INTERSPEECH 2009 BRIGHTON

ASR Corpus Design for Resource-Scarce Languages

Etienne Barnard, Marelie Davel, Charl van Heerden

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

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

Abstract

We investigate the number of speakers and the amount of data that is required for the development of useable speaker-independent speech-recognition systems in resource-scarce languages. Our experiments employ the Lwazi corpus, which contains speech in the eleven official languages of South Africa. We find that a surprisingly small number of speakers (fewer than 50) and around 10 to 20 hours of speech per language are sufficient for the purposes of acceptable phone-based recognition.

Index Terms: speech recognition, corpus design

10.21437/Interspeech.2009-727

1. Introduction

Speech recognition systems exist for only a small fraction of the languages spoken in the world. Most modern speech recognition systems use statistical models which are trained on corpora of relevant speech. 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.

The complexity of speech corpus development is strongly correlated with (a) the number of speakers that need to be canvassed and (b) the amount of speech that must be curated and transcribed. In order to minimise this complexity, it is important to have tools and guidelines that can be used to assist in designing the smallest corpora that will be sufficient for typical applications of ASR systems. The required number of speakers is of particular importance, since the logistics of gathering speech from a large number of speakers is a major challenge (especially in the developing world).

In this paper we systematically investigate the impact of both corpus size and the number of training speakers on the accuracy of a standard phone-based speech recogniser. Our experiments utilise a corpus of telephone speech in the eleven official languages of South Africa, and employ phone-recognition accuracy as a measure of quality in order to remain application independent.

2. Background: ASR corpus design

Corpus design techniques for ASR are generally aimed at specifying or selecting the most appropriate subset of data from a larger body of speech in order to optimise recognition accuracy, often while explicitly minimising the size of the selected corpus. This is achieved through various techniques that aim to include as much variability in the data as possible, while simultaneously ensuring that the corpus accurately matches the intended operating environment.

Three approaches are widely employed: (1) explicit specification of phonotactic, speaker and channel variability during corpus development, (2) automated selection of informative subsets of data from larger corpora, with the smaller subset yielding comparable recognition results, and (3) the use of active learning to optimise existing speech recognition systems.

Active and unsupervised learning methods can be combined to limit the amount of data that has to be transcribed [1]. The most informative untranscribed data is selected for a human to label, based on acoustic evidence of a partially and iteratively trained ASR system. From such work, it soon becomes evident that variation within the training data must be optimised, since randomly selected additional data does not necessarily improve recognition accuracy. By focusing on the selection (based on existing transcriptions) of a uniform distribution across different speech units such as words and phones, improvements are obtained [2]. In [3], Principal Component Analysis (PCA) is used to cluster data acoustically. These clusters then serve as a starting point for selecting the optimal utterances from a training database. As a consequence of the clustering technique, it is possible to characterise some of the acoustic properties of the data being analysed, and to obtain an understanding of the major sources of variation, such as speaker identity and gender [1].

The research described above is generally aimed at the efficient utilisation of existing speech resources. Our goal, in contrast, is to understand how to structure such a collection in the first place, when no additional language-specific data exists: primarily, how many speakers' data should be employed, and how much speech from each speaker? To our knowledge, the only direct attempt to answer this question is provided in [4], where it was found that "the number of speakers is more critical than the number of utterances for small training data sets". However, those experiments were conducted for only one target language (Japanese), and with a relatively simple recognition task which yielded asymptotic accuracy for as few as 1000 training utterances. We therefore address the same issue on a more challenging recognition task (unconstrained phone recognition), employing data from eleven different languages which belong to two significantly different language families.

3. 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 spoken dialogue systems in all eleven of South Africa's official languages, and piloted the use of these technologies in government information service delivery.

The Lwazi ASR corpus consists of annotated speech data

in the languages listed in Table 1. For the majority of these languages, no prior speech technology components or resources were available.

Language	code	# million speakers	language family
isiZulu	Zul	10.7	SB:Nguni
isiXhosa	Xho	7.9	SB:Nguni
Afrikaans	Afr	6.0	Germanic
Sepedi	Nso	4.2	SB:Sotho-Tswana
Setswana	Tsn	3.7	SB:Sotho-Tswana
Sesotho	Sot	3.6	SB:Sotho-Tswana
SA English	Eng	3.6	Germanic
Xitsonga	Tso	2.0	SB:Tswa-Ronga
siSwati	Ssw	1.2	SB:Nguni
Tshivenda	Ven	1.0	SB:Venda
isiNdebele	Nbl	0.7	SB:Nguni

Table 1: *The official languages of South Africa, their ISO 639-3:2007 language codes, estimated number of home language speakers in South Africa and language family (SB indicates Southern Bantu).*

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 a dialogue system, 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. The phonetically balanced corpus did not take tonal information into account (even though the Southern Bantu languages are tone languages), since tone is unlikely to be important for small-to-medium vocabulary applications [6]. In total, the corpus contains approximately 5 to 8 hours of speech per language, as summarised in Table 2.

Language	# total minutes	# speech minutes	# distinct phones
Afr	213	182	37
Eng (SA)	304	255	44
Nbl	564	465	46
Nso	394	301	45
Sot	387	313	44
Tsn	379	295	34
Ssw	603	479	39
Tso	378	316	54
Ven	354	286	38
Xho	470	370	52
Zul	525	407	46
Eng (N-TIMIT)	315	-	39

Table 2: *Size of the Lwazi ASR corpus. The size of the N-TIMIT corpus is provided as a comparison.*

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.

4. Phone recognition with the Lwazi corpus

The recognisers we employ are standard HMM-based systems. We use HTK 3.4 to build a context-dependent cross-word HMM-based phone recogniser with triphone models. Each model has 3 emitting states with 7 mixtures per state. (These parameter choices were determined to be optimal for phone-recognition accuracy with the complete corpora 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-specific 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.

The optimal values of parameters such as the number of mixtures and the insertion penalty (during language modeling) will in general depend on the amount of training data available. Since our values are optimised for the full corpus, our reported accuracies for reduced corpora are underestimates. Although we have not exhaustively evaluated all parameter options, we have verified that the dependencies are quite weak, and that the overall trends reported below are also observed when the parameters are adjusted.

As the initial pronunciation dictionaries were developed to provide good coverage of each 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. For the reason cited above, tone is not modelled in the system.

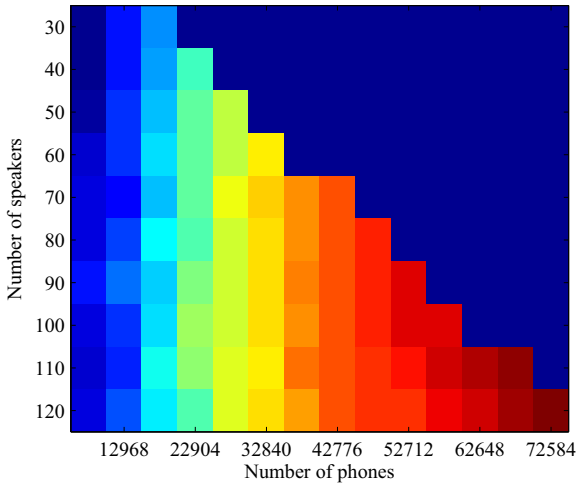
5. Results

As discussed in Section 1, the most important quantitative variables in ASR corpus design are the number of training speakers employed, and the total amount of data available. Our experiments pertinent to these variables are summarised below.

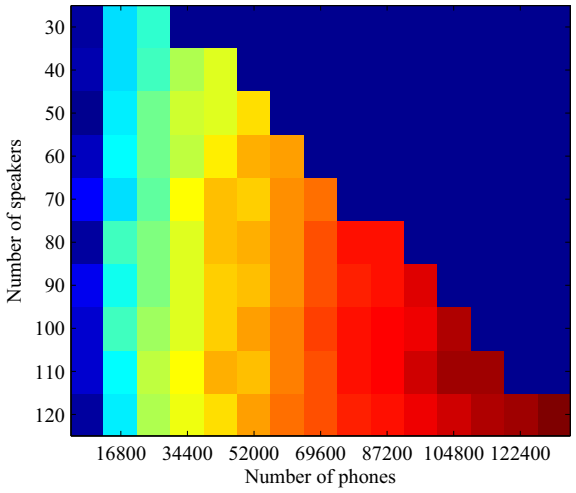
5.1. The number of training speakers

To analyse the influence of the number of training speakers on the recognition accuracy achieved, we investigate phone-recognition accuracy as a function of both the number of training speakers and the total number of phones used for training. (We use the number of phones rather than the number of words or utterances as measure of the amount of training data employed because of the significant differences in word and utterance lengths between the various languages – the phone count is therefore a better measure of the actual amount of speech employed.) The training sets are selected in such a way that the number of phones per speaker remains balanced.

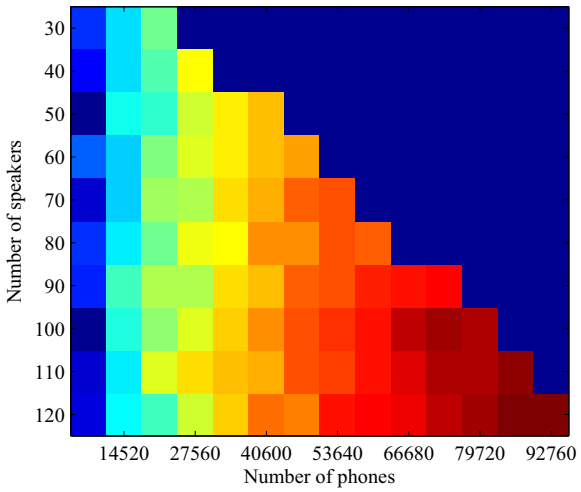
Fig. 1 shows typical results. (The solid blue triangle in the upper right-hand corner of each figure represents experiments that could not be performed because sufficient data was not available for each individual speaker.) It is clear that the number of training speakers has little or no influence on the accuracy achieved, in the range that we have investigated. Whereas the figures show systematically increasing accuracy as the number



(a) Afrikaans



(b) Sesotho



(c) isiZulu

Figure 1: *Phone accuracy as a function of both the number of speakers and the total amount of training data. The colours represent the measured accuracy, with blue being the lowest accuracy and red the highest.*

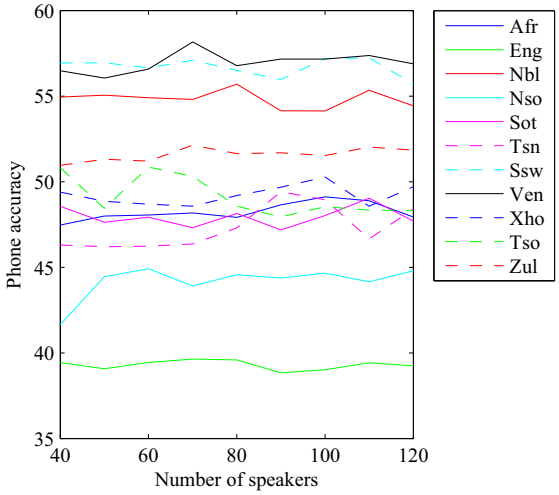


Figure 2: *Phone accuracy as a function of the number of speakers in training set. In all cases, approximately* 25% *of the available training data is used*

of training phones is increased (from left to right), increasing the number of speakers contributing to a given set of training data has little effect (top to bottom). This same behaviour is observed for all eleven languages, and is confirmed by representations such as that shown in Fig 2 (which shows the phone accuracy as a function of the number of training speakers, when about a quarter of the training data is used in each language).

5.2. The amount of training data

In Fig. 3 we show the trends of phone recognition accuracy as a function of the amount of training data, when all 120 speakers are used. Although the curves for some languages (especially Sepedi) are quite noisy, it seems clear that none of the languages is approaching asymptotic phone-recognition accuracy given the amount of training data available in our corpus. In order to obtain a rough estimate of the amount of training data required to approach such an asymptote, we employ a heuristic relationship that is expected to hold for a wide range of classifiers [8]. This relationship states that the error rate will asymptotically depend on the number of training samples (N) through the relationship $A - (B/N)$, with A and B parameters corresponding to the asymptotic error rate and the number of training samples required to approach within 1% of that error rate, respectively. We have empirically determined that this relationship provides a reasonable fit to our data for values of N greater than approximately 50,000; we have therefore used a linear least-squares fit to estimate A and B values for all our languages, including only measured accuracies for $N > 50,000$ in our analysis. Table 3 summarises the results obtained, and Fig. 4 shows a typical fit obtained in this manner. We see that quite good fits are obtained for several languages ($R^2 > 0.96$), and that the B parameter, which is related to the number of training phones required for accurate training, ranges between approximately 300,000 and 550,000 for these languages. (For $N = B$, phone accuracies within 1% of the asymptotic value are predicted.) In our corpus, the average phone duration is approximately 150 ms - hence, corpora of approximately 750 to 1,400 minutes per language are suggested.

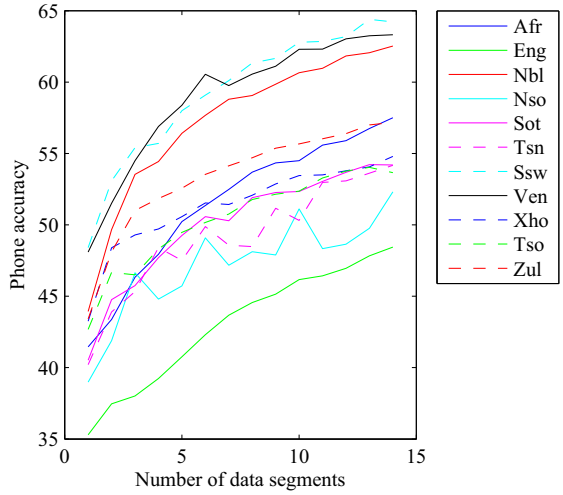


Figure 3: *Phone accuracy as a function of the amount of data in the training set, when data from all 120 training speakers is combined. The total amount of training data differs between the languages - the horizontal axis therefore indicates the number of segments used in each language, where the number of phone tokens per segment is (approximately) constant within a language, but different across languages.*

Language	A	B	R^2
Afr	64.94	549,900	0.9762
Eng	54.16	457,400	0.9650
Nbl	65.55	490,800	0.9722
Nso	55.35	380,200	0.2770
Sot	57.69	325,300	0.9201
Ssw	68.19	526,700	0.9757
Tsn	60.87	544,700	0.7975
Tso	57.26	300,100	0.8839
Ven	67.53	378,500	0.9616
Xho	57.60	331,700	0.9710
Zul	59.96	352,100	0.9636

Table 3: *Parameter values obtained by fitting measured phone-recognition rates. R^2 is the squared correlation between the estimated and actual values.*

6. Conclusion

For all the languages studied, the systematic evaluation of phone-recognition accuracy as a function of the number of training speakers and the amount of training data yields a consistent picture: Around 300,000 to 500,000 phone tokens from approximately 30-50 speakers should be sufficient to yield accuracies that are comparable to asymptotic accuracies for the type of system that we studied. The fact that these results were obtained on languages from two quite unrelated families (Germanic and Southern Bantu) is encouraging, though it would be important to investigate whether similar conclusions apply for other languages.

Our methods are still somewhat crude – for example, by carefully selecting utterances and speakers to ensure variability and enhanced coverage of "difficult" phones [9], smaller corpora may be designed to yield comparable accuracies. However, the more straightforward design employed here is representative of current standard practice. For this approach, the limited effect that additional speakers (above 50) has on system accuracy, was unexpected. High-accuracy systems with large or very large vocabularies will almost certainly require substantially more data and speakers than our estimates, but it is envisaged that such systems will grow from the more limited systems envisioned in the current research. We therefore believe that our results have great relevance for the development of speech corpora – especially for languages with limited resources.

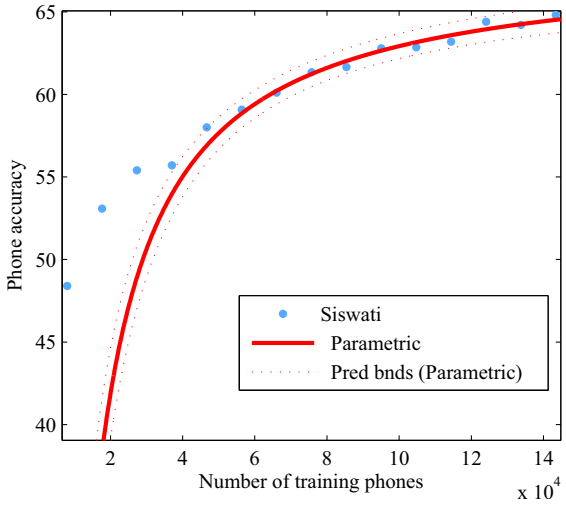


Figure 4: *Example of parametric fit (for Siswati accuracies), with 95% confidence intervals computed from the fit.*

7. References

[1] G. Riccardi and D. Hakkani-Tur, "Active and unsupervised learning for automatic speech recognition," in *Eurospeech*, Geneva, Switzerland, 2003, pp. 1825–1828.

[2] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," *ASRU workshop*, pp. 562–565, Dec 2007.

[3] A. Nagroski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," *ASRU workshop*, pp. 67–72, Nov 2003.

[4] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusumy, "An evaluation of cross-language adaptation for rapid HMM development in a new language," in *ICASSP*, Adelaide, 1994, pp. 237–240.

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

[6] 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.

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

[8] D. Schuurmans, "Characterizing rational versus exponential learning curve," *Journal of Computer and System Science*, vol. 55, no. 1, pp. 140–160, 1997.

[9] J.A.C. Badenhorst and M.H. Davel, "Data requirements for speaker independent acoustic models," in *PRASA*, 2008, pp. 147–152.