WESPAC IX 2006
The 9th Western Pacific Acoustics Conference
Seoul, Korea, June 26-28, 2006

A STUDY OF PHONE RECOGNIZER COMBINATION FOR HIGHER ACCURACY IN TIMIT PHONE RECOGNITION

Supphanat KANOKPHARA, Julie CARSON-BERNDSEN
School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland
E-mail: {supphanat.kanokphara,julie.berndsen}@ucd.ie

ABSTRACT

Generally, phone recognition system contains only a single phone recognizer. The phone set and speech representation for a recognizer are optimized for a particular task. This paper studies the effect of phone sets and speech representations for TIMIT phone recognition task. Two phone sets (TIMIT original phone set and 39 classical phone set) and two speech representations (MFCC-based and PLP-based) are tested. The phone recognizers for each phone set and speech representation are experimented and analyzed on both TIMIT training and testing sets. The results show that the 39 classical phone set with PLP speech representation phone recognizer yields highest phone accuracy. However, this best phone set works well only on a particular phone subset while the other recognizers work better on the other subsets. Therefore, the combination of several phone recognizers indicates higher phone accuracy than any single phone recognizer.

KEYWORDS: Speech Recognition, TIMIT phone set, Speech Representation, ROVER

INTRODUCTION

The motivation of this work comes from the fact that a committee of experts is typically better than a single super genius. This is better each expert is good on a particular part of information and/or able to cover error from other experts. In speech recognition, this truth has been proven in several speech recognition systems [1], [2]. In [1], the ROVER system was used to combine word recognition outputs from many word recognition systems. In [2], the system combined voting of several acoustic models trained from different acoustic-phonetic measurements.

However, we have to accept that when many experts share an idea, it is not necessary the

case that an expert will give only correct answers. Sometimes, a final correct answer can be altered by a wrong suggestion from an expert. Many researchers have been aware of this and have investigated many techniques to build robust systems able to select only correct answers from groups of recognizers. Therefore, the selection of appropriate recognizers is important in voting system design.

We believe that the appropriate recognizers for a voting system can be observed from the behavior of the recognizers on the training data. In this paper, four recognizers were tested with the well-known TIMIT corpus [3] on a phone recognition task. The four recognizers used in this paper are varied according to phone sets and speech representations. First, each recognizer is tested on the test set to see what type of phone set and what kind of speech representation are suitable for this task. Second, combinations of recognizers are tested on the training and test sets. The relation between testing on these two sets is observed.

TIMIT CORPUS PHONE SETS

The experiments use the standard TIMIT corpus consisting of 6300 sentences, 10 sentences spoken by each of 630 speakers, of which 462 are in the training set and 168 are in the test set. There is no overlap between the training and test sentences, except 2 SA sentences that were read by all speakers. The training set contains 4620 utterances and the testing set contains 1680 utterances. The core test set, which is the abridged version of the complete test set, consists of 192 utterances, 8 from each of 24 speakers. In this paper, the full training set without SA sentences is used as the training set while only the core test set is used as the test set.

Originally a phone set of 61 phones was used to annotate all phonetic transcriptions. However, another phone set which contains 39 symbols is more widely used since it gives higher accuracy and it is more suitable for speech recognition system [4]. In this paper, each recognizer is trained according to one of these phone sets.

SPEECH REPRESENTATIONS

HMMs are predominantly used as acoustic models in current speech recognition systems. The reason for this is that HMMs can normalize the time-variation of the speech signal and characterize the speech signal statistically in the optimal sense. HMM-based speech recognition systems do not deal with the speech signal directly but rather deal with a speech representation. A speech representation fulfils two roles in speech recognition systems. First, it reduces data in speech recognition processing. Second, it ideally presents only useful information from speech signal to the systems thus excluding information such as speaking style, noise, emotion, etc.

From the above reason, the choice of speech representation can also greatly affect the accuracy of the system. Nowadays, the two most frequently used speech analysis techniques are MFCC [5] and PLP [6]. The speech representations used in this paper are based on these two techniques. Normally, a practical speech representation is packed with one of these two techniques, speech energy and its dynamic features [7] (delta and double-delta). In this paper, two speech presentations (MFCC-based and PLP-based) are tested.

HMM-BASED PHONE RECOGNITION

The phone recognizers in this paper are constructed using HTK [8]. Each model contains 5 states and the covariance matrices of all states are diagonal (left-right model with no skip state). The training process starts from flat start training. Flat start training is a training strategy offered by HTK which requires no time-annotated transcription. After context-independent models have been trained, they are expanded to context-dependent models using a cross-word network. Phonetic decision trees [9] are then used to cluster the context-dependent states into classes according to the feature-table-based questions [10]. These classes are tied and trained together. From the context-dependent models, the number of model mixtures is increased by 1 until 10. All training processes are estimated by using maximum likelihood [11]. All training parameters are default. The number of training iterations after each change is determined automatically in line with [12].

The language model is trained from the training set using back-off bigrams. The language model provides phone constraints which correspond to the intra-phone-model transition probabilities. For the recognition process, the Viterbi algorithm is used without any pruning factor.

Phone recognizers are combined using ROVER. ROVER is known for allowing the combination of outputs from many recognizers. The system can be separated into two modules. The first module combines every output into a single transition network using dynamic programming alignment. The second module selects the best symbol from each transition in transition network. The selection criteria depend on the occurrence and/or the confidence score of the symbol in a transition. In this paper, the weight for occurrence is 0.7 while the weight for confidence score is 0.3. The confidence score for every null transition is fixed to 0.6.

EXPERIMENT

The experiment is separated into two phases. The first phase is to find the best individual phone recognizer as a baseline. The second phase is to find what recognizer contributes positive result to the final output in the voting system. In both phases, only four recognizers are tested, MFCC-based recognizer with 61 phone set, MFCC-based recognizer with 39 phone set, PLP-based recognizer with 61 phone set and PLP-based recognizer with 39 phone set. Even though the experiment in this paper is limited to only the TIMIT corpus and four recognizers, the structure of this experiment can be extended to support more types of recognizers and corpora. Note that for 61-phone recognizers, the recognized outputs are mapped to 39-phone outputs before comparison.

Table 1. *Single Phone recognizers' results*

	Train (%)	Test (%)
PLP&39 phones (A)	92.5	74.4
MFCC&39phones (B)	92.0	72.9
PLP&61phones (C)	90.9	73.3
MFCC&61phones (D)	90.5	73.4

Table 2. *Results of phone recognizer combination*

	Train (%)	Test (%)
A + B + C + D	92.5	75.4
A + B + D	92.8	75.7
A + B + C	93.0	75.7
A + B	91.9	74.3

Phase One. In this phase, each recognizer is trained on TIMIT training set and tested on TIMIT training and test set. Table 1 demonstrates the results of four phone recognizers. PLP-based recognizer with 39 phones yields highest accuracy and both training (92.5%) and testing sets (74.4%). The relations between training and testing set are roughly going in the same direction except for PLP&61phones and MFCC&61phones. The accuracy for the training set of PLP&61phones is higher than MFCC&61phones' while the accuracy for the test set of MFCC&61phone is higher. However, the difference is only 0.1%.

Phase Two. In this phase, every recognizer is combined using ROVER and tested on training and test set. Table 2 shows the results of combined phone recognizers. When four recognizers are combined together, the test set accuracy increases to 75.4%. This means that combined recognizers are better than any single recognizer. First, we tried to remove either C or D from the combination set because these two recognizers show almost the same accuracy in Table 1. The best result (93.0% for the training set and 75.7% for the test set) comes from the combination of three recognizers (A + B + C) on both training and testing sets. Then, we tried to remove C from the combination set. The accuracy decreases considerably (91.9% for the training set and 74.3% for the test set). The experiment was thus concluded at this point.

SUMMARY

The experiment showed interesting results. The selection of recognizers for voting systems can be observed from testing on a training set. The best result from the voting system is very promising. The accuracy of this system is even higher than the one from the combination system presented in [2]. In future work, we will evaluate this result on more corpora and additional tasks.

ACKNOWLEDGEMENTS

This material is based upon works supported by the Science Foundation Ireland for the support under Grant No. 02/IN1/I100. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

REFERENCES

1. J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)," *In Proc. IEEE ASRU Workshop*, 347-352 (1997)

2. A.K. Halberstadt and J.R. Glass, "Heterogeneous Measurements for Phonetic Classification," *In Proc. European Conf. Speech Communication and Technology*, 401-404 (1997)

3. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM* (NIST, 1993)

4. K.F. Lee and H.W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. Acoust., Speech, Signal Processing*, **37(11)**, 1641-1648 (1989)

5. S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. ASSP*, **28(4)**, 357-366 (1980)

6. H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *J. Acoust. Soc. Amer.*, **87(4)**, 1738-1752 (1990)

7. S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE-ASSP*, **34**, 52-59 (1986)

8. **http://htk.eng.cam.ac.uk/**

9. J.J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," *Ph.D. Thesis* (Cambridge University, Cambridge 1995)

10. S. Kanokphara, A. Geumann and J. Carson-Berndsen, "Accessing Language Specific Linguistic Information for Triphone Model Generation: Feature Tables in a Speech Recognition System", *In Proc. The 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 7-10 (2005)

11. B.H. Juang. "Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains," *AT&T Tech. Journal*, **64(6)**, (1985)

12. P. Tarsaku and S. Kanokphara, "A Study of HMM-Based Automatic Segmentations for Thai Continuous Speech Recognition System," *In Proc. the Symposium on Natural Language Processing*, 217-220 (2002)