Appropriate baseline values for HMM-based speech recognition

Etienne Gouws, Kobus Wolvaardt, Neil Kleynhans, Etienne Barnard

Department of Electrical, Electronic and Computer Engineering
University of Pretoria
Pretoria, South Africa

ebarnard@up.ac.za

Abstract

A number of issues related to the development of speech-recognition systems with Hidden Markov Models (HMMs) are discussed. A set of systematic experiments using the HTK toolkit and the TIMIT database are used to elucidate matters such as the number of mixtures to use for a particular training-set size, the utility of various feature sets, the value of triphone modelling, etc. These results suggest guidelines, which will be useful for those who wish to develop speech-recognition systems in new languages.

Keywords - Hidden Markov Models (HMM), Feature sets, Mixture models, Pronunciation dictionaries, Monophones, Triphones

1. Introduction

There is a growing awareness that Human Language Technologies can play a significant role in bridging the digital divide [1]. Thus, speech synthesis can be used to provide spoken output of stored information to illiterate users, speech recognition can efficiently obtain input from such users, and automatic translation can be used to provide information in a variety of languages. In each case, a key to successful application of the relevant technology is its adaptation to the home languages of the target users. Hence, it is likely that language-technology systems will be developed in a wide range of languages in the near future, and a variety of open software has been developed to support this expected trend (HTK, Sphinx, LLSTI, Festival, PublicVoiceXML). However, the availability of high-quality software is not sufficient for successful development of local-language technologies - there are also significant challenges in choosing (or developing) appropriate corpora, selecting appropriate system parameters, and so forth. It is therefore necessary to create guidelines that will assist new developers of language-technology systems in addressing these challenges.

In the current paper, we focus on speech recognition. We have undertaken a study to determine an approriate set of parameter settings for first-generation phone-based systems that use Hidden Markov Models (HMMS) [4], the dominant paradigm for current speech recognizers. Our experiments are based on the open-source toolkit HTK [5], but similar results are expected for any phone-based continuous HMM. Also, our experiments were restricted to the English-language TIMIT corpus [7]; we believe that similar results will hold for other languages, but plan to gain a better understanding of the influence of language-specific factors on the performance of speech-recognition systems in future work.

2. Experimental method

The HTK tookit (executing under the Linux operating system) was employed for all experiments. The basic steps for the development of a phone-based recognizer with HTK are as follows[5]:

1. A selected set of sound files from the TIMIT corpus are converted to a frame-synchronous feature representation (details of the training and test data used and the features studied are provided below).
2. Word-level and phone-level transcription files are constructed from the transcriptions provided with the TIMIT corpus, along with a pronunciation dictionary. (We experimented with the CMU and BEEP dictionaries - see below.)
3. Initial monophone models are created for all phones in the selected dictionaries, by using a flat start HMM prototype model (containing global means and variances of the speech data) and duplicating this prototype for each monophone to be used in training.
4. These models are then refined using the Baum-Welch algorithm to perform embedded re-estimation of all parameters [4].
5. The monophones are subsequently copied into a set of tied triphone models (we used the script driven editor ”HHEd” from the HTK toolkit for tying), and triphone models are computed using embedded re-estimation.
6. For testing purposes, a grammar is constructed as a sequence of the words that occur in the test data. Any word can follow any other word, and the sequence is of unrestricted length. This grammar is converted to a word lattice, which is used in conjunction with the same dictionary used for training and either the monophone or triphone models to assess recognition accuracy.

   HTK reports various result statistics for all recognized data. These include the percentage of test sentences that are recognized correctly, the percentage of test words recognized correctly, and the accuracy of the recognized words. (Accuracy is defined as the percentage of words correctly recognized minus the percentage of inserted words.) The results below are all expressed in terms of word accuracy.

2.1. Training and test data

All tests used the TIMIT speech corpus. TIMIT consists of 6300 utterances from 630 different speakers of American English, recorded with a high-fidelity microphone in a noise-free environment. The TIMIT data is divided into 8 sub-corpora, corresponding to speech from different dialect regions; each dialect region in turn contains a set of training speakers and a set of test speakers. We maintained this distinction between test and training speakers, and obtained our test and training data by uniform sampling across all dialect regions.

2.2. Pronunciation dictionaries

Speech recognition models are trained using the pronunciation dictionaries. These dictionaries are used to identify the phones used when words are pronounced and thus train the corosponding phone models using the corectly identified data. These dictionaries are also used during the recognition task to identify possible phone orders and the words they may result in.

The experiments concerning the 'monophone vs. triphone models' and the 'number of mixture components' were conducted twice. First using the BEEP dictionary (containing British standard English pronunciations) and secondly using the CMU dictionary (containing American standard English pronunciations customized for the data).

2.3. Comparing different feature sets

Virtually all approaches to speech recognition function by first converting the time-domain speech signal to a "feature" representation. Such a feature representation takes into account specific properties of the speech signal to represent it in a manner that is both compact and relatively invariant within a phonetic category [3]. It is clear that the details of the representation is a significant factor in the accuracy that can be achieved, and numerous representations have been proposed in the literature (e.g. Linear Predictive Coefficients (LPCs), Linear Predictive Cepstra (LPCepstra), log-scaled filterbank energies (FBANK) and Mel-Frequency Cepstral Coefficients (MFCCs), and Perceptual Linear-Predictive coefficients (PLPs) [2].

Additionally, for each representation, the change in the speech signal can be represented using temporal derivative and second-derivative information. The so-called delta and acceleration coefficients (which represent the first and second order regression coefficients respectively) are generally used for this purpose.

In order to develop guidelines on feature representations, we have experimented with LPCs, LPCepstra, filterbank energies and MFCCs; in each case the role of delta and acceleration coefficients was also studied. The number of basic coefficients (excluding delta and acceleration) was varied between 4 and 12 (by adjusting the order of the linear predictors, the number of filterbank elements, or the order of the cepstra, as appropriate). In all cases, the energy was used as an additional basic parameter – the number of basic parameters therefore varied between 5 and 13, and the total number of coefficients between 5 and 39.

2.4. The number of mixture components

Non-parametric density estimators are typically used to model the class-conditional probabilty density functions of HMMs [4]. Such estimators generally suffer from the bias-variance trade-off [5]: as increasing numbers of parameters are used to improve training-set accuracy, test-set accuracy eventually degrades (because the models overfit the training data). We use diagonal Gaussian mixture models as density estimators, and in that case the bias-variance trade-off is controlled by the number of mixture components in each density function.

Clearly, the optimal number of mixture components will depend on a variety of factors, such as the amount of training data available and the size of the input feature vector. In order to quantify this dependency for the TIMIT data, we have varied the number of mixture components between 2 and 16 for various experimental conditions. However, we use the same number of mixture components across all phonetic models in each particular run – optimizing the number of mixture components for each phone individually would probably improve accuracy a little, but would not alter the general trend that we wanted to explore.

2.5. Comparing monophones and triphones

The simplest approach to HMM-based speech recognition uses one Markov model to represent each phonetic category. However, it has long been understood ([4]) that – under appropriate circumstances – significant improvements in accuracy can be obtained by using separate models for phones in different phonetic contexts. For example, a word-initial s which is part of an str cluster has quite different acoustic properties from a word-final s which is preceded by a vowel; it therefore makes sense to devote separate models to the two cases. However, computing separate models for each context of each phone may require excessive amounts of data – especially since both the left and the right context may play a substantial role, necessitating the development of so-called $triphone$ models, which employ different models for each phone based on both its left and right context. (Thus, on the order of N^3 triphone models would be needed for a recognizer using N phones.) Techniques have therefore been developed to tie different triphone models together, based on the acoustic similarities of different contexts [5]; such tied triphone models are widely used in HMM-based speech recognition.

However, the development of triphone models has a number of drawbacks, such as the additional data required, and the knowledge and effort needed to perform successful tying. We therefore performed a number of experiments to determine how useful triphone models are (in comparison with monophone models) on various quantities of training data.

The number of TIMIT training sentences was varied between 400 and 3600 utterances (in intervals of 400), and the accuracy of both monophone and triphone models was measured for different numbers of mixture components.

3. Results

3.1. Pronunciation dictionaries

To compare the BEEP pronunciation ditionary to the CMU pronunciation dictionary (as discussed in 2.2) the results for triphone models using different mixture components was chosen. The results using both the BEEP and CMU dictionaries are plotted on one graph in Figure 1 (the red data represents the CMU dictionary and the blue data the BEEP dictionary). It can be seen that the CMU dictionary experiments outpreformed the BEEP dictionary experiments. This was expected as the TIMIT corpus contain American English speakers, thus an American pronunciation dictionary will more accurately represent what a speaker is saying and how it is being said.

3.2. Comparing different feature sets

The training and testing data used was obtained from the first four districts of the TIMIT corpus. Fourteen male and fourteen female speakers were chosen from each district, which created

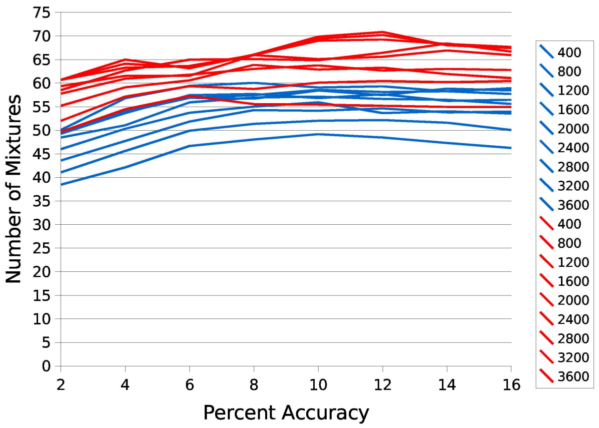

Figure 1: The accuracy of different mixture components for both the BEEP and CMU dictionaries.

the training set. The training set finally contained 713 sentences overall. For the test set, two male and two female speakers from the first four districts was chosen, which resulted in 128 test utterances in total. Our feature-set comparisons used triphone models with 8 mixture components. Results for the four feature sets discussed in Section 2.3 are shown in Figure 2. The MFCC analysis performed the best and most consistently, with no real gain in performance when the number of parameters were increased. The LPCEPSTRA analysis was the second best performing method. A gradual increased in performance was observed, which effectively stopped at total of nine parameters. The FBANK analysis accuracy measure merely decreased as the number of parameters increased but still outperformed the LPC method.

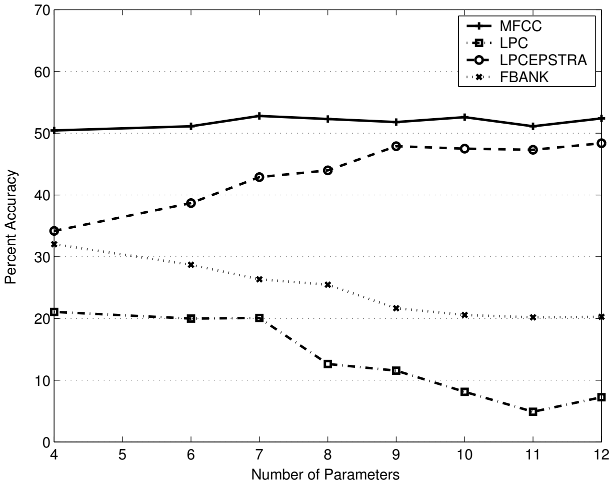

Figure 2: A graph showing the percent accuracy achieved by some of the different input speech parameterising methods used in HTK.

Knowing that the MFCC analysis gave the best results a second experiment could be performed. In figure 3 the inclusion of the delta coefficients gave a considerable increase in performance, but increases the number of parameters two fold. The acceleration coefficients also increased the performance measure but not as dramatically compared to the delta coefficients. With the inclusion of acceleration coefficients the amount of parameters increases three fold.

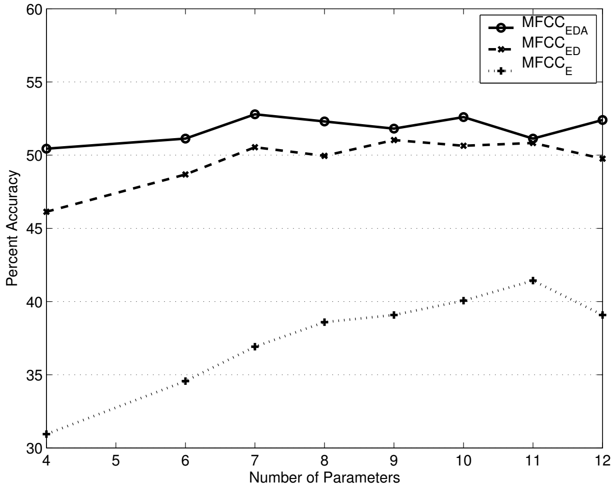

Figure 3: A graph showing the effect of including regression coefficients when parameterising the input speech signal. E = energy component, D = delta coefficient, A = accerleration coefficient.

3.3. The number of mixture components

As we discussed in Section 2.4 we varied the mixture components from 2 to 16. The results for tied triphone models are shown in figure 4. From the graph it can be seen that the optimal amount of mixture components resulting in the best recognition accuracy depends on the amount of training data. The best recognition preformance was obtained between 6 and 14 mixture components (dependant on amount of training data).

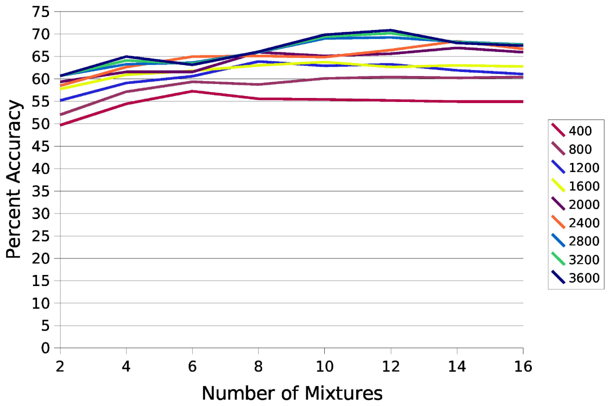

Figure 4: The accuracy of different amounts of training data as a function of increasing amounts of mixture components for tied triphone models (using the CMU pronunciation dictionary).

3.4. Comparing monophones and triphones

Figure 5 shows the word-recognition accuracy obtained with different amounts of training data, for both monophone and triphone data. (MFCC features and 1 mixture component were used.) The graph shows that triphones models outperform monophone models. Figure 6 is a repeat of figure 5 where 10 mixture models are used. The improvements gained from using 10 mixture triphones compared to 10 mixture monophones increases as the training data increases.

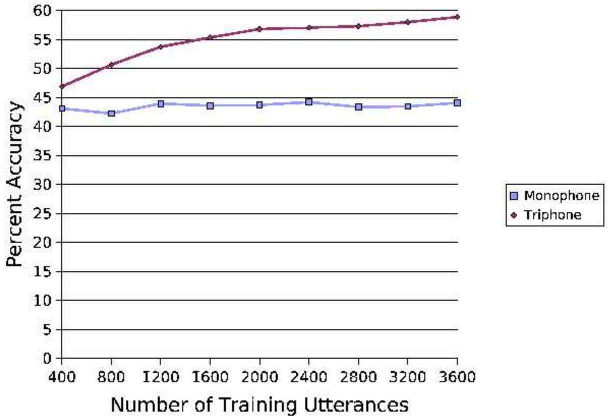

Figure 5: The accuracy of monophones and triphones (1 mixture) versus amount of data.

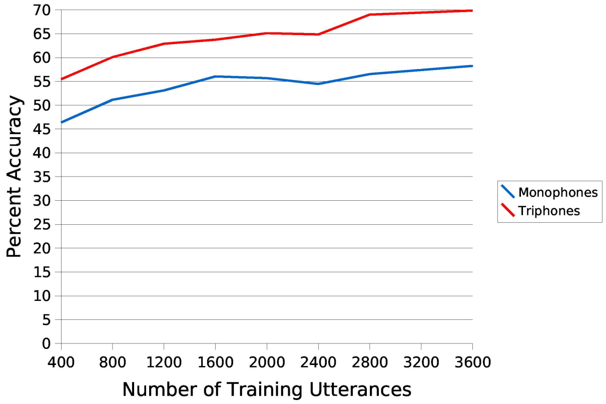

Figure 6: The accuracy of monophones and triphones (10 mixtures) versus amount of data.

4. Conclusion

In this paper, we evaluated parameter settings for first-generation phone-based systems that use Hidden Markov Models (HMMS). The parameters we evaluated includes: pronunciation dictionaries, data feature sets, mixture componenets and monophones vs. triphones. We used the HTK toolkit and the TIMIT speech corpus in our evaluations.

We compared the BEEP dictionary (British English) and CMU dictionary (American English). For the TIMIT corpus the CMU dictionary showed a 10% improvement (compared to BEEP) in recognition accuracy.

To obtain the best results from the HTK toolkit in the parameterising of the input speech signal phase the mel-frequency cepstral coefficients (MFCC) analysis should be used. It appears from the experimental results that this method extracts essential information to recognize a given set of words the best. Additional with this method's consistent performance a small amount of parameters can be used, which will give a good level of accuracy and a short amount of time needed to extract and process the parameters.

The results showed that the use of mixtures in speech recognition can considirably improve a recogniser's preformance. In single mixture recognisers, monophones reaches its maximum performance boundry with little training data and doesn't show any improvement if more training data is used. When adding mixture components, however, the performance of monophone models does increase in relation to the amount of training data. Triphones showed a simular improvement when mixture components are used, provided that the proportion between the amount of mixtures and training data is optimal.

Triphones outperform monophones when used for speech recognition. When using triphones, however, more training data is required to optimally train the recogniser than would be the case for monophones. Even when using mixture components triphones still outperform monophones by about 10%.

Even though the results discussed in this paper are restricted to the English-language TIMIT corpus, we believe that similar results will hold for other languages. Currently experiments are being conducted for other languages in order to gain a better understanding of the influence of language-specific factors on the performance of speech-recognition systems. The languages currently being investigated are IsiZulu and SePedi. The guidelines discussed in this paper are contributing to the developement of these speech-recognisers.

5. References

[1] E. Barnard, L. Cloete, H. Patel, "Language and Technology Literacy Barriers to Accessing Government Services", Lecture notes in Computer Science, Issue 2739, 2003, pp. 37 – 42

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738 – 1752, Apr. 1990

[3] L.R. Rabiner, and B.H. Juang, "Fundamentals of Speech Recognition", Pearson Education, 1993

[4] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, vol. 77, no. 2, Feb. 1989, pp. 257 – 285.

[5] S.J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.2)", Cambridge University, 2002.

[6] S.J. Young, N.H. Russell, and J.H.S. Thornton, "Token Passing: a Conceptual Model for Connected Speech Recognition Systems", CUED Technical Report F INFENG/TR38, Cambridge University, 1989.

[7] J.S. Garofolo, L.F. Lamel, "DARPA TIMIT Acoustic-phonetic Continuous Speech Corpus", U.S. Department of Commerce, 1993.