A Multi-SNR Subband Model for Speaker Identification under Noisy Environments

Kenichi Yoshida Kazuyuki Takagi Kazuhiko Ozeki

The University of Electro-Communications,
1-5-1, Chofugaoka, Chofu, Tokyo, Japan
{k-yoshida, takagi, ozeki}@ice.uec.ac.jp

Abstract

This paper describes a multi-SNR subband model for speaker identification under noisy environments. The model consists of a set of subband GMMs (Gaussian Mixture Models) trained on speech data corrupted with white Gaussian noise at several SNRs. In the recognition stage, an optimal GMM that yields the maximum accumulated likelihood on the whole input frames is selected for each subband. Then the likelihood is recombined over the subbands to give a speaker identification score. To evaluate the performance of this model, text independent speaker identification experiments were conducted under 5 different noisy environments:"bus","car","office","lobby", and "restaurant". For comparison, performance evaluation was also conducted on 3 other models: a subband model trained on clean speech, a multi-SNR fullband model, and a fullband model trained on clean speech. Results show that the multi-SNR subband model is very effective under a wide variety of noisy environments. Additional improvement was observed when an optimal GMM was selected on a short term basis instead of a whole input basis.

1. Introduction

Performance of a speaker recognition system is seriously affected by background noise. Robustness against noise is, therefore, a key issue in speaker recognition just as in speech recognition. Many methods have been proposed to improve the robustness of speaker recognition systems [1, 2, 3]. Subband methods use HMMs trained on each subband separately. By adjusting the recombination weights for subbands appropriately, it is possible to make good use of subband information for extracting speaker characteristics and suppressing noise effects [1, 2]. However, subband splitting and likelihood recombination alone are not enough to cope with a wide variety of background noises. Another well known method is one based on HMM composition [3], in which a speaker HMM is combined with a noise HMM to obtain a noisy speaker HMM. Although this method has been proved to be effective, one problem is that since there are so many types of noises, and noises are often changeable in the real world, we don't know beforehand what types of noisy HMMs we should prepare in practical applications.

In this paper, we present a multi-SNR subband model for speaker recognition. This model consists of a set of subband GMMs trained on speech data corrupted with white Gaussian noise at several SNRs. In the recognition stage, the system selects an optimal GMM for each subband that yields the maximum accumulated likelihood on the whole input frames. Then the likelihood is recombined over the subbands to give a final speaker identification score. The advantage of this model is that no assumption is made as to the type of background noise. Thus it is expected that the model can cope with a wide variety of unknown background noises.

To evaluate the recognition performance of this model, text independent speaker identification experiments were conducted under 5 different noisy environments: "bus", "car", "office", "lobby", and "restaurant". For comparison, performance evaluation was also conducted on 3 other models: a subband model trained on clean speech data, a multi-SNR fullband model, and a fullband model trained on clean speech data. To improve the tracking capability for temporal SNR changes in subbands, a slight modification of the speaker identification system was made and tested, in which an optimal GMM for each subband is selected on a short term basis instead of a whole input basis.

2. System overview

Figure 1 and 2 show the speaker identification system employed in this work. Figure 1 illustrates the training stage, and Figure 2 the identification stage.

In the training stage, a set of training data is created by adding white Gaussian noise to clean enrollment speech data at several SNRs. Then, the whole frequency band of each training data is equally divided into J subbands on mel-scale. For each subband a feature vector sequence is extracted, which is used for

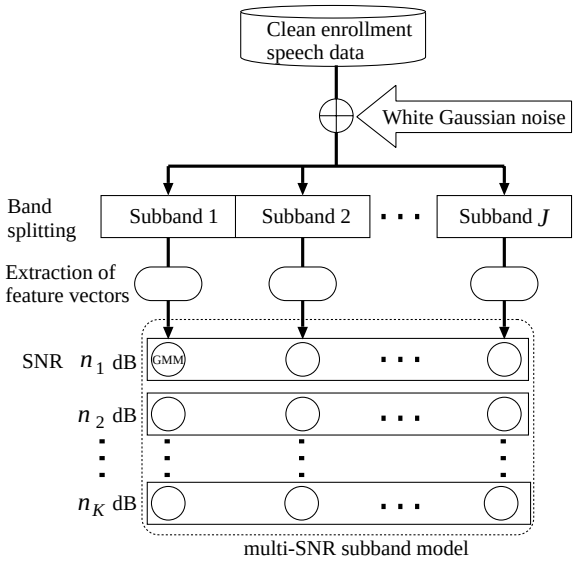

Figure 1: *Training stage.*

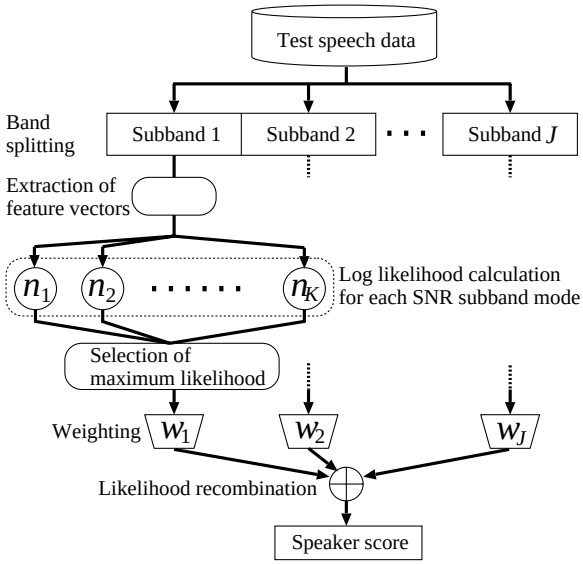

Figure 2: *Identification stage.*

training a GMM. Thus a GMM, denoted by $G(i,j,k)$, is created for the i^{th} registered speaker, the j^{th} subband, and the k^{th} SNR level. The set of GMMs thus created

$$\{G(i,j,k) \mid 1 \le i \le I, 1 \le j \le J, 1 \le k \le K\} \quad (1)$$

is referred to as a "multi-SNR subband model" here, where I is the number of registered speakers, J is the number of subbands, and K is the number of SNR levels. If i is fixed, then the set of GMMs

$$\{G(i,j,k) \mid 1 \le j \le J, 1 \le k \le K\} \quad (2)$$

is the model for the i^{th} registered speaker.

In the identification stage, a feature vector sequence is extracted from test speech data for each subband by the same way as in the training stage. As shown in Figure 2, the likelihood for each subband is calculated independently, then recombined. For a test speech data, the log likelihood score of the i^{th} registered speaker model is defined as follows:

$$Score(i) = \sum_{j=1}^{J} w_j L(i,j), \quad (3)$$

$$L(i,j) = \max_{k \in K} \sum_{t=0}^{T-1} f_{i,j,k}(t), \quad (4)$$

$$f_{i,j,k}(t) = \log P(v(j,t)|G(i,j,k)), \quad (5)$$

where $v(j,t)$ is the t^{th} frame of the feature vector sequence for the j^{th} subband, and w_j is the recombination weight for the j^{th} subband. The final decision is made as

$$Speaker = \underset{i \in I}{\text{argmax}}\ Score(i). \quad (6)$$

3. Speech material

3.1. Speech and noise database

NTT Voice Recognition Database was used in the experiments. From this database, 10 sentences were selected for enrollment data, and 10 words & 10 four-digits for test data, which were all read by 22 male and 13 female Japanese speakers in the same period.

For the noise source to create test utterances under noisy environments, Ambient Noise Database [4] was employed. From this database, 5 types of noises recorded in "bus", "car", "office", "lobby", and "restaurant" environments were selected.

3.2. Training data and test data

For each registered speaker, training data were created by adding white Gaussian noise to the clean enrollment data at 6 different SNRs (0, 10, 20, 30, 40, and ∞ dB). Also, test data were created for each registered speaker by adding the environmental noises to the clean test data. The SNRs of the test data were set at 0, 10, 20, 30, and 40dB.

3.3. Speech parameterization

Settings for speech parameterization are summarized in Table 1. The dimension of the MFCCs and the number of mel-scale filter channels will be explained in the next section.

Table 1: *Speech parameterization.*

sampling	16kHz,16bit
pre-emphasis	$1 - 0.97z^{-1}$
window type	Hamming
frame length	32ms
frame period	8ms
feature parameters	MFCCs

4. Experiments and results

4.1. Models to be compared

The speaker identification performance of the multi-SNR subband model was evaluated on the test data described in the preceding section. In addition to this model, 3 other models were also tested for comparison. Thus the following 4 models in total were tested and evaluated.

Multi-SNR subband model
Speech data of 6 SNRs as described in the preceding section were used in the training stage. The output of 56-channel mel-scale filter bank was split into 4 subbands, each having 14 channels. 10-dimensional MFCCs were derived from the output of these 14 channels, which were then used for training 16-mixture GMMs. The recombination weights were uniformly set: $w_1 = w_2 = w_3 = w_4 = 1/4$.

Clean subband model
Clean speech data was used in the training stage. Other settings were the same as in the multi-SNR subband model.

Multi-SNR fullband model
Speech data of 6 SNRs were used in the training stage as in the case of the multi-SNR subband model. The output of 56-channel mel-scale filter bank was used, without band splitting, to derive 48-dimensional MFCCs, which were then used for training 64-mixture GMMs.

Clean fullband model
Clean speech data was used in the training stage as in the case of the clean subband model. Other settings were the same as in the multi-SNR fullband model.

4.2. Results

Figure 3 (i) and (ii) show the speaker identification error rates for the 4 models under 5 noisy environments at 0dB and 10dB SNRs, respectively.

It is observed that the multi-SNR fullband model did not give improvement over the clean fullband model at 0dB SNR. At 10dB SNR, the multi-SNR fullband model performed only slightly better than the clean fullband model on the average. Thus the multi-SNR technique was not very effective when combined with fullband models.

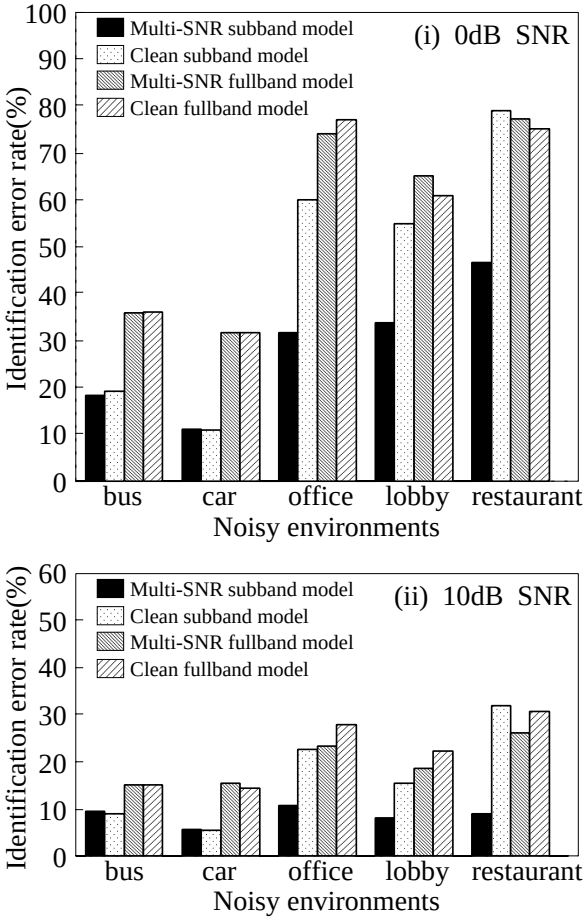

Figure 3: *Speaker identification error rates for the 4 models under 5 noisy environments at 0dB and 10dB SNRs.*

Under "bus" and "car" environments, the multi-SNR subband model and clean subband model had comparable performance. However, the multi-SNR subband model gave significant performance improvement over the clean subband model under "office", "lobby", and "restaurant" environments, where the clean subband model performed poorly. Although not shown in the graphs, these tendencies were the same at other SNRs as well.

In the "bus" and "car" cases, the noise spectrum is concentrated at the low frequency region, and the local SNR is fairly good at higher frequency regions. In the "office", "lobby", and "restaurant" cases on the other hand, the noises come from various sources such as human voices, telephones, and computers, so that the spectrum spreads over all the frequency regions. Thus the multi-SNR subband model is effec-

tive for complex, wide-band noises, which are difficult to cope with by other models.

4.3. Selection of optimal GMM on a short term basis

The short term subband SNR of speech uttered under a noisy environment will not be constant because the speech and background noises are non-stationary. Therefore, the definition of the speaker score was slightly modified so that an optimal GMM for each subband is selected on a short term basis instead of a whole input basis. The modified score for the i^{th} registered speaker on the j^{th} subband was defined as

$$L^{(M)}(i,j) = \sum_{n=0}^{N-1} \max_{k \in K} \sum_{m=0}^{M-1} f_{i,j,k}(nM+m), \qquad (7)$$

where M is the block length, and N is the number of blocks. Note that an optimal GMM for each subband is chosen based on the accumulated likelihood on each block, and the maximum likelihood is further accumulated over the blocks. The modified speaker score for the i^{th} registered speaker

$$Score^{(M)}(i) = \sum_{j=1}^{J} w_j L^{(M)}(i,j) \qquad (8)$$

was used in place of Eq.(3). Based on the modified speaker score, an additional speaker identification experiment was conducted.

Figure 4 shows the speaker identification error rates as functions of the block length M for the modified multi-SNR subband model at 0dB SNR. In the case of "bus" and "car" environments, the optimal block length turned out to be 1. In "office', "lobby", and "restaurant" cases, optimal block lengths were 20, 70, and 60, respectively. In all cases, the error rate was improved to some extent. Thus selection of optimal GMM on a short term basis is effective. However, the optimal block length depends on the type of environmental noise.

5. Conclusions

In this paper, a multi-SNR subband model was presented, and its performance in text independent speaker identification was shown under various noisy environments. It has been confirmed that the multi-SNR subband model is especially effective for complex, wide-band noises, which are difficult to cope with by other models. It has also been demonstrated that its performance can be improved by modifying the model so that an optimal GMM for each subband is selected on a short term basis. Further work is planned for automatic and dynamic determination of optimal recombination weights together with optimal block length for GMM selection under unknown

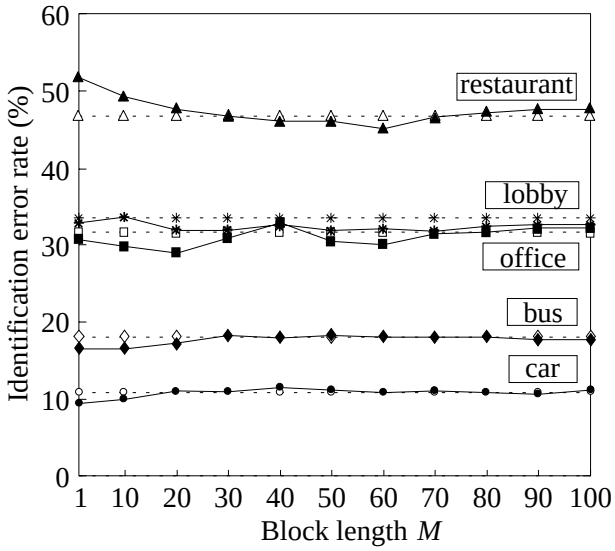

Figure 4: *Speaker identification error rates as functions of the block length M in the multi-SNR subband model at 0dB SNR. The solid lines are for the modified model, and the broken lines are for the original model.*

background noises. The multi-SNR subband model may also be applied to speaker recognition where the enrollment speech and test speech are uttered in different periods. Our future work also includes simplification of multi-SNR subband GMM generation process using the HMM composition technique [5].

6. Acknowledgement

The authors wish to thank NTT Speech Research Group for providing NTT VR Database.

7. References

[1] P. Sivakumaran, A. M. Ariyaeeinia and J. A. Hewitt, "Sub-band based speaker verification using dynamic recombination weights," *Proc. ICSLP'98*, Vol.2, pp.77–80 (1998).

[2] K. Yoshida, K. Takagi and K. Ozeki, "Speaker identification using subband HMMs," *Proc. Eurospeech'99*, Vol.2, pp.1019–1022 (1999).

[3] T. Matsui, T. Kanno and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Computer Speech and Language*, 10, pp.107–116 (1996).

[4] "Ambient Noise Database for Telephonometry 1996," NTT Advanced Technology (1996).

[5] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *Technical Report*,Cambridge University Engineering Department (1994).