SELECTIVE USE OF THE SPEECH SPECTRUM AND A VQGMM METHOD FOR SPEAKER IDENTIFICATION

Qiguang Lin, Ea-Ee Jan*, ChiWei Che, Dong-Suk Yuk, and James Flanagan*

CAIP Center, Rutgers University, Busch Campus, NJ 08855, USA
email: qlin@watson.ibm.com, cche@caip.rutgers.edu

ABSTRACT

This paper describes two separate sets of speaker identification experiments. In the first set of experiments, the speech spectrum is selectively used for speaker identification. The results show that the higher portion of the speech spectrum contains more reliable idiosyncratic information on speakers than does the lower portion of equal bandwidth. In the second set of experiments, a vector-quantization based Gaussian mixture models (VQGMMs) is developed for text-independent speaker identification. The system has been evaluated in the recent speaker identification evaluation organized by NIST. In this paper, details of the system design are given and the evaluation results are presented.

1. INTRODUCTION

The task of speech recognition is to automatically determine what has been said, while the task of speaker recognition is to automatically determine who has said a given utterance. Despite the differences in task, however, the same spectral information and the same front-end analysis techniques have traditionally been used for both speaker and speech recognition. This may have been a consequence of biased effort in speech recognition research (and also the desire to operate with telephone speech).

In the first part of this paper, it is attempted to distinguish the two tasks by exploring selective use of the speech spectrum for text-independent speaker identification. It is known that the phonetic/linguistic information (for speech recognition) is mainly confined in the frequency range of, say, 0 - 5 kHz. One example is conventional telephone speech which is bandlimited to 0.3 - 3.2 kHz. Another example is speech synthesis: Highly intelligible speech can be synthesized from formants below 5 kHz. On the other hand, personal identity information is spread over the frequency axis. For instance, speaker-specific attributes based on the glottal source are mainly confined to the low/mid frequency range, while speaker attributes based on the friction source are carried mainly in the high frequency range. The high-frequency spectrum also contains information about cross-modes of the vocal tract and the overall length of the vocal tract. It is therefore important to study what portion of the speech spectrum contributes most to speaker recognition.

From our experiments, use of the higher-frequency portion of the speech spectrum consistently produces better speaker identification performance than does use of the lower-frequency counterpart of equal bandwidth. The result is useful in developing robust speaker identification systems: 1) The approach is less susceptible to inter-session variability because most of the phonetic/linguistic information in the sound spectrum is not used; 2) It is robust against room reverberation and noise in rooms. Reverberation and noise in rooms are typically more prominent at low frequencies which are removed in the proposed approach; and 3) It is robust against impostor attacks because details of the signal spectrum at high frequencies are hard to mimic.

In the second part of the present paper, recent efforts to improve Gaussian mixture modeling (GMM) techniques are described. GMM techniques have been commonly adopted for text-independent speaker recognition and good performance has been reported (see, e.g., [1]). Conventional GMMs generate a Gaussian mixture model for each enrolled speaker. The model statistics is estimated using acoustic features covering the entire acoustic space. We argue that the statistics can be better estimated by first clustering (vector-quantizing) the acoustic space into several subspaces. Each subspace is then represented by a number of Gaussian mixture models whose parameters are determined using only those relevant acoustic features belonging to the subspace.

We refer the new system to as vector-quantization based Gaussian mixture models (VQGMMs). Our experiments show that VQGMMs surpass conventional GMMs. The system has recently been used in

*Q. Lin and E. Jan are now with IBM Human Language Technologies Group, Watson Research Center, Yorktown Heights, NY 10598, USA.

1996 NIST Speaker Identification Evaluation. From the official evaluation results [2], the system generally produces top scores among all the participating sites. For some test subsets (short utterances), the VQGMM system yields the best scores.

Details of the VQGMM system design and evaluation results are presented in the paper.

2. SELECTING THE SPECTRUM

As mentioned in the Introduction, speaker identity information is spread over the whole frequency axis. In this section, we investigate what portion of the speech spectrum contributes most to speaker identification.

2.1. Experimental setup

The New England subset of the TIMIT database is used in the experiment. The TIMIT database covers a frequency range up to 8 kHz and hence enables us to fully study various portions of the speech spectrum. In the subset, there are 38 speakers (24 males and 14 females). For each of the speakers, a Learning Vector-Quantization (LVQ [3]) codebook is generated using 5 training sentences from the speaker. During testing, the minimum L2 norm distance between the testing feature vectors and centroids of the codebook is accumulated. The person whose codebook produces the least L2 distance is considered as the identification result (closed-set, text-independent experiments). For more details on VQ-approaches to speaker identification, the reader is referred to [4, 5].

There are different ways to select a portion of the speech spectrum. In our implementation, the selection is conducted in the frequency domain after FFT computation. Figure 1 illustrates the selection process. In this figure, f_l and F_h are the lowest frequency component and highest frequency component retained in the selected spectrum, respectively. From the selected spectrum (Figure 1b), cepstral coefficients can readily be calculated by discrete Cosine transforms (DCT).

Freq. (kHz)	M = 37	M = 40	M = 43
0 - 5	5.8%	4.2%	4.2%
3 - 8	1.1%	1.6%	1.1%

Table 1: *Comparison of speaker identification error rates obtained by selecting different portions of the spectrum. Note that the two portions cover an equally-wide frequency interval, i.e., 5 kHz. M denotes the number of codewords per codebook.*

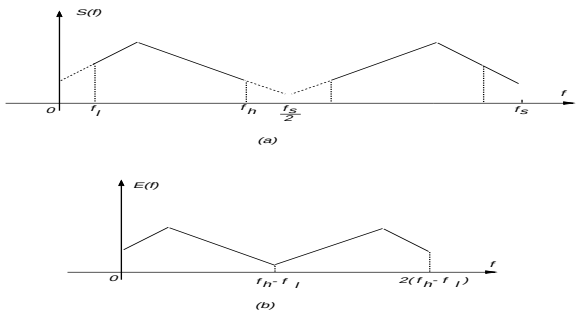

Figure 1: *Schematic diagram illustrating selective use of the sound spectrum.*

2.2. Results and discussion

In the present study, the speech spectrum is divided into two portions only: a lower-frequency portion and a higher-frequency portion. Various cut-off frequencies, dimensions of the cepstral coefficients, and sizes of the LVQ codebooks have been utilized. Consistently, it is found that use of the higher-frequency portion of the speech spectrum produces better speaker identification performance than does the lower-frequency counterpart. The experimental results for a cut-off frequency of 5 kHz and 20-dimension cepstrum coefficients are given in Table 1. The results are obtained from 190 trials (5 testing sentences $\times$ 38 speakers). It is observed from Table 1 that the new approach reduces the error rate by as much as 70% on the average.

The results are of interest to developing robust speaker recognition systems. Because the removal of the lower portion of the sound spectrum, the approach is less susceptible to inter-session variations, and less susceptible to room reverberation and ambient noise interference. Furthermore, the approach is robust against imposter attacks because the high-frequency spectrum is difficult to mimic.

The suggested selective use of the spectrum for speaker identification requires wideband acquisition of speech data. The approach can thus be deployed in applications wherever high-frequency spectral components are available, such as in secure area screening and admittance applications. The approach cannot be directly used for telephone speech where high-frequency spectral signals above 3.3 kHz are removed due to channel bandwidth limits. However, when the interval of $F_h - F_l$ (refer to Figure 1) is narrowed to 3.3 kHz, the approach may be used for telephone speech with modulation techniques [6].

Incidentally, we recently came across a paper [7]

where it was found that better speaker identification performance could be achieved when increasing the frequency range of the spectrum.

3. THE VQGMM SYSTEM

3.1. GMM for speaker id

Gaussian mixture modeling (GMM) techniques have now been successfully utilized for text-independent speaker identification [1]. In GMM techniques, a multi-variate, continuous, tied Gaussian mixture model is created for each of the enrolled speakers. The model parameters are estimated by a maximum likelihood (EM) algorithm and using all acoustic feature vectors covering the entire acoustic space. During testing, the log likelihood for each speaker Gaussian model is computed over the testing feature vectors (observations). The identified speaker is the one whose model gives the highest score.

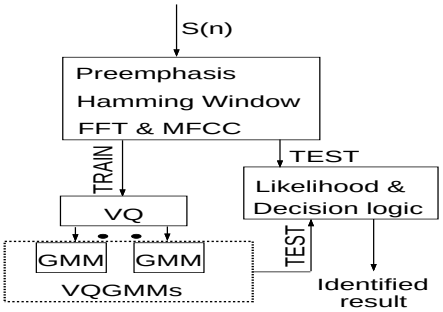

Figure 2: *Block diagram of the VQGMM system.*

3.2. VQGMM for speaker id

We argue that model parameters can be better estimated by first clustering the whole acoustic space into several subspaces. One can conceive of such a subspace representing a broad phone class. Within a subspace, the feature vectors are relatively more homogeneous. Each subspace is then characterized by a number of Gaussian mixture models whose parameters are determined using only those relevant acoustic features belonging to the subspace. In other words, feature vectors being far from the subspace are not used to estimate model parameters for that subspace.

There are different ways to divide the entire acoustic space into subspaces. We use vector quantization (binary splitting) and hence, the method is referred to as the VQGMM method.

Figure 2 depicts the training and testing procedures of the VQGMM system. During training, all features vectors from one speaker are clustered to a desired number of groups. Each group is then represented by GMMs whose parameters are determined solely by those feature vectors of the group, as described before. During testing, the likelihood scores are computed for each individual observation vector against all groups, and accumulate the attained maximum score throughout the whole utterance (a Viterbi method). The identified speaker is the one whose model gives the highest accumulated score. Note that the testing phase involves no vector quantization processes.

Compared with ergodic hidden Markov models, the present VQGMM is a simplified and effective method.

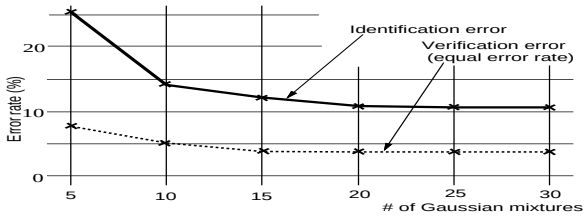

Figure 3: *Speaker id error rate in % and speaker verification equal error rate in % of the VQGMM system, as a function of the number of mixtures.*

3.3. 1996 NIST SPID Evaluation

Recently, NIST/DoD organized a speaker identification evaluation in March 1996. The evaluation task was open-set, text-independent speaker identification/detection for conversational telephone speech. More specifically, the task was to detect the presence of a claimed speaker, given a segment of conversational speech over the telephone with objectives to minimize the false alarm rate while maintaining an overall 10% miss rate. In the test set, there were 19 female and 21 male target speakers, and about 200 female and 200 male non-target speakers. Two-minute speech material was provided for each target speaker to create his/her voice model. The complete evaluation comprised 3 subsets depending on the duration of individual testing segments: 30 sec, 10 sec, and 3 sec. The speech data was collected using different handsets and unknown telephone transmission networks. Detailed specification of the evaluation is found in [2]. 9 sites from USA and other countries participated in the evaluation. The Rutgers University team used the above described VQGMM in the evaluation [8]. The system employed 19 mel-frequency cepstral coefficients (MFCCs) as the measured feature (that is, no delta or delta delta MFCCs).

First, the VQGMM is compared with GMM using

the male portion of the Devset for different numbers of mixtures. The results show that a 10% error rate reduction is achieved with a fine-tuned VQGMM versus a fine-tuned GMM system. Figure 3 plots speaker recognition performance of VQGMM as a function of the mixture number at a fixed codebook size.

Effects of cepstral mean normalization (CMN) are studied in Table 2. It can be seen that CMN is helpful for all three training conditions, except for short test utterances (3 sec). Therefore, CMN is turned off for the 3-sec testset.

Test utt. duration	*training conditions*		
	2 handsets	1 handset	1 session
30 sec	16.1/11.4	27.1/19.9	19.7/11.9
10 sec	21.8/21.8	32.0/29.8	24.6/20.8
3 sec	32.5/37.5	38.5/43.2	33.5/35.3

Table 2: *Speaker id error rate in % on the 1996 Devset without/with cepstral mean normalization. For each speaker, 2 min training speech data is provided.*

In order to determine whether a claimed speaker is present, it is necessary to normalize the likelihood score and set a threshold. We compare 3 different ways of selecting background speakers for the normalization purpose. They are: (a) using all Devset speakers as background speakers; (b) cohort selection of the all speakers in the Devset; and (c) on-line data-driven selection of the background speakers which may correspond to the top N speakers or the top N plus bottom N speakers. It is found that the third approach works best. In Figure 4, the relationships between the system performance and the size of top N background speakers, on-line determined, can be studied. It is noted that least error rates are obtained when N is between 5 and 10. This result conforms with our previous speaker verification experiments on the YOHO database [9].

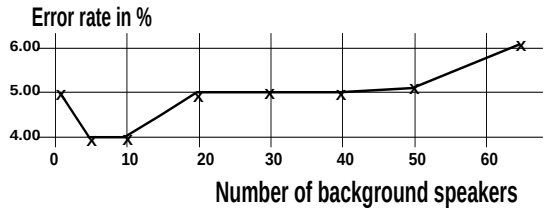

Figure 4: *Effects of the size of background speakers on speaker recognition performance.*

In the speaker identification evaluation, all systems were evaluated according to a detection cost function (DCF) defined by the evaluation organizer [10]. From the official result report [2], the Rutgers VQGMM system produced top scores. For some test subsets, especially the 3-sec set, the VQGMM system gave the best performance. An example is shown in Figure 5.

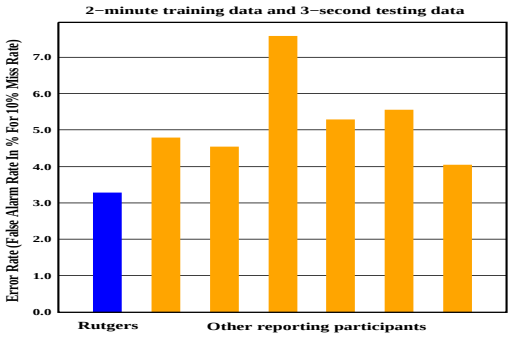

Figure 5: *Evaluation results for the 3-sec subset.*

4. Acknowledgments

This work was supported by DoD Contract MDA904-92-C-5169 and ARPA Contract DABT63-93-C-0037.

REFERENCES

[1] Reynolds, D. "A Gaussian mixture modeling approach to text-independent speaker identification," *Technical Report 967*, Lincoln Lab, MIT, 1993.

[2] Martin, A., Przybocki, M., Fiscus, J., and Pallet D., "Evaluation on SWITCHBOARD corpus selected segments," *Notebook of 1996 NIST Speaker Recognition Workshop*, Maryland, 1996.

[3] Kohonen, T. et al, "LVQ PAK: A program package for the correct application of Learning Vector Quantization algorithms", *Proc. Intern. Joint Conf. NN*, 1992, pp. 1725-1730.

[4] Soong, F., Rosenberg, A., Rabiner, L., and Juang B.-H., "A vector quantization approach to speaker recognition," *ICASSP 1985*, pp. 387-390.

[5] Lin, Q., Jan, E., and Flanagan, J., "Microphone arrays and speaker identification," *IEEE-Trans Speech & Audio Processing,* 1994, pp. 622-629.

[6] Lin, Q., Jan, E., and Flanagan, J. "Speaker recognition using selective spectrum information," Rutgers Invention Disclosure Docket #94-0309-1, Dec. 1993.

[7] Hayakawa, S. and Itakura F., "The influence of noise on the speaker recognition performance using the higher frequency band," *ICASSP 1995*, pp. 321-324.

[8] Che, C, Yuk, D.-S., Flanagan J., and Lin, Q.' "Development of the 1996 RU speaker recognition system," *Notebook of 1996 NIST Speaker Recognition Workshop*, Maryland, 1996.

[9] Che, C. and Lin, Q., "Speaker recognition using HMM with experiments on the YOHO database," *Proc. 4th Eurospeech*, Spain, 1995, pp. 625-628.

[10] Doddington, G., "Review of the evaluation plan," *Notebook of 1996 NIST Speaker Recognition Workshop*, Maryland, 1996.