INTERSPEECH 2009 BRIGHTON

Basic speech recognition for spoken dialogues

Charl van Heerden, Etienne Barnard, Marelie Davel

Human Language Technologies Research Group, Meraka Institute, CSIR, South Africa

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

Abstract

Spoken dialogue systems (SDSs) have great potential for information access in the developing world. However, the realisation of that potential requires the solution of several challenging problems, including the development of sufficiently accurate speech recognisers for a diverse multitude of languages. We investigate the feasibility of developing small-vocabulary speaker-independent ASR systems designed for use in a telephone-based information system, using ten resource-scarce languages spoken in South Africa as a case study.

We contrast a cross-language transfer approach (using a well-trained system from a different language) with the development of new language-specific corpora and systems, and evaluate the effectiveness of both approaches. We find that limited speech corpora (3 to 8 hours of data from around 200 speakers) are sufficient for the development of reasonably accurate recognisers: Error rates are in the range 2% to 12% for a ten-word task, where vocabulary words are excluded from training to simulate vocabulary-independent performance. This approach is substantially more accurate than cross-language transfer, and sufficient for the development of basic spoken dialogue systems.

Index Terms: speech recognition, limited vocabularies, technology for the developing world

1. Introduction

Recent years have seen a significant push towards the use of SDSs for information access in the developing world [1, 2, 3]. The deployment of such systems faces a number of logistical, technical and financial hurdles, but the potential benefits are immense in light of the crucial role that relevant, up-to-date information plays in improving quality of life. Application domains such as education, agriculture, health care and government services all stand to benefit from the availability of widely accessible information sources that do not require widespread computer infrastructure or computer literacy.

From the perspective of speech technology, the three most important challenges to address in this regard are the following:

- The design of spoken interfaces that are usable and friendly in diverse cultures, by users with limited or no computer literacy.
- The development of speaker-independent automatic speech recognition (ASR) systems that function reliably in the local languages of the developing world.
- The development of text-to-speech (TTS) systems that are easily understood in these same languages.

We focus on the second of these challenges. ASR systems exist for only a small fraction of the languages of the world, and for almost none of those spoken primarily in the developing world. Obtaining reasonable coverage of these languages is currently viewed as a major obstacle to widespread use of SDSs. One severe challenge is that modern ASR systems use statistical models which are trained on corpora of relevant speech (i.e. appropriate for the recognition task in terms of the language used, the profile of the speakers, speaking style, etc.) This speech generally needs to be curated and transcribed prior to the development of ASR systems, and speech from a large number of speakers is generally required in order to achieve acceptable system performance. In the developing world, where the necessary infrastructure such as computer networks, as well as first language speakers with the relevant training and experience, are limited in availability, the collection and annotation of such speech corpora is a significant hurdle to the development of ASR systems.

State-of-the-art speaker-independent ASR systems in languages such as English or Mandarin are developed with corpora containing speech from hundreds or thousands of speakers, using up to several hundred hours of speech data. It would require a very costly effort to develop such corpora in the developing world. Fortunately, the class of SDSs envisioned for information access in the developing world does generally not require the large-vocabulary, natural-language processing capabilities which necessitate such large training corpora [4, 3]. For many of these applications, dialogues can be designed that limit the active vocabulary at any point in the interaction to a dozen or fewer words. (Such limited systems are also a useful way to bootstrap systems with larger vocabularies and higher accuracies, since they can be deployed in usable applications that simultaneously perform data collection.)

We have therefore undertaken a research program to investigate the performance that can be achieved with ASR systems that use relatively small corpora (fewer than 200 speakers, less than 10 hours of speech per language). Our research utilises a corpus of telephone speech in the eleven official languages of South Africa, which is described in Section 2. In Section 3 we discuss a number of design choices that were made during our research. Section 4 describes the phone-recognition experiments that were performed to calibrate the accuracy of our ASR systems, whereas Section 5 summarises results obtained with small-vocabulary recognition tasks. In our conclusion (Section 6) we consider the broader implications of these results for ASR in the developing world.

2. The Lwazi ASR corpus

The Lwazi ASR corpus was developed as part of a project that aims to demonstrate the use of speech technology in information service delivery in South Africa [5]. Specifically, the three-year Lwazi project (2006-2009) produced the core tools and technologies required for the development of multilingual SDSs in all eleven of South Africa's official languages, and piloted the use of these technologies in government information service de-

livery.

The Lwazi ASR corpus consists of annotated speech data in the languages listed in Table 1, which also summarises the amount of speech available in each language and the number of phonemes that were used for the dictionaries described below. For the majority of these languages, no prior speech technology components or resources were available [6].

Language	code	# total minutes	# speech minutes	# distinct phonemes
isiZulu	zul	525	407	46
isiXhosa	xho	470	370	52
Afrikaans	afr	213	182	37
Sepedi	nso	394	301	45
Setswana	tsn	379	295	34
Sesotho	sot	387	313	44
SA English	eng	304	255	44
Xitsonga	tso	378	316	54
siSwati	ssw	603	479	39
Tshivenda	ven	354	286	38
isiNdebele	nbl	564	465	46

Table 1: *The official languages of South Africa, their ISO 639-3:2007 language codes, and the amount of speech contained in the Lwazi corpus*

Cost effectiveness was an important consideration during the design of the ASR corpus. In order to be able to afford the creation of resources for all the above languages the corpus was designed to be as small as possible while remaining practically usable in an SDS, thereby enabling the development of seed ASR systems that are able to support more extensive data collection efforts.

The ASR speech corpus consists of approximately 200 speakers per language (2,200 speakers in total), producing read and elicited speech, recorded over a telephone channel. Each speaker produced approximately 30 utterances; 16 of these were randomly selected from a phonetically balanced corpus and the remainder consist of short words and phrases: answers to open questions, answers to yes/no questions, spelt words, dates and numbers.

As general text corpora were not readily available for the selection of sets of phonetically balanced sentences (for any of the languages apart from English), a text corpus was developed for each language from data obtained directly from publishers supplemented with data crawled from the Internet. A 5,000-word pronunciation dictionary per language was created from high frequency words using bootstrapping. Grapheme-to-phoneme rules extracted from these dictionaries [7] were used to phoneticise the large text corpus, and a phonetically balanced corpus selected based on triphone coverage. The phonetically balanced corpus did not take tonal information into account (even though the Southern Bantu languages – which include nine of our languages – are tone languages), since tone is unlikely to be important for small-to-medium vocabulary applications [8].

The speaker population was selected to provide a balanced profile with regard to age, gender and type of telephone (mobile or landline). Only first language speakers were recorded. All speech was digitised as 8kHz, 16-bit wav files.

Calls were supervised by an operator who asked the relevant (elicited speech) questions and guided the speakers in reading through the (read speech) sentences distributed beforehand. While operators were expected to verify the quality of recordings and repeat questions where necessary, the operators for the different languages performed this task at varying levels of effectiveness.

The recorded speech was annotated with orthographic transcriptions and markers indicating background noise, speaker noise and partial words. In order to accommodate first language transcribers with limited or no experience in corpus development, the transcription protocol was kept as simple as possible. Apart from minimal punctuation and simple indicators for proper nouns and spelt words, transcribers were requested to simply transcribe exactly what they heard (guided by the prompt sheets). This was a surprisingly difficult task for many transcribers, and the various utterances were re-transcribed a number of times before an acceptable level of quality was achieved.

The Lwazi ASR corpus is freely available under an open content license, and can be obtained directly from the Lwazi website [5].

3. Designing ASR systems with limited training data

When limited training data is available, a fundamental choice concerns the use of whole-word rather than phone-based ASR systems. However, for the class of information-access systems that are the goal of the current research, whole-word systems are not a viable option: application developers require the flexibility in vocabulary design that is achievable with phone-based recognisers.

Another option is to pool resources across languages, either by sharing data or by bootstrapping from a well-trained recogniser [9]. The South African languages would appear to be a suitable target for such sharing, given the family relationships between these languages. However, earlier work with the South African languages achieved limited benefit from cross-language sharing [10, 11], and in our pilot experiments information sharing has also not yielded significant gains. Although we believe that cross-language sharing will prove beneficial for the Lwazi corpus in the long run, we have therefore not included such sharing in the results reported here.

The recognisers employed are consequently fairly standard HMM-based systems. We use HTK 3.4 to build a context-dependent cross-word HMM-based phoneme recogniser with triphone models. Each model had 3 emitting states with 7 mixtures per state. (This combination was determined to be optimal for phone-recognition accuracy during pilot experiments.) 39 features are used: 13 MFCCs together with their first and second order derivatives. Cepstral Mean Normalisation (CMN) as well as Cepstral Variance Normalisation (CVN) are used to perform speaker-independent normalisation. A diagonal covariance matrix is used; to partially compensate for the implicit assumption of feature independence, semi-tied transforms are applied. A flat phone-based language model is employed for phone recognition, and deterministic grammars are used for the small-vocabulary experiment (as described below).

As the initial pronunciation dictionaries were developed to provide good coverage of the language in general, these dictionaries did not cover the entire ASR corpus. Grapheme-to-phoneme rules are therefore extracted from the general dictionaries using the Default&Refine algorithm [7] and used to generate missing pronunciations.

4. Phone recognition with the Lwazi corpus

Given the limitations of our training data, both in terms of quality and size, we have assessed the capabilities of our recognisers in two ways. The first measure, basic phone recognition, will be described in this section, while a small-vocabulary recognition experiment across languages is described in Section 5.

For phone recognition, we divided the data into a test set, which consists of 30 randomly selected speakers in each language, and a training set (the remaining speakers, approximately 170 per language). The recogniser for each language was built using all the training data for that language, using the recognition architecture described in Section 3. These recognisers were then evaluated by performing a Viterbi search with a language model that allows unrestricted transitions between any pair of phonemes. Dynamic programming was used to match the resulting phoneme strings against the strings that result from automatic phonemic transcription of the orthographic transcriptions of the test utterances. The resulting accuracies are summarised in Table 2. The table also lists the phonotactic perplexity of each language – that is, the perplexity that is measured if a bigram model is used to model the phoneme sequences that occur in the training set.

Language	% Corr	% Acc	Ave # phones	Phone ppl
Afrikaans	71.76	63.14	16.55	14.45
SA English	62.51	54.26	14.61	15.80
isiNdebele	74.21	65.41	28.66	10.26
isiXhosa	69.25	57.24	17.79	10.67
isiZulu	71.18	60.95	23.42	11.20
Tshivenda	76.37	66.78	19.53	9.99
Sepedi	66.44	55.19	16.45	11.54
Sesotho	68.17	54.79	18.57	10.40
Setswana	69.00	56.19	20.85	11.15
siSwati	74.19	64.46	30.66	10.38
Xitsonga	70.32	59.41	14.35	10.34
NTIMIT	64.07	55.73		

Table 2: *Phone-recognition correctness ("Corr") and accuracy ("Acc") achieved for each of the languages in the Lwazi corpus. "Ave # phones" refers to the average number of occurrences of each phone for each speaker, and the final column lists the phonotactic perplexity of each language in our corpus. NTIMIT results from [12] are provided for comparative purposes.*

The only language in our corpus for which comparable results are available is English. Since our data are collected over the telephone, we include recently published results [12] for the NTIMIT corpus in the table. We see that our English recogniser is somewhat less accurate than the system in [12] on NTIMIT. This is probably a consequence of the practical issues described in Section 2, as well as variability in the telephone channels and acoustic environments that occur in our data.

Interestingly, the correctness and accuracy of all other languages are higher than that of English, despite the fact that most languages have more phonemes than English. One possible explanation for this observation is the fact that English has fewer phonotactic constraints than any of the other languages, as can be deduced from the perplexity values in Table 2. (The Southern Bantu languages employ CV (consonant-vowel) or V syllable structures predominantly.) Overall, however, phonotactic perplexity does not correlate well with correctness or accuracy in our results, so other explanations for the relative accuracies should also be investigated.

5. Small-vocabulary speech recognition with the Lwazi corpus

Phone recognition is a useful benchmark to employ for recognition in new languages, since extensive intuition exists on phone-recognition accuracies achieved on standard corpora. However, initial applications of ASR in the developing world will in practice require accurate small-vocabulary recognition (as described in Section 1). We therefore describe experiments aimed at estimating our performance on such tasks next.

During the collection of the Lwazi ASR corpus, callers were asked several questions, of which some resulted in only a small set of responses. These included the following:

- Are you married?
- Are you speaking on a landline or a cellphone?
- What is your gender?
- What is your mother tongue?
- Where do you live? / Where were you born?

Since these same questions were asked of all speakers across all languages, they form a suitable basis for small-vocabulary experiments. Mother tongue speakers were then asked to label all answers that were semantically equivalent. In this fashion, answers such "Egoli" (isiZulu name for Johannesburg, meaning "place of gold") and "Johannesburg" were considered equivalents.

This resulted in 10 distinct semantic concepts for each language, with approximately one to three different lexical items corresponding to the same concept in a language. Because of the similar questions, similar meanings are attached to the matching concepts in each language, except for minor variations because of cultural differences. (For example, the majority of English and Afrikaans speakers would simply answer "yes" or "no" to the first question. In contrast, the majority of Xitsonga, Sepedi and Tshivenda speakers would answer the question in different ways depending on their gender. A Xitsonga man would for example say "ni tekile / a ni tekangi" (I have taken / I haven't taken), whereas a woman would say "ni tekiwile / a ni tekiwangi" (I have been taken / I haven't been taken).) Our small vocabulary task was constructed by removing all utterances that contain any of the phrases corresponding to any of these concepts from the training set, since such vocabulary-independent performance is the realistic goal for application in SDSs. For testing purposes, all utterances that contain only these phrases were employed; recognition was deemed correct if the phrase was placed into the correct semantic category. Because of the relatively small set of test utterances (five or fewer per speaker), we employed ten-fold cross validation to estimate recognition accuracy.

A vocabulary of ten words (actually, concepts) is a good test of typical recognition tasks in an SDS which is aimed at Interactive Voice Response (IVR) applications, where the dialogue is structured to contain mostly menu items and command words. Common tasks such as yes/no recognition require even smaller vocabularies, and larger tasks with highly distinctive vocabularies may in fact give comparable accuracies to those achieved with our artificially-constructed grammar.

As a baseline for comparison, we have also measured the accuracies that can be achieved with the cross-language transfer procedure described, for example, in [2, 3]. That procedure,

which is often a starting point for resource-scarce languages, utilises a well-trained recogniser in a world language such as English. All the words in the recognition task are transcribed using the phonemes of this well-trained recogniser, mapping the phonemes in the actual target language to the closest world-language phonemes where necessary. This cross-language dictionary is then used for recognition. Three English recognisers were investigated for our baseline, namely recognisers trained on the NTIMIT and Wall Street Journal corpora (the latter band-limited and downsampled to match our telephone corpus), and one trained on the English part of the Lwazi corpus.

These baseline systems are compared with the language-specific recognisers in Table 3. (We were not able to carry out this experiment for isiNdebele, for lack of access to a mother-tongue speaker who could perform the semantic mappings.)

Language	Lwazi models	Lwazi eng model	Ntimit	WSJ
isiZulu	90.53	80.00	37.57	69.19
isiXhosa	95.29	77.78	34.34	61.28
Afrikaans	96.11	90.35	60.36	79.15
Sepedi	89.49	83.72	54.91	43.41
Setswana	87.66	76.95	39.02	52.09
Sesotho	97.14	79.48	30.65	50.65
SA English	91.94	91.94	82.86	81.95
Xitsonga	97.90	77.58	54.99	60.60
siSwati	96.62	77.01	46.46	61.09
Tshivenda	97.74	66.37	57.56	52.14

Table 3: *Small vocabulary word recognition accuracies for* 10 *languages. Each system is required to distinguish between ten different semantic categories, with each category represented by one to three different lexical items.*

We see that accuracies above 90% are achieved in all languages except Sepedi and Setswana. With careful dialogue design [13], this should be sufficient for a useable SDS. Of the three baseline systems that use phoneme mappings, the Lwazi English model is easily the most accurate. This is to be expected, since the acoustic conditions of NTIMIT and WSJ are somewhat dissimilar to those in Lwazi; however, the magnitude of the differences in accuracy is somewhat surprising. Even this best baseline system is, however, much less accurate than the language-specific acoustic models in most cases. Excluding English, the languages with the smallest absolute difference between baseline and trained models are Afrikaans, which is linguistically quite similar to English, and Sepedi, which also performed worst in the phone-recognition experiments (Section 4).

Given the similarities between the semantic categories in the different languages, it it interesting to compare the accuracies achieved in this task across languages (with error rates ranging between 2.1% and 12.3%). The quality of the phone recognisers partially explains these differences – in particular, the relatively poor performance of Sepedi and Setswana at both phone recognition and small-vocabulary word recognition is notable. However, Sesotho (with relatively accurate word recognition) and isiZulu (relatively accurate phone recognition) point to other relevant factors, such as the acoustic confusibility of words in semantically distinct classes that happens to occur in some languages but not others.

6. Conclusion

We have shown that relatively small corpora can be used to develop phone-based speech recognition systems that are usable in SDSs. Although our corpora are relatively unsophisticated (using non-experts for operators and transcribers, and suffering from many acoustic imperfections as a result of difficulties in canvassing sufficiently many callers), speech-recognition systems built with these corpora perform reasonably well at phone-recognition and small vocabulary word-recognition tasks. In particular, the observed word-recognition accuracy is generally much higher than that achieved by phone mapping to a well-trained system in another language.

During this process, we have uncovered many of the challenges that must be addressed for speech recognition to become useable in much of the developing world. To compensate for a shortage of expertise in linguistics and speech technology, extensive error checking is required, and even then widely variable quality may result (as reflected by our word-recognition results). However, as more experience is gained with these tasks across the developing world, and more data from linguistically related languages becomes available, the deployment of useful speech-recognition systems in the languages of the developing world will become increasingly common.

7. References

[1] R. Tucker and K. Shalonova, "The Local Language Speech Technology Initiative," in *SCALLA Conf.*, Nepal, 2004.

[2] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld, "Healthline: Speech-based access to health information by low-literate users," in *Proc. IEEE Int. Conf. on ICTD*, Bangalore, India, 2007, pp. 131–139.

[3] J. Sherwani, S. Palijo, S. Mirza, T. Ahmed, N. Ali, and R. Rosenfeld, "Speech vs. touch-tone: Telephony interfaces for information access by low literate users," in *Proc. IEEE Int. Conf. on ICTD*, Doha, Qatar, 2009, pp. 447–457.

[4] M. Plauche, U. Nallasamy, J. Pal, C. Wooters, and D. Ramachandran, "Speech recognition for illiterate access to information and technology," in *Proc. IEEE Int. Conf. on ICTD*, Berkeley, USA, pp. 83–92.

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

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

[7] M. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

[8] S. Zerbian and E. Barnard, "Phonetics of intonation in South African Bantu languages," *Southern African Linguistics and Applied Language Studies*, vol. 26, no. 2, pp. 235–254, 2008.

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

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

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

[12] N. Morales, J. Tejedor, J. Garrido, J. Colas, and D.T. Toledano, "STC-TIMIT: Generation of a single-channel telephone corpus," in *LREC*, Marrakech, Morocco, 2008, pp. 391–395.

[13] M.H. Cohen, J.P. Giangola, and J. Balogh, *Voice User Interface Design*, Addison-Wesley, 2004.