Soft-clustering Technique for Training Data in Age- and Gender-independent Speech Recognition

Daisuke Enami*, Faqiang Zhu*, Kazumasa Yamamoto* and Seiichi Nakagawa*
* Department of Computer Science and Engineering, Toyohashi University of Technology, Japan
E-mail:{enami, faqiang, kyama, nakagawa} @slp.cs.tut.ac.jp

***Abstract*—In this paper, we propose approaches for the Gaussian mixture model (GMM) based soft clustering of training data and the GMM- or/and hidden Markov model (HMM)-based cluster selection in age and gender-independent speech recognition. Typically, increasing the number of speaker classes leads to more specific models in speaker-class-dependent speech recognition, and thus better recognition performance. However, the amount of data for each class model is reduced by the increase in the number of classes, which leads to unreliable model parameters. To solve the problem of the reduction of training data, we propose a GMM-based soft clustering method that allows overlap, and a selecting method for selecting a speaker model using a GMM or/and HMM. In an experiment, we obtained a 5.0% absolute gain for word error rate (WER), and a 24.9% gain for the relative WER over an age- and gender-dependent baseline.**

I. Introduction

Recently, speech processing technology performs reasonably well, and speech-related systems are regarded as favorable human-machine interfaces. To allow use by many users, speaker-independent speech recognition systems have been developed and incorporated in many products.

In speech recognition, since a system can not identify the speaker and speech environment in advance, there is a problematic reduction in speech recognition performance owing to the mismatch of the input speech and the acoustic model training data. To attain the performance required for a recognition system, an acoustic model that can take into consideration various speakers and speech environments is essential[1]. Against this background, high-quality acoustic modeling and speaker adaptation were carried out within a speech recognition system.

The clustering of training data has commonly been employed for high-quality acoustic modeling. Sankar[2] and Kosaka[3] et al. clustered training data using a cluster tree. More recently, the i-Vector-based approach was reported by Zhang et al[4]. There are also techniques that build a more detailed model by increasing the number of model classes. Increasing the number of model classes reduces the amount of data for each model, and the reliability of the model is thus reduce. To solve this problem, Jouvet et al. presented a margin classification method[5], and a variety of methods of clustering-based data sampling have been proposed[6].

As adaptation methods for model parameters, maximum likelihood estimation (MLE), maximum a posteriori (MAP) estimation[7], and maximum likelihood linear regression (MLLR)[8] have commonly been used. The effectiveness of these adaptation methods depends on the amount of adaptation data, and the methods often used in combination through interpolation. In [9], Gomez et al. proposed an unsupervised speaker adaptation method based on HMM sufficient statistics using linear interpolation and speaker selection. Although this method is a novel approach, it is difficult to prepare many speaker-dependent HMMs.

In this paper, we propose a GMM-based soft-clustering technique for training data that allows overlap to avoid a reduction in the amount of data when increasing the number of classes, and a technique for selecting a class-dependent model using the GMM or/and HMM. In the experiment, we obtained a 5.0% absolute gain for the WER, and a 24.9% grain for the relative WER over an age- and gender-dependent baseline.

The remainder of the paper is organized as follows; In the next section, the database and baseline for our experiment are described. A soft-clustering technique for training data is then introduced in Section III. In Section IV, we explain the experimental setup and results. Finally, Section V presents our conclusions and some future works.

II. Database and Baseline

For an age- and gender-independent speech recognition system, we used three types of corpora. The data for the elder class is the Senior-Japanese Newspaper Article Sentences (S-JNAS) [1] database consisting of 151 male and 150 female speakers aged 60 to 90 years. The data for the adult class is the ASJ+JNAS[2] database consisting of each 153 male and female speakers aged 20 to 60 years. The data for the child class is the CIAIR-VCV[3] database recorded by NAGOYA University. CIAIR-VCV consists of 145 male and 143 female speakers aged 6 to 12 years.

As the baseline of our experiment, we first classified all data introduced above into six classes on the basis of age and gender. Each corpus contains male and female speech data; hence, we divide the training data into those for elder-male (E-M), elder-female (E-F), adult-male (A-M), adult-female (A-F), child-male (C-M), child-female (C-F); i.e., a total of six classes. We then trained GMM and HMM acoustic models for the six classes as for the baseline models.

[1] http://db.ciair.coe.nagoya-u.ac.jp/dbciair/koureisha_files/index.htm (in Japanese)
[2] http://www.mibel.cs.tsukuba.ac.jp/_090624/jnas/instruct.html
[3] http://db.ciair.coe.nagoya-u.ac.jp/dbciair/dbciair2/kodomo.htm (in Japanese)

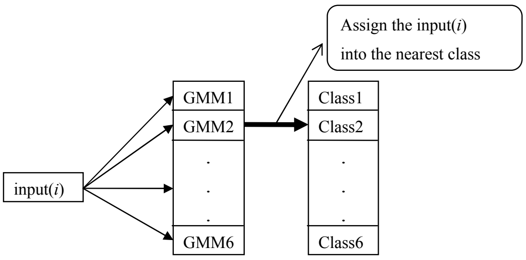

Fig. 1. Re-classification method

III. Clustering of training data

A. Re-classification

The method that only considers the information of age and gender used in Section II does not directly relate to speaker characteristics. For example, the acoustic characteristics of some 60-year-olds (Elder-class) may be similar to those of a 40-year-olds (Adult-class). To obtain more appropriate classes, we use the feature-based method to re-classify the training data into six clusters. Employing this method, we initially use the GMM models classified/trained in Section II. For each sentence or speaker in the training, we calculate the similarity (likelihood) of the six types of GMM, and then choose the best one as the target class.

Figure 1 shows the re-classification method. We re-train the six types of GMM and HMM using these classified data using the same method used for the initial six-class clustering.

B. Increase to 20 classes

In this section, we describe how we increase the number of classes from 6 to 20. First, we cluster the training data to 20 classes according to age, gender and microphone type. Considering a balance data quantity, we classify the data into packages. In the child corpus, we classify the data into three packages (6–7, 8–9, 10–12 years old) according to age for males and females. In the adult corpus, we also classify the data into three packages (10–20, 30, 40–50 years old), and for the elder corpus, we classify the data into four packages (60–70, 70–90 years old, and each microphone headset (HS) and desktop (DT)) accoding to age and the two type of HS and DT for males and females. We then randomly select 800 sentences from each class to train GMMs that are used to represent the characteristics of the class, and we refer to these GMMs as initial 20-GMMs.

However, the quantity of training data classified in the manner described above is insufficient for HMM training in some classes. To resolve this problem, we use a soft clustering technique that permits overlapping between classes. We adopt three parameters (rs, t_{min}, t_{max}) to control the overlap among classes, where rs denotes the relative difference in the similarity score among the best target class and the candidate target classes (the similarity score is calculated with the initial 20-GMMs), and t_{min} and t_{max} represent the minimum and maximum numbers of the assigned target classes, respectively. We construct the overlap clustering method as follows:

I: the number of training data.
n: the number of candidate target classes.
rs: relative score.
sc: similarity score between current training data and each GMM.

1) For $i = 1$ to I, set $n = 0$; Execute steps 2) 3) 4).
2) Calculate the similarity score of initial 20-GMMs and sort them from best to worst.
3) For $j = 2$ to 20
   if $sc_{(j)} - sc_{(1)} < rs$, $n = n + 1$;
4) If $n < t_{min}$, assign training data x_i from class 1 to class t_{min};
   if $n > t_{max}$, assign training data x_i from class 1 to class t_{max};
   if $t_{min} \leq n \leq t_{max}$, assign training data x_i from class 1 to class n.
5) Train the 20-class GMMs and HMMs

In this process, we set $rs = 0.5$, $t_{min} = 2$ and $t_{max} = 5$ according to the results of preliminary experiments. Note that in our experiment, we classify the training data not only by sentence but also by person. The latter means that we classify all the sentences of one person to the same class according to the average similarity score at one time. Although we also set $rs = 0.2$, $t_{min} = 3$ and $t_{max} = 4$, we found that the set of $rs = 0.5$, $t_{min} = 2$ and $t_{max} = 5$ and the classification of the training data based on every sentence but not every person performed better. After this soft clustering, we obtained a satisfactory data quantity for each class. Using these data, we re-trained the 20-class HMMs using all assigned data and GMMs with 800 randomly selected sentences in each class.

C. Further increase to 30 classes

To further explore the effect of increasing the number of classes using soft clustering, in this section, we extend classification of the training data to 30 classes employing the following method.

1) By calculating the similarity score using the initial 20-class GMMs and the corresponding 20 class initial training data. We choose the 10 worst GMMs and divide their mean vectors into 20 mean vectors to make 20 GMMs using the following equations.

$$y_{\Phi-1}(i,j) = y_{\Phi}(i,j) * 1.05(i = 1, \ldots, I, j = 1, \ldots, J)$$

$$y_{\Phi-2}(i,j) = y_{\Phi}(i,j) * 0.95(i = 1, \ldots, I, j = 1, \ldots, J)$$

   where y denotes the mean value, Φ denotes the index of the GMM, and i and j denote the mixture number and dimension, respectively. In this experiment, I and J are set to 128 and 12.

2) Repeat re-training of the divided GMMs until convergence.
   For $k = 1$ to 10
   a) Use the divided Φ_{k-1} and Φ_{k-2} to classify the initial training data of Φ_k. We use $\Phi_{(E-M60-70HS)-1}$ and $\Phi_{(E-M60-70HS)-2}$ to classify the initial 60–70 year old elder-male data recorded by the headset microphone described in section III-B and use newly classified data to train new $\Phi_{(E-M60-70HS)-1}$ and $\Phi_{(E-M60-70HS)-2}$.

b) Use the new Φ_{k-1} and Φ_{k-2} to classify the initial training data of Φ_k.
c) Repeat step 2b) until the parameters of Φ_{k-1} and Φ_{i-2} converge.

3) Set the 20-class GMMs trained in step 2) and the original 10-class GMMs undivided in step 1); in total, there are 30-class GMMs as the initial 30-class GMMs.
4) Use the same overlapping method presented in Section III-B to obtain the overlapped clustering. In this process, rs, t_{min} and t_{max} are set to 0.8, 2 and 8 according to results obtained by experiment.
5) Train 30-class GMMs and HMMs

D. Cluster Selection Method

For a sequence of T test vectors, $X = x_1, x_2, \cdots, x_T$, the standard approach is to calculate the GMM likelihood (similarity score) in the logarithmic domain as

$$L(X|\lambda) = \log p(X|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda)$$

The class that the GMM has the highest likelihood of belonging to is regarded as the target; the target class is then used to select the corresponding HMMs. We also used a speaker selection technique using the direct calculation of HMM based decoding likelihood. For on-time or real-time recognition, we use the first 20 frames and 50 frames of the sentences, and also use all frames for comparison.

IV. Experiment

A. Setup

In our evaluation experiment, we employed six types of test data denoted E-M, E-F, A-M, A-F, C-M and C-F as described in Section II. Each type of test data has 100 sentences, giving a total of 600 sentences. To obtain reasonable and reliable results in the cluster selection stage, we first processed the test data with simple voice active detection (VAD). The training set in the S-JNAS corpus consists of 48,160 and 48,096 sentences uttered by males and females, respectively, and the numbers for the ASJ+JNAS corpus are 20,333 and 25,059. The CIAIR-VCV corpus consists of sentences and word, 4140 sentences and 7391 words uttered by males, and 5200 sentences and 6454 words uttered by females.

We extracted 38 dimensional Mel-frequency cepstral coefficients (MFCCs) comprising 12 MFCCs and their first and second derivatives, and the first and second derivatives of the logarithmic power. The speech was analyzed using a 25 ms Hamming window with pre-emphasis coefficient 0.97 and shifted with a 10 ms fixed frame advance. GMMs and HMMs were trained with HTK Toolkit[10]. Each HMM contains 4 states and each state contains 4 Gaussian mixture distributions. Each Gaussian mixture distribution has 38 dimensions and their covariances represented by full matrices. In the context-independent 116 syllable-based HMM training, we used the EM algorithm, and 928 context-dependent syllable-based HMMs were trained using the MAP estimation algorithm and taking the context-independent HMMs as the initial models. We only updated the means, transitions, and mixture weights in the context-dependent HMM training stage, and the covariance matrices were not updated. Each GMM for the speaker clustering contains 128 Gaussian mixture distributions and each Gaussian mixture distribution has 12 dimensions and their covariance represented by a diagonal matrix. The language model (LM) used in this experiment is a trigram trained using a Japanese newspaper (75 months, a vocabulary of about 20,000 words). Here, since the children's CIAIR-VCV corpus consists of speech obtained in reading a fairy tale, the language model is out-of-domain. As the decoder for automatic speech recognition (ASR), we used the in-house Large Vocabulary Continuous Speech Recognition (LVCSR) system, SPOJUS++ (SPOken Japanese Understanding System)[11], which has many novel features including a dynamic expansion of a linear dictionary, a likelihood index for the efficient handling of inter-word dependency and one pass decoding.

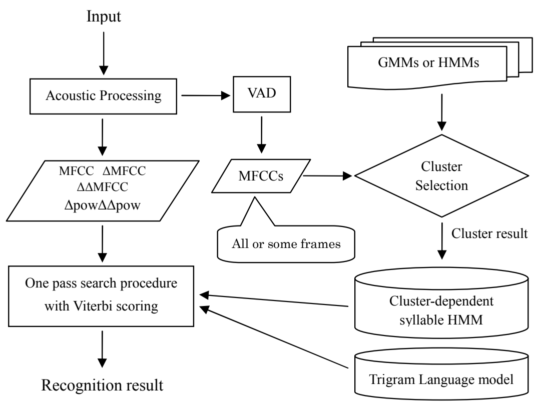

Fig. 2. Diagram of processing in our system

B. System Overview

The overall system is shown in Figure 2. Our system has two main parts: speaker cluster selection and speech recognition. Speaker cluster selection is performed either by GMMs consisting of 128 Gaussian components and 12 MFCCs or by HMMs and 38 MFCCs using likelihoods of recognition. Speech recognition is performed by a one-pass procedure with Viterbi scoring based on syllable-based HMMs and trigram language models.

C. Results

For our experimental results, we first show the results for increasing the number of speaker classes. Second, we compare the methods of selecting speaker models using GMMs or HMMs. Third, we compare the difference in clustering for each sentence and each person.

1) Effect of increasing the number of classes/models: Figure 3 shows the results for increasing the number of speaker classes/recognition models with speaker cluster selection using HMMs in terms of the average WER for a total of 600 sentences. In the case of using the acoustic model trained as the baseline and cluster selection of a speaker with all frames of the input utterance (baseline), the WER was 27.5%. To obtain more appropriate classes, we used the feature-based method to re-classify the training data. The case (re-classification) yielded a WER of 25.7%. Furthermore, to create

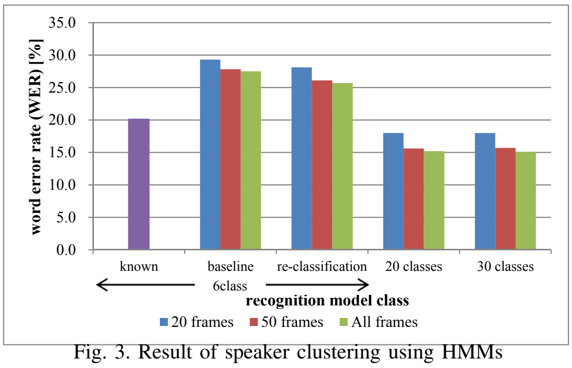

Fig. 3. Result of speaker clustering using HMMs

acoustic models close to the speaker of any input, we increased the number of classes/models. Using 20 class HMMs yielded a WER of 15.2%, and using 30 classes yielded a WER of 15.1%. Finally, we obtained a 5.0 % absolute gain in the WER, and 24.9 % gain in the relative WER over the age- and gender-dependent baseline WER of 20.1% (*a priori* known class). Using the first 20 or 50 frames of the input utterance, with 30 class models, achieved 11.3% or 12.1% gains over the baseline, respectively. By increasing the number of classes/models, we obtained significant improvement by re-classifying 6 classes to 20 classes. However, increasing from 20 classes to 30 classes resulted in only a small improvement. In each speaker class, the improvement for the children's class is especially large relative to the baseline, on average for males and females (i.e. using all frames), there was a 13.7% absolute improvement. In addition, when using all frames, better results were obtained for all speaker classes. Thus, using more classes, the variation in age and gender could be suppressed.

2) Comparison of the HMM with GMM: We compared the methods of speaker cluster selection using GMMs and HMMs. Tables I presents the results. In comparison with the case of an *a priori* known class, we obtained absolute improvement of 0.3%, 1.7% and 1.1% (relative improvement of 1.0%, 8.5% and 4.5%) when using only 20 frames, only 50 frames and all frames, respectively. Hence, we can say that cluster selection based on the HMM is better than that based on the GMM. Finally, the cluster selection was performed by using combination of HMM and GMM. This combination was very effective for a speaker identification task[12]. From Table I, we found that the combination was better than using only HMM or GMM in the case of 20 and 50 frames.

3) Variation in class assignment: Table II gives the average percentage of the most frequent class out of 20 or 30 classes occupied by the model selected most often for each speaker. Selecting a speaker class with many frames increases the percentage of times that is classified in the same class. However, since the rate if belonging to the same class is about only 50%–80%, we can say it is beneficial to select a class for every utterance, but not for every speaker. By deciding the class using only the first utterance of the same speaker, we obtained a WER of 15.3% (worth than 15.1% in Table I). These results indicate that the acoustic features vary even for the same speaker.

V. Conclusion and Future Works

In this paper, we proposed a GMM-based soft clustering technique for training data that allows overlap to reduce reduction in the amount of data per class, when increasing the number of age- and gender-dependent classes, and a method for selecting a speaker model using a GMM or/and HMM. By increasing the model classes, we obtained a 5.0 % absolute gain in the WER, and a 24.9% gain in the relative WER for the speaker age&gender-dependent baseline (class known in advance). Additionally, we found that cluster selection based on the HMM is better than that based on the GMM, and these combination was more better in the case of 20 and 50 frames..

In the future, we will employ a soft-clustering technique for training data using HMMs, and consider a combination of speaker clustering and speaker normalization/adaptation techniques, such as vocal tract length normalization(VTLN)[13].

TABLE I
LVCSR results using 30 class models by cluster selection using HMM, GMM and HMM&GMM (WER) [%]

speaker class		E-M	E-F	A-M	A-F	C-M	C-F	ave.
known		10.9	9.2	9.5	7.6	47.4	35.9	20.1
20 frames	HMM	12.2	8.6	11.7	9.2	36.6	29.6	18.0
	GMM	12.4	7.7	11.0	10.8	37.2	30.5	18.3
	HMM&GMM	9.6	9.0	11.1	8.5	34.9	29.1	17.0
50 frames	HMM	11.0	7.6	9.9	7.7	32.7	25.2	15.7
	GMM	10.4	8.0	11.0	10.2	36.8	28.1	17.4
	HMM&GMM	10.0	7.7	9.8	7.5	32.5	25.0	15.4
all frames	HMM	10.5	8.0	9.3	7.1	31.8	24.2	15.1
	GMM	9.5	6.5	9.9	9.6	33.9	27.9	16.2
	HMM&GMM	9.8	8.6	9.4	7.2	31.9	24.0	15.1

TABLE II
Variation in class assignment of data when selecting speaker class [%]

		20 class			30 class		
	frames	20	50	all	20	50	all
speaker class	E-M	33.0	42.0	62.0	24.0	33.0	55.0
	E-F	39.0	40.0	49.0	35.0	36.0	47.0
	A-M	51.5	60.7	74.1	38.5	52.0	66.1
	A-F	53.7	73.3	86.7	33.9	52.4	75.0
	C-M	53.9	67.9	89.0	53.9	67.9	89.1
	C-F	46.9	68.6	81.7	15.9	68.6	81.7

References

[1] M. Benzeghiba, R. de Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi & C. Wellekens, "Automatic speech recognition and variability: a review", *Speech Communication*, vol. 49, pp. 763-786, 2007.

[2] A. Sankar, F. Beaufays, V. Digalakis, "Training data clustering for improved speech recognition", in *Proc. EUROSPEECH*, pp.503-506, 1995.

[3] T. Kosaka, S. Matsunaga, S. Sagayama, "Speaker-independent speech recognition based on tree-structured speaker clustering", in*Proc. Computer Speech and Language*, Vol. 10, pp.55-74, 1996.

[4] Yu Zhang, Jian Xu, Zhi-Jie Yan, Qiang Huo , "An i-vector based approach to training data clustering for improved speech recognition", in *Proc. INTERSPEECH*, pp.789-792, 2011.

[5] D. Jouvet, N. Vinuesa, "Classifivation margin for improved class-based speech recognition performance", in *Proc. ICASSP*, pp.4285-4288, 2012.

[6] Xin. Chen, Yunxin. Ahao, "Data sampling ensemble acoustic modelling in speaker independent speech recognition", in *Proc. ICASSP*, pp.5130-5133, 2010.

[7] Jean-luc Gauvain , Chin-hui Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", IEEE *Trans. Speech and Audio Process*, Vol.2, pp.291-298, 1994.

[8] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, Vol.9, pp.171-185, 1995.

[9] R.Gomez, T. Toda, H. Saruwatari, K. Shikano, "Improving Rapid Unsupervised Speaker Adaptation Based on HMM-Sufficient Statistics', *Proc. ICASSP*, pp.1001-1004, 2006.

[10] HTK Toolkit, U.K. http://htk.eng.cam.ac.

[11] Y. Fujii, K. Yamamoto, and S. Nakagawa, "Large vocabulary Speech Recognition System: SPOJUS++", MUSP, pp.110-118, 2011.

[12] S. Nakagawa, W. Zhang, M. Takahashi, "Text independent/text-prompted speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM", IEICE Trans, Vol.E89-D, pp.1058-1064, 2006.

[13] E. Eide, H. Gish, "Parametric approach to vocal tract length normalization", in *Proc. ICASSP*, pp.346-348, 1996.