Phoneme- and Word-based Language Identification of South African Languages using Lwazi

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*—We present baseline language identification results for South African languages using the Lwazi corpus. Four different strategies to perform 11-way classification are implemented: phoneme recognition followed by language modelling (PRLM), parallel PRLM, and two analogous approaches based on word-level recognition. The optimal system uses a parallel word recognition followed by language modelling approach with trigram language models, and obtained an ID error rate of 0.418, a decision cost of 0.182, and a confusion of 2.92. West Germanic languages were easiest to identify.**

I. Introduction

Automatic language identification (LID) is a growing application of speech processing technology. LID has many practical uses, which range from pre-selecting a speech recognition system for a certain language to routing phone calls to human emergency line operators depending on the language of the speaker. In South Africa, this is especially relevant, since it is common for different languages to be spoken in the same settings and locations. Accurate language identification would therefore be particularly helpful in making many resources and services more easily and publically available.

Until recently, almost no South African data was freely available for the development of speech technology. The Lwazi corpus provides new opportunities in this respect. This paper presents the results of a baseline language identification system developed using the Lwazi corpus.

Two standard strategies to perform LID are investigated: phoneme recognition followed by language modelling (PRLM) and parallel PRLM (PPRLM) [1], [2]. In addition, we use two similar strategies based on word rather than phoneme recognition: word recognition followed by language modelling (WRLM) and parallel WRLM [3].

II. The Lwazi ASR corpus

The Lwazi speech corpus provides orthographically annotated speech data for all eleven South African languages. It was originally developed with the goal of speaker-independent speech recognition in mind. It consists of telephone recordings of approximately 200 speakers per language. Each speaker provides approximately 30 utterances, consisting of phrases from phonetically balanced corpora and answers to open and yes/no questions [4]. The utterances contain no code-switching or non-language words. From a LID perspective, the corpus is relatively small, consisting of between 3 and 8 hours of speech per language. It is freely available under an open content license [5].

The different languages in the Lwazi corpus, their language families, and the size of their vocabularies and phone sets are shown in Table I. The speech data is annotated orthographically, and accompanied by baseline pronunciation dictionaries which are an expansion of the original Lwazi dictionaries [6] created using grapheme-to-phoneme prediction strategies [7]. Aside from the phone types in Table I, the corpus includes phone markers for silence and speaker noise.

A. Data Sets

The corpus was split into training, development and evaluation sets. The evaluation sets used for this research coincide with the fixed Lwazi evaluation sets of [7], and consist of a total of approximately 1,200 utterances from 40 randomly selected speakers per language. The development sets consist of approximately 600 utterances from a further 20 speakers per language, also randomly chosen. The training sets contain the

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

TABLE II
DESCRIPTION OF THE LWAZI TEST SETS.

Lang.	Development			Evaluation		
	Phone tokens	Word tokens	Dur. (mins)	Phone tokens	Word tokens	Dur. (mins)
afr	14,080	3,574	26	28,724	7,165	51
eng	16,757	4,224	31	33,061	8,373	63
nbl	27,327	4,097	62	54,428	8,185	121
ssw	26,477	4,035	64	53,154	8,010	128
xho	21,162	3,521	51	45,275	7,408	105
zul	22,603	3,597	53	46,682	7,568	115
nso	23,379	5,948	58	47,657	1,1893	127
sot	20,082	5,003	44	38,280	9,628	85
tsn	17,739	4,799	43	38,280	9,891	85
tso	19,032	4,607	51	37,560	9,067	92
ven	17,383	4,143	42	34,140	8,143	87
Total	226,021	47,548	525	455,954	95,331	1,068

TABLE III
DESCRIPTION OF THE LWAZI TRAINING SET.

Lang.	Phone tokens	Word tokens	Dur. (mins)
afr	98,427	24,754	179
eng	112,388	28,302	207
nbl	193,987	29,229	432
ssw	180,143	27,456	442
xho	166,502	27,711	406
zul	151,270	24,363	360
nso	151,680	38,682	385
sot	136,536	34,007	300
tsn	131,737	35,694	339
tso	146,736	35,615	378
ven	119,140	28,817	305
Total	1,588,546	334,630	3,733

TABLE IV
SIZES OF THE LANGUAGE MODEL TRAINING SETS, AFTER REMOVAL OF UTERRANCES OVERLAPPING WITH THE DEVELOPMENT OR EVALUATION SETS.

Lang.	Phonemes		Words		Utts
	tokens	% remaining	tokens	% remaining	% remaining
afr	16,575	16.8%	4,735	19.1%	22.4%
eng	17,989	16.0%	5,279	18.7%	22.9%
nbl	59,118	30.5%	9,882	33.8%	34.8%
ssw	52,572	29.2%	8,921	32.5%	36.0%
xho	69,063	41.5%	12,136	43.8%	43.7%
zul	67,166	44.4%	11,399	46.8%	48.1%
nso	39,039	25.7%	9,808	25.4%	33.8%
sot	39,788	29.1%	10,224	30.1%	32.1%
tsn	44,769	34.0%	12,229	34.3%	37.6%
tso	40,693	27.7%	10,110	28.4%	30.5%
ven	33,551	28.2%	8,046	27.9%	30.4%
Total	480,323	30.2%	102,769	30.7%	33.9%

remaining data, consisting of roughly 4,000 utterances from approximately 140 speakers per language. The utterances are fairly short, with an average length of 5 seconds each. The total duration and number of phonemes in each of the test sets is shown in Table II, and those in the training set are shown in Table III.

The Lwazi corpus contains many phonetically rich utterances. Answers to questions are generally short and formulaic, and the phonetically balanced sentences which speakers were asked to read are often repeated up to 15 times throughout the corpus. As a result, many of the utterances in the test sets also occur verbatim in the training set. To avoid positive bias in language identification, these utterances were removed from the training set prior to language model training.

The prompts used to obtain phonetically balanced utterances are not publically available, so recurring sentences were detected using exhaustive pairwise string alignment between the training and test set word transcriptions. Training utterances longer than five words were removed if they resulted in a match accuracy of at least 60% with any of the test sentences. Training utterances consisting of five or fewer words were removed if a perfect match (disregarding silence and noise markers) occurred in one of the test sets.

The amount of training data available for language modelling was significantly reduced by this process, as shown in Table IV. Not every language is affected equally: Afrikaans and English suffer the greatest reductions, whereas isiXhosa and isiZulu are affected the least. This results in language model training sets of disparate sizes.

III. BASELINE SYSTEM DESCRIPTION

In the first two experiments, LID is performed through PRLM and PPRLM, two standard approaches to language identification [1], [2]. PRLM uses a single-language phoneme speech recognition system to tokenise the training and test data for each language. One n-gram language model is then trained on the resulting training set phoneme sequences for each language. Identification is performed by using these language models to score the phoneme sequence of a test utterance. The language model which produces the highest log likelihood score is selected as the most likely candidate for identification.

It is not necessary for the front end phone recognisers to be trained on speech in the languages to be identified for LID to be feasible. For this reason, speech recognition is usually carried out using acoustic models trained on the CALLFRIEND corpus [8]. However, since telephone speech data are available for each of the 11 South African languages in the Lwazi corpus, our experiments will use acoustic models trained on these data.

PPRLM functions similarly to PRLM, but includes a set of single-language phoneme recognisers that tokenise the input utterance in parallel. Language models are obtained for the tokenised training sets of each language, using each available tokeniser. Hence, this strategy employs $N_L \times N_T$ language models, where $N_L = 11$ is the number of languages, and N_T

is the number of tokenisers, in our case also 11. For each language, the log likelihoods of language models resulting from all acoustic models are considered, and the global optimum identified. This has the added benefit of incorporating language-specific phonotactic constraints during recognition, so that acoustic as well as phonotactic likelihood plays a role in identification. PPRLM has been shown to be an accurate and reliable method of LID as compared to other methods [1].

The methods used for word-based language identification are analogous with the PRLM and PPRLM methods, with the exception that single-language word recognisers are used in place of phoneme recognisers. WRLM uses a single-language word recogniser on the training and test sets of each language, and trains n-gram word language models on the word sequences obtained from the training sets. Identification of an utterance is then performed by scoring its word transcription using each language model, and selecting the language with the highest log likelihood, as in PRLM.

PWRLM functions the same as PPRLM, but obtains a global optimum from the log likelihoods of the WRLM word language models using the acoustic models, instead of the PRLM language models. Although WRLM and PWRLM are not standard approaches, word-based identification methods have proven successful in prior studies [3].

A. Acoustic models

A set of HMM acoustic monophone models was trained using the full training set for all languages. Speaker-independent diagonal covariance monophone models with three states per model and one Gaussian mixture per state were obtained through embedded Baum-Welsh re-estimation, and normalised using cepstral mean normalization. Each speech frame was parameterised as a 39-dimensional feature vector which consisted of 13 Mel-frequency cepstral coefficients (MFCCs) and their first and second differentials.

Once a first set of acoustic models had been obtained from a flat start using the first pronunciation of each training word in the dictionaries, a Viterbi word recogniser was used to re-align the training set to account for pronunciation variants. The models were re-aligned and re-trained a maximum of 8 times, after which the transcriptions were seen to stabilise.

The monophone models were expanded to crossword triphone models through decision tree state clustering [9]. Noise and silence models were used as context phonemes but not expanded to triphones. The triphone models were then improved by gradually increasing the number of Gaussian mixtures, and performing four iterations of embedded re-estimation after each increase. This procedure was continued until the models had 8 mixtures per state, after which phoneme and word recognition results on the development set no longer improved significantly.

Using the full training set in acoustic modelling should not result in a positive bias in identification results, since the recognisers are only used to obtain transcriptions of the same training sets for language model training. The acoustic models are trained on the training set, and the phonotactic and word sequence language models used for LID are trained on the strings recognised in that same set, so the error rate on the training data is expected to be lower than that on the test data. Any resulting bias in LID scores would thus be negative.

B. Phoneme and word recognition

For phoneme recognition, we trained backoff bigram models on the reduced training sets described in Table IV. These are not the same language models used for subsequent language identification. Language model probabilities were calculated with an absolute discount ratio of 1.0. Phoneme recognition for PRLM purposes is commonly performed using flat language models, but we found that using bigram models led to improved identification results. Word recognition was similarly performed using backoff bigram language models, trained on the word transcriptions of the same reduced training set, with an absolute discount ratio of 1.0.

Both phoneme and word recognition was performed using the HTK hidden Markov model (HMM) based speech recogniser, which performs a time-synchronous beam search using the Token-Passing procedure [10]. The same recognition parameters were used for all languages. For both phoneme and word recognition, a word insertion penalty of -10 was used. A language model scaling factor of 6 was used for phoneme recognition, and a factor of 10 for word recognition. These values were found to produce optimal overall recognition error rates on the development set.

Eleven single-language phoneme and word recognisers were obtained in this way, one of each for each language in the Lwazi corpus. Recognition was performed on the training and test sets of all languages using each recogniser. This resulted in 121 sets of phone transcriptions and an equal number of word transcriptions, each split up into a training, development, and evaluation set.

C. Phonotactic and word sequence language models

For each of the eleven languages, eleven phoneme and word language models were trained on the transcriptions of the reduced training set obtained using the corresponding set of acoustic models. Both phoneme and word language models were optimised on the development sets obtained using the same acoustic models. Optimisation led to a constant discount value of 0.7 for all four systems.

The language model order was gradually increased until identification results no longer improved on the development set. For PRLM and WRLM, each individual recogniser was optimised in this way. For PPRLM and PWRLM, all language models were trained with the same order, and an optimum was determined for the overall system.

D. Gaussian Back-end

Using multiple acoustic phoneme or word recognition systems in parallel, the information of these needs to be fused. This is accomplished by combining all scores for a particular test segment (in our case 121, from $N_{\mathrm{PR}} = 11$ acoustic PR outputs each modeled with $N_L = 11$ phone sequence LMs) in a Gaussian back-end model [2], e.g., Linear Discriminant

Analysis. The parameters of this model (one mean vector per language and a shared covariance matrix) need to be trained; we use the development test segments for that. The Gaussian back-end produces for each test segment x one log-likelihood l_i per hypothesis language i. Together with a prior over languages π_i these likelihoods can be converted to posterior probabilities $P(i \mid x)$ using

$$P(i \mid x) = \frac{\pi_i \exp l_i}{\sum_j \pi_j \exp l_j}. \quad (1)$$

E. Evaluation metrics

We evaluate our language identification systems using the independent evaluation set of segments, using several different evaluation metrics. The first simply is the language identification error rate. The language ID error rate can be computed using

$$E_{\text{LID}} = \frac{1}{N_L} \sum_{i=1}^{N_L} \frac{1}{||\mathcal{T}_i||} \sum_{j \in \mathcal{T}_i} 1 - \delta(i, \arg\max_k P(k \mid x_j)), \quad (2)$$

where $\delta(i, j)$ is the Kronecker delta function counting equal language indices, j indexes the trials in set $\mathcal{T}_i$ of language i, and $||\cdot||$ is the cardinality operator counting trials. Effectively, E_{LID} is the average language identification error rate compensated for evaluation priors. Note that when assuming flat priors $\pi_i = 1/N_L$ the maximum posterior language probability is equal to the maximum language likelihood, and (1) is not needed. The language identification error rate is sensitive to a per language log-likelihood bias. If the recogniser gives relatively higher log likelihoods to one language than another, the confusion matrix will become asymmetric which typically leads to more errors. Although ID error rate is not a standard measure in LID, it is a meaningful evaluation metric when classification is performed by maximum likelihood and evaluation is performed with equal priors, as is the case in our research.

One may argue that the assumption of equal priors is not very realistic, so we would also like to probe the prior distribution space at several other points. The way this happened in NIST Language Recognition Evaluations [8] from 2005–2009 was by using a varying prior distribution, and evaluating using a decision-cost function paradigm, by computing the "average cost" C_{avg} over these distributions. In our terminology, for every test segment, the prior is probed at N_L points, with k running over the target languages $1, \ldots, N_L$

$$\pi_i = \begin{cases} P_{\text{tar}} & \text{for } i = k \\ (1 - P_{\text{tar}})/(N_L - 1) & \text{for } i \neq k, \end{cases} \quad (3)$$

where $P_{\text{tar}} = \frac{1}{2}$ represents the prior of the target language. For a given test segment j, a posterior distribution is computed for every prior distribution, and decisions are made by thresholding the posterior $P(k \mid x_j)$ at $\frac{1}{2}$. These decisions can lead to 'false alarms' if $j \neq k \wedge P(k \mid x_j) > \frac{1}{2}$ or 'misses' if $j = k \wedge P(k \mid x_j) < \frac{1}{2}$. The false alarm proportions P_{FA}^{jk} are computed for every target language k and test language j, and averaged over j and k,

$$P_{\text{FA}} = \frac{1}{N_L(N_L - 1)} \sum_{jk} P_{\text{FA}}^{jk}. \quad (4)$$

Similarly, the miss proportions P_{miss}^k are computed for each target language k, and averaged

$$P_{\text{miss}} = \frac{1}{N_L} \sum_k P_{\text{miss}}^k. \quad (5)$$

The final NIST evaluation metric is the cost function

$$C_{\text{avg}} = (1 - P_{\text{tar}}) C_{\text{FA}} P_{\text{FA}} + P_{\text{tar}} C_{\text{miss}} P_{\text{miss}}, \quad (6)$$

where the cost parameters $C_{\text{FA}} = C_{\text{miss}} = 1$.

The way we computed C_{avg} from the language posteriors and the true language labels already makes this metric calibration sensitive. Another measure that is calibration sensitive is the multiclass cross entropy,

$$H_{\text{mc}} = \sum_{i=1}^{N_L} \frac{\pi_i}{||\mathcal{T}_i||} \sum k \in \mathcal{T}_k - \log P(i \mid x_k) \quad (7)$$

which measures the uncertainty in the posterior of the true language. This metric can be translated into a *perplexity*

$$\text{Perplexity} = \exp H_{\text{mc}} \quad (8)$$

which is the average number of languages you would still have to choose from after consulting the language recogniser. We use a slightly different version of this metric, which was first used in language recognition in the Albayzin language recognition evaluation in 2012 [11], the *confusion*

$$C = \text{Perplexity} - 1 \quad (9)$$

All three evaluation measures described above are calibration sensitive, although no calibration was performed for this particular research. Performances are not reported in equal-error-rate (EER) because it is an ill-defined metric for language recognition purposes [12].

IV. Results

A. PRLM

PRLM results were obtained using the language models trained on the transcriptions of one speech recogniser at a time. The optimal n-gram order for the obtained language models and the identification results we obtained using this method are reported in Table V.

There is a clear distinction between the systems trained on West Germanic and Southern Bantu acoustic data. LID using the transcriptions of Afrikaans and English recognisers performs worst overall, and increasing the order of the language model does not have a beneficial effect on the error rates and detection costs. This most likely has to do with the small sizes of the training sets used to generate the phonotactic language models for these languages, although the different phone sets could also be a part of the cause. The results of every individual Southern Bantu recogniser, as well as their optimal n-gram orders, are very similar to each other.

TABLE V
LANGUAGE IDENTIFICATION RESULTS USING DIFFERENT PHONEME RECOGNISERS FOR PRLM

Recognition Language	Optimal n-gram order	E_{LID}	C_{avg}	Confusion
afr	3	0.719	0.329	6.72
eng	3	0.716	0.329	6.75
nbl	5	0.645	0.290	5.61
ssw	4	0.648	0.292	5.63
xho	5	0.643	0.294	6.68
zul	5	0.634	0.286	5.46
nso	5	0.667	0.304	6.05
sot	4	0.668	0.300	6.03
tsn	5	0.670	0.307	6.17
tso	4	0.664	0.294	5.83
ven	5	0.671	0.303	6.08
Average		0.610	0.303	6.09

TABLE VI
LANGUAGE IDENTIFICATION RESULTS USING DIFFERENT WORD RECOGNISERS FOR WRLM

Recognition Language	Optimal n-gram order	E_{LID}	C_{avg}	Confusion
afr	2	0.808	0.403	8.62
eng	4	0.798	0.397	8.54
nbl	2	0.739	0.350	7.51
ssw	2	0.744	0.353	7.60
xho	3	0.718	0.338	7.10
zul	2	0.711	0.334	7.11
nso	2	0.741	0.346	7.51
sot	3	0.742	0.349	7.53
tsn	3	0.735	0.348	7.50
tso	2	0.738	0.347	7.36
ven	3	0.743	0.355	7.59
Average		0.747	0.356	7.63

B. WRLM

WRLM results were obtained by using word recognisers trained on one language at a time. The optimal n-gram order and the identification results for each language model using this method are shown in Table VI.

WRLM performs worse than PRLM in all instances. The effect of the language model on the results is very similar to that observed in the PRLM results. The language models trained on the data in Western Germanic languages produce worse results than those trained on the Southern Bantu languages, whereas the scores for the Bantu recognisers are very similar to each other.

Unlike in PRLM, there is no clear relation between optimal n-gram order and language family. The optimal n-gram order is mostly lower than in PRLM, as would be expected, since there are much fewer word tokens. English is an exception to this, with optimal performance using an n-gram order of 4. It should be noted, however, that increasing the n-gram order generally had very little effect on the WRLM identification scores. Overall, none of the WRLM results are very good.

C. Parallel PRLM

PPRLM results were obtained using the same language models used for PRLM, but this time optimising over a combination of all the log likelihoods. The LID results of language model n-gram order optimisation are shown in Table VII. Optimal results were obtained when all language models used were the same order. The results do not substantially improve after n-gram order 5.

The effects of combining the language models is made clearer by the progression shown in Table VIII. The phoneme transcriptions of different recognisers are gradually included (in random order). As the number of languages increases, the results gradually improve for all evaluation metrics. Any individual addition is beneficial to the overall result.

A confusion matrix for the 5-gram PPRLM language identification system is shown in Table IX. A clear trend across language families is visible, with confusion mostly occurring within the same language families. The West Germanic languages are in general correctly identified more frequently than the Southern Bantu languages. Because the Southern

TABLE VII
LANGUAGE IDENTIFICATION RESULTS WITH DIFFERENT ORDER n-GRAMS USING PPRLM AND PWRLM.

	PPRLM			PWRLM		
Order	E_{LID}	C_{avg}	Conf.	E_{LID}	C_{avg}	Conf.
1	0.620	0.275	5.33	0.557	0.252	4.61
2	0.571	0.247	4.46	0.420	0.186	2.95
3	0.503	0.215	3.55	0.417	0.184	2.92
4	0.459	0.195	3.09	0.416	0.184	2.92
5	0.446	0.190	2.96		—	
6	0.445	0.190	2.95		—	

TABLE VIII
A COMPARISON OF THE RESULTS OF 5-GRAM PPRLM AND 3-GRAM PWRLM USING AN INCREASING NUMBER OF PARALLEL TRANSCRIPTIONS.

	PPRLM			PWRLM		
No. of langs	E_{LID}	C_{avg}	Conf.	E_{LID}	C_{avg}	Conf.
1	0.728	0.340	6.96	0.805	0.402	8.64
2	0.674	0.307	5.99	0.757	0.366	7.79
3	0.581	0.252	4.52	0.683	0.319	6.51
4	0.540	0.233	4.02	0.619	0.285	5.44
5	0.522	0.224	3.76	0.577	0.254	4.93
6	0.497	0.212	3.47	0.541	0.243	4.42
7	0.486	0.208	3.40	0.541	0.243	4.42
8	0.476	0.200	3.25	0.492	0.217	3.81
9	0.464	0.197	3.15	0.471	0.207	3.56
10	0.452	0.193	3.05	0.444	0.196	3.24
11	0.446	0.190	2.96	0.417	0.184	2.92

TABLE IX
A CONFUSION MATRIX OF THE OPTIMAL 5-GRAM PPRLM LID SYSTEM. THE PERCENTAGE OF THE EVALUATION SET CLASSIFIED AS EACH LANGUAGE IS SHOWN. THE COLUMNS DENOTE THE TEST LANGUAGES AND THE ROWS THE HYPOTHETICAL LANGUAGES.

	afr	eng	nbl	ssw	xho	zul	nso	sot	tsn	tso	ven	**Overall**
afr	71.1	18.3	5.4	5.4	4.7	5.7	4.0	5.3	6.9	6.5	4.8	12.6
eng	17.3	66.2	4.9	5.5	6.3	9.3	4.5	7.2	7.4	7.7	5.7	12.9
nbl	0.0	0.2	52.6	4.2	4.5	5.4	1.0	1.5	0.7	2.8	2.7	6.9
ssw	0.3	0.4	5.6	54.3	2.6	3.5	1.2	2.9	1.2	2.5	1.6	6.9
xho	0.3	0.4	3.9	2.5	55.5	5.7	0.7	0.8	0.4	1.3	1.1	6.5
zul	0.9	1.7	7.0	9.0	10.5	52.1	2.4	2.7	3.1	3.8	3.5	8.7
nso	0.8	0.6	3.6	3.1	1.4	1.9	49.8	6.0	11.4	4.4	5.3	8.0
sot	3.8	4.6	4.3	6.9	4.7	6.7	15.0	54.2	16.0	8.3	8.0	12.0
tsn	1.8	1.6	2.2	1.9	3.7	2.7	13.2	10.8	45.9	3.0	4.2	8.3
tso	0.9	1.3	3.3	3.0	2.3	2.4	2.6	3.6	2.6	50.5	5.5	7.1
ven	2.8	4.8	7.5	4.3	3.9	4.6	5.5	5.1	4.5	9.1	57.5	10.0

TABLE X
A CONFUSION MATRIX OF THE OPTIMAL 3-GRAM PWRLM LID SYSTEM. THE PERCENTAGE OF THE EVALUATION SET CLASSIFIED AS EACH LANGUAGE IS SHOWN. THE COLUMNS DENOTE THE TEST LANGUAGES AND THE ROWS THE HYPOTHETICAL LANGUAGES.

	afr	eng	nbl	ssw	xho	zul	nso	sot	tsn	tso	ven	**Overall**
afr	69.3	22.6	12.1	10.2	11.5	15.0	10.0	11.4	14.0	12.1	9.7	18.0
eng	13.8	63.1	5.6	6.8	8.9	9.1	5.7	8.5	9.9	8.4	5.3	13.2
nbl	0.2	0.2	60.8	2.2	1.0	1.7	1.6	1.3	1.5	1.6	0.8	6.6
ssw	0.6	0.4	2.0	59.2	1.3	3.7	1.8	1.0	0.8	0.9	0.7	6.6
xho	0.7	0.3	1.3	1.6	62.3	2.2	1.3	1.3	0.8	0.7	0.7	6.6
zul	1.7	0.8	1.6	3.4	2.3	51.0	1.3	2.6	1.7	2.8	1.3	6.4
nso	1.3	1.8	1.9	2.3	1.4	2.7	52.4	3.9	5.3	2.5	3.7	7.2
sot	5.8	3.8	3.2	4.4	4.7	4.0	8.9	54.6	9.8	4.4	4.9	9.9
tsn	1.9	2.4	2.1	2.8	1.3	2.3	7.6	5.6	47.9	2.9	4.4	7.4
tso	0.9	1.1	3.0	3.5	2.6	3.6	3.2	2.8	2.0	56.5	4.2	7.6
ven	3.8	3.5	6.4	3.7	2.8	4.7	6.1	7.0	6.4	7.2	64.4	10.6

TABLE XI
THE RESULTS OF 5-GRAM PPRLM AND 3-GRAM PWRLM AS A FUNCTION OF THE DURATION OF THE TEST UTTERANCE.

Strategy	Duration	E_{LID}	C_{avg}	Confusion
PPRLM	0s-2.5s	0.454	0.204	3.22
	2.5s-5.5s	0.423	0.176	2.68
	5.5s+	0.462	0.194	3.86
PWRLM	0s-2.5s	0.440	0.207	2.92
	2.5s-5.5s	0.391	0.180	2.76
	5.5s+	0.421	0.181	2.92

Bantu languages are phonetically close to one another, but distinct from the West Germanic languages, identification of these is easier. Compared to the West Germanic and the Nguni language families, the Sotho-Tswana languages exhibit increased confusion among themselves. Not every language is selected as often by the identification system, as it has a moderate bias toward Afrikaans, English and Sesotho.

D. Parallel WRLM

PWRLM results were obtained by combining the log likelihoods of all language models used in WRLM, analogous to the method used in PPRLM. The LID results using PWRLM with different n-gram order word language models are shown in Table VII. PWRLM results do not substantially improve after n-gram order 3.

Although none of the individual WRLM systems perform better than their PRLM counterparts, the optimal PWRLM results are better than the optimal PPRLM results. Optimal PWRLM results are obtained by using 3-grams, but these are only marginally better than those obtained using 2-grams, with which the optimal PPRLM system is also outperformed.

The effect of a gradual increase in language models is shown in Table VIII. Like in PPRLM, each additional language model improves the overall performance of the system. As the number of used language models increases, the gap between the PPRLM and PWRLM results narrows, but the latter does not start producing better results (in terms of average cost and confusion) until all language models are considered.

A confusion matrix for the 3-gram PWRLM LID system is shown in Table X. Compared to the optimal PPRLM system shown in Table IX, Afrikaans and English are correctly identified slightly less often, but the system performs better for all Southern Bantu languages, with the exception of isiZulu. The amount of confusion within the Sotho-Tswana and Nguni language families is substantially reduced. However, despite not performing as well for Afrikaans and English as the optimal PPRLM system, PWRLM results in a much stronger bias toward these languages. Utterances are classified as Afrikaans much more often than any of the other individual languages.

E. Utterance duration

The results reported thus far are an average over all utterances in the Lwazi evaluation set. Table XI shows the results for test utterances depending on the length of the utterances, for test utterances shorter than 2.5s, between 2.5s and 5.5s, and longer than 5.5s.

Contrary to our expectations, performance does not improve with the length of the utterance. The best results are obtained from the utterances between 2.5s and 5.5s, regardless of the strategy used. This is likely a result of the structure of the Lwazi corpus: shorter utterances are likely to be answers to specific questions, and therefore more formulaic.

V. DISCUSSION

This paper presents a set of baseline language identification results using the Lwazi data. We used twenty-two single-language automatic speech recognisers trained on the different languages in the Lwazi corpus to obtain sets of phoneme and word transcriptions for each of the languages. To obtain LID results, we trained phoneme and word language models on these transcriptions using various orders of n-gram.

We show results for four different LID strategies. PRLM uses language models trained on the phoneme transcriptions of each of the Lwazi languages, and selects the most likely language by comparing the language model log likelihood.

WRLM uses word transcriptions for the same task. PPRLM runs eleven phoneme recognisers in parallel, and determines a global optimum based on a combination of the phoneme language models trained on their results. PWRLM does the same, but uses word language models trained on the word transcriptions.

Although PRLM obtained better results for identification than WRLM, the optimal PWRLM outperformed the optimal PPRLM system. The best results were obtained using PWRLM with 3-gram language models, resulting in a language identification error rate of 0.418, a decision cost of 0.184, and a confusion of 2.92. This is not exceptionally good, but it should be borne in mind that the Lwazi corpus is small in size, and that the utterances were just 5 seconds long on average. To improve future results using the Lwazi corpus, GMM-based LID could be considered as an alternative, since this strategy has been shown to perform better than PPRLM on short utterances [2]. However, since no prior results for South African language identification have been reported, these results can be considered a new benchmark.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," in *IEEE transactions on speech and audio processing*, vol. 4, no. 1, 1996, pp. 31–44.

[2] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and R. A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification," in *Proc. Eurospeech*, 2003, pp. 1345–1349.

[3] M. A. Zissman and K. Berkling, "Automatic language identification," *Speech Communication*, vol. 35(1-2), pp. 115–124, 2001.

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

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

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

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

[8] "The 2007 NIST language recognition evaluation plan," http://www.nist.gov/speech/tests/lang/2007/, 2007.

[9] 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.

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

[11] L. J. Rodríguez-Fuentes, N. Brümmer, M. Penagarikano, A. Varona, M. Diez, and G. Bordel. (2012, Nov.) The Albayzin 2012 language recognition evaluation plan. [Online]. Available: http://iberspeech2012.ii.uam.es/images/PDFs/albayzin_lre12_evalplan_v1.3_springer.pdf

[12] D. A. van Leeuwen and K. P. Truong, "An open-set detection evaluation methodology applied to language and emotion recognition," in *Proceedings of INTERSPEECH 2007*. Antwerp: ISCA, August 2007, pp. 338–341.