Baseline Speech Recognition of South African Languages using Lwazi and AST

Daan Henselmans and Thomas Niesler
Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
drhenselmans@gmail.com, trn@sun.ac.za

David van Leeuwen
Centre of Language and Speech Technology
Radboud University Nijmegen, the Netherlands
d.vanleeuwen@let.ru.nl

***Abstract*—This paper presents baseline speech recognition results using the Lwazi and AST corpora. Phoneme and word recognition are performed in all eleven South African languages. For four languages, the AST and Lwazi data were merged together to create more elaborate acoustic models. Phoneme recognition results were found to be similar to previously published figures. Word recognition results were similar across all languages, and relatively poor due to small language model training set sizes. The addition of AST data was shown to lead to limited improvements in both phoneme and word error rates.**

I. Introduction

Telephone-based language technology can be used as an aid in building up a technological infrastructure in developing countries, where other information sources are often scarce. Until very recently, almost no South African data was available for the development of automatic speech recognition (ASR) and associated systems. The Lwazi corpus was developed to change this [1], [2]. Prior to the Lwazi corpus, the AST corpora were developed for the same goal, although they are comprised of just five South African languages [3], [4].

This paper aims to further explore the possibilities of both phoneme and word recognition for all South African languages. The Lwazi corpus has been used for phoneme recognition before, but its potential for word recognition has only been investigated in a small-vocabulary task [2]. Although Lwazi and AST are similar in goal and scope, the possibility of combining the two corpora to achieve better ASR results has never been investigated. We present a set of baseline speech recognition experiments based on the Lwazi corpus, and similar experiments based on a merged corpus composed of both Lwazi and AST data.

II. The Lwazi Corpus

The Lwazi ASR corpus was developed between 2006 and 2009, as part of a government project which aimed to demonstrate the benefits of speech technology in South Africa. The corpus was compiled to investigate the feasibility of speaker-independent speech recognition using limited resources, and was shown to achieve adequate performance in phoneme recognition and small vocabulary word-recognition tasks [2]. The corpus consists of speech data for each of the 11 official South African languages, annotated orthographically. It is accompanied by baseline dictionaries obtained by expanding the original Lwazi dictionaries [5] using grapheme-to-phoneme prediction [6]. The Lwazi corpus is freely available under an open content license [7].

TABLE I
Phone sets and vocabulary sizes for the eleven Lwazi datasets (excluding silence and noise markers).

Code	Language	Language Family	Phone types	Word types
afr	Afrikaans	West Germanic	37	1,585
eng	SA English	West Germanic	44	2,112
nbl	isiNdebele	Nguni	47	4,751
ssw	siSwati	Nguni	41	5,092
xho	isiXhosa	Nguni	50	4,727
zul	isiZulu	Nguni	45	5,376
nso	Sepedi	Sotho-Tswana	30	3,276
sot	Sesotho	Sotho-Tswana	29	2,568
tsn	Setswana	Sotho-Tswana	34	2,980
tso	Xitsonga	Tswa-Ronga	52	2,747
ven	Tshivenda	Venda	40	2,441
Overall			110	31,757

For each language, 200 first language speakers were recorded over a telephone channel, each providing approximately 30 utterances. The utterances include phrases randomly selected from phonetically balanced corpora developed specifically for this task, as well as short words and phrases, such as the answers to open and yes/no questions, spelt words, dates, and numbers [2]. The orthographic annotations include markers for speaker and background noise, as well as for partial words. The number of words and phones in the various data sets are shown in Table I. Aside from the types in this table, each of the languages included one phone and one word marker to denote silence, one phone and one word marker to denote speaker noise, and one word marker to denote fillers.

TABLE II
DESCRIPTION OF THE LWAZI TRAINING SETS.

Lang.	No. of phone tokens	No. of word tokens	No. of speakers	No. of utts	Dur. (mins)
afr	98,427	24,754	140	4,186	179
eng	112,388	28,302	136	4,048	207
nbl	193,987	29,229	140	4,216	432
ssw	180,143	27,456	136	4,050	442
xho	166,502	27,711	150	4,474	406
zul	151,270	24,363	139	4,018	360
nso	151,680	38,682	130	3,854	385
sot	136,536	34,007	142	4,223	300
tsn	131,737	35,694	143	4,199	339
tso	146,736	35,615	154	4,624	378
ven	119,140	28,817	138	4,142	305
Total	1,588,546	334,630	1,548	46,034	3,733

TABLE III
DESCRIPTION OF THE LWAZI DEVELOPMENT SETS.

Lang.	No. of phone tokens	No. of word tokens	No. of speakers	No. of utts	Dur. (mins)
afr	14,080	3,574	20	598	26
eng	16,757	4,224	20	598	31
nbl	27,327	4,097	20	603	62
ssw	26,477	4,035	20	593	64
xho	21,162	3,521	20	586	51
zul	22,603	3,597	20	582	53
nso	23,379	5,948	20	596	58
sot	20,082	5,003	20	605	44
tsn	17,739	4,799	20	574	43
tso	19,032	4,607	20	601	51
ven	17,383	4,143	20	596	42
Total	226,021	47,548	220	6,532	525

TABLE IV
DESCRIPTION OF THE LWAZI EVALUATION SETS.

Lang.	No. of phone tokens	No. of word tokens	No. of speakers	No. of utts	Dur. (mins)
afr	28,724	7,165	40	1,199	51
eng	33,061	8,373	40	1,197	63
nbl	54,428	8,185	40	1,194	121
ssw	53,154	8,010	40	1,195	128
xho	45,275	7,408	40	1,182	105
zul	46,682	7,568	40	1,185	115
nso	47,657	1,1893	40	1,190	127
sot	38,280	9,628	40	1,199	85
tsn	36,993	9,891	40	1,197	94
tso	37,560	9,067	40	1,201	92
ven	34,140	8,143	40	1,201	87
Total	455,954	95,331	440	13,140	1,068

TABLE V
WORDS AND PHONES IN THE LANGUAGE MODEL TRAINING SETS, AFTER REMOVAL OF UTERRANCES OVERLAPPING WITH THE DEVELOPMENT OR EVALUATION SETS.

Lang.	No. of phone tokens	No. of word tokens	% word tokens remain-ing	No of word types	% eval words not in reduced train set
afr	16,575	4,735	19.1%	486	76.6%
eng	17,989	5,279	18.7%	604	74.2%
nbl	59,118	9,882	33.8%	2,124	61.4%
ssw	52,572	8,921	32.5%	2,324	60.8%
xho	69,063	12,136	43.8%	3,097	42.6%
zul	67,166	11,399	46.8%	3,050	51.6%
nso	39,039	9,808	25.4%	1,483	57.2%
sot	39,788	10,224	30.1%	1,309	52.8%
tsn	44,769	12,229	34.3%	1,629	48.8%
tso	40,693	10,110	28.4%	1,292	56.9%
ven	33,551	8,046	27.9%	1,193	56.5%
Total	480,323	102769	30.7%	26,221	

A. Training and test sets

In order to develop speech recognition systems, the data for each language was split into training, development and evaluation sets. The evaluation sets employed in this work coincide with the fixed evaluation subsets for the Lwazi corpus, which were obtained by the selection of 20 male and 20 female speakers per language [6]. The development sets consisted of a further 20 randomly chosen speakers. The number of phone and word types and tokens in each of these sets is shown in Tables II, III, and IV.

B. Removing phonetically rich utterances

The Lwazi corpus includes a large number of phonetically rich utterances. Many of those are repeated up to 15 times throughout the training and test sets. To avoid positive bias in our recognition results, all utterances which occurred in both the training and test sets were removed from the training set for language modelling purposes. Since the prompts for the Lwazi utterances are not publically available, such recurrences were detected by exhaustive pairwise string alignment between the training and test set transcriptions. All training utterances which were at least six words long and resulted in a match accuracy of at least 60% with any of the test sentences were removed. Training utterances which were less than six words long were removed only when they were identical to test utterances, since a match of three words did not necessarily imply a phonetically rich utterance in these cases.

This procedure significantly reduced the size of the training set available for language modelling, as shown in Table V. The extent of its impact varies between the languages. The largest overlap between test and training utterances was detected for Afrikaans and English. Since these languages already had the smallest training sets in terms of phone tokens, the resulting differences in the language model training sets are large. The contrast is stark in comparison with the Nguni languages, which had the largest training sets, but suffered the smallest reductions after removal of overlapping utterances. These smaller training sets were used only to train language models; acoustic model training continued to use the full training set. There is a possibility that this will have an effect on the reported error rates, but the degree of overlap is too large for acoustic modelling to be viable on the reduced training set.

Aside from simply reducing the amount of data available to train the language models, many words which occurred in the full training set are no longer present in the reduced set at all. The scale at which the vocabulary is reduced ranges between 55 and 80% depending on the language, as shown in Table V. In the cases where many words are removed, this would lead to many out-of-vocabulary words during recognition. To avoid word recognition results being dominated by OOV errors, the language models used closed vocabularies.

III. Acoustic models

A set of HMM acoustic monophone models was trained for each of the 11 languages using the full training sets. Embedded Baum-Welsh re-estimation was used to obtain speaker-independent diagonal-covariance monophone models with three states per model and one Gaussian mixture per state. Each speech frame was parameterised as a 39-dimensional feature vector consisting of 12 Mel-frequency ceptral coefficients (MFCCs) and their first and second differentials. The models were normalised using cepstral mean normalization.

Model development proceeded from a flat start using the first pronunciation of each training word found in the dictionary. Once a first set of acoustic models had been obtained, the training set was re-aligned using a Viterbi word recogniser to allow pronunciation variants to be accounted for. Such realignment followed by re-training was performed 8 times, after which the phone-level training transcriptions and monophone recognition results on the development set were seen to stabilise.

Cross-word triphone models were obtained by decision tree state clustering [8]. The AST decision tree questions were used for this purpose [3]. The noise and silence models were used as context phonemes, but were not expanded to triphones themselves. The results of state clustering on the acoustic models is summarised in Table VI, which shows the number of possible triphones in each language, the number of observed triphones in the training set, and the number of distinct states remaining after clustering.

TABLE VI
Description of triphone acoustic models after decision-tree state-clustering

Language	No. possible triphones	No. distinct triphones	No. clustered states
afr	56,279	5,276	908
eng	93,106	7,447	1,013
nbl	12,849	5,371	1,419
ssw	75,811	5,597	1,090
xho	135,202	5,830	979
zul	99,407	5,980	1,397
nso	30,722	4,395	1,049
sot	27,871	3,812	1,285
tsn	44,066	4,827	1,117
tso	151,634	5,583	1,376
ven	70,562	4,669	1,190

The triphone models were then improved by gradually increasing the number of Gaussian mixtures. After each increase, four iterations of embedded re-estimation were performed. The reported results are for models with 8 mixtures per state, which were found to be optimal on the development set.

For speech recognition, the HTK hidden Markov model (HMM) based speech recogniser was used to perform a time-synchronous beam search using the Token-Passing procedure [9]. Optimal values for the word insertion penalty and language model scaling factor were determined by optimising the results for Afrikaans, Sesotho and isiZulu on the development sets, as representatives of the major language families. Each of these languages was shown to have the same optimal parameters. This optimization was performed for phoneme and word recognition independently. For the word insertion penalty, the optimum was -10. For the language model scaling factor, the optimal values were 6 and 10 for phoneme and word recognition respectively.

IV. Language models

Unigram language models were obtained using phone and word transcriptions of the reduced training set for each language (see Section II-B). Backoff bigram language models [10] were obtained using the same training sets, with the language model probabilities calculated using an absolute discount of 1.0 [11]. Language model perplexities were calculated on the evaluation sets for both unigram and bigram models. Perplexity can be considered an indication of the difficulty of predicting the next phoneme or word in a sequence.

A. *Phoneme language models*

The unigram phoneme language models give an indication of the diversity of the phone sets of the different languages. Evaluation set perplexities for phoneme unigram and bigram models are shown in Table VII. There are consistently 4 more unigrams than there are phonemes in the lexicon, since sentence boundary markers, noise, and silence are included.

The West Germanic languages, Afrikaans and English, show a higher bigram perplexity than the Southern Bantu languages. This implies that phoneme sequences in Southern Bantu languages are more predictable. A likely explanation for this is that Bantu languages conventionally use open syllables, making their phoneme sequences more predictable than those in Afrikaans and English. The bigram phoneme perplexities are somewhat higher than those observed in earlier analysis of the Lwazi corpus [2], which is likely due to the different language model training sets used.

TABLE VII
PHONEME UNIGRAM AND BIGRAM LANGUAGE MODELS AND THEIR EVALUATION SET PERPLEXITIES.

Language	No. of unigrams	Unigram perplexity (eval)	No. of bigrams	Bigram perplexity (eval)
afr	41	28.67	654	21.55
eng	48	34.92	961	24.15
nbl	51	22.99	698	12.29
ssw	45	22.66	733	12.20
xho	54	24.08	846	12.00
zul	49	24.27	772	12.42
nso	34	20.42	544	10.92
sot	33	20.62	519	10.69
tsn	38	21.48	645	11.95
tso	56	25.35	772	12.33
ven	44	23.36	609	12.30

TABLE VIII
WORD UNIGRAM AND BIGRAM LANGUAGE MODELS AND THEIR EVALUATION SET PERPLEXITIES.

Language	No. of unigrams	Unigram perplexity (eval)	No. of bigrams	Bigram perplexity (eval)
afr	1,589	509.31	1,086	491.11
eng	2,116	590.34	1,311	525.96
nbl	4,755	1,234.27	2,515	791.14
ssw	5,096	1,005.18	2,403	689.44
xho	4,731	1,067.72	3,493	499.85
zul	5,380	1,058.82	3,258	634.02
nso	3,280	366.96	2,915	256.31
sot	2,572	304.83	2,688	177.16
tsn	2,984	347.36	3,543	211.74
tso	2,751	386.50	2,677	238.18
ven	2,445	451.46	2,422	327.46

B. Word language models

The perplexities of the word unigram and bigram models are shown in Table VIII. The unigram perplexities are highest for the closely related Nguni languages (isiZulu, isiXhosa, isiNdebele and Siswati), despite the large size of their language model training sets. The larger vocabulary size is very likely the reason for these higher perplexities. Afrikaans and English also show high unigram perplexities, but this is most likely due to the smaller size of their training sets.

These trends continue in the bigram perplexities, but not as strongly. The perplexities of the Afrikaans and English word language models do not decrease much, which would support the notion that the high perplexities are due to their small training set. The Nguni language perplexities decrease much more, especially isiXhosa, which is now average compared to the other languages. The other languages have their language model perplexities reduced by varying degrees.

TABLE IX
PHONEME RECOGNITION ERROR RATES FOR EACH LANGUAGE. THE CURRENT RECOGNITION RESULTS ARE COMPARED TO THE PRIOR LWAZI RESULTS IN [2]

Language	Unigram PER	Bigram PER	2009 PER
afr	45.06	43.29	36.86
eng	52.66	50.54	45.74
nbl	38.17	35.04	34.59
ssw	36.92	34.06	35.54
xho	38.75	34.70	42.76
zul	44.38	41.47	39.05
nso	36.33	33.47	44.81
sot	38.54	34.86	45.21
tsn	39.70	36.88	43.81
tso	39.06	34.64	40.59
ven	37.10	34.00	33.22
Average	40.61	37.54	40.20

V. RESULTS USING LWAZI

Since the Lwazi corpus includes only orthographic transcriptions, phonetic transcriptions were generated using iterative re-alignment with the dictionary. Hence, the most appropriate measure of performance for the acoustic models is the error rate of a word-based speech recognition system. However, in the context of low-resource languages, word recognition has limited diagnostic value, due to the small and constrained nature of the acoustic and language model training corpora. Hence, the results for both phoneme recognition and word recognition are reported. It should be borne in mind, however, that the former are not based on manual phonetic transcripts.

A. Phoneme recognition

Phoneme recognition error rates are shown in Table IX. In all cases, bigram language models give better results than unigram models, as would be expected. However, the difference is not particularly large for the West Germanic languages, which had phoneme language models with higher bigram perplexities and smaller language model training sets.

The Germanic phoneme recognisers exhibit higher error rates, corresponding to the higher perplexities of their language models. The remaining languages show similar performance, although isiZulu fares noticably worse than other South African languages, despite its relatively low perplexity and comparatively large training set. This is unexpected, considering the close similarity of isiZulu and isiXhosa, which would lead one to expect comparable results.

Prior research on phoneme recognition using the Lwazi corpus is available [2]. Compared to our current results, this 2009 system performs notably better on Afrikaans and English. This may be due to differences in the language model training sets (see Section II-A). The current Southern Bantu recognisers generally have similar or better results than their 2009 counterparts, with the exception of isiZulu, which

TABLE X
WORD RECOGNITION ERROR RATES FOR EACH LANGUAGE.

Language	Unigram WER	Bigram WER
afr	53.52	52.52
eng	60.85	58.88
nbl	48.48	48.15
ssw	53.72	53.60
xho	58.77	55.02
zul	65.25	62.99
nso	61.09	57.47
sot	59.47	52.71
tsn	60.42	55.32
tso	56.24	50.74
ven	58.81	56.70
Average	57.87	54.92

performs slightly worse. On average, the phoneme recognition error rates reported here are comparable to those in [2].

A comparison with the phoneme error rates obtained for the similar AST corpora [3] is difficult due to the different phone sets employed. Section VI will consider this further.

B. Word recognition

Word recognition error rates are shown in Table X. Since removing repeated phrases from the training sets greatly reduced their sizes, word recognition error rates are high, as would be expected due to their high perplexities. Using bigram language models is better in all languages, although the improvements over unigram results are inconsistent.

IsiZulu, which has the highest word language model perplexity, has the highest error rates. Curiously, its close sister isiXhosa fares noticeably better. Afrikaans performs better than average, despite its larger perplexity and the smaller size of its training set. We note that there are no clear trends based on language family, as were observed in phoneme recognition. On average, however, the error rates are similar across all languages.

Recognition performance may improve if a cross-validation framework was employed to preserve language model training data while also avoiding the test/train overlap. There are no prior benchmarks available on word recognition using the Lwazi or AST corpora.

VI. MERGING LWAZI AND AST

For five South African languages, namely isiZulu, isiXhosa, Afrikaans, English, and Sesotho, corpora are also available in the previously developed AST databases [4]. These corpora contain more data for these languages than their respective counterparts in the Lwazi corpus. Considering the minimal size of the Lwazi corpus, this additional data should provide a means to improve results. English Lwazi and AST data have previously been successfully combined to improve acoustic models for a call routing system [12]. This lead us to believe merging AST training data with that of their Lwazi counterpart could be beneficial to recognition results.

TABLE XI
DESCRIPTION OF THE TRAINING SET OBTAINED BY MERGING THE AST AND LWAZI TRAINING SETS.

Language	No. of AST phone types	No. of merged phone types	No. of merged word tokens	No. of merged phone tokens	Merged dur. (mins)
afr	81	37	72,137	268,872	550
eng	50	44	76,240	280,619	566
xho	101	50	64,381	343,851	825
zul	94	45	72,894	433,847	1,012
Overall	125	82	285,652	1,327,189	2,955

TABLE XII
PHONEME AND WORD RECOGNITION ERROR RATES FOR SYSTEMS WITH MERGED AST AND LWAZI DATA.

Language	Unigram PER	Bigram PER	Unigram WER	Bigram PER
afr	43.30	41.80	52.25	51.40
eng	52.12	49.67	59.32	57.98
xho	37.94	33.48	54.39	50.23
zul	44.48	41.37	67.93	65.70
Average	16.17	15.12	21.26	20.48

Table XI describes the effects of merging the Lwazi and AST corpora for four different AST languages. The two corpora have different annotation styles, and different phone sets are used for the same language. In order to merge training data, the AST phonemes were mapped to their closest Lwazi counterparts, which significantly reduced the size of the phone set for Afrikaans, isiXhosa and isiZulu. The Lwazi dictionaries were extended with entries from the AST dictionaries for words which did not occur in the Lwazi corpus. The resulting training sets are two to three times larger than the Lwazi training sets in terms of phone and word tokens and duration.

A. Results using Lwazi and AST

A new set of acoustic models was trained on the resulting extended training sets and dictionaries, using the methods described in Section III. Phoneme and word recognition were performed using the same language models and parameters used in Section V. The results for these experiments are shown in Table XII.

Adding the AST data proved to be slightly beneficial for phoneme recognition, with improvements for all four languages regardless of the type of language model. However, the improvements were small despite the large increase in training set size. This could be ascribed to either differences in acoustic conditions, to the difference in annotation style between the two corpora, or to a combination of both. Furthermore, the language models are still trained on very small amounts of Lwazi data.

For word recognition, the results are mixed. While performance is slightly better for Afrikaans, English, and isiXhosa, the results are worse for isiZulu. Again, we ascribe this to the differing transcription styles of the corpora.

We did attempt to supplement the Lwazi language model training data with AST data. This did not lead to improved results, however.

VII. Conclusion

We have presented a set of baseline phoneme and word speech recognition results for the Lwazi corpus. We have also presented results for a system combining Lwazi with AST data. The results for phoneme recognition obtained using only the Lwazi data are similar to those reported earlier by other researchers, although our language model training sets were much smaller. The word recognition results are poor, although it should be borne in mind that in this case, the language model training sets were especially limited and that there is no prior benchmark using Lwazi data to perform this task. To improve results in the future, more language model training data would be beneficial.

Merging the AST and Lwazi data proved beneficial for phoneme recognition, but less so for word recognition. Differences in annotation style between AST and Lwazi necessitated the mapping of the larger AST onto the smaller Lwazi phone set for acoustic model training. To better merge the two corpora it would be advantageous to obtain a pronunciation dictionary using a single uniform style for all languages.

Acknowledgements

The authors would like to thank Febe de Wet for assistance with the Lwazi utterance groups and dictionaries and structuring the Lwazi training and test sets. This research has been carried out in the project *Elftal* which received financial support from the Dutch Language Union and the South African Department of Arts and Culture.

References

[1] J. Badenhorst, C. van Heerden, M. Davel, and E. Barnard, "Collecting and evaluating speech recognition corpora for nine southern bantu languages," in *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages*, 2009, pp. 1–8.

[2] C. van Heerden, E. Barnard, and M. Davel, "Basic speech recognition for spoken dialogues," in *Proceedings of INTERSPEECH 2009*, Brighton, United Kingdom, 2009.

[3] T. Niesler and P. Louw, "Comparative phonetic analysis and phoneme recognition for Afrikaans, English and Xhosa using the African Speech Technology telephone speech databases," *The South African Computer Journal*, vol. 32, pp. 3–12, 2004.

[4] J. C. Roux, P. H. Louw, and T. R. Niesler, "The african speech technology project: An assessment," in *LREC, European Language Resources Association*, 2004.

[5] M. H. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proceedings of INTERSPEECH 2009*, Brighton, UK, 2009, pp. 2851–2854.

[6] C. van Heerden, J. Badenhorst, F. de Wet, and M. Davel, "Lwazi asr evaluation," CSIR, Tech. Rep., May 2013.

[7] Meraka-Institute. (2009) Lwazi ASR corpus. [Online]. Available: http://www.meraka.org.za/lwazi

[8] P. Woodland, J. Odell, V. Valtchev, and S. Young, "Large vocabulary continuous speech recognition using HTK," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, 1994, pp. 125–128.

[9] S. Young, "Token passing, a simple conceptual model for connected speech recognition systems," Cambridge University, Tech. Rep. TR38, 1989.

[10] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recogniser," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35(3), pp. 400–401, 1987.

[11] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modelling," *Computer, Speech and Language*, vol. 8, pp. 1–38, 1994.

[12] N. Kleynhans, R. Molapo, and F. de Wet, "Acoustic model optimisation for a call routing system," in *Proc. Pattern Recognition Association of South Africa (PRASA 2012)*, 2012.