Speaker Recognition for DSR

Mohamed Abdel Fattah, Fuji Ren, Shingo Kuroiwa
Faculty of Engineering, the University of Tokushima
2-1 Minamijosanjima
Tokushima, Japan 770-8506
(mohafi, ren, kuroiwa) @is.tokushima-u.ac.jp

Abstract—**Due to the coexistence of different compression algorithms in the fixed and mobile telephone networks, it is impossible to predict which combination of coders and channels the speech has undergone before arriving to the server. To overcome the previous mentioned problem, the European Telecommunication Standards Institute (ETSI) has standardized a front-end for Distributed Speech Recognition (DSR). But once again, the distortion added due to feature compression in the front-end side increases the variance flooring effect that increases the identification error rate. The penalty incurred in reducing the bitrate is degradation in speaker recognition performance. In this paper we present a non traditional solution for the previous mentioned problems. To reduce the bitrate, speech signal is segmented at client and the most effective phonemes for speaker recognition are selected to be sent to the server. Speaker recognition is occurred at server. Applying this approach on YOHO corpus, we could achieve 0.05% identification error rate (ER) using an average segment of 20.4% of the testing utterance for recognition. This result outperforms previously published results on the speaker identification task from error rate (ER) point of view as well as the minimum speech segment required for speaker identification.**

I. INTRODUCTION

A "client-server" system, where speech features are extracted at the client (device), then compressed and transmitted to a remote server hosting the speaker recognizer performs better than when encoded speech is used for speaker recognition. However there is some recognition degradation when compared to clean (uncompressed) feature vectors. Furthermore, due to the coexistence of different compression algorithms in the fixed and mobile telephone networks, it is impossible to predict which combination of coders and channels the speech has undergone before arriving to the server. The consequent mismatch between speech used in training the recognition system and speech to be recognized is another significant source of performance degradation [1]. To overcome the previous mentioned problems, the European Telecommunication Standards Institute (ETSI) has standardized a front-end for Distributed Speech Recognition (DSR) where speech feature is coded (compressed) in the mobile phone, transmitted over the cellular network, and recognition is performed in the server side by using the decoded speech feature [2]. But the distortion added due to feature compression in the front-end side is a drawback. This problem increases the variance flooring effect that increases the identification error rate when training GMM. Moreover (ETSI) standard is using a fixed feature parameter vector (MFCC (0-12), log-power) for speech data which deprives researchers from using different feature parameter vectors. To overcome these problems, we propose a non traditional approach to reduce the bitrate for the transmitted utterance without feature compression and decrease the identification error rate as well. First we investigated the phoneme effect on speaker recognition system. We found that some phonemes have strong effect on speaker identification. By segmenting the most effective phonemes for speaker recognition task from a speaker utterance, we could decrease the system complexity and the recognition time. Moreover, this technique is very useful to speed up the authentication process through wire/wireless communication systems. This paper is concerned with improving the performance of speaker recognition systems in two areas: decreasing the identification error rate and decreasing the utterance part required for identification task.

There are many research papers for speaker recognition using DSR approach based on different speech databases. Qin Jin, Alex Waibel used the naive de-lambing method based on NIST 1999 Speaker Recognition database [3]. S. Grassi, M. Ansorge, F. Pellandini, P.-A. Farine used Gaussian Mixture Models (GMM) classifiers based on TIMIT database [1], where Chin-Hung Sit, Man-Wai Mak, and Sun-Yuan Kung used SPIDRE corpus [4].In order to compare our results with previous works; it is convenient to have comparisons with the researches which used YOHO database. For example, D. Reynolds could achieve error rates as low as 0.7% using Gaussian mixture models (GMM's) for speaker identification using YOHO corpus [5], while B.L. Pellom reported the same error rate with reduction of the time to identify a speaker by a factor of 140 [6]. All previous mentioned researches could not achieve as low error rate as required. Moreover these researches used all speaker utterance for speaker recognition task which increased the recognition time and increased the transmitted data in the case of wire/wireless communication systems.

Some other researchers used different techniques for speaker recognition. Dominique Genoudy, used neural-network acoustic models of a hybrid connectionist-HMM speech recognizer to adapt a speaker-independent network by performing a small amount of additional training using data from the target speaker, giving an acoustic model specifically tuned to that speaker [7]. O. Thyes, used "eigenvoice" approach, in which client and test speaker models are confined to a low-dimensional linear subspace obtained previously from a different set of training data. He reported 5% ER for

Eigenvoice dimension of 70 using YOHO database [8]. Wan, reported identification error rate of 4.5% using polynomial order of 10 for Support Vector Machines approach when applied on YOHO corpus [9]. Campbell, used Polynomial Classifiers for Text-Prompted Speaker Recognition. His best identification error rate was 0.38% using second order Polynomial Classifiers for YOHO database [10]. However the different techniques used for all previous mentioned researches, it is strongly required to achieve as low ER as possible and also decrease the required speaker utterance part for recognition.

André G, segmented speech to 5 classes. Unvoiced segment class in addition of 4 different classes based on rising and falling of energy and fundamental frequency (f0) [11]. Alex Park, segmented speech to eight phonetic classes and used several approaches for speaker identification task based on YOHO corpus. His best identification error rate was 0.25% when he used multiple classifiers (phonetically structured GMM + speaker adaptive) [12]. However André G and Alex Park segmented the speech, they did not take the advantage of the whole effect of all phonemes in it.

Although the goal of text independent speaker recognition has led to an increased focus on global speaker modeling, it is well known that some phones have better speaker distinguishing capabilities than others [13, and 14]. For instance, in [13] vowels and nasals were found to be most discriminating phoneme groups. Global speaker modeling techniques like the GMM approach are not able to take optimal advantage of the acoustic differences of diverse phonetic events. No doubt that taking the advantage of speech segmentation is enhancing the identification error rate as well as decreasing the required speech segments for speaker identification task. This advantage was not taken into account for the traditional speaker recognition models.

In this paper we investigate the phoneme effect on speaker recognition task. Our targets are:

1- Decrease the required speech segment for speaker identification task to reduce the bitrate, decrease the system complexity and speed up the speaker identification process.

2- Decrease the identification error rate.

In order to achieve the above targets, we have investigated the speaker phonemes effect on the speaker identification task. Then we selected the most effective phonemes for speaker identification. Using this technique, we could reduce the bitrate and decrease the identification error rate as well. Our results outperform all previously published results on the speaker ID from the precision point of view as well as minimum speech segment required for identification process.

II. YOHO DATABASE

The data consists of 138 speakers - 106 males and 32 females recorded in a span of 3 months. To record the data, a high quality telephone handset was used. For each speaker, both training, also referred to as enrollment, and testing, or verification, sessions have been created. The enrollment sessions consist of four sessions each containing 24 utterances while the verification data has 10 sessions of 4 utterances each. Each speaker has the same training data set where testing data are different for each speaker. Each utterance consists of "combination lock" phrases which are each a set of three doublets of digits, for example "23-42-91" pronounced as "twenty three, forty two, ninety one". The sampling rate for the speech files is 8 kHz, and the sample coding is 12-bit linear (stored as 16-bit words). The total number of pronounced phoneme types in YOHO database is 18 phoneme types.

III. THE PROPOSED SYSTEM

Figure 1 shows a block diagram of the processing stages for the proposed DSR system.

At the terminal the speech signal is sampled and parameterized to construct feature vectors. These feature vectors are then segmented to phonemes using HMM speaker independent phoneme model that constructed by using all speakers training data. Then select the most effective phonemes for speaker recognition to obtain a lower data rate for transmission. Transmit these phonemes besides the associated label for each phone. Before transmission, frame structure & error protection occurred. Error detection & correction occurred at the server DSR back-end. Also server feature processing may occur at server to generate more features from the received feature vectors such as delta and acceleration coefficients. At the server side, we have phoneme based GMM speaker dependent model for each speaker for each phoneme. Using the label associated with each transmitted phoneme to direct the phoneme segment to the correct phoneme based GMM speaker dependent model for recognition.

IV. IMPLEMENTATION

The system consists of the following modules:

1- HMM speaker independent phoneme model constructed by using all speakers training data. This HMM phoneme model is used to segment each speaker training and testing utterances into phoneme segments.

2- Segmenting each testing data utterance to phoneme segments using the previous constructed HMM speaker independent phoneme model. Then using each speaker dependent GMM phoneme model to calculate the identification error rate as a function of each phoneme as in section sec. A. Taking the contribution of the testing utterance phonemes of the same type to calculate the identification error rate as a function of each phoneme as in section sec. B. Taking the contribution of some testing utterance phonemes to calculate the identification error rate as in sec. C.

A. Phonemes effect on speaker identification

The probability density function for a feature vector $\vec{z}$ is a weighted sum, or *mixture*, of *k* class-conditional Gaussian distributions. For a given phone of a certain speaker, s_p, the probability of observing $\vec{z}$ is given by

$$p(\vec{z} \mid s_p) = \sum_{k=1}^{K} w_{s_p,k} N(\vec{z}; \vec{\mu}_{s_p,k}, \sum\nolimits_{s_p,k}) \quad (1)$$

Where $w_{s_p,k}$, $\vec{\mu}_{s_p,k}$, $\sum\nolimits_{s_p,k}$ are the mixture weight, mean, and covariance matrix, respectively, for the i-th component, which has a Gaussian distribution given by:

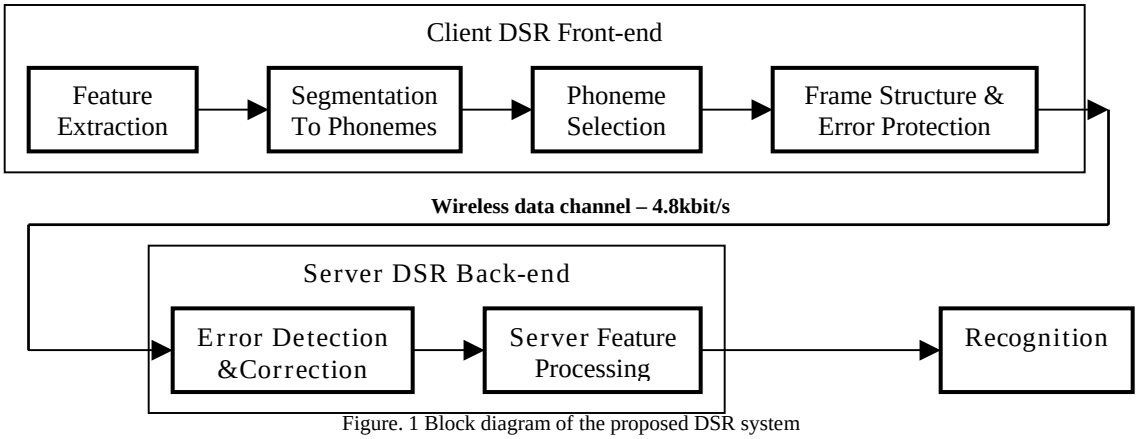

Figure. 1 Block diagram of the proposed DSR system

$$N(\vec{z};\vec{\mu},\sum)=\frac{1}{\sqrt{(2\pi)^n\,|\sum|}}e^{-\frac{1}{2}(\vec{z}-\vec{\mu})'\sum^{-1}(\vec{z}-\vec{\mu})} \quad (2)$$

Where n is the dimension of $\vec{z}$. We used $\sum$ as diagonal covariance matrices. Given a set of training vectors of a certain phoneme of a certain speaker, an initial set of means is estimated using the *k*-means clustering. The mixture weights, means, and covariances are then iteratively trained using the expectation maximization (EM) algorithm. After segmenting each speaker training utterance to phoneme segments, we used equation (1) to construct phoneme model for each speaker. So each speaker utterance was represented as: (sil < ($phonemes)> [sp] < ($phonemes)> [sp] < ($phonemes)> sil). Since ($phonemes) is some phoneme combination of the 18 phonemes of YOHO database represented as:

$phonemes = ah | ao |ay | eh |er | ey |f | ih | iy | k |n | r |s | t |th | uw | v | w;

Using this approach, we constructed speaker dependent model for each phoneme except 2 phonemes which are ("r" and "er") since the system failed to construct them for some speakers because the frequencies of these 2 phonemes are low. After that we used each phoneme model for each speaker to test each separate speaker phoneme (obtained after segmenting the testing utterance using HMM speaker independent phoneme model) for speaker identification task using maximum likelihood of each phoneme. Table 1 illustrates the identification error rate for each speaker phoneme.

From table 1, ER depends on phoneme type and the frequency of the phoneme in the training data. ER is inversely proportional to the phoneme frequency of the training data since as the phoneme frequency increases the GMM phone based model accuracy increases too. ER is low in the case of vowels and nasal phones. Vowels like "ih", "uw", "ah", and "ao" give good speaker identification results where "eh" and "iy" do not. Diphthong phones like "ay" and "ey" give very bad results. It is common for speaker identification task to calculate the ER using the whole utterance. In the next section, we take the contribution of all phonemes of the same type for a certain utterance to calculate the identification error rate.

B. Identification error rate using the contribution of utterance phonemes of the same type

We conducted the above experiment, but we took the contribution of all phonemes of the same type in each utterance into account to calculate the ER. Table 2 illustrates the identification error rate for each speaker phoneme when taking all utterance phonemes of the same type into account.

It is clear that the ER improved in general. The effect of testing data phoneme frequencies on speaker identification task is very strong as shown in table 2. In table 2, although phonemes "iy" has high frequency for training and testing data, it has the highest ER value, whereas the nasal phoneme "n" gives the best identification result. Once again nasal and vowels like "ih" give the best identification results and "ay", "ey" and "iy" still give the worst identification results.

C. The effect of combining phonemes on speaker identification

In table 2, Phonemes "n" has the best speaker identification result then phoneme "ih" and so on. Taking the contributions of successive phonemes to calculate the speaker identification error rate must achieve better results than using each one alone. When we conducted the same experiment as above but we calculated ER using the contributions of phonemes "n" and "ih", the ER became 0.46%. Moreover the estimated average duration time of all "n" and "ih" segments in the whole YOHO testing database was as follows:

("n" + "ih" segments duration time in all testing phrases)/ (total testing data time) = 8.8%

Using the same approach for all remaining phonemes, we could achieve the results of table 3.

It is very clear from table 3 that as taking the contributions of more phonemes, the identification error rate decreases until a certain phoneme combinations then it increases again. The best result is ER = 0.05% when using (n + ih + f + uw + ah + th) combination and the total required speech segment for recognition is 20.4013% of the whole testing utterance in the average. When adding the

Table 1: Identification error rate using separate phone

Phone	ih	uw	ah	ao
ER	1.20%	1.50%	1.70%	2.50%
Phone	th	f	n	k
ER	2.60%	2.70%	4%	5.60%
Phone	eh	v	t	s
ER	6.20%	9%	14.80%	16.90%
Phone	w	ay	ey	iy
ER	18%	32%	49.40%	55.50%

Table 2: Identification error rate using utterance phones of the same type

Phone	n	ih	f	uw
ER	0.90%	0.97%	1.30%	1.50%
Phone	ah	th	ao	k
ER	1.80%	2%	2.40%	4.80%
Phone	eh	t	v	s
ER	4.90%	5%	7.30%	12.60%
Phone	w	ay	ey	iy
ER	15%	31.80%	49.30%	55%

Table 3: Identification error rate for phonemes combination & phonemes segment ratio

Phone	ER	Seg. ratio
n	0.90%	5.888398%
n + ih	0.46%	8.809841%
n + ih + f	0.22%	13.36078%
n + ih + f + uw	0.14%	15.88506%
n + ih + f + uw + ah	0.14%	17.66334%
n + ih + f + uw + ah + th	0.05%	20.4013%
n + ih + f + uw + ah + th + ao	0.14%	23.96435%
n + ih + f + uw + ah + th + ao + k	0.13%	25.62414%
n + ih + f + uw + ah + th + ao + k + eh	0.09%	26.89874%
n + ih + f + uw + ah + th + ao + k + eh + t	0.07%	30.75527%
n + ih + f + uw + ah + th + ao + k + eh + t + v	0.11%	34.03236%
n + ih + f + uw + ah + th + ao + k + eh + t + v +s	0.09%	39.14136%
n + ih + f + uw + ah + th + ao + k + eh + t + v +s + w	0.09%	40.73546%
n + ih + f + uw + ah + th + ao + k + eh + t + v +s + w + ay	0.37%	45.27926%
n + ih + f + uw + ah + th + ao + k + eh + t + v +s + w + ay + ey	0.48%	46.64604%
n + ih + f + uw + ah + th + ao + k + eh + t + v +s + w + ay + ey + iy	0.78%	51.86819%

contributions of the phonemes that give bad identification value such as “ay”, ”ey” and ”iy”, the ER increases as shown in table 3. Most of the results appear in table 3 outperform all previous published result from the identification error rate point of view as well as the minimum speech segment required for identification task.

IIV. CONCLUSIONS

In this paper we tried to reduce the bitrate using segmentation approach which avoided the previous mentioned problems. Our system is flexible, for example we can compromise the ER value with the required bitrate. For instance we can accept a low value of ER = 0.9% to send an average speech segment of 5.8% of the whole testing utterance. Also we may have to send an average speech segment of 20.4% of the whole testing utterance to achieve 0.05% identification error rate.

In the future work we will investigate the rest phonemes effect on speaker recognition using other speech corpora.

ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 14350204 and 14380166, in 2004, Hoso-Bunka Foundation (HBF) and International Communications Foundation (ICF).

REFERENCES

[1]- S. Grassi, M. Ansorge, F. Pellandini, P.-A. Farine, "Distributed Speaker Recognition Using the ETSI AURORA Standard", Proc. of 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication, Budapest, Hungary, Oct. 11-12, 2002, pp.120-125.

[2]- David Pearce, “Enabling New Speech Driven Services for Mobile Devices: An overview of the ETSI standards activities for Distributed Speech Recognition Front ends”, AVIOS 2000: The Speech Applications Conference, San Jose, CA, USA Enabling New Speech, May 22-24, 2000.

[3]- Qin Jin, Alex Waibel, “A Naïve De-lambing Method for Speaker Identification”, in ICSLP2000.

[4]- C.H. Sit, M.W. Mak and S.Y. Kung, "Maximum likelihood and maximum a posteriori adaptation for distributed speaker recognition systems”, International Conference on Biometric Authentication (ICBA'04), Hong Kong.

5- D. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” Speech Communication., vol. 17, 1995, pp. 91–108.

6- B.L. Pellom, J.H.L. Hansen, "An Efficient Scoring Algorithm for Gaussian Mixture Model based Speaker Identification," IEEE Signal Processing Letters, vol. 5, no. 11, Nov. 1998, pp. 281-284.

7- Dominique Genoud, Dan Ellis, Nelson Morgan “Combined speech and speaker recognition with speaker-adapted connectionist models” ASRU-99, Keystone CO, December 1999.

8- O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua, “SPEAKER IDENTIFICATION AND VERIFICATION USING EIGENVOICES”, International Conference on Spoken Language Processing 2000 (ICSLP 2000)

9- Wan, V. and Campbell, W. M., “Support Vector Machines for Speaker Verification and Identification”, Neural Networks for Signal Processing X, 2000, pp. 775-784.

10- Timothy J. Hazen, Eugene Weinstein, Ryan Kabir, Alex Park, Bernd Heisele “MULTI-MODAL FACE AND SPEAKER IDENTIFICATION ON A HANDHELD DEVICE” In Proc. of Workshop on Multimodal User Authentication, December 11-12, 2003, Santa Barbara, California, pp. 113–120.

11- André G. Adami1, Hynek Hermansky, “Segmentation of Speech for Speaker and Language Recognition”, EUROSPEECH 2003 – GENEVA.

12- Alex Park and Timothy J. Hazen, "ASR dependent techniques for speaker identification," Proceedings of the International Conference on Spoken Language Processing, Denver, Colorado, September, 2002.

13- Eatock, J.P. and Mason, J.S., “A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes”, Proc. ICASSP’94, Adelaide, 1994, pp. 133-136.

14- Nolan, F., “The Phonetic Bases of Speaker Recognition”, Cambridge CUP, 1983.