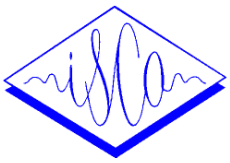

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SELECTIVE TRAINING OF HMMS BY USING TWO-STAGE CLUSTERING

S. Sato, T. Imai, H. Tanaka, and A. Ando

NHK (Nippon Hoso Kyokai; Japan Broadcasting Corp.)
Science and Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku
Tokyo 157-8510, JAPAN

ABSTRACT

This paper proposes a method of constructing acoustic models from training data clustered in two stages. In the first stage, training data from a target task are clustered and generate GMMs for each cluster. The second stage uses the GMMs to select training data from a large-scale database based on the GMM likelihood. MAP estimation adapts an acoustic model for each cluster using the selected training data. In decoding, the best acoustic model is selected from all acoustic models based on the GMM likelihood using some initial frames of an input utterance. Broadcast news transcription experiments showed that the proposed models achieved a word error reduction of 20% and a processing time reduction of 22%, compared with a non-clustered model.

1 INTRODUCTION

From March 2000, NHK (Japan Broadcasting Corporation) has been subtitling live news programs using a real-time speech recognition system. Several techniques were developed for the transcription system. (1) A time-dependent language model (TDLM)[1] is adapted to recent broadcast news. (2) A progressive 2-pass decoder [2] outputs the latest available results by making a quick decision during speech while maintaining high accuracy. (3) Gender-dependent state-tied HMMs were trained with a large-scale news speech database.

Currently, the automatic subtitling is restricted to sections of speech read by specified anchorpersons, because the use of common gender-dependent acoustic models to handle other speakers, such as reporters, would rise computational costs and reduce word accuracy to unacceptable degree. To extend captioning to reporters, acoustic models must be improved to handle a wider variety of speakers.

In order to adapt an acoustic model to a new speaker, multiple HMMs each adapted to a speaker cluster can be used. If a large-scale database with speaker labels (unlike our speech database for broadcast news) were available for training, speaker-cluster-dependent HMMs [3] could easily be created. If off-line decoding were allowed, all test data could be clustered to create the cluster-dependent HMMs [4]. In our broadcast news transcription system, however, real-time and on-line processing is required. One proposed on-line adaptation method [5][6] is to detect turn taking and incrementally adapt an acoustic model to the test utterance, which is relatively small. On the other hand, if a wide variety of adaptation utterances in our large-scale news speech database could be clustered efficiently, this would make acoustic models more accurate.

We propose new speaker-cluster-dependent acoustic models which deal with a large-scale database with no speaker labels and work on-line. They are based on 2-stage clustering and model adaptation. Section 2 describes the proposed method. An experimental setup for Japanese broadcast news recognition is described in Section 3. Evaluation results of the proposed method are described and discussed in Section 4 with an analysis of applicability to our real-time system.

2 TWO-STAGE CLUSTERING AND ADAPTATION

2.1 Overview

To select adaptation data, we adopt a 2-stage clustering using a Gaussian Mixture speaker Model (GMM) [7] that represents utterances in each cluster. From among the HMMs adapted by the data in each cluster, a decoder selects the best acoustic model using the initial part of input utterances. Figure 1 gives an overview.

Figure 2 illustrates the timescale of data used. Data (a) is the target news program to be recognized. The first stage clusters a small database (b) gathered from past editions of the same program as (a) and generates corresponding GMMs for each cluster. The second stage clusters a large-scale database (c) of any TV programs using these GMMs. The adaptation model for each cluster is trained by the utterances in the cluster.

2.2 First Stage of Clustering

The small database (b) used in the first stage consists of the same TV program as the target program, compiled over a period of several days. Speech data in the programs are chopped into sentences and they are acoustically analyzed into feature vectors as training data on a sentence basis. The clustering procedure is followed by the repetition of two operations: aligning each datum to the most probable cluster and estimating GMM parameters. The detailed procedure is as follows.

(1) Assign each datum to any cluster randomly. The number of mixtures in the initial GMM is set to 1.

(2) Estimate the GMM parameters (means and variances) by the maximum likelihood algorithm with the data assigned to the cluster.

(3) Reassign each datum to the cluster which gives the highest GMM likelihood among all the clusters.

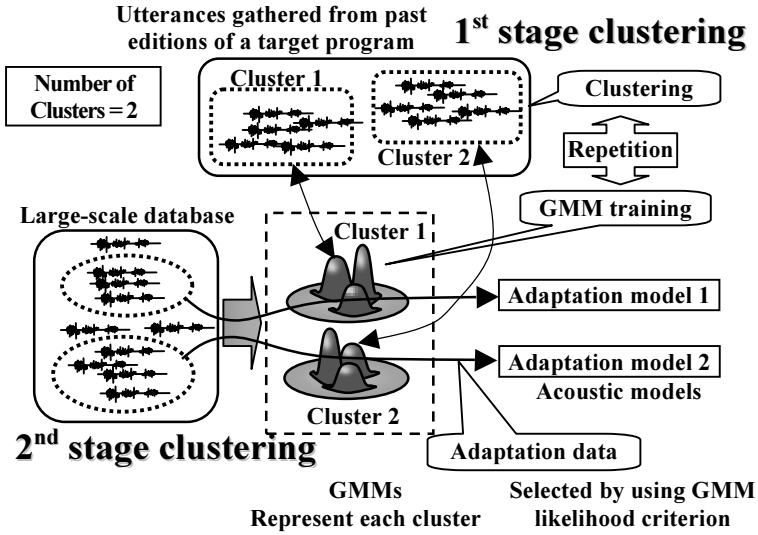

Figure 1: Overview of the proposed method. Clustered adaptation data were selected from a large-scale database in two stages. The first stage generates GMMs from a small database and the second stage clusters a large-scale database.

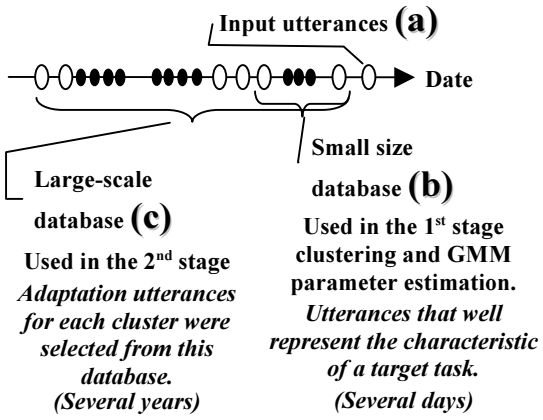

Figure 2: Timescale of each database used for adaptation.

(4) Repeat the procedure (2)-(3) until no migration of data among the clusters occurs or a specified number of iterations is reached.

(5) Increment the number of mixtures and repeat the procedure (2)-(4). When the number of mixtures reaches the desired figure, terminate the first stage of clustering.

2.3 Second Stage of Clustering

The second stage uses the large-scale database (c), compiled from many types of programs including the target program to cover a wide variety of speakers. The utterances in the large-scale database are also analyzed into feature vectors on a sentence basis. Each one is merged into the cluster which gives the highest GMM likelihood among all the clusters, if the likelihood exceeds a threshold. The threshold is set to μ_c-$2\sigma_c$, where μ_c is the sentence-averaged likelihood and σ_c is the standard deviation of cluster *c*. It corresponds to a 97% interval for each cluster. Due to the threshold, data which are unnecessary for adapting speaker-cluster-dependent HMMs are discarded. The clustering is performed faster than conventional clustering methods, because it requires only the likelihood computation of each datum against each GMM.

2.4 Adaptation

After selecting training data for each cluster, gender-dependent HMMs are adapted to each cluster by MAP [8] estimation. There are two reasons to adopt the MAP adaptation. One is that there may not be enough data to train HMMs in a maximum likelihood way when the number of clusters is larger. The other is that the same tying structure is advantageous when switching to the most appropriate cluster-dependent HMM in decoding, because it requires only the change of Gaussian parameters (means, variances and weights) without changing a tree-structured phoneme network of tied HMMs.

2.5 Decoding

Our broadcast news transcription system is required to output a result almost instantaneously. Since the adapted cluster-dependent HMM is expected to decode input sentences faster than the original model, the time taken to select the most appropriate model before starting recognition will be compensated for to some extent. We use a short (Δt) fragment of the data at the beginning of the input utterances to determine the acoustic model. The fragment must, however, be long enough to maintain the accuracy of the model selection. The selected HMM is the one whose cluster gives the highest GMM likelihood for the short part of the input utterances. The model selection using GMMs has a low computational cost.

3 EXPERIMENTS

3.1 Setup

We carried out Japanese broadcast news transcription experiments. For the evaluated speech data ((a) in Figure 2) we used 335 sentences consisting of 8,489 words in total chosen from four news programs aired on Sep. 30, 1998. They were uttered by 7 anchormen and 7 male reporters. The first stage of clustering used 3,091 utterances in the same four news programs from Sep. 20 through 29 in 1998 ((b) in Figure 2). The number of mixtures was increased to 32 for each GMM. The number of clusters was set to 2, 4, 8, 12 or 16. As a large-scale database for the second stage clustering ((c) in Figure 2) we used 68,101 utterances in news programs broadcast from June 1996 through Sep. 29, 1998.

All the data used in the experiments were analyzed into 39 parameters (12 MFCCs with log-power and their first- and second-order regression coefficients) every 10msec after digitization at 16kHz and 16bits with a Hamming window of 25msec width. Baseline models were trained by using the same data in the second stage. They were state-tied triphone HMMs constructed by tree-based clustering with 42 Japanese phonemes. The numbers of HMMs were 5,898 logical and 3,441 physical, with 2,787 states for a lexicon of 20K words. We also used this model as a prior density of the MAP estimation.

We used our 2-pass decoder [9] as follows. The first pass, using a bigram language model and triphone HMMs, generates a word lattice time-synchronously by Viterbi beam search. At a sentence-end the word lattice is recursively traced back to get N-best sentences. The second pass rescores the N-best sentences by a trigram language model to decide and output the best sentence. We also used a TDLM (Time Dependent Language Model) as a language model trained from NHK's news scripts extending back over 7 years with latest news more heavily weighted, after Japanese morphological analysis. Four different TDLMs were trained for the four different news programs in the test set. The perplexities of the trigram models for the evaluated sentences were 11 to 54. The rate of words outside of the vocabulary was 0.75% to 1.13%. The evaluation was executed on an Alpha-21264 500MHz machine with 1GB memory.

First, to find the optimum number of clusters, we tried values of 2, 4, 8, 12 and 16 (experiment 1). We selected the best acoustic model for each whole utterance. Since the model selection is done after the sentence end, the recognition does not work in real-time. It is probably the least likely case that the model selection would make an error. The recognition results were compared with the baseline acoustic model whose word error rate (WER) was 7.95% and the real time factor (RTF), the ratio of processing time to the length of input speech, was 0.96.

Second, we tried to decrease the length of data for the model selection in order to reduce the total time of the selection and decoding so as to obtain real-time operation. We varied the fragment length (Δt) at the beginning of input utterances from 0.1 to 2.0 sec (experiment 2). Decoding starts after the selection of HMMs corresponding to the most probable GMM according to the Δt fragment. In this experiment, a real-time factor (RTF) was evaluated by the processing time calculated as the sum of Δt and the time required for decoding.

3.2 Results

The result of experiment 1 is illustrated in Figure 3. White dots denote WER on the left axis and black dots denote RTF on the right axis. The result showed that WER was reduced as the number of the clusters increased from 2 to 12, but the 16-cluster system showed worse WER than the 12-cluster system. We could conclude that the best number of clusters tested was 12 in this condition, with a 20% reduction of WER compared with the baseline method. The result for RTF showed the same tendency as the result for WER. The most effective number of clusters for RTF was also 12, with a 23% reduction. An HMM which reduced WER also improved RTF because the beam search was based on a likelihood threshold and a better HMM reduced the number of confusable words within the threshold.

WER and RTF in experiment 2 are illustrated in Figure 4. The fragment length (Δt) used in the model selection is displayed in the abscissa axis. White dots denote WER on the left axis. Black dots denote RTF on the right axis. It should be noted that RTF displayed here includes the time required to get a fragment (Δt) and the time required for decoding. The RTF linearly increased as the length increased. The WER did not change much when the fragment length was between 0.5 sec. and 2.0 sec. compared with the case when the selection was done on the whole sentence. But a fragment length of less than 0.5 sec. degraded WER due to mis-selection of acoustic models. Therefore, we conclude that the best fragment length was 0.5 sec. which gave 20% reduction of WER and 22% reduction of RTF compared with the baseline method.

Figure 5 illustrates WER on a varied fragment length for different numbers of clusters from 2 to 16. The 12-cluster system showed the best performance at all fragment lengths. Stable selection of an appropriate HMM was possible from shorter fragments as the number of clusters smaller.

4 CONCLUSION

We proposed a new acoustic model adaptation method using a fast 2-stage clustering based on GMMs. The 2-stage clustering algorithm runs fast on a large-scale database. At the second stage, using the GMM likelihood threshold, undesired adaptation utterances in the database were discarded to improve the accuracy of each acoustic model. In decoding, because the utterances in the clusters are modeled by GMMs, the appropriate acoustic model could be determined fast using only a short fragment of an input utterance.

We applied the proposed adaptation model in a broadcast news transcription experiment. The result showed that a 12-cluster system showed the best performance; the appropriate acoustic model could be determined with a 0.5 sec fragment of the input utterance, and reductions of 20% in word error rate (WER) and 22% in real time factor (RTF) were achieved relative to the baseline.

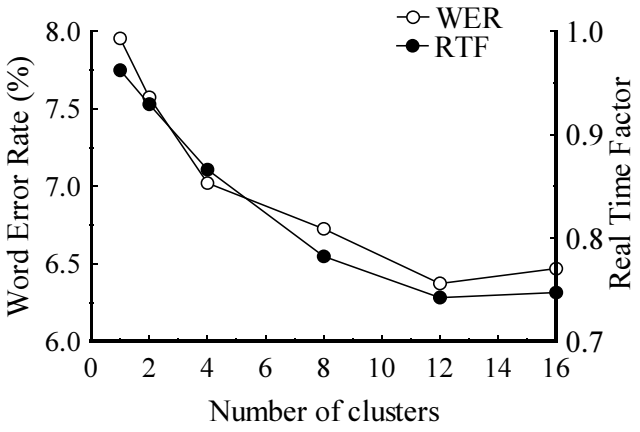

Figure 3: WER and RTF in experiment 1. The best performance was observed for the 12-Cluster system. WER was reduced to 80% and RTF reduced to 77% compared with the baseline (i.e. the values for one cluster).

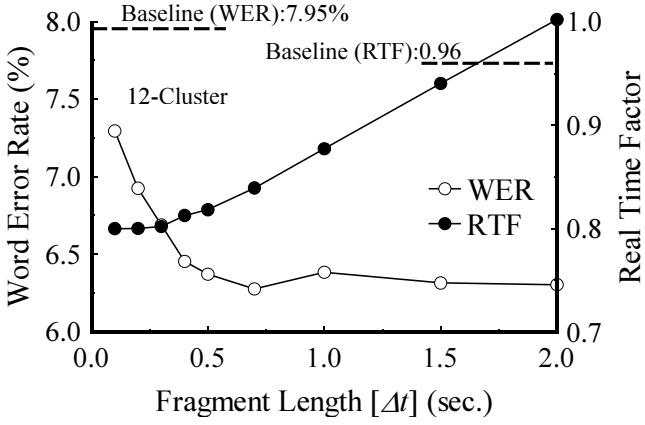

Figure 4: WER and RTF in experiment 2 (12-cluster). In the case Δt=0.5, WER was reduced by 20% and RTF by 22% relative to the baseline and WER did not change much when Δt>0.5.

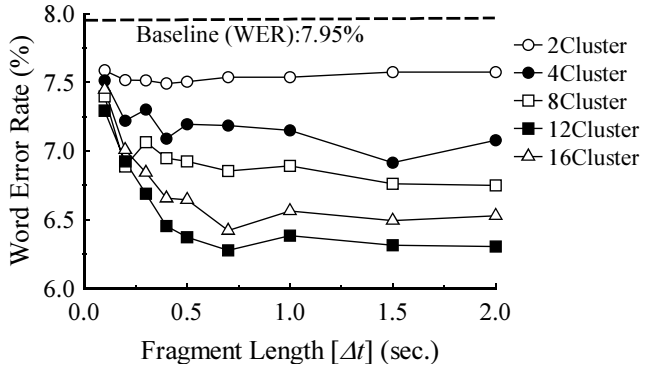

Figure 5: WER parameterized by number of clusters in experiment 2. Under almost all conditions the 12-cluster system showed the best performance.

5 REFERENCES

[1] A. Kobayashi, K. Onoe, T. Imai, and A. Ando, "Time Dependent Language Model for Broadcast News Transcription and Its Post-Correction", *Proceedings of International Conference on Spoken Language Processing*, pp.2435-2438, Dec. 1998.

[2] T. Imai, A. Kobayashi, S. Sato, H. Tanaka, and A. Ando, "Progressive 2-Pass Decoder for Real-time Broadcast News Captioning", *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 1937-1940, Jun. 2000.

[3] Y. Gao, M. Padmanabhan and M. Pichey "Speaker Adaptation Based on Pre-clustering Training Speakers" *Proceedings of Eurospeech'97*. vol.4, pp.2091-2094

[4] S. E. Johnson, P. C. Woodland. "Speaker Clustering using Direct Maximization of the MLLR-adopted Likelihood", *Proceedings of International Conference on Spoken Language Processing*, vol.5. pp. 1775-1778, Dec. 1998.

[5] K. Ohtsuki, S. Furui, N. Sakurai, A Iwasaki and Z. Zhang, "Language Modeling and Acoustic Modeling for Automatic Transcription of Japanese Broadcast-News Speech", *Technical Report of the Institute of Electronics, Information and Communication Engineers*, SP98-108 Dec.1998 (in Japanese).

[6] N. Murai and T. Kobayashi, "Dictation of Multiparty Conversation Using Statistical Turn Taking Model and Speaker Model", *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pp. 2908-2911. Jun. 2000.

[7] Reynolds D.A and Rose R.C, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Trans. S.A.T* vol.3 No.1 pp72-83 Jan.1995.

[8] J.L Gauvian, C.H.Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains", *IEEE Trans. S.A.P.* vol.2, No.2 pp291-298, 1994.

[9] T. Imai, K. Onoe, A. Kobayashi, and A. Ando, "A Decoder for Broadcast News Transcription", *Proceedings of Autumn Meeting of the Acoustical Society of Japan*, 2-1-12, Sep. 1998 (in Japanese).