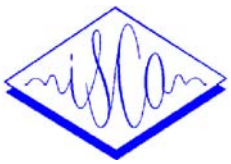

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SPEAKER IDENTIFICATION USING SUBBAND HMMS

K.Yoshida, K.Takagi and K.Ozeki

The University of Electro-Communications,
1–5–1, Chofugaoka, Chofu, Tokyo, Japan
Email: {yoshida, takagi, ozeki}@achilleus.cs.uec.ac.jp

10.21437/Eurospeech.1999-249

ABSTRACT

This paper is concerned with optimum band splitting and optimum recombination weights in subband HMM-based speaker identification. In the first experiment, the full frequency band (8kHz) was split into two subbands, and speaker identification rate was measured for various splitting frequencies and recombination weights. It was found that subbands 0-2kHz and 2-8kHz with equal recombination weights gave the best identification rate, outperforming a baseline method without band-splitting. In the second experiment, the full-band was split into three subbands with various splitting frequencies. Splitting into 0-2kHz, 2-6kHz, and 6-8kHz gave the best result, slightly outperforming the two-subband case. Finally, four-subband experiment was conducted, the result of which suggests that the speaker information and the phonemic information are complementary to a considerable degree in the spectral domain.

1. INTRODUCTION

Speaker characteristics extend over high frequency regions of speech signals that are irrelevant to phonemic information[1], though speaker recognition is still possible without using such high frequency regions as in the case of telephone speech. This means that speaker characteristics spread over the whole spectral domain. However, the contribution of spectral information to speaker recognition is not uniform over the frequency regions. Some frequency regions may contain phonemic information more than speaker characteristics, while others may contain only speaker characteristics. Moreover, when the speech signal is contaminated with noise, the S/N may vary depending on the frequency region. These observations lead to the idea of subband methods[3],[4], in which the full frequency band is split into multiple subbands for separate modeling and processing. The log-likelihood scores from the subband models are then recombined to give a final score for speaker recognition. Similar approaches have also been proposed for the task of speech recognition[5],[6].

Crucial issues in subband methods are how to split the full frequency band, and how to recombine likelihood scores from the subband models. In this work, three text-independent speaker identification experiments were conducted. In the first experiment, the full-band was split into two subbands, and the speaker identification rate was measured for various splitting frequencies and recombination weights to find out the best ones. In the second experiment, the full-band was split into three subbands with various splitting frequencies. In the third experiment, four subbands with equal bandwidths were employed to investigate the effectiveness of finer splitting, and the contribution of each frequency region to speaker identification.

2. SYSTEM OVERVIEW

Figure 1 shows the speaker identification system employed in this work.

In the enrollment stage, input speech data is first segmented into phones, and labeled accordingly by using speaker-independent HMMs. The phone segments are then band-split and analyzed by means of the selective linear predictive analysis method[2] to obtain feature vectors for each subband. Finally phone HMMs are formed by the standard training procedure using subband feature vectors. Thus, for each registered speaker, one HMM is created for each phone and for each subband. A speaker model comprises the set of these HMMs.

In the identification stage, test speech data is analyzed in the same way as in the enrollment stage to obtain sequences of feature vectors, one sequence for each subband. Then, for each registered speaker, the sequence of feature vectors for a subband is aligned with a free concatenation of phone HMMs corresponding to the same subband using the Viterbi algorithm. A weighted sum of the resulting log-likelihood scores for the subbands is calculated to give the score for the registered speaker. Finally, the registered speaker that attains the maximum score is selected to decide who the speaker is. It should be noted that this system performs text-independent speaker identification, because the alignment is done with a free concatenation of phone HMMs.

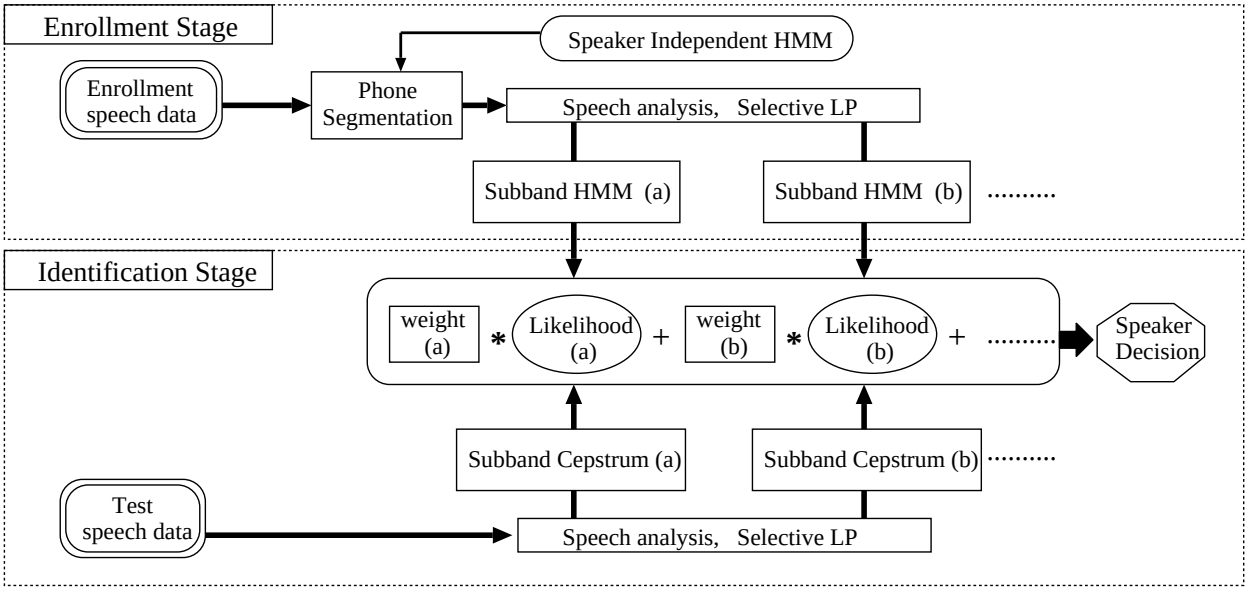

Figure 1: Subband-based speaker identification scheme.

3. SPEECH DATA AND SPEAKER MODELS

3.1. Speech Databases

An ATR Speech Database (Set C)[7] was used to form speaker-independent HMMs for phone segmentation. For speaker identification experiments, an NTT Voice Recognition Database was used. This database contains various items such as speaker's names, monosyllables, words, digits, four-digits, sentences, which were read by 23 male and 13 female speakers with three different speaking rates: slow, normal, and fast. The material is all in Japanese.

From the NTT-VR Database, ten sentences were selected for enrollment, and ten words + ten four-digits for test, which were read by 22 male and 13 female speakers in the same period.

3.2. Speech Analysis Conditions

Speech analysis conditions are summarized in Table 1. The dimension of a feature vector is the same as the LP analysis order.

3.3. Speaker Models

Six phone categories (five vowels and one non-vowel) were employed. The five vowels were /a/, /i/, /u/, /e/, and /o/, and the non-vowel category contained all the speech sounds that were not labeled as vowels. Each phone was represented, separately for each subband, by an HMM having 3-state, left-to-right, no skip transition structure with 4 Gaussian mixture components per state.

Table 1: Speech analysis conditions.

sampling	16kHz,16bit	
pre-emphasis	$1 - 0.97z^{-1}$	
window type	Hamming	
frame length	32ms	
frame period	8ms	
LP analysis	autocorrelation method	
LP analysis order	subband	16
	full-band	32
feature vector	LP mel-cepstrum	

Besides the subband speaker HMMs, full-band speaker HMMs of the same structure as subband HMMs were formed to provide a baseline system. The speaker-independent HMMs for phone segmentation in the enrollment stage were also of the same structure.

3.4. Likelihood Recombination

The log-likelihood scores from subband HMMs were recombined in the following manner:

$$P(x|\lambda^s) = \sum_k w_k P(x_k|\lambda_k^s)$$

where x is the test speech, and x_k is the sequence of feature vectors for the k^{th} subband. The s^{th} speaker model is represented by λ^s, λ_k^s being the model for the k^{th} subband. $P(x_k|\lambda_k^s)$ represents the log-likelihood obtained by aligning x_k, using the Viterbi algorithm, with a free concatenation of phone HMMs belonging to λ_k^s, and w_k is the weight for the k^{th} subband. Thus alignment was done for each subband independently

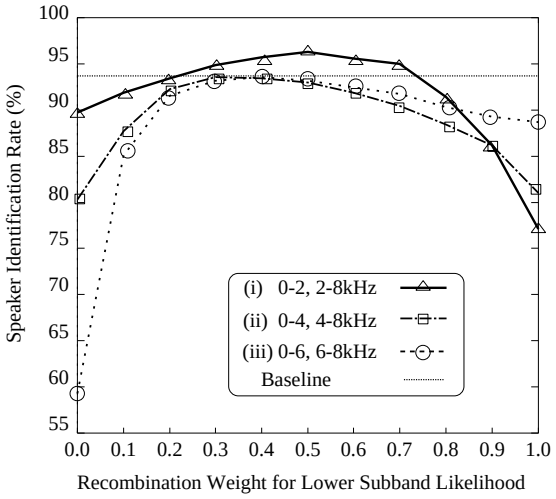

Figure 2: Speaker identification rates for three manners of subband splitting with various recombination weights in the two-subband case.

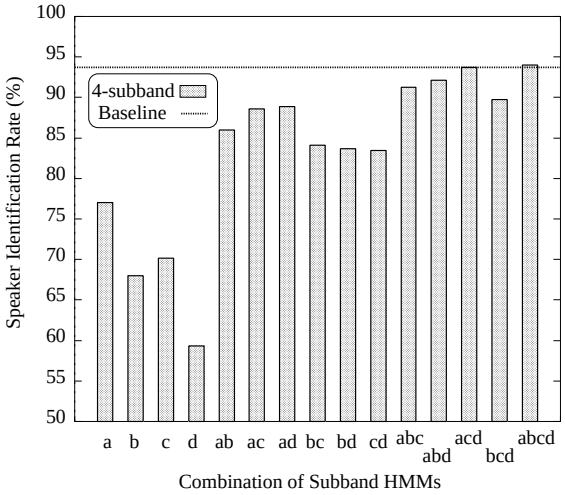

Figure 3: Speaker identification rates for subband combinations in the four-subband case.

and asynchronously of other subbands. The final decision was made by finding

$$s_{final} = \arg\max_{s} P(x|\lambda^s)$$

4. EXPERIMENTS AND RESULTS

4.1. Experiments

4.1.1. Two-subband case

The full-band (0-8kHz) was split into two subbands in three ways:
(i) 0-2kHz and 2-8kHz, (ii) 0-4kHz and 4-8kHz, and (iii) 0-6kHz and 6-8kHz.
The lower subband weight w_l, together with the higher subband weight $w_h = 1 - w_l$, was varied from 0 to 1 in steps of 0.1.

4.1.2. Three-subband case

The full-band was split into three subbands in three ways:
(i) 0-2kHz, 2-4kHz, and 4-8kHz, (ii) 0-2kHz, 2-6kHz, and 6-8kHz, and (iii) 0-4kHz, 4-6kHz, and 6-8kHz.
The subbands were given equal recombination weights.

4.1.3. Four-subband case

The full-band was split into four subbands with equal bandwidths of 2kHz:
(a) 0-2kHz, (b) 2-4kHz, (c) 4-6kHz, and (d) 6-8kHz.
Then, to investigate the importance of frequency regions, every combination taken out of these subbands was tried. When the subbands (a), (b), and (c) were used, for example, the recombination weights w_a, w_b, w_c, and w_d for (a), (b), (c), and (d) were set as $w_a = w_b = w_c = 1/3, w_d = 0$.

Table 2: Result of three-subband experiment.

Subbands (kHz)	Identification Rate(%)
0-2, 2-4, 4-8	95.4
0-2, 2-6, 6-8	96.6
0-4, 4-6, 6-8	91.7

4.2. Results

Figure 2, Table 2 and Figure 3 show the speaker identification rates for the two-subband case, the three-subband case, and the four-subband case, respectively. The baseline performance was obtained by using the full-band HMMs with 32th order of LP analysis as described in Section 3. Increasing the order up to 64 did not improve the baseline performance.

5. DISCUSSION

In the two-subband case, it is seen from Figure 2 that the best performance is obtained for the splitting frequency of 2kHz with equal recombination weights for the lower and the higher subbands. The identification rate 96.3% in this case outperforms the baseline rate 93.7%. For other splitting frequencies, 4kHz and 6kHz, the best performance is obtained when a smaller weight is given to the lower subband. The performance for these splitting frequencies, however, never exceeds the baseline. Thus, when two-subband splitting is employed, the best splitting frequency is

around 2kHz, and the subbands thus obtained have equal importance with respect to speaker identification.

In the three-subband case, Table 2 shows that when the recombination weights are uniform over the subbands, the best performance is attained for subband splitting into 0-2kHz, 2-6kHz, and 6-8kHz. In this case, the identification rate 96.6% is slightly higher than that in the two-subband case. Thus splitting into narrow subbands in the lower and the higher frequency regions, and a wide subband in the middle, seems to be effective.

In the four-subband case, as is seen from Figure 3, the best performance is naturally obtained when all the subbands are combined. However, the performance hardly exceeds the baseline. Therefore, finer subband splitting with equal recombination weights does not necessarily yield a better result. From the subband elimination result in Figure 3, the order of importance of subbands is estimated to be: (a) 0-2kHz > (d) 6-8kHz > (c) 4-6kHz > (b) 2-4kHz. It is interesting to note that very little performance degradation is observed by eliminating the subband 2-4kHz, which is considered to bear much phonemic information. Also, the second most important is the subband 6-8kHz, which contains little phonemic information. Thus, the phonemic information and the speaker information are complementary to a considerable degree in the frequency domain.

Considering the order of subband importance in the four-subband case, together with the best subband splitting in the two and three-subband case, the following splitting rule might be effective: assign a narrow band-width for an important frequency region, and a wide band-width for a less important frequency region.

6. CONCLUSIONS

In the subband-based speaker identification experiments, the following facts were observed:

1. In the two-subband case, the best splitting frequency is around 2kHz. The best performance, which exceeds the full-band baseline, is obtained when the lower subband likelihood and the higher subband likelihood are combined with equal weights.
2. In the three-subband case, when equal recombination weights are given to the subbands, the best performance is obtained for the subbands 0-2, 2-6, 6-8kHz, which slightly exceeds the performance in the two-subband case.
3. In the four-subband case, combining all the subband with equal weights hardly improves the performance. The most important subband is 0-2kHz, while the least important is 2-4kHz. The subband 6-8kHz is second most important.

From the order of importance of subbands in the four-subband case, together with the best subband splitting in the two and three-subband cases, the following splitting rule seems to be effective: assign a narrow subband for an important frequency region, and a wide subband for a less important frequency region. The order of subband importance also suggests that the phonemic information and speaker information are complementary to a considerable degree in the frequency domain.

The present work is only preliminary. Our future work includes determination of the optimum number of subbands, the optimum assignment of the subband widths, and the optimum values of recombination weights. Speaker recognition in noisy environments is another important application area.

ACKNOWLEDGMENT

The authors would like to thank NTT Speech Research Group for providing the NTT VR Database.

REFERENCES

[1] S.Hayakawa and F.Itakura, "Speaker recognition using speaker individual information in the higher frequency band," *The Journal of the Acoustical Society of Japan*, Vol.51, No.11, pp.861-868, 1995.

[2] J.Makhoul, "Spectral Linear Prediction: Properties and Applications," *IEEE Trans. Acoustics Speech and Signal Processing*, Vol.23, No.3, pp.283-296, June 1975.

[3] R.Auckenthaler,and J.S.Mason, "Equalizing subband error rates in speaker recognition," *Proc. Eurospeech'97*, pp.2303-2306.

[4] 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.

[5] H.Bourlard and S.Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. ICSLP'96*, Vol.1, pp.426-429.

[6] H.Hermansky, S.Tibrewala and M.Pavel, "Towards ASR on partially corrupted speech," *Proc. ICSLP'96*, Vol.1, pp.462-465.

[7] K.Takeda *et al.*, "Speech Database User's Manual," ATR Interpreting Telephony Research Laboratories, TR-I-0028, TR-A-0026, 1988.