA_ACtrain but using SA pronunciations. SA_AM2 and SA_AM3 were chosen to have approximately the same num-

Table 5: WERs (%) measured on SA_test and US_test when exchanging SA and US pronunciation dictionaries between systems.

Configuration			SA_test	US_test
AM	LM	Dictionary		
1. SA_AM1	SA_LM40k	SA_dict40k	28.1	62.6
5. SA_AM1	SA_LM40k	US_dict40k	36.1	64.8
6. US_AM1	US_LM40k	SA_dict40k	62.3	47.8
4. US_AM1	US_LM40k	US_dict40k	58.7	30.9

Table 6: WERs (%) measured on SA_test with systems employing SA acoustic models (AMs) trained using SA and US pronunciations, respectively labelled SA_AM2 and SA_AM3. These two AMs were trained on a subset of SA_ACtrain. SA_AM4 was trained on all of SA_ACtrain using US pronunciations partially determined by letter-to-sound rules.

Configuration			SA_ test
AM	LM	Dictionary	
7. SA_AM2	SA_LM40k	SA_dict40k	29.1
8. SA_AM3	SA_LM40k	US_dict40k	29.8
9. SA_AM4	SA_LM40k	US_dict40k	28.6

ber of parameters (approximately 1800 states). The performance of these acoustic models is presented in Table 6.

The 1% absolute drop in WER between configurations 1 and 7 in Tables 5 and 6, respectively, is a result of training on a smaller dataset. The results show that the 8% performance drop in WER on SA_test between configurations 1 and 5 (Table 5) is much larger than the 0.7% drop between configurations 7 and 8 (Table 6). Hence we see that, when using a recognition pronunciation dictionary that is consistent with the training dictionary, the performance degradation caused by the dictionary mismatch is dramatically reduced.

Since a considerable amount of data is lost in the above procedure, a further experiment was conducted in which the missing US pronunciations were generated using letter-to-sound rules (Wan et al., 2008). All the SA acoustic training material could thus be utilised to train an SA acoustic model using US pronunciations (SA_AM4). Using this acoustic model with SA_LM40k

Table 7: WERs (%) measured on SA_test and US_test when exchanging SA and US acoustic models (AMs) between systems.

	Configuration			SA_test	US_test
	AM	LM	Dictionary		
1.	SA_AM1	SA_LM40k	SA_dict40k	28.1	62.6
10.	US_AM1	SA_LM40k	SA_dict40k	54.0	53.3
11.	US_AM2	SA_LM40k	SA_dict40k	48.7	-
12.	SA_AM1	US_LM40k	US_dict40k	41.9	60.0
4.	US_AM1	US_LM40k	US_dict40k	58.7	30.9

and US_dict40k, a WER of 28.6% was achieved on SA_test (configuration 9, Table 6). This figure should be compared with the SA baseline of 28.1% WER (configuration 1, Table 5). Hence, a drop of only 0.5% in performance is observed compared to the 8% drop incurred when training the acoustic model using SA pronunciations (configuration 5). We can conclude that, as long as we ensure that the training and recognition pronunciation dictionaries correspond, the adoption of US pronunciations in an SA system leads to a relatively small deterioration in performance.

5.4. Acoustic model swaps

Aside from repeating the 40k baselines first presented in Table 4, Table 7 indicates the performance of an SA system using a US acoustic model (configuration 10) as well as a US system using an SA acoustic model (configuration 12). It is evident that exchanging acoustic models between domains leads to severe performance degradation. This is also observed consistently in the results reported in Tables 4 and 5.

However, the preceding section has highlighted the interdependence of acoustic models and pronunciation dictionaries. With this in mind, a fairer estimate of the performance degradation that can be expected when using US acoustic data in an SA system can be found by following a procedure analogous to that used to train SA_AM3. A new US acoustic model (US_AM2) was trained on the 54% of US_ACtrain which was covered by the 8998 words for which SA pronunciations were available in SA_dict60k. A system employing this US model with SA_LM40k and SA_dict40k achieved a WER of 48.7% (configuration 11, Table 7). By comparing this system to configuration 7 in Table 6 (which employs an SA acoustic model trained on a similar amount

Table 8: Summary of WERs (%) and absolute penalties (%) measured on SA_test when sacrificing SA system components for US alternatives. All configurations use the same 40k vocabulary.

	Configuration	WER	Penalty
1.	SA 40k baseline	28.1	-
11.	Replace SA with US acoustic model	48.7	20.6
2.	Replace SA with US language model	32.8	4.7
9.	Replace SA with US pronunciations	28.6	0.5

of SA data), it is apparent that an acoustic model mismatch still results in a big penalty even when consistency between the training and recognition dictionaries is ensured.

5.5. Detailed summary and conclusion

Section 5 has presented experiments in which language, pronunciation and acoustic models were exchanged between comparable systems developed for the SA and US domains. Experiments focused on the usability of US system components in the SA system in order to determine the penalty involved when importing components from this variety.

A summary of the most important results is given in Table 8. Note that the penalties shown in this table are not cumulative, but indicate the impact of swapping the three different components of the recognition system. Acoustic differences were found to contribute most to degradation, with substantial deterioration in all cross-domain tests. In particular, a 20.6% penalty is incurred when US acoustic data is used to train acoustic models for the SA system. Experiments in which language models were exchanged indicated a drop of 4.7% in WER when sacrificing the SA language model for its US alternative. Although the exchange of SA and US dictionaries initially indicated big penalties, further investigation revealed that when US pronunciations are used consistently during the training of SA acoustic models as well as during recognition, the deterioration in WER is just 0.5%.

We conclude that, from an SA perspective, pronunciations from the better-resourced US variety of English can be used at a relatively small cost. Language modelling data can also be used, but at a slightly higher cost. However, a substantial penalty is paid when using acoustic data from the US domain.

6. Augmenting SA with US resources

In Section 5 we considered the case in which an SA language, pronunciation or acoustic model was assumed to be unavailable and consequently a US counterpart was used *instead*. In this section we consider the case in which the US resources are available *in addition to* the SA training material. The aim is to determine whether performance improvements can be obtained by augmenting the available SA resources with US resources during system development.

6.1. Models and experimental setup

For the experiments in this section, the larger US corpus US_AC130h was used to train a new US acoustic model (US_AM130h) with 4518 states following the procedure described in Section 4. Since the focus of this section is to achieve the best possible recognition accuracy on the SA test set, we use the full SA_dict60k pronunciation dictionary (Section 3.4) and associated 60k vocabulary. Trigram language models were trained on SA_LMtrain and SA_ACtrain using this 60k vocabulary. The two resulting models were linearly interpolated in a procedure analogous to that described in Section 5.1 to yield SA_LM60k. This new baseline language model achieves a perplexity of 139.9 on SA_test with an OOV rate of 1.02%.

6.2. Augmenting SA with US acoustic training data

While the HUB-4 1996/1997 training set US_AC130h contains approximately 130 hours of US acoustic data (Table 2), the SA training set SA_ACtrain contains just 17 hours (Table 1). In the following we determine whether the much larger US dataset can be used to improve the performance of an SA speech recogniser.

Table 9 summarises the recognition performance of systems employing different acoustic models trained on the SA and US sets. In all cases decoding of SA_test was performed using SA_LM60k and SA_dict60k. Configuration 13 is the 60k baseline SA system and employs SA_AM1 as acoustic model (Section 5.1) which was trained exclusively on SA_ACtrain. This 60k system shows an absolute improvement of 3.5% in WER over the 40k SA baseline (28.1%, configuration 1, Table 4). This improvement may be ascribed to the larger vocabulary of the system, and associated lower OOV rate (1.02% instead of 3.78% on SA_test). Configuration 14 employs the larger US baseline acoustic model US_AM130h. As before, a large mismatch is seen between the

US acoustic model and the SA test set, with an increase of 20.5% absolute in WER relative to the SA baseline.

Configuration 15 employs an acoustic model (4653 states) trained on the combination of US_AC130h and SA_ACtrain by straightforward pooling. In this case, the US and SA training pronunciation dictionaries (Section 3.4) were used respectively for the alignment of the US and SA data during training. This model outperforms the US acoustic model used in configuration 14 but is still clearly inferior to the baseline SA acoustic model (configuration 13). This finding is in contrast to several studies in which improved performance was achieved by pooling data from different varieties of the same language (Chengalvarayan, 2001; Caballero et al., 2009; Despres et al., 2009), although it agrees with the findings in (Fischer et al., 1998) and some of the findings in (Kamper et al., 2012b). As before our results emphasise that the SA and US speech data are acoustically quite different and that an acoustic mismatch leads to an appreciable penalty.

For configuration 16, a number of iterations of maximum likelihood (ML) retraining on SA_ACtrain was performed using embedded Baum-Welch re-estimation, using the acoustic model of configuration 15 as starting point. Although WER performance improves from the 29.0% to 24.8%, this figure is still slightly inferior to the SA baseline of 24.6%.

Finally, maximum a posteriori (MAP) adaptation (Gauvain and Lee, 1994) was performed on the acoustic model of configuration 15, using SA_ACtrain as adaptation material. The performance of the resulting acoustic model is indicated by configuration 17 in Table 9. In comparison to the SA baseline (configuration 13), an improvement of 0.3% absolute in WER is observed. Using bootstrap confidence interval estimation (Bisani and Ney,

Table 9: WERs (%) measured on SA_test in the evaluation of several acoustic models when using US_AC130h in addition to SA_ACtrain. SA_LM60k and SA_dict60k were used in all cases during recognition.

	Acoustic model	WER
13.	SA_AM1: SA baseline trained on SA_ACtrain	24.6
14.	US_AM130h: US models trained on US_AC130h	45.1
15.	Models trained on US_AC130h and SA_ACtrain	29.0
16.	ML retraining of models in configuration 15	24.8
17.	MAP adaptation of models in configuration 15	24.3

2004), this improvement was found to be statistically significant only at the 80% level.

6.3. Reducing target domain acoustic training data

In configuration 17 the entire SA_ACtrain corpus was used for adaptation. Although SA_ACtrain is small compared to US_AC130h, it is still substantial in terms of what may be available in an under-resourced setting. In the following we consider the trends that emerge when adaptation is performed on increasingly smaller amounts of SA acoustic data.

Several subsets of SA_ACtrain were extracted, starting with the chronologically most recent data and systematically adding more data until all of SA_ACtrain was included. For each such subset, two acoustic models were trained: an SA-only acoustic model trained exclusively on the subset (analogous to SA_AM1 in configuration 13); and an acoustic model obtained by first pooling the data from that subset with US_AC130h and then applying MAP adaptation using the same subset (analogous to configuration 17).

Figure 1 shows the performance of systems using the SA-only and MAP-adapted acoustic models. SA_LM60k and SA_dict60k were used in all cases. The WER is shown as a function of the amount of SA acoustic data used for training or adaptation. The rightmost points on the two curves correspond to configurations 13 and 17 respectively (Table 9). Both curves indicate that the performance improvement obtained by using more data tapers off when more than approximately ten hours of SA data is available. Furthermore, the MAP-adapted models consistently outperform the corresponding SA-only models[1]. Initially this improvement is quite large, with a decrease of 2.5% absolute in WER when using just one hour of SA acoustic data. The improvement falls to 0.6% when six hours of SA data is available and reaches 0.3% when all 17 hours of data in SA_ACtrain is used. The average performance improvement of the MAP-adapted models over the SA-only models for systems using more than ten hours of SA data is 0.3% absolute in WER. We conclude that, when less in-domain SA acoustic data is available, we stand to gain more from also incorporating data from the well-resourced US domain.

[1] These observations are over the limited SA dataset (17 hours) and the particular training/test set split used. It is possible that, in a different setup, the two values shown in Figure 1 would converge

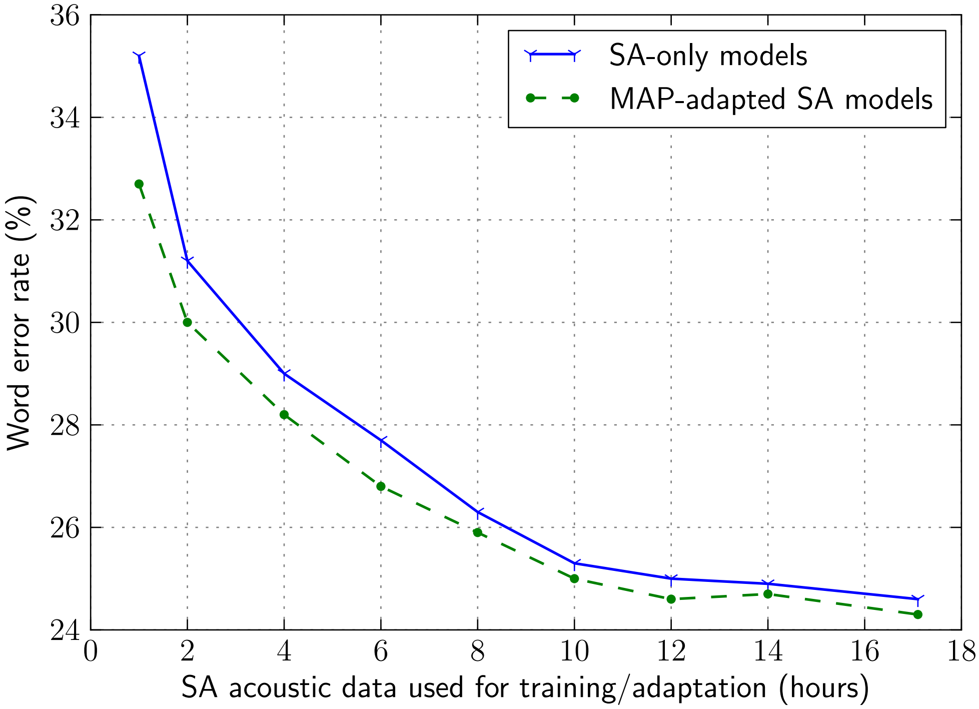

Figure 1: WERs measured on SA_test in a comparison of SA-only acoustic models (trained on SA data alone) and MAP-adapted acoustic models (trained by adapting a US-based model) when using different amounts of SA acoustic data. SA_LM60k and SA_dict60k were used in all cases during recognition.

Figure 1 also gives an indication of the amount of in-domain data that the 130 hours of out-of-domain data is "worth". For instance, the performance of an SA system trained exclusively on nine hours of SA data (achieving a WER of almost 26%) can also be achieved by using eight hours of SA data for the MAP adaptation of a US-based model. Hence, for an SA training set size of eight hours, the additional 130 hours of US data is worth one hour of SA data. For the systems incorporating less than ten hours of SA data, the "worth" of the US data fluctuates between one and two hours. For ten hours and more, it grows larger. In order to achieve the SA-only baseline performance of 24.6% WER, which requires 17 hours of SA data, approximately twelve hours of SA adaptation data is required in addition to the 130 hours of US data. Hence, for an SA training set size of twelve hours, the additional US data is worth five hours of SA data. Using this interpretation, the additive improvements afforded by the additional US data become increasingly valuable in terms of the additional amount of in-domain data that would have to be collected to achieve the same improvement.

6.4. Augmenting SA with US and UK language modelling training data

Using the SA_LMtrain, US_LMtrain and UK_LMtrain text sets (Section 3.3) as well as the transcriptions of the SA_ACtrain acoustic data (Table 1), several language models were trained and evaluated. First, trigram language models were trained separately on each of the four sets using the vocabulary of SA_dict60k. Next, these four language models were linearly interpolated in various combinations. Interpolation weights were optimised on the SA development set described in Section 5.1 using the SRILM toolkit (Stolcke, 2002). The SA_dict60k vocabulary was used throughout and decoding of SA_test was performed using SA_AM1 and SA_dict60k. Recognition results, language model perplexities, and language model interpolation weights are presented in Table 10.

First we wanted to determine, were no SA language modelling data to be available at all, whether it is better to incorporate US or UK material than to rely exclusively on the transcriptions of SA_ACtrain. Configuration 18 employs a language model trained exclusively on SA_ACtrain while configurations 19 and 20 interpolate this model with US and UK background data, respectively. Both configurations 19 and 20 outperform configuration 18, indicating that the use of either US or UK background material is advantageous. It appears that the UK-based model in configuration 20 is better matched to the SA domain than the US-based model in configuration 19.

Table 10: Language model interpolation weights, perplexities and WERs (%) measured on SA_test for the evaluation of several trigram language models when using US and/or UK background material in addition to SA material. SA_AM1 and SA_dict60k were used in all cases during recognition.

	US_	UK_	SA_	SA_	SA_test	
	LMtrain	LMtrain	LMtrain	ACtrain	Perplexity	WER
18.	-	-	-	1.00	328.9	33.9
19.	0.45	-	-	0.55	189.3	27.3
20.	-	0.55	-	0.45	174.0	26.8
13.	-	-	0.70	0.30	139.9	24.6
21.	0.12	-	0.61	0.27	134.5	23.9
22.	-	0.32	0.47	0.22	131.4	23.9
23.	0.04	0.30	0.46	0.21	129.8	23.8

This is despite the fact that UK_LMtrain (30M words) is much smaller than US_LMtrain (130M words).

Configuration 13 is the SA baseline first introduced in Table 9 and employs the SA_LM60k language model used in the experiments in Section 6.2. This model was obtained by interpolating language models trained on SA_LMtrain and on SA_ACtrain. By including a US language model in the interpolation, the perplexity improves and WER decreases by 0.7% absolute (configuration 21). By including a UK language model instead, an even lower perplexity is achieved but no further WER improvements are observed (configuration 22). In addition to the lower perplexity, a comparison of configurations 21 and 22 shows that that the UK language model is assigned a higher interpolation weight (0.32) than the US language model (0.12), again indicating that the UK data is better matched to the SA domain than the US data. Finally, by interpolating all four models from the US, UK and SA domains, the lowest perplexity and a WER of 23.8% is achieved (configuration 23). This represents an absolute improvement of 0.8% in WER over the SA baseline. When using this best overall language model in combination with the best overall acoustic model (used in configuration 17, Table 9), an overall best WER of 23.7% is achieved.

6.5. Reducing target domain language modelling data

In the preceding experiments we have used the entire SA_LMtrain corpus of approximately 109M words, which is substantial. In an under-resourced setting, this amount of language modelling material might not be available. We therefore also wanted to determine how system performance would change in a scenario where a substantial set of out-of-domain background language modelling material was assumed to be available (US and UK material in this case) but only a small set of in-domain (SA) material.

For this purpose, several SA background language modelling subsets of increasing size were extracted from SA_LMtrain. Using each of these, a trigram language model was trained using the vocabulary of SA_dict60k. Each of these subset language models was subsequently used to produce two new interpolated language models: an SA-only language model obtained by interpolating the subset language model with the language model trained on SA_ACtrain; and an SA+US+UK language model obtained by interpolating the subset language model with language models respectively trained on SA_ACtrain, US_LMtrain and UK_LMtrain. In all cases interpolation weights were optimised on the SA development set described in Section 5.1.

Figure 2 shows the perplexities and recognition performance achieved when using the SA-only and the SA+US+UK language models. SA_AM1 and SA_dict60k were used during decoding in all cases. The rightmost points correspond respectively to configurations 13 and 23 in Table 10. The WER is shown as a function of the size of the SA background language modelling set.

Consider first the curves for the SA+US+UK language models. An improvement of 0.9% absolute in WER is observed when using 10M words of SA data compared to the case where no SA data is used. As the training data is increased from 10M to 70M words, a commensurate decrease in both perplexity and WER is observed. The WER improves by 1.3% absolute over this range at a rate of approximately 0.22% WER per 10M words. After 70M words, recognition performance improvement seems to flatten out despite further improvements in perplexity. The only further improvement in recognition performance is observed when increasing the SA language modelling text from 100M to 109M words.

In contrast, the curves for the SA-only language models indicate that perplexity and recognition performance improve steadily as the SA language model training data increases. It is evident that the SA+US+UK language models consistently outperform the corresponding SA-only models. Initially

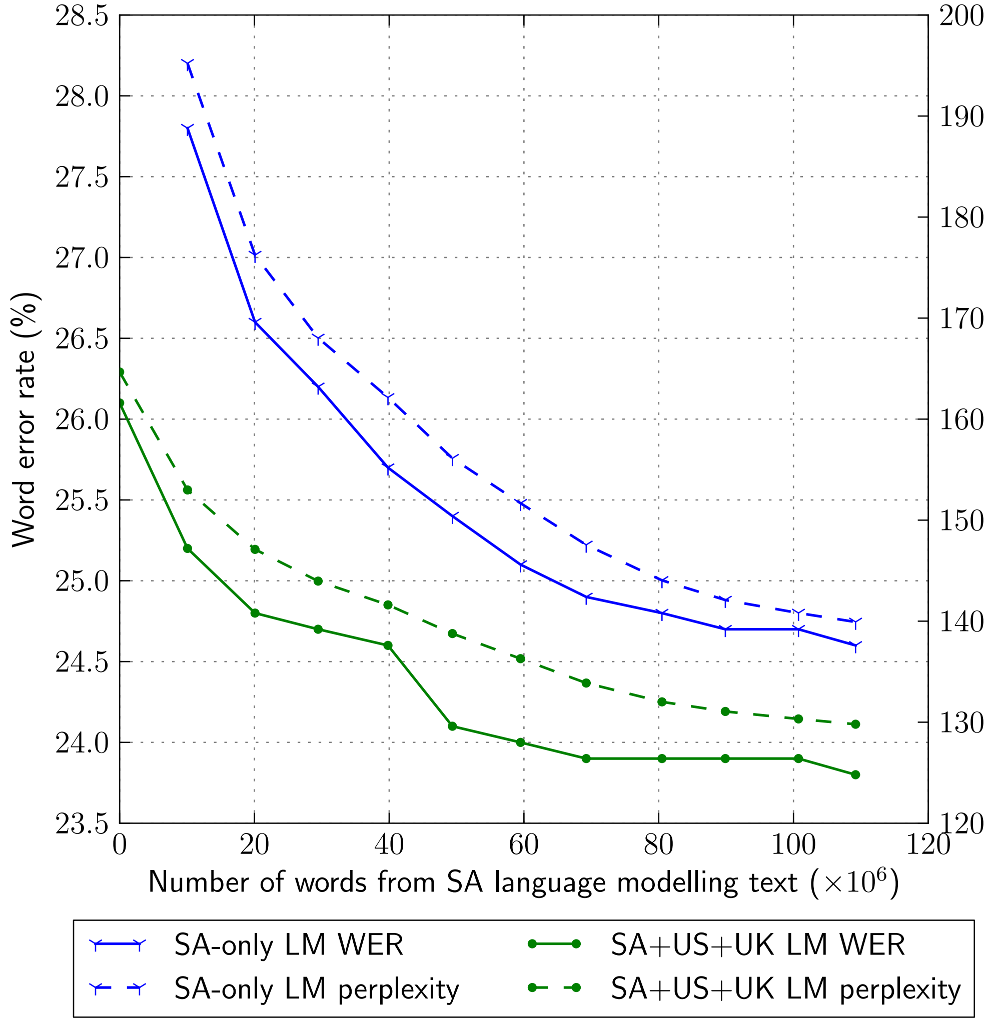

Figure 2: WERs measured on SA_test in the evaluation of language models when using US and UK background material in addition to different amounts of SA language modelling material. SA_AM1 and SA_dict60k were used in all cases during recognition.

this improvement is relatively high with a decrease of 2.6% absolute in WER when 10M words of SA language modelling data is available. This improvement falls to 1.3% at 50M words and finally to 0.8% when using the full 109M words in SA_LMtrain. In order to achieve a performance equal to the SA-only baseline of 24.6% WER (configuration 13), an SA background language modelling text of approximately 40M words is required in addition to the US and UK sets of respectively 130M and 30M words.

6.6. Detailed summary and conclusion

Section 6 has considered whether improved speech recognition performance can be achieved by supplementing the existing SA resources with US and/or UK data. A summary of the key results is given in Table 11. The incorporation of an additional 130 hours of US data into the SA system by MAP adaptation led to an absolute improvement of 0.3% in terms of WER relative to a system trained exclusively on the 17 hours of SA data. The incorporation of an additional 130M words of US and 30M words of UK language modelling data led to an absolute improvement of 0.8% in WER relative to a system using only 109M words of SA data. By supplementing both acoustic and language modelling data, an absolute improvement of 0.9% was achieved relative to a system trained exclusively on SA resources. Hence, US acoustic and language modelling resources can be used to improve the performance of an SA baseline system by almost one percent in WER. In terms of both acoustic and language modelling, when less SA data is available, we stand to gain progressively more from additional US and UK data. However, in both cases the benefit afforded by the additional US and UK data also becomes increasingly valuable in terms of the additional SA resources that would have to be compiled to achieve the same improvement.

Table 11: Summary of WERs (%) and absolute improvement (%) over the SA baseline, measured on SA_test, when including US and/or UK resources in the development of an SA system.

	Configuration	WER	Improvement
13.	SA baseline	24.6	-
17.	Include US acoustic data using MAP	24.3	0.3
23.	Use the SA+US+UK language model	23.8	0.8
24.	MAP-adapted AM and SA+US+UK LM	23.7	0.9

7. Overall conclusions

We have presented an experimental evaluation of the use of North American (US) resources in the development of a South African (SA) large vocabulary speech recognition system.

Speech recognition results showed that a US recognition system in its unmodified form is not suitable for use within the South African domain. Directed experiments indicated that differences between the two domains are present in language modelling data, in pronunciations, as well as in acoustic modelling data. In a scenario where certain SA resources were assumed to be completely absent, it was found that the incorporation of a US pronunciation dictionary into an SA system led to the smallest performance penalty (0.5% absolute in terms of word error rate). Larger accompanying penalties (between 2% and 5%) were observed when using language modelling data from the US or UK domains instead of corresponding data from the SA domain. Despite this degradation, when no SA language modelling data is available at all, it is better to incorporate material from the US or UK domains than to rely exclusively on the transcriptions of the SA acoustic data. The most severe penalty (more than 20% absolute) was observed when an acoustic model trained on SA data was replaced by a model trained on US data.

In a set of adaptation experiments we showed that SA acoustic models can be improved slightly but consistently (approximately 0.3% absolute) by incorporating a large corpus of US acoustic data in addition to the SA data. In this regard, maximum a posteriori (MAP) adaptation clearly outperformed straightforward pooling of the SA and US acoustic data which led to deteriorated performance. The addition of out-of-domain language modelling data from the US and UK domains also led to consistently better performance. These improvements (in the order of 0.8%) were larger than those attained by incorporating additional acoustic data.

Although the incorporation of additional US and/or UK data were beneficial, we found that an increase in the size of the SA training corpora remains the dominant driver for improved recognition performance. In terms of acoustic modelling data, our experiments indicated that at least ten hours of in-domain SA acoustic data should be compiled since performance begins to level off once this point is reached. In terms of language modelling data, the corresponding figure is approximately 50M words of text from the SA domain.

From the system developer's perspective it is clear that resource development efforts in an under-resourced setting such as ours should prioritise the compilation of acoustic training data above pronunciation and language modelling material. Additional acoustic and especially language modelling resources from other varieties of English can subsequently be used to further improve performance.

8. Acknowledgements

This research was supported financially by the Royal Society and the South African National Research Foundation (NRF) under a South Africa – UK Science Network grant. Parts of this work were executed using the High Performance Computer (HPC) facility at Stellenbosch University. The authors would like to thank Alison Wileman, for her hard work on the South African English pronunciation dictionary and transcriptions, and Matt Gibson, for his helpful comments and suggestions.

References

Abberley, D., Renals, S., Cook, G., 1998. Retrieval of broadcast news documents with the THISL system, in: Proc. ICASSP, Seattle, WA. pp. 3781–3784.

Arısoy, E., Sak, H., Saraclar, M., 2007. Language modeling for automatic Turkish broadcast news transcription, in: Proc. Interspeech, Antwerp, Belgium. pp. 2381–2384.

Bisani, M., Ney, H., 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation, in: Proc. ICASSP, Montreal, Quebec, Canada. pp. 409–412.

Bowerman, S., 2004. White South African English: phonology, in: Schneider, E.W., Burridge, K., Kortmann, B., Mesthrie, R., Upton, C. (Eds.), A Handbook of Varieties of English, Mouton de Gruyter, Berlin, Germany. pp. 931–942.

Caballero, M., Moreno, A., Nogueiras, A., 2009. Multidialectal Spanish acoustic modeling for speech recognition. Speech Commun. 51, 217–229.

Cettolo, M., 2000. Segmentation, classification and clustering of an Italian broadcast news corpus, in: Proc. RIAO, Paris, France. pp. 372–381.

Chen, S.F., Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. Comput. Speech Lang. 13, 359–394.

Chengalvarayan, R., 2001. Accent-independent universal HMM-based speech recognizer for American, Australian and British English, in: Proc. Eurospeech, Aalborg, Denmark. pp. 2733–2736.

Davel, M.H., Van Heerden, C., Kleynhans, N., Barnard, E., 2011. Efficient harvesting of internet audio for resource-scarce ASR, in: Proc. Interspeech, Florence, Italy. pp. 3153–3156.

De Vries, N.J., Badenhorst, J., Davel, M.H., Barnard, E., De Waal, A., 2011. Woefzela – an open-source platform for ASR data collection in the developing world, in: Proc. Interspeech, Florence, Italy. pp. 3177–3180.

Despres, J., Fousek, P., Gauvain, J.L., Gay, S., Josse, Y., Lamel, L., Messaoudi, A., 2009. Modeling Northern and Southern varieties of Dutch for STT, in: Proc. Interspeech, Brighton, UK. pp. 96–99.

Fischer, V., Gao, Y., Janke, E., 1998. Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer, in: Proc. ICSLP, Sydney, Australia. pp. 787–790.

Fiscus, J., Garofolo, J., Przybocki, M., Fisher, W., Pallett, D., 1998. 1997 English broadcast news speech (HUB4). Linguistic Data Consortium, Philadelphia, PA.

Gales, M.J.F., Kim, D.Y., Woodland, P.C., Chan, H.Y., Mrva, D., Sinha, R., Tranter, S.E., 2006. Progress in the CU-HTK broadcast news transcription system. IEEE Trans. Acoust., Speech, Signal Process. 14, 1513–1525.

Gauvain, J.L., Adda, G., Adda-Decker, M., Allauzen, A., Gendner, V., Lamel, L., Schwenk, H., 2005. Where are we in transcribing French broadcast news?, in: Proc. Interspeech, Lisbon, Portugal. pp. 1665–1668.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2, 291–298.

Graff, D., Garofolo, J., Fiscus, J., Fisher, W., Pallett, D., 1997. 1996 English broadcast news speech (HUB4). Linguistic Data Consortium, Philadelphia, PA.

Hain, T., Burget, L., Dines, J., Garner, P.N., El Hannani, A., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., 2010. The AMIDA 2009 meeting transcription system, in: Proc. Interspeech, Makuhari, Japan. pp. 358–361.

Hecht, R., Riedler, J., Backfried, G., 2002. German broadcast news transcription, in: Proc. ICSLP, Denver, CO. pp. 1753–1756.

Imseng, D., Dines, J., Motlicek, P., Garner, P.N., Bourlard, H., 2012. Comparing different acoustic modeling techniques for multilingual boosting, in: Proc. Interspeech, Portland, OR. pp. 1910–1913.

Kamper, H., De Wet, F., Hain, T., Niesler, T.R., 2012a. Resource development and experiments in automatic South African broadcast news transcription, in: Proc. SLTU, Cape Town, South Africa. pp. 102–106.

Kamper, H., Muamba Mukanya, F.J., Niesler, T.R., 2012b. Multi-accent acoustic modelling of South African English. Speech Communication 54, 801–813.

Kirchhoff, K., Vergyri, D., 2005. Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. Speech Commun. 46, 37–51.

Le, V., Besacier, L., 2009. Automatic speech recognition for under-resourced languages: application to vietnamese language. IEEE Trans. Speech Audio Process. 17, 1471–1482.

Loots, L., Niesler, T.R., 2010. Automatic conversion between pronunciations of different english accents. Speech Commun. 53, 75–84.

MacIntyre, R., 1998. 1996 CSR HUB4 language model. Linguistic Data Consortium, Philadelphia, PA.

NIST, 2009. Speech Recognition Scoring Toolkit (SCTK). Online available at: http://www.nist.gov/speech/tools.

Qian, Y., Xu, J., Povey, D., Liu, J., 2011. Strategies for using MLP based features with limited target-language training data, in: Proc. ASRU, Waikoloa, HI. pp. 354–358.

Rabiner, L., Juang, B., 1993. Fundamentals of Speech Recognition. Prentice Hall, New Jersey, USA.

Schultz, T., Waibel, A., 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Commun. 35, 31–51.

Stolcke, A., 2002. SRILM – An extensible language modeling toolkit, in: Proc. ICSLP, Denver, CO. pp. 901–904.

Swietojanski, P., Ghoshal, A., Renals, S., 2012. Unsupervised cross-lingual knowledge transfer for DNN-based LVCSR, in: Proc. IEEE SLT, Miami, FL. pp. 419–422.

Van Leeuwen, D., Kessens, J., Sanders, E., Van Den Heuvel, H., 2009. Results of the N-Best 2008 Dutch Speech Recognition Evaluation, in: Proc. Interspeech, Brighton, UK. pp. 2571–2574.

Vu, N., Breiter, W., Metze, F., Schultz, T., 2012. An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on ASR performance, in: Proc. Interspeech, Portland, OR. pp. 2586–2589.

Vu, N., Kraus, F., Schultz, T., 2011. Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training, in: Proc. Interspeech, Florence, Italy. pp. 3145–3148.

Wan, V., Dines, J., El Hannani, A., Hain, T., 2008. Bob: A lexicon and pronunciation dictionary generator., in: Proc. SLT, Goa, India. pp. 217–220.

Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J., 1997. Broadcast news tanscription using HTK, in: Proc. ICASSP, Munich, Germany. pp. 719–722.

Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Liu, X., Moore, G.L., Odell, J.J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C., 2009. The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department.

Zhang, X., Demuynck, K., Van Compernolle, D., Van Hamme, H., 2012. Subspace-GMM acoustic models for under-resourced languages: feasibility study, in: Proc. SLTU, Cape Town, South Africa. pp. 1–4.