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 cannot 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 reduced. 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% gain 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+JNAS2 database consisting of each 153 male and female speakers aged 20 to 60 years. The data for the child class is the CIAIR-VCV3 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.

2http://www.mibel.cs.tsukuba.ac.jp/\_090624/jnas/instruct.html
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 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)) according to age and the two type of HS and DT.

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, \( \Phi \) 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 \( \Phi_{k-1} \) and \( \Phi_{k-2} \) to classify the initial training data of \( \Phi_k \). 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 \( \Phi_{(E-M60-70HS)-2} \).
models. We only updated the means, transitions, and mixture
algorithm and taking the context-independent HMMs as the initial
based HMMs were trained using the MAP estimation algo-
used the EM algorithm, and 928 context-dependent syllable-
context-independent 116 syllable-based HMM training, we
and their covariances represented by full matrices. In the
shifted with a 10 ms fixed frame advance. GMMs and HMMs
the logarithmic power. The speech was analyzed using a 25
ficients (MFCCs) comprising 12 MFCCs and their first and
6454 words uttered by females.
VCV corpus consists of sentences and word, 4140 sentences
for the ASJ+JNAS corpus are 20,333 and 25,059. The CIAIR-
uttered by males and females, respectively, and the numbers
the S-JNAS corpus consists of 48,160 and 48,096 sentences
in the cluster selection stage, we first processed the test data
data denoted E-M, E-F, A-M, A-F, C-M and C-F as described
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
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 coeffi-
cients (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 distribu-
tions. 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 algo-
rithm 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 auto-
matic 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.

B. System Overview
The overall system is shown in Figure 2. Our system has
two main parts: speaker cluster selection and speech recogni-
tion. 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
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].

**REFERENCES**


