A SUPPLEMENTARY FEATURE FOR SPEAKER VERIFICATION IN CONSTRAINED ROOM ACOUSTICS

Sung-Wan Yoon, Samuel Kim and Hong-Goo Kang

MCSP Lab, Dept. of Electrical and Electronic Engineering,Yonsei University,
134 Shinchon-dong, Sudaemun-gu, Seoul 120-749, Korea
yocello@mcsp.yonsei.ac.kr

ABSTRACT

This paper describes a preliminary experiment to find a supplementary feature for speaker verification. In conventional speaker verification systems, spectral features such as the mel frequency cepstral coefficient (MFCC) are used universally for all speakers. For some specific applications demanding high security, however, the system also needs to adopt speaker specific supplementary features. Assuming that verification is done in specified room environment, we analyze the effect of the room response to speaker verification. Simulation results show that the additional supplementary feature is crucial for improving the system performance.

1. Introduction

The problem of speaker verification task is essentially a hypothesis testing or binary classification. Both a target speaker model and an impostor model are necessary for making a decision. The Gaussian mixture model (GMM) has been widely used as a probabilistic model in most state-of-the-art speaker verification systems [1], and the MFCC is one of the most widely used features in speaker or speech recognition [2].

In some specific applications of speaker recognition field such as the forensics or requiring high security the system should be adapted to verify only one-person with high confidence. Though the conventional MFCC-based verification system performs very well in general, we would like to further improve the quality using additional features that are related to the target speaker only. The supplementary features can be another type of features or some kind of environmental information that the claimed speaker locates. This paper focuses on the environmental information, especially for the room acoustics.

In general, it is very difficult to determine what kind of supplementary features improve the performance of the verification system and how much the performance will be improved. We even do not know which parameters will be good for verification performance. However, if there is some con-strained condition that only the specific speaker has the discriminative feature, to which the others do not have, it will be a meaningful experiment.

In this paper we conduct preliminary experiments to search supplementary features that are highly related to the target speaker. The type of supplementary features could be the filtering operation depending on the claimed speaker's pitch information or the spatial information of the room.

The desirable characteristics of the supplementary feature combined with the original spectral features are summarized as follows:

- The distance or mismatch between the original model and the newly trained one is reduced or at least similar to the target speaker utterance.

- In the case of anti-speaker, impostor, the mismatch should be increased. Thus, when some kind of distortion by the supplementary feature is added to the impostor case, the likelihood verification score of the impostor be-comes decreased.

To conduct the experiment, at first, we set up the scenario with the room acoustics environment. If the verification for the true speaker is performed in the specific office room, the contribution of the room impulse response will be convolved with the utterance from the speaker. In this case, the room impulse response will be the supplementary feature for the true speaker. Assuming that impostors cannot enter the room, there is no contribution from the room impulse response to their utterance.

A preliminary result shows that the supplementary feature with the restriction of room entrance shows noticeably good performance. Consequently, we can argue that supplementary features which represent the specific speaker's environment will work well to some appropriate applications having reasonable constraints.

In the next section, the baseline recognition system will be described. In section 3, the conventional recognition and the scenario of experiments for searching the supplementary feature are discussed. We will describe the experiment setup using the YOHO database in section 4 and the preliminary results of our verification system in section 5. And conclusion is followed in section 6.

2. Baseline Recognition System

The widely used speaker verification systems utilize GMM trained by expectation-maximization (EM) algorithm [3][4]. A Gaussian mixture density represents the acoustic distribution of each claimant speaker given by

$$p(\vec{x}_t|\lambda_k) = \sum_{j=1}^{M} p_i^k b_i^k(\vec{x}_t) \tag{1}$$

where p_i^k and b_i^k are the mixture weight and the Gaussian density for the i-th mixture out of M for speaker k, respectively. Each component density D is a variate Gaussian function of the form

$$b_i(\vec{x}_t) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(\vec{x}_t - \vec{\mu}_i)^t \Sigma_i^{-1}(\vec{x}_t - \vec{\mu}_i)\} \tag{2}$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$ [3]. Thus, the complete Gaussian mixture density is parameterized by three factors, such as the mean vectors, covariance matrices and mixture weights from all component densities. Thus, the probability of specific speaker k is modeled by these parameters with the notation

$$\lambda_k = \{p_i, \vec{\mu}_i, \Sigma_i\}, i = 1, ..., M \tag{3}$$

In the training process of the modeling the above GMM parameters which best matches the distribution of training features, the maximum likelihood (ML) is estimated using EM algorithm.

In the verification process, the average log-likelihood of a claimant speaker given an utterance $X = \{\vec{x}_1 ... \vec{x}_T\}$ is computed as

$$LL(X|\lambda_k) = \frac{1}{T}\sum_{t=1}^{T} \log p(\vec{x}_t|\lambda_k) \tag{4}$$

This log-likelihood score is very sensitive to variations in text, speaking behavior, and recording conditions, especially from the impostors' utterances. The task of threshold determination for acceptance or reject is a very difficult one. In order to overcome this problem, many normalized score methods are used such as cohort model [5], universal back-ground model [7], etc. As a result, the verification likelihood score applied normalized term is

$$LL(X|\lambda_k) = \log p(X|\lambda_k) - \log p(X|\lambda_{BGM}) \tag{5}$$

where λ_{BGM} is background model is trained by speakers uttering general text-independent utterances or the T user's phrase [5].

3. Speaker-Dependent Supplementary Feature

3.1. Homomorphic Analysis

Figure 1 shows the front-end system of recognition system. In this figure, we assume that the speaker is forced to record his or her voice in the specific room. So, the speech signal, $s(t)$, combines with the additive noise, $d(t)$, and room impulse response, $h(t)$.

Time-frequency representation of speech signal can be represented as follows.

$$X(n,\omega) = \sum_{t=-\infty}^{\infty} x(t) w_n(t) e^{-j\omega t} \tag{6}$$

where n is frame index and $w_n(t)$ is the analysis window which is assumed to be non-zero only in the interval $[T_n, T_n + N_w - 1]$. Considering the speech production model and channel condition, it can be represented as

$$X(n,\omega) = [\{S(n,\omega) + D(n,\omega)\} \cdot H_r(n,\omega)] * W_n(\omega) \tag{7}$$

where $D(n,\omega)$, $H_r(n,\omega)$, and $W_n(\omega)$ are frequency representation of additive noise, room impulse response, and analysis window at the corresponding frame index, respectively.

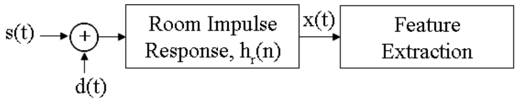

Fig. 1. Front-End Process of recognition system

By inverse-Fourier transform after logarithmic operation, we can extract a cepstral feature for the windowed speech signal.

$$c(n,\omega) = \frac{1}{2\pi}\int_0^{2\pi} \log X(n,\omega) e^{j\omega n} d\omega \tag{8}$$

Under the assumption that we can sufficiently eliminate the additive noise by speech enhancement preprocessing be-fore extracting features [6], and approximate the effect of the analysis window, we can rewrite the equation as follows.

$$\begin{aligned} \hat{c}(n,\omega) &\approx \frac{1}{2\pi}\int_0^{2\pi} \log S(n,\omega) H_r(n,\omega) e^{j\omega n} d\omega \\ &= c_S(n,m) + c_{h_r}(n,m) \end{aligned} \tag{9}$$

Without any channel compensation, the cepstral coefficient is represented by a combination of speech vocal tract information, $c_S(n,m)$, and the room information, $c_{h_r}(n,m)$.

3.2. The conventional channel normalization method

The cepstral mean subtraction (CMS) [9], the most widely used channel normalization method, improves the perform-ance of the speaker verification system by compensating the mismatch between the training and test environment, where the channel conditions of one case is different from the others. The CMS method removes the overall mean of the feature vectors from the each component over the entire enrolled speakers in the verification system. However, it cannot assure the improvement of the performance, because CMS may also remove the specific characteristics of the speaker.

3.3. Supplementary feature for verifying a specific speaker

In this paper, we perform preliminary experiments to search the supplementary features for speaker verification. In conventional speaker verification systems, we assume that the enrolled speakers have the same environmental condition. In other words, they are imposed to use the same microphone or assume to be located in the same place.

In some applications of speaker recognition such as forensic or requiring high security for verifying or identifying the claimants, the target speaker should tolerate some inconvenience to succeed the correct verification. Thus, we will assume that we know the target speaker's environment as well as the supplementary features discriminating the claimed speaker from the anti-speakers. Nevertheless we do not know the exact kind or form of the supplementary features. The examples of the process can be filtering operation using the pitch information of the target speaker or some kind of spatial information in verification environment.

The scenario of our verification experiment is follows:

- To enroll the own ID in the verification system, the user has to be placed in the specified environments such as his/her own office or living room.

- In the verification test, only true speaker can enter the places but impostors can not.

When the target speaker uttered using a microphone located far from speakers, the room impulse response of the specified environment will be applied. So, in this situation, the characteristic of target speaker who enters the place has the discriminated feature from the others. Consequently, the room environment plays a role in providing a supplementary feature to target speaker, this is equivalent to the $c_{h_r}(n,m)$ in Eq. (9). Thus, the new cepstral coefficient has a form of the addition between the cepstral coefficient of the original speech and the room impulse response.

4. Experiment

4.1. Experiment Setup

The GMM-based speaker verification is performed on the YOHO corpora supporting the text-independent mode. The particular vocabulary employed in this collection consists of two-digit numbers (”thirty-four”, ”sixty-one”, etc), spoken continuously insets of three (e.g. ”36-45-89”). For the YOHO database, there are 138 speakers (108 male and 30 female); for each speaker, there are 4 enrollment sessions of 24 utterances each, and 10 verification sessions of four utterances each, for a total of 136 utterances in 14 sessions per speaker. All waveforms are low-pass filtered at 3.8 kHz and sampled at 8kHz. To extract the feature vector, 12-features are computed from each 20 ms window (50The features are the 12-th order (DC component removed) MFCC.

To compute the supplementary feature, we use the room impulse response measured in the normal office room. Figure 2 shows the magnitude and phase response of the used room impulse function. And then, the supplementary feature related to the room impulse response is extracted through Eq. (6) (9). For the experiment performed with the YOHO corpora, each speaker is modeled by 32 mixture GMM trained using 24 training utterance in the enrollment session 1 only. Using all four enrollment sessions results in error rates that are too low to allow meaningful comparisons between the experiments [3][8]. For constructing of the background model to normalize the verification score, we use a simple universal background model [7].

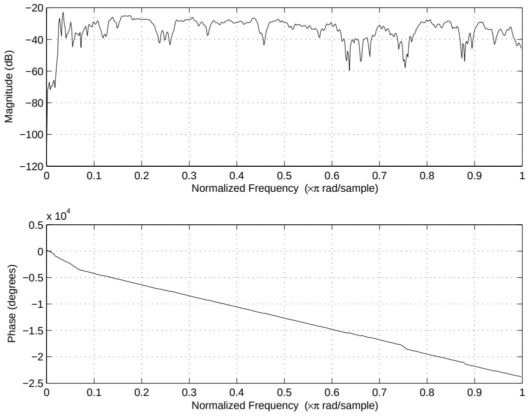

Fig. 2. Magnitude and phase spectrum of room impulse response

4.2. Verification experiment

A series of conducted experiments are summarized as follows:

To obtain the baseline system performance, the verification experiment is performed in a normal condition with or without the CMS method.

After adding the supplementary feature, the verification in the room environment is performed.

When the supplementary feature applies, the environments for the uttering the speech can be different from each other (true and impostor). Since the goal of our experiment is verifying the only one specific speaker (target), the likelihood probability of verification score is computed from different environments depending on whether the claimant is a true speaker or an impostor. In other words, the only true speaker can enter the place, but the impostors cannot. The related front-end processing of the verification system is shown in Figure 3.

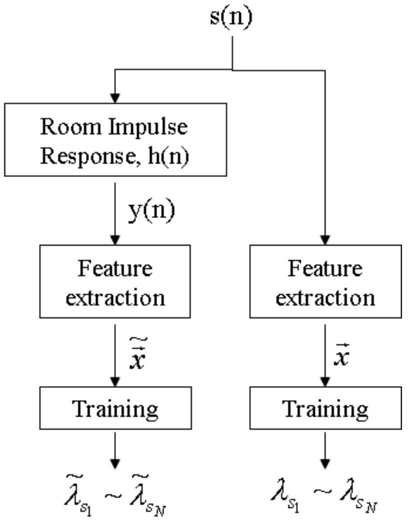
s(n)
Room Impulse Response, h(n)
y(n)
Feature extraction
Feature extraction
$\tilde{\vec{x}}$
$\vec{x}$
Training
Training
$\tilde{\lambda}_{s_1} \sim \tilde{\lambda}_{s_N}$
$\lambda_{s_1} \sim \lambda_{s_N}$

Fig. 3. Block diagram of the experiment setup

If a true speaker k claims to be verified, the verification score will be following,

$$LL_{target} = p(\tilde{\vec{x}}_k)|\tilde{\lambda}_k) - p(\tilde{\vec{x}}_k)|\lambda_{BGM}) \tag{10}$$

where $\tilde{\vec{x}}_k$, $\tilde{\lambda}_k$, and λ_{BGM} is an acoustic feature vector combined with supplementary feature, a speaker model trained with $\tilde{\vec{x}}_k$, and universal background model, respectively.

In the case of verifying an impostor claiming for speaker k, the verification score becomes

$$LL_{impostor} = p(\vec{x}_k)|\tilde{\lambda}_k) - p(\vec{x}_k)|\lambda_{BGM}) \tag{11}$$

where $\vec{x}_k$ is an acoustic feature vector extracted from normal condition, using closed microphone, i.e. does not include room acoustics because the impostor is assumed to be not able to enter the room.

5. Results

Figure 4 shows an average DET curves from verification results. Each curve corresponds to normal or room condition, and with or without CMS method.

We noticed interesting verification results. In the normal condition, the verification score is obtained using Eq. (5). As a well-known property, the CMS improves the overall performance of the verification system as shown in Figure 4. However, in the room condition, CMS does not improve the performance in comparison to the case without CMS. The reason of this result is caused from the removal of speaker dependent feature, i. e. room acoustics by CMS. The room impulse response plays a kind of specific supplementary-feature added to the original MFCC. It can be induced that the speaker dependent feature, crucial for verifying, is included in the room environment. The performance of the system in room condition without CMS represents noticeably good performance.

In Table 1, Equal-Error-Rate (EER) for each environment is shown. The EER value for the supplementary feature with-out CMS is the lowest compared to the others, corresponding to the DET curve.

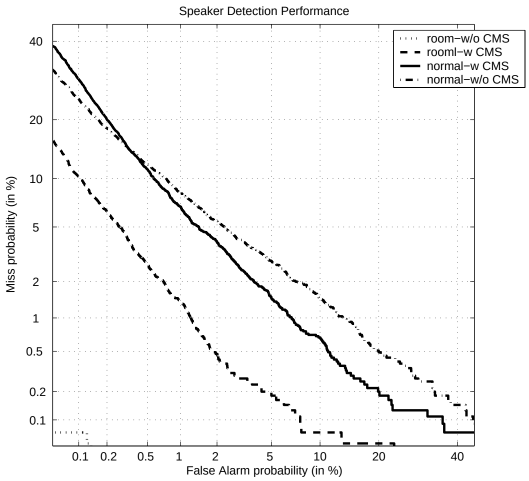

Fig. 4. Plot of DET curves from the different environment - normal or room condition, and with or without CMS method.

Table 1. EER for different condition

Environment	CMS	EER(%)
Normal	No	3.690
Normal	Yes	2.780
Room	Yes	1.142
Room	No	0.072

6. Conclusions

We have presented the results of preliminary experiment to search the supplementary feature for speaker recognition. Though, we do not know the exact form or component of optimal supplementary feature to the specific speaker, various kinds of approaches can be introduced to the appropriate ap-plication. Since we provided the constraint that only true speaker could enter the room, the true speaker is easily discriminated from the impostors. Simulation results showed that the performance of the proposed method was superior to the conventional feature. Further research on the supplementary features that are applicable only to the specific speakers is a challenging one.

7. Acknowledgments

This work was supported by the Biometric Engineering Research Center (KOSEF).

8. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[2] T. F. Quatieri, *Discrete time speech signal processing*, Prentice Hall, 2002.

[3] D. A. Reynolds, “Speaker identification and verification using gaussian mixture speaker models,” *Speech Communication*, vol. 17, pp. 91-108, August 1995.

[4] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Royal Stat. Soc.*, vol. 39, pp. 1-38, 1977.

[5] A. E. Rosenberg, J. DeLong, C-H Lee, B-H Juang, and F. K. Soong “The use of cohort normalized scores for speaker verification,” *Proc. ICSLP92*, pp. 599-602, Nov. 1992.

[6] R. J. McAulay and M. L. Malpass, “peech enhancement using a soft-decision noise suppression filter,” *IEEE Trans. on acoustic, speech and signal processing*, vol. ASSP-28, pp. 137-145, April, 1980.

[7] D. A. Reynolds, “Comparison of background normalization methods for text-independent speaker verification,” *Proc. Eurospeech 97*, vol. 2, pp. 963-966, 1997.

[8] D. Tran and M. Wagner, “A proposed likelihood transformation for speaker verification,” *Proc. ICASSP 2000*, vol. 2, pp.1069-1072, June, 2000.

[9] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. on Acoust. Speech and Signal Processing*, vol. 29, No. 2, pp. 254-272, April, 1981.