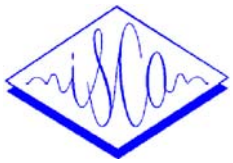

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

HMM STATE CLUSTERING ACROSS ALLOPHONE CLASS BOUNDARIES

Ze'ev Rivlin, Ananth Sankar, and Harry Bratt

Speech Technology And Research Laboratory
SRI International
Menlo Park, California 94025
U.S.A.
{zev,sankar,harry}@speech.sri.com

ABSTRACT

We present a novel approach to hidden Markov model (HMM) state clustering based on the use of broad phone classes and an allophone class entropy measure. Most state-of-the-art large-vocabulary speech recognizers are based on context-dependent (CD) phone HMMs that use Gaussian mixture models for the state-conditioned observation densities. A common approach for robust HMM parameter estimation is to cluster HMM states where each state cluster shares a set of parameters such as the components of a Gaussian mixture model. In all the current state clustering algorithms, the HMM states are clustered only within their respective allophone classes. While this makes some intuitive sense, it prevents the clustering of states across allophone class boundaries, even when the states are acoustically similar. Our algorithm allows clustering across allophone class boundaries by defining broad phone groups within which two states from different allophone classes can be clustered together. An allophone class entropy measure is used to control the clustering of states belonging to different allophone classes. Experimental results on three test sets are presented.

10.21437/Eurospeech.1997-59

1. INTRODUCTION

Most state-of-the-art large-vocabulary speech recognizers are based on context-dependent (CD) phone hidden Markov models (HMMs) that use Gaussian mixture models for the state-conditioned observation densities. Typically, such systems contain on the order of thousands of triphone models, and thus the amount of data for certain models often becomes too small to allow robust parameter estimation.

Approaches to robust estimation of the CD HMM parameters include Bayesian smoothing techniques [1], clustering of HMM models [2], and HMM state clustering [3, 4, 5]. The idea behind HMM state clustering is to cluster acoustically similar states and then train a separate set of Gaussians for each state cluster. The states in each cluster either share the same mixture Gaussian distributions [3, 4] or only share the same Gaussians but use different mixture weights for each state [5]. Sharing of parameters across acoustically similar states reduces the number of parameters to be estimated for the same amount of training data and level of acoustic resolution. The result is more robust parameter estimation.

In all current HMM state clustering algorithms, the states of triphones from different allophone classes (i.e., different middle phone) are separately clustered. This is motivated by the assumption that HMM states of phones from different allophone classes will not be acoustically similar and should thus not be allowed to cluster across allophone class boundaries. However, we believe that some states in different allophone classes are acoustically similar. For example, in a 3-state HMM system, we might expect the 3rd state of the vowel [aw], as in the word "house (h[aw]se)", to be similar to the 3rd state of the vowel [ow], as in the word "boat (b[ow]t)", because, phonetically, both are diphthongs with similar articulatory targets. We present an approach that permits clustering across allophone class boundaries by defining broad phone classes each comprised of a set of allophone classes. The decision of whether or not to carry out the cross-allophone-class clustering for a given pair of phone state clusters is based on a tradeoff between acoustic similarity and allophone class membership differences.

In Section 2 we introduce the new state clustering algorithm, in Section 3 we provide experimental results, and in Section 4 we present our conclusions.

2. ALGORITHM DESCRIPTION

The new state clustering algorithm is implemented within the context of SRI's DECIPHER^{TM} speech recognition system. This system currently uses a bottom-up agglomerative state clustering algorithm. The state clusters share a set of Gaussian distributions (or codebook) referred to as a "Genone" [5]. Each state within a cluster has a unique set of mixture weights (mixture weight distribution) for these shared Gaussians. In the current DECIPHER^{TM} system (our baseline), state clustering is permitted only within allophone classes. We define 38 allophone classes. The HMM states within each allophone class are clustered independently of the states in other allophone classes. Therefore, the resulting HMM state clusters only contain states from the same allophone class. This is also the case in other state clustering algorithms [3, 4]. However, as pointed out in Section 1, we believe that some states in different allophone classes are acoustically similar and thus should be allowed to cluster together.

The simplest way to allow cross-allophone-class clustering is to create a single class containing all the HMM

states and then perform agglomerative state clustering as in [5]. However, since the clustering algorithm needs to compute $O(N^2)$ distances, where N is the number of states, increasing the class size dramatically increases the time to compute the state clusters. To achieve clustering across allophone class boundaries within reasonable time, we defined three broad phone classes corresponding to the obstruents, sonorants, and the silence model. Each of these three broad phone classes consists of a set of allophone classes. Thus, clustering across allophone class boundaries is permitted without any restrictions within a given broad phone class. This scheme allows for increased robustness in parameter estimation due to combining of acoustically similiar states from different allophone classes. However, there is a potential for added confusability in recognition between allophone classes when phone states from different classes now share parameters. We address this issue below.

In the 38-class and 3-class systems described above, we use the weighted-by-counts increase in Gaussian mixture weight distribution entropy as defined in [5] to determine the distance between two state clusters S_a and S_b for potential merging to create the combined state cluster S. The distance measure is

$$d_1(S_a, S_b) = (n_a + n_b)H_1(S) - n_a H_1(S_a) - n_b H_1(S_b), \quad (1)$$

where $H_1(S)$ is the entropy of the Gaussian mixture weight distribution [5], and n_a and n_b are the counts from state clusters S_a and S_b, respectively.

In the case of cross-allophone-class clustering, there is a tradeoff between the robustness gained in parameter sharing for acoustically similar phone states and the potential confusion in phone recognition introduced due to merging of phone states belonging to different allophone classes. To adjust the tradeoff, we apply a penalty for cross-allophone-class clustering that is the weighted-by-counts increase in allophone class entropy:

$$d_2(S_a, S_b) = (n_a + n_b)H_2(S) - n_a H_2(S_a) - n_b H_2(S_b), \quad (2)$$

where

$$H_2(S) = -\sum_{r \in R} P(r|S) log P(r|S). \quad (3)$$

$H_2(S)$ is the entropy of the allophone class distribution $[P(r|S), r \in R]$, where R is the set of 38 allophone classes. The probability $P(r|S)$ is estimated as the fraction of the total counts for a given state cluster S from the allophone class r. $H_2(S)$ is a measure of allophone class mixing within a cluster S. Thus, minimizing the increase in $H_2(S)$ due to clustering penalizes the combination of states from different allophone classes.

Interpolating $d_1(S_a, S_b)$ and $d_2(S_a, S_b)$, we arrive at the combined distance measure:

$$d(S_a, S_b) = \alpha d_1(S_a, S_b) + (1 - \alpha) d_2(S_a, S_b), \quad (4)$$

where α is a linear interpolation coefficient optimized to provide the desired tradeoff between gained robustness in parameter estimation and added confusability between allophone classes. For example, when $\alpha = 1$, we combine the two state clusters whose combination introduces the minimum increase in Gaussian mixture weight distribution entropy, ignoring the allophone class membership of the phone states. At the other extreme, when $\alpha = 0$, we combine the two state clusters whose combination introduces the minimum increase in allophone class mixing. Values of α between 0 and 1 can be interpreted as combining state clusters based on acoustic similarity but applying a penalty for mixing states from different allophone classes.

In addition to the baseline acoustic model which used 38 allophone classes for state clustering, we built two more acoustic models, one with the three broad phone classes with unrestricted cross-allophone-class clustering, and one with the three broad phone classes with restricted clustering across allophone class boundaries via the allophone class entropy. In both cases, the states in each broad phone class were clustered using the agglomerative clustering algorithm, and separate Gaussian densities were trained for each state cluster as in [5]. All three acoustic models had roughly 2000 genones.

Experiments were performed using these acoustic models on a test set derived from the Wall Street Journal (WSJ) corpus [6] and gave encouraging preliminary experimental results. Further experimentation was then performed on another WSJ test set and a test set derived from the North American Business News Corpora (NABN), but improvement in recognition performance was not observed for these tests. Results are reported in Section 3.

3. EXPERIMENTAL RESULTS

Preliminary recognition experiments were performed on a 10-male-speaker subset of WSJ, each speaker uttering about 23 sentences, for a total of 230 sentences. A 20,000-word bigram language model was used for this test. This test set is denoted as 'WSJ-A'. We ran recognition experiments by rescoring word lattices [7] that were generated using the bigram language model and a previously trained acoustic model. The lattices were rescored to determine the best recognition hypothesis with the baseline acoustic model and the two new acoustic models.

The column denoted 'WSJ-A' of Table 1 shows the recognition performance on the WSJ-A test set using the baseline state clustering algorithm where states were clustered only within the 38 allophone classes (Baseline), the algorithm with three broad phone classes (Broad phone class), and the broad-phone-class approach with the use of allophone class entropy (Broad phone class with class entropy). Since the first two approaches do not consider allophone class entropy, this is equivalent to an α value of 1 for Equation 4, as indicated in Table 1. When class entropy was introduced for the WSJ-A experiment, the optimal value of α was found to be 0.7. Table 1 also indicates what percentage of the clusters consists of phone

Clustering Method	# Allophone Classes	WSJ-A	WSJS0-DEV94	NABN-DEV95
Baseline	**38**	**20.4%** ($\alpha = 1.0$, 0% mixed)	**10.5%** ($\alpha = 1.0$, 0% mixed)	**21.3%** ($\alpha = 1.0$, 0% mixed)
Broad phone class	**3**	**19.9%** ($\alpha = 1.0$, 28% mixed)	**11.1%** ($\alpha = 1.0$, 28% mixed)	**21.2%** ($\alpha = 1.0$, 28% mixed)
Broad phone class with class entropy	**3**	**19.3%** ($\alpha = 0.7$, 2% mixed)	**10.9%** ($\alpha = 0.7$, 2% mixed)	**21.2%** ($\alpha = 0.86$, 5% mixed)

Table 1. Word Error Rates for different HMM-state clustering algorithms

states from more than one allophone class (i.e., phone state clusters with nonzero allophone class entropy). As α decreases, the penalty for merging clusters with phone states from different allophone classes effectively increases, and there is less such merging. For example, when α decreased from 1.0 to 0.7 in the WSJ-A experiment, the percentage of mixed-allophone-class phone state clusters (% mixed) decreased from 28% to 2%.

The 5.4% reduction in word error rate (from 20.4% to 19.3%) is significant based on both the Matched Pairs test and the Wilcoxon Signed Rank test at the 5% level of significance.[1]

Based on these encouraging preliminary results, we conducted experiments with two more test sets to see if the performance improvement was generalizable. The first of these two is a 10-speaker male subset of the 1994 WSJ S0 development set [8] containing 209 sentences. A 5000-word bigram language model was used for this task. This test set is denoted 'WSJS0-DEV94'. The other test set is the Sennheiser microphone portion of the 1995 HUB3 NABN development set [9] denoted 'NABN-DEV95'. As seen in Table 1, we did not get an improved performance for either of these other two test sets.

One possible reason for this is the following. Introducing the allophone class entropy can be thought of as imposing a penalty for clustering across allophone class boundaries. Ideally, we would like to penalize cross-allophone-class clustering when the allophones involved are confusable in recognition. However, the penalty we use here depends on differences in allophone class membership of phone states from the two state clusters considered for merging, rather than considering actual phone confusion in recognition (e.g., in some cases, phone states of different allophone classes that share Gaussian distributions may nonetheless *never* be confused in recognition).

4. CONCLUSIONS

We have introduced an algorithm for HMM phone state clustering that allows phone states from different allophone classes to share Gaussian mixture distributions. This is achieved by defining three broad phone classes within which HMM states from different allophone classes can cluster. Acoustic similarity of states is measured by the weighted-by-counts increase in Gaussian mixture weight distribution entropy and an allophone class entropy is used to control clustering of states from different allophone classes. We interpolate the two measures of entropy to trade off consideration of acoustic similarity and allophone class mixing in making the clustering decision. In contrast to our system, the algorithms presented in [3, 4, 5] do not allow cross-allophone-class clustering.

Preliminary results on a WSJ test set were promising, but these results did not generalize to other test sets with which we experimented. We are presently looking into an improvement on this algorithm that uses a penalty for cross-allophone-class clustering which depends on actual phone confusability in recognition.

[1] Significance tests were conducted using the NIST Speech Recognition Scoring Package (SCORE), Version 3.6.2

ACKNOWLEDGMENTS

Support for this work from DARPA through the Naval, Command, Control and Ocean Surveillance Center Contracts #N66001-94-C-6046 and #N66001-94-C-6048 is gratefully acknowledged. The Government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Government funding agencies. We also gratefully acknowledge fruitful discussions with Vassilios Digalakis.

REFERENCES

[1] J. Gauvain and C.-H. Lee, "Bayesian Learning for Hidden Markov Models with Gaussian Mixture State Observation Densities," *Speech Communication*, vol. 11, 1992.

[2] K. F. Lee, *Automatic Speech Recognition – The Development of the SPHINX System*. Kluwer Academic Publishers, 1989.

[3] M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting Unseen Triphones With Senones," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. II–311 – II–314, 1993.

[4] P. Woodland, J. Odell, V. Valtchev, and S. Young, "Large Vocabulary Continuous Speech Recognition Using HTK," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. II–125 – II–128, 1994.

[5] V. Digalakis, P. Monaco, and H. Murveit, "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 281–289, 1996.

[6] G. Doddington, "CSR Corpus Development," in *Proc. DARPA SLS Workshop*, pp. 363–366, 1992.

[7] H. Murveit, J. Butzberger, V. Digalakis, and M. Weintraub, "Large-Vocabulary Dictation Using SRI's DECIPHER(TM) Speech Recognition System: Progressive-Search Techniques," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. II–319–II–322, 1993.

[8] D. S. Pallet, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, A. Martin, and M. A. Przybocki, "1994 Benchmark Tests for the ARPA Spoken Language Program," in *Proceedings of the Spoken Language Systems Technology Workshop*, pp. 5–36, 1995.

[9] R. Stern, "Specification of the 1996 Hub4 Broadcast News Evaluation," in *Proceedings of the DARPA Speech Recognition Workshop*, (Chantilly, VA), 1997.