Speaker Identification by Combining Various Vocal Tract and Vocal Source Features

Yuta Kawakami[1], Longbiao Wang[1], Atsuhiko Kai[2], and Seiichi Nakagawa[3]

[1]Nagaoka University of Technology, Japan,
[2]Shizuoka University, Japan,
[3]Toyohashi University of Technology, Japan
{wang@vos,s123118@stn}.nagaokaut.ac.jp

Abstract. Previously, we proposed a speaker recognition system using a combination of MFCC-based vocal tract feature and phase information which includes rich vocal source information. In this paper, we investigate the efficiency of combination of various vocal tract features (MFCC and LPCC) and vocal source features (phase and LPC residual) for normal-duration and short-duration utterance. The Japanese Newspaper Article Sentence (JNAS) database was used to evaluate our proposed method. The combination of various vocal tract and vocal source features achieved remarkable improvement than the conventional MFCC-based vocal tract feature for both normal-duration and short-duration utterances.

Keywords: Speaker identification, phase, LPC residual, LPCC, GMM

1 Introduction

For the speaker identification task, many feature parameters had been used [1]. Mel-Frequency Cepstral Coefficients (MFCCs) [2] are basic feature for general speech processing. Linear Predictive Coding (LPC) [3,4] based features like the LPC Cepstral Coefficients (LPCC) [5] are also used. Line Spectral Frequencies (LSFs) [5] are coefficients in frequency domain, which are equivalent of LPC coefficients, this method is used for speech coding. Perceptual Linear Prediction (PLP) coefficients [6] consider psychophysics by using some human auditory-based filters, this also uses LPC method. These methods perform good also for speaker identification. Wang et al. used MFCC-based Gaussian Mixture Model (GMM) and LPCC-based Hidden Markov Model (HMM) for the distant speaker recognition, which worked well [7,8]. However, these feature parameters contain much vocal tract characteristics than that of vocal source. Vocal source characteristics are considered to be important for the speaker identification.

To catch vocal source characteristics, Markov et al. proposed a GMM-based speaker identification system that integrates pitch and the LPC residual with the LPC-derived cepstral coefficients [9]. Their experimental results show that using pitch information is most effective when the correlation between pitch and the cepstral coefficients is taken into consideration. Zheng et al. proposed Wavelet Octave Coefficients of Residues (WOCOR) which are based on LPC residual. They reported the improvement of speaker recognition performance by combining WOCOR with MFCC [10]. Recently, group

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 382–389, 2014.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

delay-based phase information has been used [11]- [13]. Group delay is defined as the negative derivative of the phase of the Fourier transform of a signal. Hedge et al. reported the improvement of the speaker recognition performance by combining MFCC with group delay [11]. However, the group delay based phase also contains power spectrum information, therefore, the complementary nature of the power spectrum-based MFCC and group delay phase was not sufficient enough.

Previously, Wang et al. proposed phase related features which is directly extracted from the Fourier transform of the speech wave for speaker recognition [14]- [21]. The phase information is valid for speaker identification, because it captures rich vocal source information. The combination of MFCC and phase information outperformed than the MFCC because the complementary nature of the power spectrum-based MFCC (vocal tract information) and phase spectrum-based feature (vocal source information) was used. However, the sufficient performance could not be achieved especially for short-duration utterance. That seems to be improved by combining various vocal tract and vocal source features which have complementary speaker information. In this paper, we investigate the efficiency of combination of various vocal tract features (MFCC and LPCC) and vocal source features (phase and LPC residual) for normal-duration and short-duration utterance.

The rest of this paper is organized as follows: Section 2 presents feature extraction method for the phase information and the LPC residual based feature. Section 3 describes combining method for two features. Section 4 discusses experimental setup and speaker identification results. Section 5 gives our conclusions.

2 Feature Extraction Method

In this section, we introduce two vocal source-based feature extraction methods, phase related feature and LPC residual-based feature.

2.1 Phase related features

The spectrum $S(\omega, t)$ of a signal is obtained by DFT of an input speech signal sequence

$$\begin{aligned} S(\omega,t) &= X(\omega,t) + jY(\omega,t) \\ &= \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times e^{j\theta(\omega,t)}. \end{aligned} \tag{1}$$

However, the phase changes, depending on the clipping position of the input speech even at the same frequency ω. To overcome this problem, the phase of a certain basis frequency ω is kept constant, and the phases of other frequencies are estimated relative to this. For example, by setting the phase of basis frequency ω to 0, we obtain

$$S'(\omega,t) = \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times e^{j\theta(\omega,t)} \times e^{-j\theta(\omega,t)}, \tag{2}$$

whereas for the other frequency $\omega' = 2\pi f'$, the spectrum becomes

$$\begin{aligned} S'(\omega',t) &= \sqrt{X^2(\omega',t)+Y^2(\omega',t)} \times e^{j\theta(\omega',t)} \times e^{-j\frac{\omega'}{\omega}\theta(\omega,t)} \\ &= \tilde{X}(\omega',t) + j\tilde{Y}(\omega',t). \end{aligned} \tag{3}$$

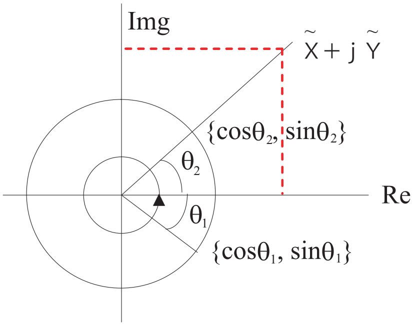

Fig. 1. Modified phase information

In this way, the phase can be normalized. Then, the real and imaginary parts of (3) become

$$\tilde{X}(\omega', t) = \sqrt{X^2(\omega', t) + Y^2(\omega', t)} \times \cos\Big(\theta(\omega', t) - \frac{\omega'}{\omega}\theta(\omega, t)\Big) \tag{4}$$

$$\tilde{Y}(\omega', t) = \sqrt{X^2(\omega', t) + Y^2(\omega', t)} \times \sin\Big(\theta(\omega', t) - \frac{\omega'}{\omega}\theta(\omega, t)\Big), \tag{5}$$

and the phase information is normalized as follows:

$$\tilde{\theta}(\omega', t) = \theta(\omega', t) - \frac{\omega'}{\omega}\theta(\omega, t). \tag{6}$$

In the experiments described in this paper, the basis frequency ω is set to $2\pi \times 1000Hz$. In a previous study, to reduce the number of feature parameters, we used phase information in a sub-band frequency range only. However, a problem arose with this method when comparing two phase values. For example, for two values $\pi - \tilde{\theta}_1$ and $\tilde{\theta}_2 = -\pi + \tilde{\theta}_1$, the difference is $2\pi - 2\tilde{\theta}_1$. If $\tilde{\theta}_1 \approx 0$, then the difference $\approx 2\pi$, despite the two phases being very similar to each other. Therefore, we modified the phase into coordinates on a unit circle [19], like fig. 1, that is,

$$\tilde{\theta} \rightarrow \{\cos\tilde{\theta}, \sin\tilde{\theta}\}. \tag{7}$$

In addition, we used the pseudo pitch synchronize method when clipping input speech. This method searches peak positions of the signal, and the positions are used as the center positions of the clipping window. Fig. 2 shows the overview of the method. We had confirmed the improvement of the speaker identification performance even in noisy environments by using pseudo pitch synchronization [20,21].

2.2 LPC residual based features

Linear Predictive Coding (LPC) is a basic method to get vocal tract characteristics, and its cepstram coefficients (LPCC) are generally used as the feature parameters [5,8]. From

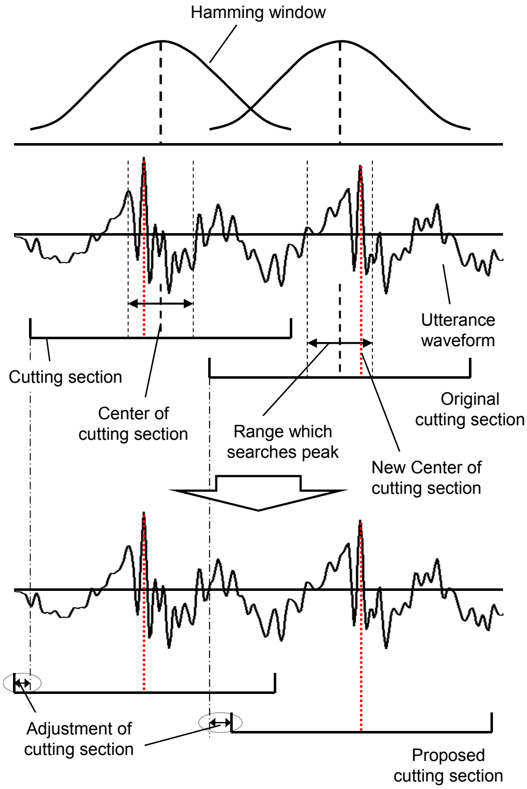
Hamming window
Utterance waveform
Cutting section
Center of cutting section
Range which searches peak
Original cutting section
New Center of cutting section
Adjustment of cutting section
Proposed cutting section

Fig. 2. Overview of the pseudo pitch synchronize method

LPC coefficients, the source signal is approximated as $\hat{s}(n)$ by a weighted sum of past samples

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k), \tag{8}$$

where p is the order of prediction, and a_k are LPC coefficients. The LPC residual signal, or prediction error $e(n)$ is calculated as the difference between the source signal and the predicted signal

$$\begin{aligned} e(n) &= s(n) - \hat{s}(n) \\ &= s(n) - \sum_{k=1}^{p} a_k s(n-k). \end{aligned} \tag{9}$$

The LPCC contains vocal tract characteristics, in contrast, the LPC residual is interpreted as the vocal source information [9,10]. In this work, the obtained LPC residual signal is transformed into cepstral coefficients using the standard mel-frequency filter-bank analysis technique by following steps [9].

1. Framing and windowing the LPC residual signal with the same rate and length as the original speech.
2. Obtaining the magnitude spectrum with FFT.
3. Forming filter banks in the mel scale.
4. Computing the log filter-bank amplitudes.
5. Calculating cepstral coefficients from the filter-bank amplitude using DCT.

3 Combination of various likelihoods based on different features

In this paper, the likelihood of GMMs based on one feature is combined with the likelihoods of GMMs based on other features. When a combination of the multiple methods is used to identify the speaker, the likelihoods are linearly coupled to produce a new score L^n_{comb} given by

$$L^n_{comb} = \sum_{i=1}^{I} \alpha_i L^n_{feat_i}, \quad n = 1, 2, \cdots, N, \quad \sum_i \alpha_i = 1, \tag{10}$$

where $L^n_{feat_i}$ are the likelihoods produced by the n-th speaker model based on i-th feature. N is the number of speakers registered, I is the number of feature and α_i donates the weighting coefficients for i-th feature, which are determined empirically. The speaker (or speaker model) with maximum likelihood is judged to be the target speaker.

4 Experiments

4.1 Experimental setup

We conducted speaker identification experiments for the JNAS (Japanese Newspaper Article Sentence) database [22] which contains 135 males and 135 females, about 100 clean utterances per person. The input speech was sampled at 16 kHz. Each utterance had about 6 seconds on average.

Speakers were modeled by GMMs with 128 mixtures from scratch. Each speaker models were trained by 5 utterances for 4 features (MFCC, LPCC, phase, LPC residual). The feature extraction conditions are shown in Table 1. We used the rest of database for the test, the number of test utterance was 23,160 (about 85 utterance per person). GMM likelihoods for multiple features were coupled as combination score. In this work, for the test data, we also used short utterance by cutting whole utterances into 2, 1 and 0.5 seconds, in addition to the whole one.

Table 1. Feature extract conditions

Feature	Vocal tract features		Vocal source features	
	MFCC	LPCC	Phase	LPC residual
LPC order		14		14
Frame length	25 ms	25 ms	12.5ms	25 ms
Frame shift	10 ms	10 ms	5 ms	10 ms
Dimensions	25 12 MFCCs, 12 Δs and a Δ power	25 12 LPCCs, 12 Δs and a Δ power	24 $\sin\tilde{\theta}$, $\cos\tilde{\theta}$ of the first 12 $\tilde{\theta}$ s of the phase spectrum (60-750 Hz range)	25 12 MFCCs, 12 Δs and a Δ power of the residual signal

Table 2. Speaker identification rates by single feature (%)

feature	whole	2 sec	1 sec	0.5 sec
MFCC	95.1	89.8	81.6	66.2
LPCC	95.3	90.1	82.3	68.0
Phase	90.8	79.3	64.0	46.2
LPC residual	94.5	87.2	78.5	63.7

4.2 Experimental results

Speaker identification rates by single features are shown in Table 2, and combination results are indicated in Table 3. Comparing with the vocal tract features, the LPCC performed better than the MFCC, and with the vocal source features, the LPC residual was better than the phase, for any length. Nevertheless, in the combinations of two features, the rates by “MFCC+Phase” exceeded that of “LPCC+LPC residual”. This means MFCC and Phase information has better complementarity than other combinations.

By combining various vocal tract and vocal source feature (combination of 4 features), the best identification rates were obtained. The results verify that performance of “MFCC+phase” or “LPCC+LPC residual” is not sufficient. For shorter utterances, identification rates were degraded. However, combination of all features achieved 48.0 % relative error reduction (66.2 % to 82.4 %) from the MFCC only, for 0.5 seconds utterances. This means we can get complementary information from the utterances by each feature, and the combination of them is effective for the short-utterance speaker identification.

5 Conclusions and Future Work

In this paper, we confirmed the efficiency of the vocal source information for speaker identification. Then the combination of MFCC and Phase performed the best in two features combinations. By combining 4 features, the identification rates were improved and the best performance was obtained. The results indicate that various vocal tract and vocal source features have complementarity for speaker recognition. In addition, the

Table 3. Speaker identification rates by multiple features (%), The combination coefficients are fine tuned for each test

feature	whole	2 sec	1 sec	0.5 sec
MFCC	95.1	89.8	81.6	66.2
MFCC+Phase	98.4	96.0	91.1	79.8
LPCC+LPC residual	96.3	92.7	86.7	74.5
Vocal tract feature (MFCC+LPCC)	96.5	93.0	87.1	75.5
all features	**98.4**	**96.3**	**92.2**	**82.4**

combination method was effective for short-utterance speaker identification. However, for short utterances, the improvement of the identification rates might be insufficient. Hereafter, we address this problem by improving the feature extraction method.

References

1. T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors", Speech Communication Vol. 52, No. 1, pp. 12-40, (2010).
2. S. Davis, B. Santa and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 28, Issue 4, pp. 357-366 (1980).
3. J. Makhoul, B. Bolt and Newman, "Linear prediction: A tutorial review", Proc. of IEEE Vol.63, Issue 4, pp.561-580 (1975).
4. R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach", IEEE Signal Processing Magazine, 13, Sept. pp. 58-71 (1996).
5. X. Huang, A. Acero and H. W. Hon, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development", Prentice-Hall, New Jersey (2001).
6. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", The Journal of the Acoustical Society of America 87 (4), pp. 1738-1752
7. L. Wang, N. Kitaoka and S. Nakagawa, "'Robust Distant Speaker Recognition Based on Position Dependent Cepstral Mean Normalization", Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech'2005-Eurospeech), pp.1977-1980 (2005)
8. L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speaker recognition based on position dependent CMN by combining speaker-specific GMM with speaker-adapted HMM", Speech Communication 49, pp. 501-513, (2007).
9. K.P. Markov and S. Nakagawa, "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition", Jour. ASJ (E), Vol.20, No. 4, pp. 281-291 (1999).
10. N. Zheng, T. Lee and P.C. Ching, "Integration of complementary acoustic features for speaker recognition", IEEE Signal Processing Letters, Vol. 14, No. 3, pp. 181-184 (2007).
11. R. M. Hedge, H. A. Murthy, and G. V. R. Rao, "Application of the modified group delay function to speaker identification and discrimination", Proc. ICASSP 2004, Vol. 1, pp.517-520 (2004).
12. R. Padmanabhan, S. Parthasarathi and H. Murthy, "Robustness of phase based features for speaker recognition", Proc. Interspeech, pp.2355-2358 (2009).

13. J. Kua, J. Epps, E. Ambikairajah and E. Choi, “LS regularization of group delay features for speaker recognition”, Proc. Interspeech, pp. 2887-2890 (2009).
14. S. Nakagawa, K. Asakawa, and L. Wang, “Speaker recognition by combining MFCC and phase information”, Proc. InterSpeech, pp. 2005-2008, (2007).
15. L. Wang, S. Ohtsuka and S. Nakagawa, “High improvement of speaker identification and verification by combining MFCC and phase information”, Proc. ICASSP, pp.4529-4532, (2009).
16. L. Wang, K. Minami, K. Yamamoto and S. Nakagawa, “Speaker identification by combining MFCC and phase information in noisy environments”, Proc. ICASSP, pp.4502-4505, (2010).
17. L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, “Speaker recognition by combining MFCC and phase information in noisy conditions”, IEICE Transactions on Information and Systems, Vol. E93-D, No. 9, pp. 2397-2406 (2010).
18. Y. Hirano, L. Wang, A. Kai, and S. Nakagawa, “On the Use of Phase Information-based Joint Factor Analysis for Speaker Verification under Channel Mismatch Condition”, Proc. of APSIPA ASC 2012, (4 pages) (2012).
19. S. Nakagawa, L. Wang and S. Ohtsuka, “Speaker Identification and Verification by Combining MFCC and Phase Information”, IEEE Trans. on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1085-1095 (2012).
20. K. Shimada, K. Yamamoto and S. Nakagawa, “Speaker identification using pseudo pitch/synchronized phase information in voiced sound”, Proc. APSIPA ASC 2011, pp.1-6 (2011).
21. Y. Kawakami, L. Wang and S. Nakagawa, “Speaker Identification Using Pseudo Pitch Synchronized Phase Information in Noisy Environments”, Proc. APSIPA ASC 2013, (5 pages) (2013).
22. K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano and S. Itahashi, “JNAS:Japanese speech coupus for large vocabulary continuous speech recognition research”, J. Acoust. Soc. Jpn. (E), Vol. 20, No. 13, pp. 199-206 (1999).