Data requirements for speaker independent acoustic models

Jacob A. C. Badenhorst[1], Marelie H. Davel[2]

[1]School of Electrical, Electronic and Computer Engineering,
North-West University, Potchefstroom, South Africa
[2]HLT Research Group, Meraka Institute, CSIR, South Africa
jbadenhorst@csir.co.za, mdavel@csir.co.za

Abstract

When developing speech recognition systems in resource-constrained environments, careful design of the training corpus can play an important role in compensating for data scarcity. One of the factors to consider relates to the speaker composition of a corpus: finding the appropriate balance between the number of speakers and number of speaker-specific utterances. We define a model stability measure based on the Bhattacharyya bound and apply this to analyse inter- and intra-speaker variability of a training corpus. We find that the different phone groups exhibit significantly different behaviour across groups, but within groups similar trends are observed. We demonstrate that, at a predictable point, additional data from one speaker does not contribute further to modelling accuracy and demonstrate the trends that can be expected when additional speakers are added. We also note that inter- and intra-speaker variability are independent effects, with some phone groups requiring more speaker-specific data, and others more cross-speaker data. More complex models require more training data, but exhibit similar overall trends to a simple Gaussian model.

1. Introduction

When building speech recognition systems for the languages of the developing world, it is often necessary to create new speech recognition corpora with limited resources. It is therefore important to design a speech corpus carefully in order to compensate to the extent possible for the scarcity of data. For example, even though the Lwazi corpus [1] is currently the most comprehensive speech recognition corpus available for South African languages, it contains only approximately 2 hours of annotated audio for each of the 11 languages – significantly less resources than typically used in the construction of a speech recognition system.

When designing a speech corpus, we are interested in the interplay between the number of speakers and number of utterances per speaker on the estimation accuracy of acoustic models for different phone types. Adding additional utterances from one speaker is more cost-efficient than adding additional speakers. How should the variety of speakers and utterances per speaker be balanced? Can we estimate whether the cost of adding additional data will be justified?

In this paper, we address these questions in the context of standard Gaussian Mixture Models (GMMs) as employed in a Hidden Markov Model (HMM) based speech recognition system. Specifically, we utilise a Monte Carlo estimation of the Bhattacharyya bound to characterise the similarity of two models, and use the stability of this measure when estimated for different subsets of the same data set to characterise the estimation accuracy that can be obtained with a specific type of acoustic model, using that data set. The effect of an increasing number of speakers and utterances is then analysed using this technique for acoustic models of different types of phones, and some interesting trends are observed.

The similarity technique we define here also allows us to understand the similarity between different phones, for example, the same nominal phone across languages. This is useful when combining training data across languages in order to compensate for a lack of sufficient training data, a useful strategy in resource-scarce environments. By evaluating model stability we can better understand whether the measured differences between models stem from an actual variance in the data, or from variability introduced by estimation errors, and also estimate whether different models are similar enough to support data sharing.

The paper is structured as follows: In Section 2 we discuss related work and provide some background on the Bhattacharyya bound. In Section 3 we describe the general technique we use for the analysis of model similarity and stability. In Section 4 we use this technique to analyse our data set, specifically with regard to the effect of an increasing number of speakers and utterances for different types of phones and types of acoustic models, and discuss the trends observed. Section 5 contains some concluding remarks.

2. Background

Data selection strategies for speech recognition purposes typically focus on selecting informative subsets of data from large corpora, with the smaller subset yielding comparable results [2]; or the use of active learning to improve the accuracy of existing speech recognition systems [3]. Both techniques provide a perspective on the sources of variation inherent in a speech corpus, and the effect of this variation on speech recognition accuracy.

In [2], Principle 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 different speakers and genders. Interestingly, the effect of utterance length has also been analysed as a main source of variation [3].

Active and unsupervised learning methods can be combined to circumvent the need for transcribing massive amounts of data [3]. 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 the optimisation of the amount of variation

inherent to training data is needed, 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 phonemes, improvements are obtained [4].

In the current work, the separability of two probability density functions is measured by a widely-used upper bound of the Bayes error, namely the Bhattacharyya bound [5]. If the Bayes error is given by

$$\epsilon = \int min[P_1 p_1(X), P_2 p_2(X)]dX \quad (1)$$

(with P_i and $p_i(X)$ denoting the prior probability and class-conditional density function for class i, respectively), the upper bound of the integrand can be determined by making use of the fact that

$$min[a, b] \leq a^s b^{1-s} \quad 0 \leq s \leq 1 \quad (2)$$

and is called the Chernoff bound, with s a parameter to be estimated through optimisation (Eq. 2 states that the geometric mean of two positive numbers is larger than the smaller one). If the condition for selection of an optimal s is relaxed by choosing $s = 0.5$, this simplified bound is referred to as the Bhattacharyya bound:

$$\epsilon = \sqrt{P_1 P_2} \int \sqrt{p_1(X) p_2(X)} dX \quad (3)$$

When both density functions are Gaussian with mean M_i and covariance matrix Σ_i, integration of ϵ leads to a closed-form expression for ϵ

$$\epsilon = \sqrt{P_1 P_2} e^{-\mu(1/2)} \quad (4)$$

where

$$\mu(1/2) = \frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (M_2 - M_1) + \frac{1}{2} ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \quad (5)$$

is referred to as the Bhattacharyya distance. For complex distributions, the Bhattacharyya bound can be estimated via Monte Carlo simulation.

3. Approach

We approach the task of analysing model estimation accuracy by first defining an appropriate similarity measure [6] and then defining a measure of model estimation stability based on this similarity measure. These two techniques are described below.

3.1. Measuring model similarity

From Eq. 3 and using the sample value of the expectation of the integral, we derive an estimator for the Bhattacharyya bound of two Gaussian Mixture Models. In practice we calculate:

$$\epsilon = \sqrt{P_1 P_2} \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} \sqrt{\frac{p_2(x_i)}{p_1(x_i)}} + \sum_{i=1}^{n_2} \sqrt{\frac{p_1(x_i)}{p_2(x_i)}} \right] \quad (6)$$

where x_i are the actual samples and both n_1 and n_2 are the number of samples with regard to each of the two probability densities respectively. For our purpose we assume that the prior values $P_1 = P_2 = 0.5$ and utilise equal numbers of samples drawn from each distribution. We ensure that we utilise a sufficient number of samples by first selecting a set of model pairs that cover a range of similarity values, and then evaluating the variance observed in the estimated bound between these model pairs over various runs (initiated with different sampling seed values) using an increasing number of samples per run. The number of samples is then selected where the variance across different runs falls below an acceptable threshold.

Note that the ϵ error bound can easily be converted to a distance measure using Eq. 4 but we find it more intuitive to work with the bound directly.

3.2. Measuring model estimation stability

In order to estimate the stability of an acoustic model, we separate the training data for that model into a number of disjoint subsets. All subsets are selected to be mutually exclusive with respect to the speakers they contain. For each subset, a separate acoustic model can be trained, and the Bhattacharyya bound between each pair of models is calculated. By calculating both the mean of this bound and the standard deviation of this measure across the various model pairs, a statistically sound measure of model estimation stability is obtained.

4. Analysis and results

4.1. Data and experimental setup

We use the November 1992 ARPA Continuous Speech Recognition Wall Street Journal Corpus as training data for our analysis. The dataset consists of 102 speakers recorded over the same channel. This enables us to experiment with up to 20 hours of data, which is comparable to the amount of data contained in the Lwazi corpus. In order to be able to control the number of phone observations used to train our acoustic models, we first train a speech recognition system and then use forced alignment to label all of the utterances.

We perform speech recognition using standard HMMs with three emitting states, tied across models, each containing up to 12 GMMs trained on 39-feature MFCC-based vectors (13 MFCCs, deltas and double-deltas with cepstral mean subtraction). Similar feature vectors are utilised in our analysis.

Using the process discussed in section 3.1, we estimate the number of samples required for our Bhattacharyya estimator and find that 20,000 samples are sufficient for our purpose. Table 1 summarises the number of samples required to keep standard deviations below a threshold of 0.0100 for the various model comparisons. With 20,000 samples, the standard deviation among different estimations of bounds between GMMs containing up to 6 mixtures are below 0.0020 for very similar phones and below 0.0061 for quite dissimilar phones. (Model pairs with Bhattacharyya bounds of approximately 0.5 and 0.1 respectively). We also find that with 20,000 samples and a single GMM, these estimates are within 0.0002 and 0.0020 from the corresponding analytically calculated values. We separate our data set into 5 disjoint subsets and estimate the mean of the 10 distances obtained between the various model pairs.

4.2. Initial analysis

During our initial analysis we develop speaker-and-utterance three-dimensional plots for acoustic models of different phone types at two levels of model complexity: a simple single Gaussian model (GMM with 1 mixture) and a complex 6-mixture

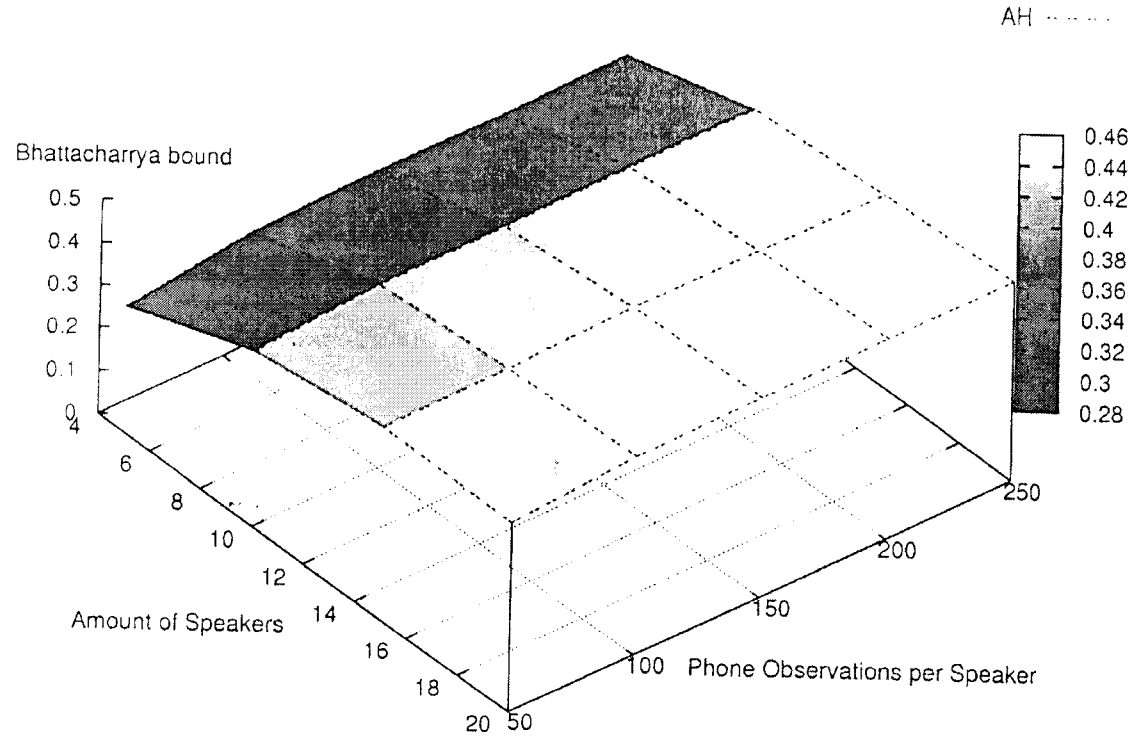

Figure 1: *Speaker-and-utterance three-dimensional plot for the phone /ah/*

Num mixtures	Samples required	ϵ	σ_s
1	5,000	0.1	0.0100
		0.3	0.0073
		0.5	0.0031
2	10,000	0.1	0.0062
		0.3	0.0060
		0.5	0.0018
4	20,000	0.1	0.0045
		0.3	0.0020
		0.5	0.0017
6	20,000	0.1	0.0061
		0.3	0.0045
		0.5	0.0020

Table 1: *Number of samples required for accurate estimation of bounds.*

GMM. (The choice to utilise a 6-mixture GMM was made to balance high speech recognition accuracy for our data set with computational requirements during bound estimation.) Each plot indicates the value of the Bhattacharyya mean, as described in Section 3.2, as a function of both the number of speakers in the training corpus and the number of phone occurrences per speaker. As the mean value shown is an estimate of the Bhattacharyya bound, this value should approach 0.5 once a model is fully trained on an optimal set of data. An example of such a plot for the phone /ah/ is shown in Figure 1.

From this analysis the following was observed: (1) A specific number of speakers and phone occurrences result in significantly different results for the different phones. (2) While phones from the different broad phone categories (such as vowels, plosives or fricatives) exhibit varying learning behaviour, phones within a specific phone group follow remarkably similar trends. (3) Similar trends are observed when utilising either the more simple or the more complex acoustic model.

These initial observations are explored further in the following sections for a number of broad phone categories. For each broad category, a number of representatives are selected to illustrate the trends observed. Specifically, the following phones are selected: /ah/ and /ih/ (vowels), /n/ (nasals), /l/ and /r/ (liquids), /d/ (voiced plosives), /t/ and /p/ (unvoiced plosives), /z/ (voiced fricatives) and /s/ (unvoiced fricatives), after verifying that these phones are indeed representative of the larger groups. Given (3) above, the next two sections first discuss trends obtained using the simpler model, before the effect of moving towards a more complex model is discussed in Section 4.5

4.3. Number of phone occurrences required per speaker

In this section we aim to understand whether a saturation point is reached after which additional examples of phones by a specific speaker no longer improve the accuracy of the speaker independent acoustic model for that phone. We therefore take a

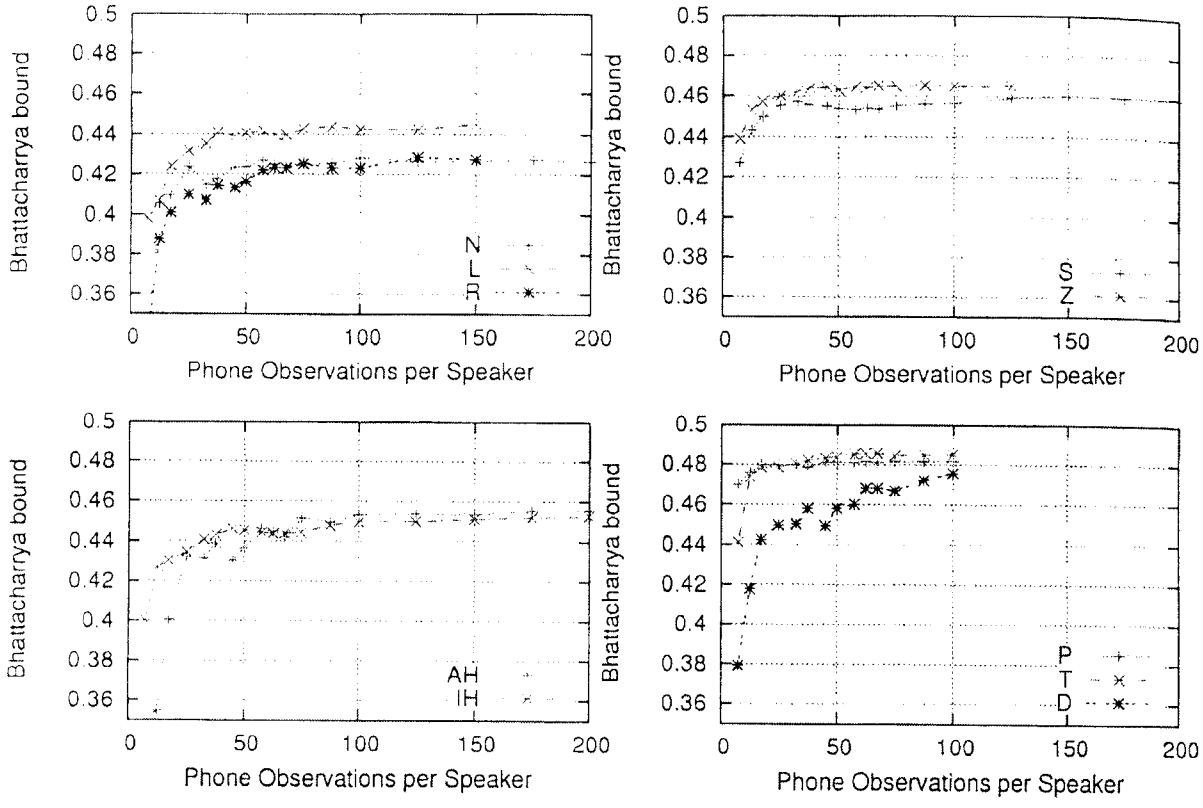

Figure 2: *Effect of number of phone utterances per speaker on mean of Bhattacharyya bound for different phone groups using data from 20 speakers*

cross-section of the 3-D plot in Figure 1, for a specific number of speakers (20) and evaluate the effect of increasing the phone observations per speaker. As clearly demonstrated in Figure 2, the means all reach an asymptote quite quickly and for 20 speakers, this asymptote does not yet approach the ideal 0.5 level for most of the phone types. When this experiment is repeated with 50 speakers, even fewer phone observations per speaker are required to reach the asymptote, and all the asymptotes are also nearer to the ideal level of 0.5. Interestingly though, the total numbers of phone observations necessary for the model of a phone to reach the asymptote are comparable for the 20 and 50 speaker cases. [1]

For the different phone types we observe that vowels are the slowest to reach the saturation point (at approximately 100 phone observations per speaker in the 20-speaker case) while unvoiced plosives and fricatives stabilise the most quickly, reaching this point at only 35 phone observations for /s/, 45 phone observations for /z/ and 25 phone observations for /p/ or /t/. There is a clear difference between the unvoiced and voiced versions of the plosives, with voiced versions taking significantly longer to stabilise (compare /d/ at 85 phone observations with /t/ at 25 phone observations). For most phones, those that saturate more quickly achieve a higher bound (closer to the ideal 0.5). However for some phones, such as /d/, a large number of phone occurrences are required per speaker, but the higher bound indicates that fewer speakers are required to obtain an accurate estimate. Similarly, the fricatives (/s/ and /z/) reach their asymptote very quickly, but this asymptote is fairly low, indicating low intra-speaker but high inter-speaker variability for this phone.

[1] Note that in order to be able to evaluate the effect at 50 speakers, only 2 models could be trained and 1 distance estimated, in comparison to the 5 models and 10 distances possible at the 20-speaker level.

4.4. Number of speakers required per phone

In this section we aim to understand the effect of adding additional speakers to a training corpus during acoustic model construction. We select a number of phone observations per speaker (100) where the asymptote has already been reached for all phones if 20 training speakers are employed. We construct a training set where we systematically add 100 observations for each new speaker. The results of this experiment are shown in Figure 3.

This time, the asymptote is not reached, and it is clear that additional speakers would improve the modelling accuracy for all phone types. On theoretical grounds we expect that the means should in all cases approach 0.5, and this expectation is supported by the observed trends. Again we observe that the unvoiced plosives and fricatives quickly achieve high values for the bound (close to the ideal 0.5). Low inter-speaker variability for the phone /d/ is also confirmed with high bound values. The high inter-speaker variability of the fricative phones

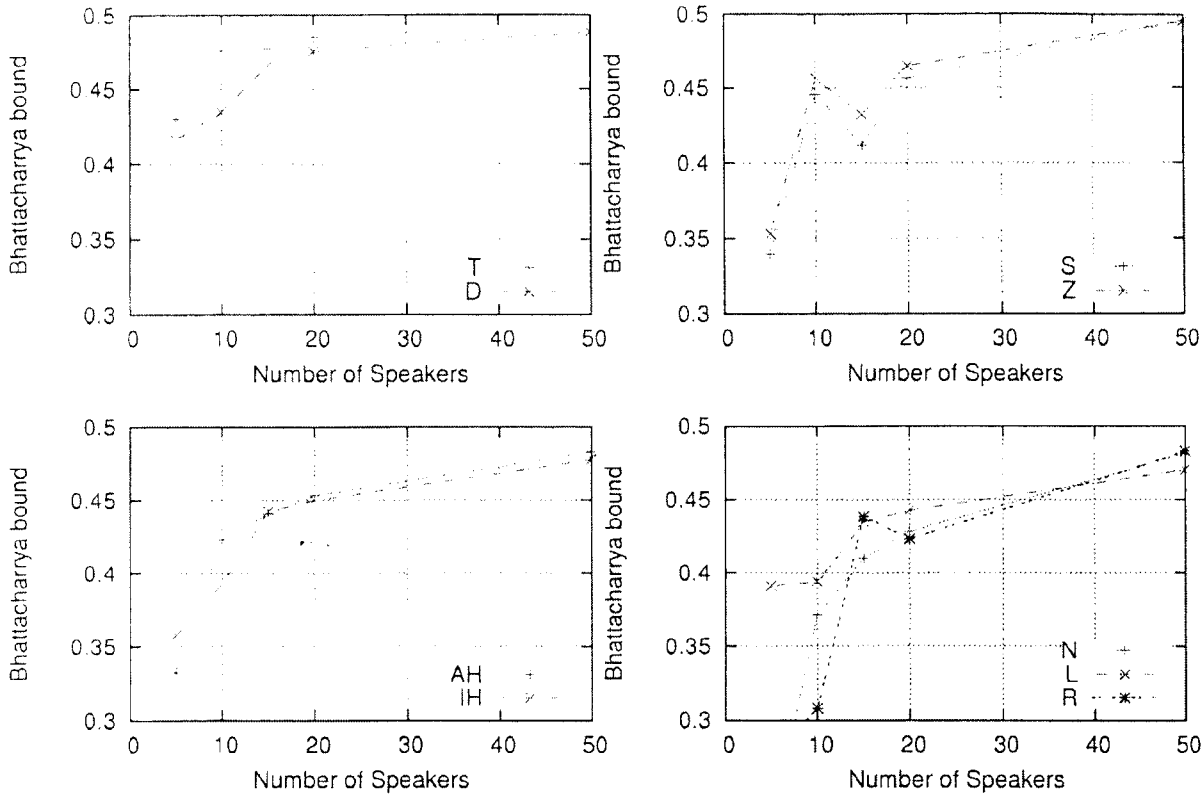

Figure 3. *Effect of number of speakers on mean of Bhattacharyya bound for different phone groups using 100 utterances per speaker*

are apparent in the unstable behaviour they exhibit (varying between 0.4 to 0.45 up to 20 speakers). Interestingly the vowels are not the slowest to reach large bound values as the speakers are increased: the phones /n/ (nasals) and /r/ (liquids) converge more slowly, signalling a higher inter-speaker variability for this group.

These results confirm the results obtained in Section 4.3 and comparative behaviour for the different phone types is summarised in Table 2.

Phone type	Inter-speaker variability	Intra-speaker variability
Unvoiced plosives	low	low
Voiced plosives	low	high
Unvoiced fricatives	medium	low
Voiced fricatives	medium	low
Vowels	medium	high
Nasals	high	medium
Liquids	high	medium

Table 2. *Comparative inter- and intra-speaker variability for different phone types.*

4.5. Effect of model complexity

The numerical values of the Bhattacharyya bound for different model types cannot be compared directly, since factors such as the existence of local minima during training increase the apparent variability of more complex models. We therefore compare such models by studying the observed bound values as a fraction of the observed asymptotic values. While the more complex model requires additional samples before the asymptote is reached, the same trends across phone groups are observed for more complex models. This is illustrated in Figure 4 where this fraction is shown, as the number of phone occurrences per speaker is increased. In these figures, data from 20 speakers is shown for both the simple single Gaussian model as well as the more complex 6-mixture GMM.

5. Conclusions

We have introduced a systematic approach that enables us to study the resource requirements for speech-recognition systems, based on the mean Bhattacharyya bound between models trained on different subsets of the data. We find that the different broad categories of phones have significantly different data requirements: whereas as few as 20 speakers and fewer than 50 samples per speaker are sufficient for the plosives /t/ and /d/, even 100 samples per speaker from each of 50 speakers do not describe the vowels, liquids or nasals adequately. Overall, the number of speakers for even a basic speaker-independent re-

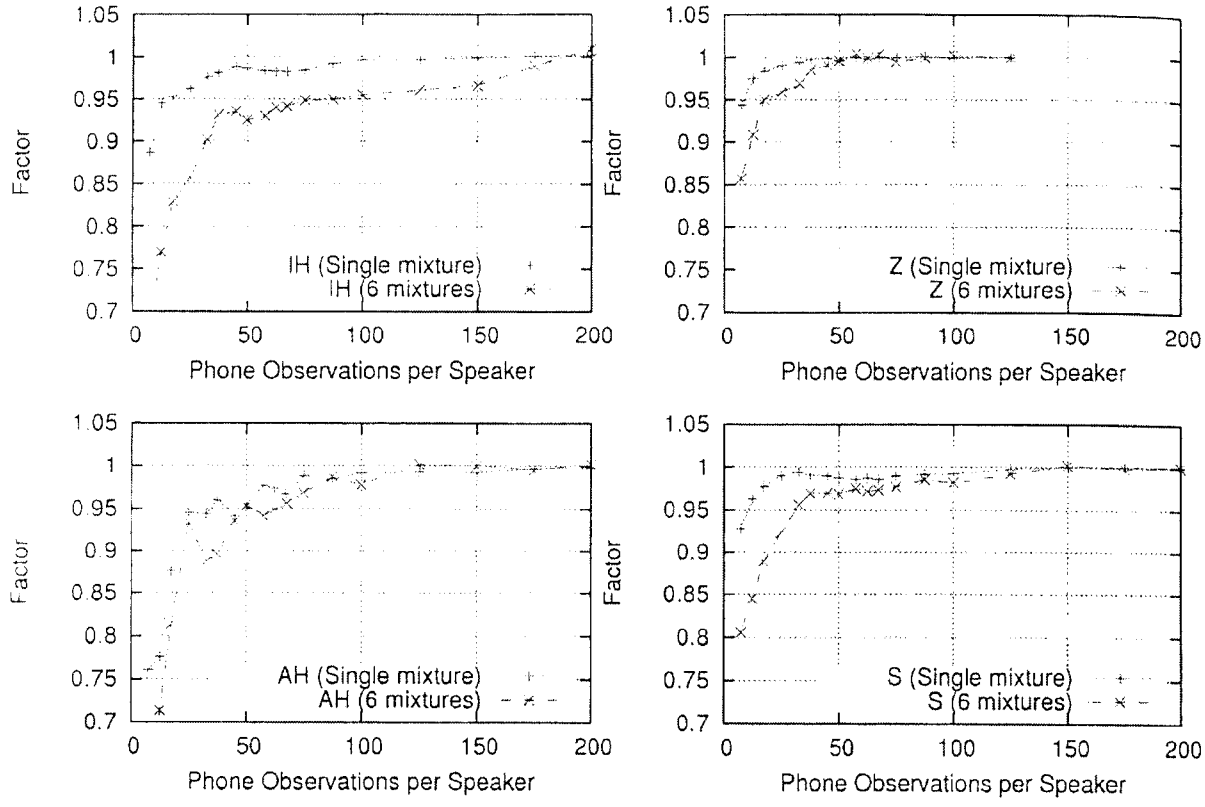

Figure 4: *Comparing the effect of model complexity on the relative distance to asymptote for two phone groups*

source collection therefore needs to contain significantly more than 50 speakers. (We are not able to suggest a reasonable lower bound based on the data used in this study.)

We found similar trends for simple and more complex models, with the more complex models requiring somewhat more speakers and phone occurrences to stabilise. Our work has focused on simple models, and can be extended in various directions. It would be interesting to see whether robust asymptotes are achieved as the number of speakers is increased; other variables, such as gender or speaking style should also be studied along with more complex models (e.g. context-specific models, multistate models such as HMMs and more complex density estimators). In our current work we are also investigating how the measurements described here relate to actual speech recognition performance obtained.

These insights are likely to play an increasingly important role as the reach of speech processing systems extends beyond the major languages of the world.

6. References

[1] "Lwazi ASR corpus," 2008, http://www.meraka.org.za/lwazi.

[2] A. Nagroski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, pp. 67–72, 30 Nov. - 3 Des 2003.

[3] G. Riccardi and D. Hakkani-Tur, "Active and unsupervised learning for automatic speech recognition," in *Proceedings of EUROSPEECH*, Genéva, Switzerland, 2003.

[4] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," *Automatic Speech Recognition and Understanding, 2007. ASRU. IEEE Workshop on*, pp. 562–565, 9 - 13 Des 2007.

[5] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Inc., 2nd edition, 1990.

[6] P.A. Olsen and J.R. Hershey, "Bhattacharyya error and divergence using variational importance sampling," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007, pp. 46–49.