AUTOMATIC TRANSCRIPTION OF SOUTH AFRICAN BROADCAST NEWS

Herman Kamper[1], *Febe de Wet*[1,2], *Thomas Hain*[3], *Thomas Niesler*[1]

[1]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[2]Human Language Technology Competency Area, CSIR Meraka Institute, Pretoria, South Africa
[3]Department of Computer Science, University of Sheffield, UK

kamperh@sun.ac.za, fdw@sun.ac.za, t.hain@dcs.shef.ac.uk, trn@sun.ac.za

ABSTRACT

We present a description of the development and evaluation of a first South African broadcast news transcription system. We describe a number of speech resources which have been collected in the resource-scarce South African environment for system development purposes: a 20 hour corpus of South African English (SAE) broadcast news; a 109M word corpus of South African newspaper text collected for language modelling purposes; and a 60k word SAE pronunciation dictionary. The development of our system is based on similar state-of-the-art broadcast news transcription systems, and uses cross-word triphone HMMs, MF-PLP features and per-segment cepstral mean and variance normalisation. Our final system achieves a word error rate of 24.6%. We find that, for newsreader data, Indian and Black South African English accents are recognised more accurately than the speech by White English mother tongue speakers. However, for the spontaneous speech found in interviews and crossings to other locations, the latter accent is associated with the best results, although for this speech the error rates are high overall. Finally, we consider the recognition of MP3-compressed audio and show that performance only deteriorates at high compression levels.

Index Terms— Broadcast news transcription, South African English, under-resourced languages, English accents

1. INTRODUCTION

The transcription of broadcast news has a long history within the field of automatic speech recognition. Significant progress has been made in a variety of specialist areas ranging from the segmentation of the raw audio to acoustic and language modelling and adaptation [1, 2]. Although broadcast news transcription research initially focussed on North American English, work has since been extended to several other languages and accents, including British English [3], Italian [4], German [5], French [6] and Turkish [7]. Studies that deal specifically with different accents of a particular language in the broadcast news domain include [8], in which a cross-evaluation for systems trained on the Bavarian and standard dialects of German were considered, as well as [9] and [10], in which broadcast news systems for the Northern and Southern varieties of Dutch were compared.

The broadcast news domain provides both a ready source of speech audio data, as well as a variety of speech styles and quality, ranging from carefully produced newsreader speech to spontaneous interviews over noisy telephone channels. Furthermore, the broadcast news task allows useful benchmarking and comparisons between systems. Finally, broadcast news systems can form components for subsequent speech technologies such as information retrieval and pronunciation training systems.

We present a description and evaluation for a first speech recognition system dealing with specifically South African broadcast news (SABN). SABN is particular in several aspects. Most notably, the presence of several prevalent accents complicates system development (Sections 2 and 3). Furthermore, South African English (SAE) is considered an under-resourced variety of English because exceedingly little annotated speech data are available for the development of speech recognition systems.

2. ACCENTS OF ENGLISH IN SOUTH AFRICA

The South African constitution recognises eleven official languages. Of these languages, English is the lingua franca, as well as the language of government, commerce and science. Despite this, only 8.2% of the population speak it as a first language and hence English is used predominantly by non-mother-tongue speakers. This results in a large number of accents, which are reflected in our SABN corpus.

Five major varieties of SAE are identified in the literature [11]: Afrikaans English (AE), Black South African English (BE), Cape Flats English (CE), White South African English (EE), and Indian South African English (IE). While these labels are not intended to reflect Apartheid classifications, there is still an undeniable correlation between the different varieties of English used in South Africa and the various ethnic groups. Table 1 gives an indication of the proportion of the South African population speaking each of these accents. The table demonstrates clearly that non-mother-tongue variants of English (spoken by AE, BE and some CE speakers) are used by the overwhelming majority of the South African population.

Accent	Ethnic group and first language	Speakers
AE	White Afrikaans speakers	5.7%
BE	Black speakers of an official Black language	77.8%
CE	Coloured Afrikaans or English speakers	8.8%
EE	White English speakers	3.8%
IE	Indian or Asian English speakers	2.3%
-	Other	1.7%

Table 1. Percentage of the population falling into specific speaker groups, indicating the proportion of speakers of the five South African English accents [12].

3. SOUTH AFRICAN BROADCAST NEWS

The work presented here is based on a corpus of SABN which has recently been compiled at Stellenbosch University. The corpus consists of approximately 20 hours of audio recordings from one of the country's main radio news channels, SAFM. Bulletins were broadcast between 1996 and 2006 and are a mix of newsreader speech, interviews, and crossings to reporters. These varying channel conditions have been annotated for each utterance in the corpus as RD (newsreader), SI (studio speech), NST (non-studio telephone speech) or NS (wideband, non-studio speech). Table 2 summarises these classifications and also indicates the closest classic Hub-4 channel condition. Audio was sampled at 16 kHz and stored with 16-bit precision. The corpus was manually transcribed and speaker identity and accent were annotated for each utterance. Word fragments were annotated to indicate what was said as well as what the speaker intended to say. Silences, filled pauses and speaker noises were also labelled.

The data were divided into training and test sets as indicated in Tables 3 and 4, respectively. The first chronological 17.10 hours of data (extending up to March 2005) were used for training, and the last 2.65 hours (April 2005 to March 2006) for testing. A more thorough breakdown of the training set, indicating the amount of speech audio data separately for each accent and channel condition, is given in Table 5. There are 27 newsreaders in the training set of whom 8 are male and 19 are female. Of these speakers, 11 use the BE variety (34.7% of the RD speech data), 9 use EE (46.2%) and 7 use IE (19.1%). The EE accent is represented much more strongly among our newsreaders than the figure of 3.8% in Table 1 would suggest.

Of the 508 speakers in the training set involved in interviews or crossings, 392 are male and 116 are female. All five SAE accents are represented by these speakers. In contrast to the newsreader data, the most prevalent accent used during interviews and crossings is BE, constituting 29.5% of the SI, NST and NS speech data. The next most common accent is EE, which accounts for a further 15.7%. A number of foreign English accents are also present in this portion of the training set, most notably British English (UKE, 14.9%) and American English (USE, 10.4%). The test set is similar in composition to the training set.

SABN	Hub-4	Description
RD	F0	Newsreader speech.
SI	F1	Other studio speech. Fairly spontaneous, not read.
NST	F2	Telephone speech. Usually interviews with non-reporters. Highly spontaneous.
NS	F4	Wideband non-studio speech. Includes reporters on location, often in very unfavourable noise conditions. Fairly spontaneous, not read.

Table 2. Definition and description of the SABN audio channel conditions and corresponding classic Hub-4 labels.

	RD	SI	NST	NS	Total
Segments	7205	302	956	684	9147
Words	139241	6684	22419	16244	184588
Speakers	27	61	262	208	535
Speech (h)	12.90	0.60	2.07	1.54	17.10

Table 3. Composition of the SABN training set.

	RD	SI	NST	NS	Total
Segments	1000	60	223	129	1412
Words	18683	1247	4360	2574	26864
Speakers	11	11	56	33	107
Speech (h)	1.86	0.12	0.41	0.25	2.65

Table 4. Composition of the SABN test set.

4. SYSTEM DEVELOPMENT

4.1. Language modelling

A corpus of newspaper text was compiled 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, a language model training set consisting of approximately 109M words and including material from January 2000 to March 2005 was compiled. Using the SRILM toolkit [13], a trigram language model was trained on this dataset. Additionally, a trigram language model was trained on the acoustic training set transcriptions (185k words, Table 3). The two were subsequently linearly interpolated to yield the language model used for the experiments described in the remainder of this paper. All language models used the same 60k vocabulary (described in Section 4.2) as well as Kneser-Ney smoothing within a Katz backoff [14] structure. The perplexities achieved by the three language models on the acoustic test set transcriptions (Table 4) are indicated in Table 6. These results show that the improvement in perplexity resulting from the interpolation of the two component language models is substantial.

Accent	RD	SI	NST	NS	Total
AE	-	1.3	12.2	3.1	16.5
BE	268.3	2.0	40.5	32.1	342.9
CE	-	1.4	16.2	6.7	24.3
EE	357.9	3.7	31.3	4.7	397.6
IE	147.7	6.3	8.5	4.1	166.5
UKE	-	12.2	5.4	20.0	37.7
USE	-	8.2	6.8	11.3	26.3
OE	-	0.7	3.3	10.5	14.5
Total	773.9	35.8	124.1	92.4	1026.2

Table 5. Amount of audio speech data (in minutes) shown separately for each accent and channel condition in the SABN training set. Dashes indicate the absence of training data. UKE and USE refer to British and American English Accents respectively, while OE indicates all other English accents.

Language model	Perplexity
Trained on 109M word newspaper text corpus	162.9
Trained on acoustic training set transcriptions	328.9
Linear interpolation of the above two models	139.9

Table 6. Language model perplexities measured on the acoustic test set transcriptions.

4.2. Pronunciation dictionary

A training pronunciation dictionary for the 14 622 unique words in the acoustic training data (Table 3) was developed by a phonetic expert. Subsequently, pronunciations for the most frequent words in the language model training data (Section 4.1) were determined by the same phonetic expert to obtain a recognition dictionary with 60 698 words and on average 1.25 pronunciations per word. The majority of dictionary entries reflect typical EE pronunciations and were recorded using an IPA-based phoneset developed to describe the languages of Southern Africa [15]. These pronunciations were subsequently converted to use 45 ARPABET phones by means of a mapping based on the closest IPA symbol. The 60k words in our dictionary achieve an out-of-vocabulary rate of 1.02% on the acoustic test set transcriptions.

4.3. Acoustic modelling

Decision-tree state-clustered cross-word triphone HMMs with a three-state left-to-right model topology and 16 mixtures per state were employed as speech models in our SABN system. Decoding experiments made use of the HTK HDecode decoder using the first-best output [16]. All word error rates (WERs) were computed using the NIST Scoring Toolkit (SCTK) [17].

Initially, training and test audio data were parametrised as a stream of 39 dimensional mel-frequency cepstral coefficients (MFCCs). A first alignment of the parametrised training set was performed with broadcast news models trained on approximately 100 hours of data collected in North America (1996 and 1997 Hub-4 data). Starting from this first alignment, initial triphone HMMs (shown at the top of Figure 1) were trained on the SABN training set using a standard HTK model training strategy and configuration.

Using these initial HMMs, we set out to compare MFCC features to the alternative mel-frequency perceptual linear prediction (MF-PLP) parametrisation, which has been applied in other broadcast news transcription systems [1]. After parametrising the training set as a stream of 39 dimensional MF-PLP feature vectors, single-pass retraining[1] [16] was performed using the initial triphone HMMs to obtain MF-PLP HMMs, as illustrated in Figure 1. WERs for the two comparable systems, respectively employing MFCC and MF-PLP parametrisations, are given in Table 7. When compared with its MFCC counterpart, the system employing MF-PLP parametrisation yields an absolute improvement of 1.2% in WER. MF-PLPs were therefore used in all subsequent experiments.

[1] In this study, each single-pass retraining step was followed with re-alignment and a single iteration of two-model re-estimation [16]. This was done to ensure a fair comparison of the different features.

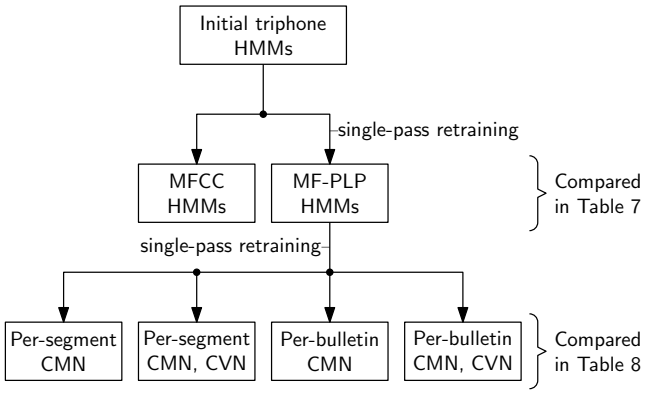

Fig. 1. The acoustic model development procedure, indicating how different features and feature normalisation procedures were compared.

Features	RD	SI	NST	NS	Overall
MFCC	14.7	27.6	68.4	66.1	28.9
MF-PLP	14.4	25.9	65.8	61.1	27.7

Table 7. WERs (%) for systems employing MFCC and MF-PLP parametrisation, respectively.

Normalisation	RD	SI	NST	NS	Overall
Per-segment CMN	13.8	20.6	58.2	53.1	25.1
Per-seg. CMN & CVN	13.6	19.5	57.3	52.0	**24.6**
Per-bulletin CMN	13.8	21.6	65.1	60.2	26.9
Per-bul. CMN & CVN	13.4	21.6	64.0	59.9	26.4

Table 8. WERs (%) for systems employing different cepstral feature normalisation approaches.

The next step was to determine the best feature normalisation approach. In a procedure similar to that described in the preceding paragraph, we performed single-pass retraining, this time using the MF-PLP HMMs, in order to obtain models using feature vectors for which:

1. cepstral mean normalisation (CMN) was performed on a per-segment basis,
2. CMN was performed on a per-segment basis together with cepstral variance normalisation (CVN),
3. CMN was applied on a per-bulletin basis, and
4. CMN and CVN were applied together, both on a per-bulletin basis.

The four resulting systems are shown at the bottom of Figure 1 and their performance is compared in Table 8. In all cases, CVN was applied on a per-bulletin level. A comparison of Tables 7 and 8 reveals that the improvement due to feature normalisation is substantial. In particular, by comparing the upper and lower halves of Table 8, it is evident that performing CMN on a per-segment level is superior to performing CMN on a per-bulletin level. Furthermore, the improvement afforded by CVN is observed by comparing the first to the second system in Table 8, as well as comparing the third to the fourth system.

5. EXPERIMENTAL RESULTS

5.1. System performance

The final SABN system employs an acoustic model set consisting of 2624 states and uses MF-PLP features calculated with per-segment CMN and CVN. This system achieved a WER of 24.6%, as indicated in Table 8. In order to gain more insight into the impact of the speaker accent on this overall performance, the WER for this system was also calculated on a per-accent basis, as presented in Table 9.

Table 9 indicates that the error rates for newsreader (RD) and studio (SI) speech are reasonable, while for the non-studio channel conditions (NST and NS) error rates are very high. This is not clearly related to the accent as all five South African accents since AE, BE,

Accent	RD	SI	NST	NS	Overall
AE	-	-	60.7	67.0	63.3
BE	13.7	19.6	64.3	56.9	29.4
CE	-	-	61.7	-	61.7
EE	14.1	-	54.1	41.6	17.2
IE	12.7	-	59.2	-	16.6
UKE	-	17.7	22.7	32.2	23.8
USE	-	39.3	-	50.5	48.0
Other	-	-	63.0	66.7	65.3
Overall	13.6	19.5	57.3	52.0	24.6

Table 9. WERs (%) measured separately for each accent and channel condition. Dashes indicate the absence of test data.

MP3 bitrate	RD	SI	NST	NS	Overall
128 kbps	13.6	18.9	57.0	51.9	24.6
64 kbps	13.4	18.8	57.8	52.3	24.6
32 kbps	14.3	20.8	58.7	50.7	25.3

Table 10. System performance in terms of WER (%) when decoding MP3 audio compressed at various bitrates.

CE, EE and IE show similarly poor performance. This may be due to the small amount of data available for these channel conditions, as indicated in Tables 3 and 5. Nevertheless, even for the BE accent, for which more NST and NS data are available, performance is very poor (64.3% and 56.9% WER for NST and NS respectively). Interestingly, for the NST channel condition the WER is lowest for UKE. This is attributed to the informal observation that many of the utterances in this category consist of fairly well-prepared speech, such as statements by international politicians.

A further interesting observation is that the error rates for the newsreader (RD) speech indicate that IE is the easiest of the accents to recognise, followed by (very surprisingly) BE and then EE. In contrast, for the NST and NS channel conditions, the WERs for EE are lower than those for BE and IE. We speculate that these differences are due to a tendency for BE and IE newsreaders to speak particularly carefully when presenting prepared speech.

5.2. MP3 audio compression

Online resources are an invaluable source of speech audio data which can be especially important for under-resourced languages. Many online speech audio resources are, however, only available in a compressed format. An experimental evaluation was therefore performed to determine the effect of such compression on our final SABN system's performance. To achieve this, the original test set audio data (Table 4) were converted to the popular MP3 format at various bitrates. System performance for these different degrees of compression is shown in Table 10. Only at very high compression levels (32 kbps) can a deterioration relative to baseline performance be observed.

6. SUMMARY AND CONCLUSIONS

We have described the development of a first broadcast news transcription system for South African English (SAE). This included the compilation of audio as well as language model training data. It also involved the development of a suitable pronunciation dictionary for SAE. The accent of each speaker was recorded, and the channel conditions for each utterance marked using a classification similar to that employed in Hub-4 systems. Acoustic material consisted of 17.10 hours of training and 2.65 hours of test material. Trigram language models were trained on approximately 109M words of newspaper text, as well as on the acoustic training set transcriptions. The pronunciation dictionary contained pronunciations for 60k words. We compared MFCC and MF-PLP parametrisation and found that MF-PLP was superior by approximately 1.2% in word error rate (WER). We also compared different feature normalisation approaches and found that per-segment cepstral mean normalisation together with per-segment cepstral variance normalisation resulted in best performance. Our best system achieved an overall WER of 24.6%, despite very poor performance on spontaneous and telephone speech. Finally, we demonstrated that for MP3-compressed audio our system maintains best performance except at bitrates below 64 kbps.

The presence of several accents in our corpus presents the opportunity for future investigations into accent-robust and multi-accent South African English automatic broadcast news transcription. At present we are using a single pronunciation dictionary. Future systems may incorporate more than one dictionary as well as more sophisticated acoustic models. We also plan to contrast the performance our our South African broadcast news system with similar British (UK) and American (US) English systems, and investigate any differences we may find. In particular we would like to identify how resources from the well-resourced UK and US varieties of English can be used in the development of systems for the poorly-resourced South African variety and what penalties are incurred.

7. 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.

8. REFERENCES

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

[2] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 14, no. 5, pp. 1513–1525, 2006.

[3] D. Abberley, S. Renals, and G. Cook, "Retrieval of broadcast news documents with the THISL system," in *Proc. ICASSP*, Seattle, WA, 1998, pp. 3781–3784.

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

[5] K. McTait and M. Adda-Decker, "The 300k LIMSI German broadcast news transcription system," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 213–216.

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

[7] E. Arısoy, H. Sak, and M. Saraclar, "Language modeling for automatic Turkish broadcast news transcription," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2381–2384.

[8] R. Hecht, J. Riedler, and G. Backfried, "German broadcast news transcription," in *Proc. ICSLP*, Denver, CO, 2002.

[9] D. van Leeuwen, J. Kessens, E. Sanders, and H. van den Heuvel, "Results of the N-Best 2008 Dutch Speech Recognition Evaluation," in *Proc. Interspeech*, Brighton, UK, 2009.

[10] J. Despres, P. Fousek, J. L. Gauvain, S. Gay, Y. Josse, L. Lamel, and A. Messaoudi, "Modeling Northern and Southern varieties of Dutch for STT," in *Proc. Interspeech*, Brighton, UK, 2009.

[11] E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds., *A Handbook of Varieties of English*, Mouton de Gruyter, Berlin, Germany, 2004.

[12] Statistics South Africa, "Census 2001: Primary tables South Africa: Census 1996 and 2001 compared," 2004.

[13] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. ICSLP*, Denver, CO, 2002, pp. 901–904.

[14] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Comput. Speech Lang.*, vol. 13, pp. 359–394, 1999.

[15] T. R. Niesler, P. Louw, and J. Roux, "Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases," *South Afr. Ling. Appl. Lang. Stud.*, vol. 23, no. 4, pp. 459–474, 2005.

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

[17] National Institute of Standards and Technology (NIST), "Speech Recognition Scoring Toolkit (SCTK)," Online available at: http://www.nist.gov/speech/tools/.