Speaker Identification Using Harmonic Structure of LP-Residual Spectrum

Shoji Hayakawa, Kazuya Takeda and Fumitada Itakura

Nagoya University, Furo-cho 1 Chikusa-ku, Nagoya 464-01 JAPAN
hayakawa@itakura.nuee.nagoya-u.ac.jp
takeda@nuee.nagoya-u.ac.jp
ita@nuee.nagoya-u.ac.jp

Abstract. The harmonic structure of LP-residual spectrum is different in speakers. Therefore the harmonic structure may be useful for speaker recognition. In order to prove this hypothesis, *Power Difference of Spectra in Subband* (PDSS) is proposed as a new feature parameter to extract information of the harmonic structure of the linear prediction residual spectrum. VQ-based text-independent speaker identification experiments for 25 male and 25 female speakers are conducted to investigate the speaker identification ability of PDSS. Experimental results show that PDSS alone provides 66.9% maximal identification. In addition, it was found that the LPC cepstrum combined with PDSS results in a 41.2% reduction in identification errors compared with using only the LPC cepstrum. Moreover, a 52.4% reduction of identification errors over using only LPC cepstrum is attained by combining the LPC cepstrum with both delta cepstrum and PDSS. It is shown that PDSS can compensate for the LPC cepstrum and delta cepstrum for improving speaker identification performance.

1 Introduction

The spectral envelope has been used not only for speech recognition, but also for speaker recognition[10]. A powerful method to estimate the spectral envelope is linear prediction analysis[4][7]. Spectrum of the LP-residual signal, which can be obtained by applying the inverse LPC filter to a speech signal, is flatter and the formant information is almost removed. Therefore the LP-residual spectrum tends to be independent of phonetic information, and will have speaker individual information related to source periodicity. Several researchers have already examined the usefulness of the linear prediction residual signal in speaker recognition experiments[3][12][5]. They use cepstral representation of the LP-residual signal or the information related to fundamental frequency (F_0). The harmonic structure of LP-residual spectrum has not been used previously for speaker recognition.

The LP-residual spectra of the vowel /a/ uttered by 6 male speakers are shown in Figure 1. We can see that the dynamic range of the harmonic structure depends on speaker. The harmonic structure is expected to be effective as a speaker individual information.

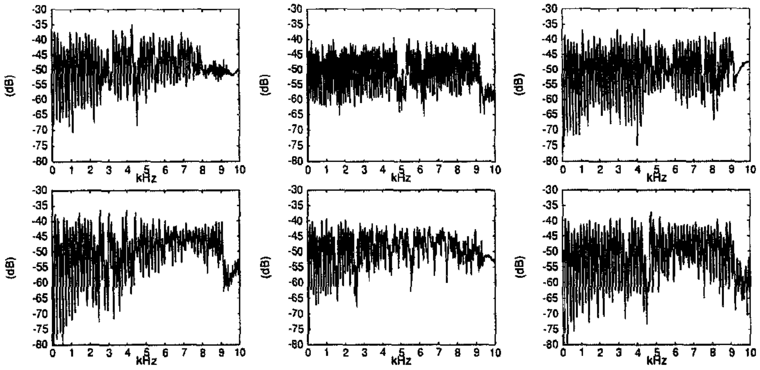

Fig. 1. Power spectra of LP-residual signal of the vowel /a/, uttered by 6 Japanese male speakers. Each power spectrum is normalized by the total power of the LP-residual signal.

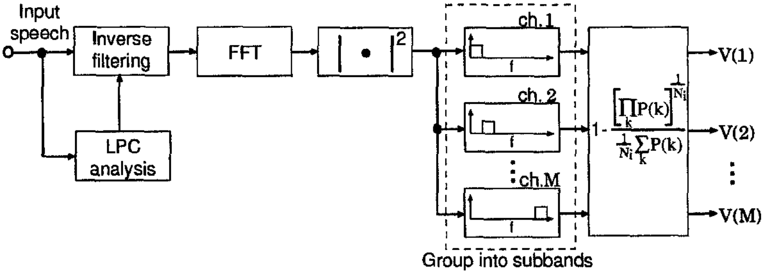

Fig. 2. Block diagram of used to calculate PDSS parameters.

In this paper, the effectiveness of using the harmonic structure for speaker recognition is demonstrated.

2 Feature extraction from the LP-residual signal

In order to parameterize the information in the harmonic structure of the LP-residual spectrum, we propose a new feature parameter, *Power Difference of Spectrum in Subband* (PDSS). This method utilizes the property that the larger the periodicity of LP-residual spectrum is, the bigger the power difference between peak and dip of the spectrum. PDSS is obtained as follows:

1. Calculate the LP-residual signal using the p-order linear prediction coefficients.
2. Append sufficient zeros to the LP-residual signal to increase the frequency resolution and calculate the power spectrum.
3. Group power spectrum into M subbands.

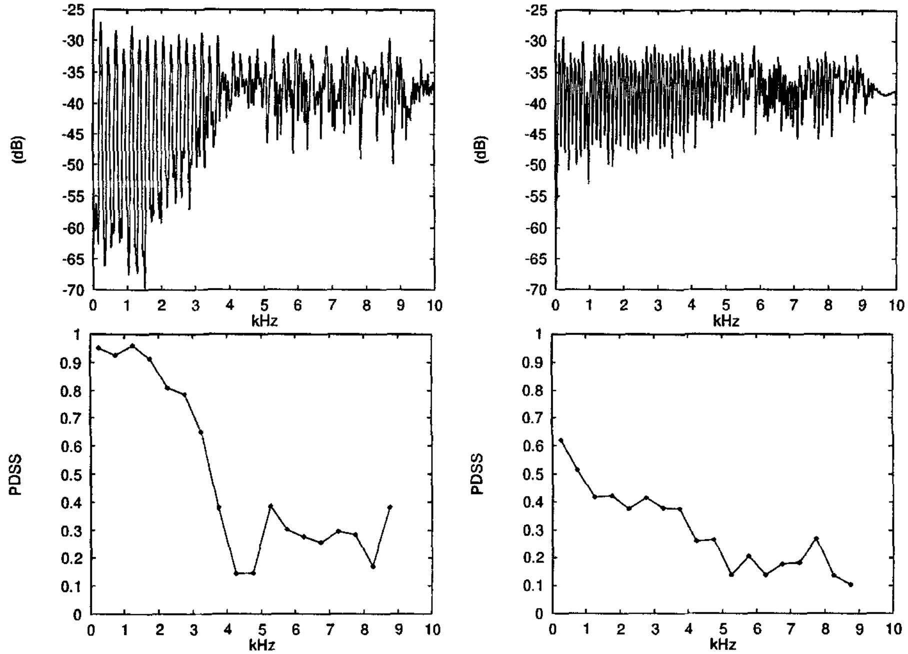

Fig. 3. LP-residual spectra of Japanese vowel /a/ uttered by a female (upper right-hand) and a male (upper left-hand) speakers and corresponding PDSSs (lower right-hand and left-hand).

4. Calculate the ratio of the geometric to the arithmetic mean of the power spectrum in ith subband, and subtract it from 1.0, and PDSS V(i) is obtained.

$$\mathrm{V}(i) = 1.0 - \frac{\left[\prod_{k=L_i}^{H_i} P(k)\right]^{1/N_i}}{\frac{1}{N_i}\sum_{k=L_i}^{H_i} P(k)}, \tag{1}$$

where $N_i = H_i - L_i + 1$ is the sample number of frequency points in ith subband and L_i, H_i is the lower and upper limit of frequency in ith subband respectively. A block diagram of this processing is shown in Figure 2.

Examples of PDSS analysis are shown in Figure 3. We can see that the higher the periodicity of the LP-residual spectrum, the closer to 1.0 the PDSS is, and the lower the periodicity, the closer to 0.0.

PDSS is interpreted as the subband version of *Spectral-flatness measure* which is introduced by Gray et al. [2][8] for quantifying the flatness of signal spectrum.

Table 1. Analysis conditions

Sampling frequency f_s	20kHz
Cutoff frequency f_c	9kHz
Frame length	32ms
Frame shift	8ms
Window	Hamming
LPC analysis order	32
LPC cepstral order	28
Delta cepstral window	72ms
Bandwidth of subband (PDSS)	500Hz
FFT points (PDSS)	8192

Table 2. Codebook size for each feature parameter.

PDSS	for voiced frame	64
LPC cepstrum	for voiced frame	64
	for unvoiced frame	64
Delta cepstrum	for voiced frame	64
	for unvoiced frame	64

3 Speaker identification experiments

3.1 Experimental conditions

The database consists of Japanese sentences uttered by 25 male and 25 female speakers every 3 months over 1 year. In the first recording session, 15 sentences are uttered for training, and in the other 4 sessions 18 sentences which are of different content from the first session, are uttered for test.

Analysis conditions are tabulated in Table 1. The LPC cepstrum and the delta cepstrum [1] are used as feature parameters of spectral envelope in Sec.3.3. Cepstral order is optimized through preliminary experiments, and 28 LPC cepstral and 28 delta cepstral coefficients are used. Although 20 subbands are obtained for a sampling frequency of 20kHz, since the bandwidth of each subband of PDSS is 500Hz, both 19th and 20th subbands are discarded due to the influence of the cutoff frequency $f_c = 9.0$kHz.

Codebooks of each speaker are made by the LBG algorithm [6] using reference speech data uttered in the first session. We use two kinds of codebooks for each speaker: codebooks for voiced frame and unvoiced frame[9]. Since PDSS uses the information in the harmonic structure of LP-residual spectrum, only codebook for voiced frame are used. Codebook sizes of each feature parameter are tabulated in Table 2. A block diagram of the speaker identification system is shown in Figure 4. Voiced and unvoiced frame decision is performed by the modified autocorrelation technique[4]. Distance measure is a weighted cepstral distance using the reciprocal of the intraspeaker cepstral variance[13].

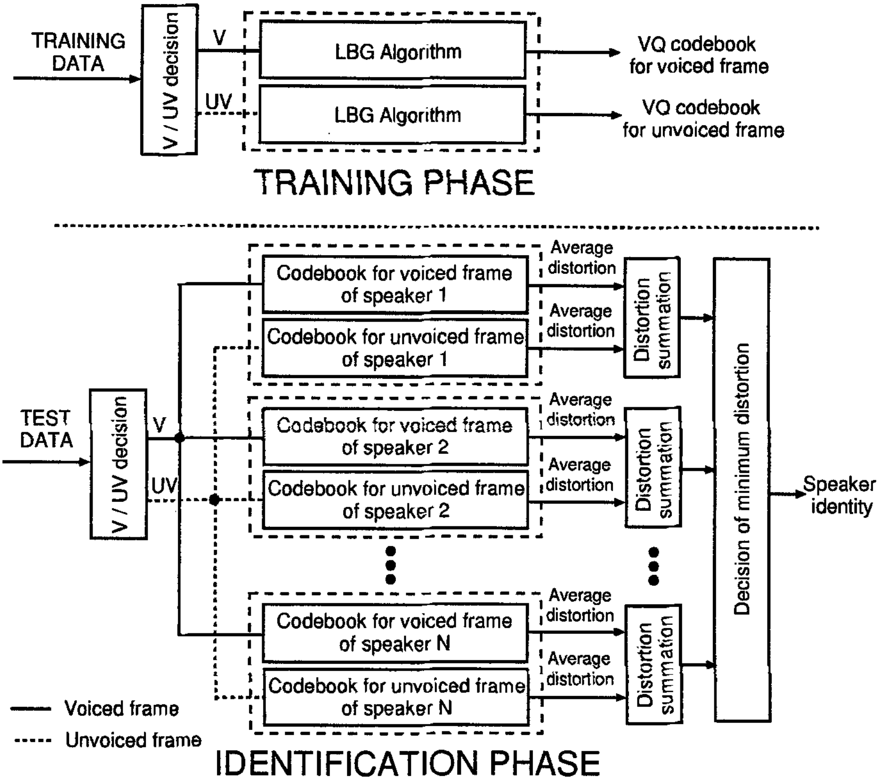
TRAINING DATA
V / UV decision
V
UV
LBG Algorithm
LBG Algorithm
VQ codebook for voiced frame
VQ codebook for unvoiced frame
TRAINING PHASE
TEST DATA
V / UV decision
V
UV
Codebook for voiced frame of speaker 1
Codebook for unvoiced frame of speaker 1
Codebook for voiced frame of speaker 2
Codebook for unvoiced frame of speaker 2
Codebook for voiced frame of speaker N
Codebook for unvoiced frame of speaker N
Average distortion
Distortion summation
Decision of minimum distortion
Speaker identity
Voiced frame
Unvoiced frame
IDENTIFICATION PHASE

Fig. 4. Speaker identification system using VQ distortions.

3.2 Stand alone recognition performance

First, we investigate the speaker recognition performance of PDSS alone. Unvoiced frames are discarded as we are interested in the speaker recognition performance using the information of the harmonic structure. Experiment results for the highest frequency band for PDSS are shown in Figure 5. Since the bandwidth of the subbands is fixed to 500Hz, the horizontal axis scale divided by 2 corresponds to the upper frequency used (in kHz) of PDSS. The best identification rate of 66.9% demonstrates that the harmonic structure of the LP-residual spectrum definitely contains useful information for identifying speakers.

3.3 The effects of combining the information of harmonic structure with the spectral envelope information.

It is expected that the PDSS tends to be independent of the phonetic information because the PDSS is obtained from the LP-residual spectrum which is almost flat. Therefore a better speaker identification performance will be obtained by

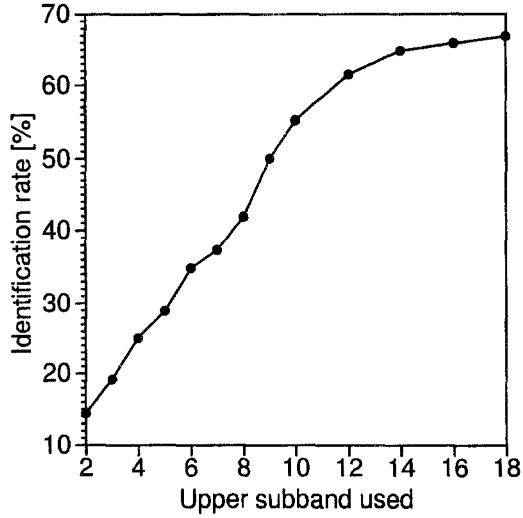

Fig. 5. Speaker identification rate of PDSS as a function of the number of subbands used (starting from the first subband).

Table 3. Summary of identification error rates in the case of single use of each feature parameter, and combining with some feature parameters (%).

PDSS (18 subbands)	33.1
PDSS (8 subbands)	74.9
LPC cepstrum	1.8
Delta cepstrum	2.6
LPC cepstrum + Delta cepstrum	1.3
LPC cepstrum + PDSS (8 subbands)	1.0
LPC cepstrum + Delta cepstrum + PDSS (8 subbands)	0.8

combining the features of spectral envelope with those of PDSS. Distortions of two and three feature parameters are linearly combined by means of equations (2) and (3) respectively[11].

$$d = \alpha \cdot \frac{d_1}{\overline{d_1}} + (1-\alpha) \cdot \frac{d_2}{\overline{d_2}} \tag{2}$$

$$d = \alpha \cdot \frac{d_1}{\overline{d_1}} + \beta \cdot \frac{d_2}{\overline{d_2}} + (1-\alpha-\beta) \cdot \frac{d_3}{\overline{d_3}} \tag{3}$$

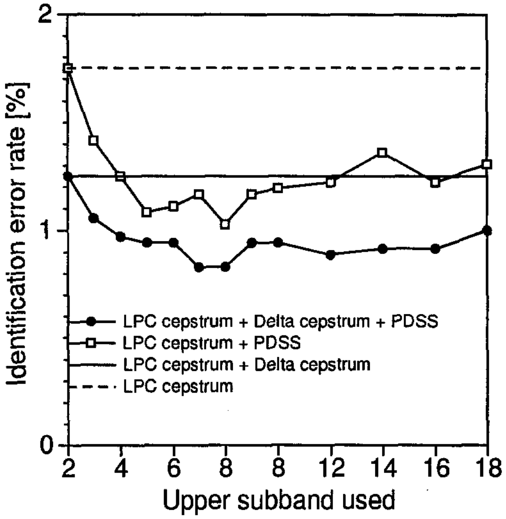

Fig. 6. Identification error rates obtained with optimal weighting.

d_1 : Distortion of 1st feature parameter,
$\overline{d_1}$: Average of pooled intra-speaker distortion of 1st feature parameter,
d_2 : Distortion of 2nd feature parameter,
$\overline{d_2}$: Average of pooled intra-speaker distortion of 2nd feature parameter,
d_3 : Distortion of 3rd feature parameter,
$\overline{d_3}$: Average of pooled intra-speaker distortion of 3rd feature parameter,
α, β : Combination factor $(0 \leq \alpha \leq 1, 0 \leq \beta \leq 1, \alpha + \beta \leq 1)$.

Experimental results when optimal weighting of α, β are shown in Figure 6. Identification error rates when using each feature parameter alone, and combining with some feature parameters are summarized in Table 3. Combining LPC cepstrum with PDSS gives maximal 41.2% reduction of identification errors of the LPC cepstrum alone when used subbands of PDSS are 8 subbands (0-4kHz). Moreover, a 52.4% reduction of identification errors over using only LPC cepstrum is attained by combining the LPC cepstrum with both delta cepstrum and PDSS. It is shown that PDSS is complementary to both the LPC cepstrum and delta cepstrum for improving speaker identification performance.

4 Conclusions

Power Difference of Spectra in Subband (PDSS) is proposed as a new feature parameter to extract speaker information from the harmonic structure of LP-residual spectrum. VQ-based text-independent speaker identification experiments for 25 male and 25 female speakers were conducted to investigate the speaker identification ability of PDSS. The obtained results can be summarized as follows:

1. PDSS alone gives identification rates of 66.9% for maximal.
2. Linear combination of the distortions in conjunction with both the LPC cepstral coefficients and the PDSS gives a better performance of 99.0% than using only the LPC cepstrum (98.2%).
3. The highest identification rate of 99.2% is attained by combining with LPC cepstrum, delta cepstrum and PDSS.

It is concluded that an additional amount of speaker individual information is contained in the harmonic structure of LP-residual spectrum, and PDSS is useful for speaker recognition.

References

1. Furui, S. : "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., & Speech, Signal Process.*, **ASSP-29**, No.2, pp.254–272 (1981).
2. Gray, A. H. Jr. and Markel, J. D. : "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans.Acoust.,Speech, & Signal Process.* **ASSP-22**, No.3, pp.207–217 (1974).
3. He, J., Liu, L. and Palm, G. : "On the use of features from prediction residual signals in speaker identification," *ESCA Proc. EUROSPEECH*, pp.313–316, (1995).
4. Itakura, F. and Saito, S. : "Analysis synthesis telephony based upon the maximum likelihood method," *Reports of 6th Int. Cong. Acoust.*, ed. by Y. Kohasi, C-5-5, pp.17–20 (1968).
5. Kashiwagi, H., Nakamura, S. and Takanashi, M. : "Speaker identification by spectral envelope of linear prediction residual," *IECE Trans. A* **J68-A**, No.7, pp.702–703 (1985). (in Japanese)
6. Linde, Y., Buzo, A. and Gray, R. M. : "An algorithm for vector quantizer design," *IEEE Trans. Comm.* **COM-28**, No.1, pp.84–95 (1980).
7. Makhoul, J. : "Linear prediction: A tutorial review," *Proc. of IEEE.* **63**, No.4, pp.561–580 (1975).
8. Markel, J. D. and Gray, A. H. Jr. : *Linear prediction of speech*, Springer-Verlag (1976).
9. Matsui, T. and Furui, S. : "Text-independent speaker recognition using vocal tract and pitch information," *Proc. ICSLP*, Vol.**1**, pp.137–140 (1990).
10. Rosenberg, A. E. and Soong, F. K. : "Recent Research in Automatic Speaker Recognition," *Advances in Speech Signal Processing,* ed.by S. Furui and M. M. Sondhi, pp.701–738, Marcel Dekker, New York, (1992).
11. Soong, F. K. and Rosenberg, A. E. : "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., & Speech, Signal Process.*, **ASSP-36**, No.6, pp.871–879 (1988).
12. Thèvenaz, P. and Hügli, H. : "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communication*, **17**, pp.145–157 (1995).
13. Tohkura, Y. : "A weighted cepstral distance measure for speech recognition," *IEEE Trans. Acoust., & Speech, Signal Process.*, **ASSP-35**, No.10, pp.1414–1422 (1987).