2004 8th International Conference on
Control, Automation, Robotics and Vision
Kunming, China, 6-9th December 2004

CELP-based Speaker Verification: An Evaluation under Noisy Conditions

Yasushi YAMAZAKI†, Yusuke FUJITA††, and Naohisa KOMATSU††

†Dept. of Information and Media Sciences, The University of Kitakyushu
1-1 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka, 808-0135 Japan
E-mail: yamazaki@env.kitakyu-u.ac.jp
††Dept. of Computer Science, Waseda University
3-4-1 Ohkubo, Shinjuku-ku, Tokyo, 169-8555 Japan
E-mail: {fujita,komatsu}@kom.comm.waseda.ac.jp

Abstract

We propose a text-independent speaker verification method based on a speech coding scheme. The proposed method utilizes CELP parameters which are used in speech coding schemes for mobile communication systems, and verifies a speaker only with the encoded speech information. The reliability of the proposed method under noisy conditions is mainly discussed with some simulation results.

1 Introduction

With the recent advances in information and communication systems, the need to protect one's privacy has been increasing more than ever. Especially, the development of an identity verification scheme to verify specific users is one of the key technologies to provide user-oriented network services in the future. In many verification systems, a user is verified by something she/he knows or possesses. It is true that these kinds of parameters are easy to handle at a low cost. However, at the same time, these parameters have the problem concerning human errors such as forgetting passwords or losing ID cards. To solve this problem, many approaches to identify oneself using biometrics have been proposed[1]. Especially, speech is one of the typical biometric parameters and is also used generally in person-to-person communications. In other words, speech contains semantic information as well as singular information.

In the case of digital communication systems, speech information is encoded for transmission. It should be noted that the encoded speech also contains semantic information and singular information. Therefore, we can expect to realize a speech or speaker recognition system by using only the encoded speech. However, sufficient research has not been carried out on the use of encoded speech for speaker recognition.

Taking the above suggestions into account, we have introduced a speaker verification method based on a speech coding scheme in digital transmission systems[2]. In the proposed method, we utilize CELP (Code Excited Linear Prediction) parameters which are used in speech coding schemes for mobile communication systems or IP networks. The merits of the proposed method are as follows: (1) Speaker verification is easily realized in the current mobile terminals or network systems by adding a little function. (2) Since CELP parameters contain a speaker's characteristics of articulation, text-independent speaker verification is realized. However, in the proposed method, there is a problem that verification accuracy is degraded under the condition in which speech and noise coexist. Therefore, in this paper, we propose a noise-robust speaker verification method to improve our previous one. Moreover, we discuss the reliability of the proposed method with some simulation results using noise-mixed speech data.

2 Process of speaker verification

In this section, we describe the proposed speaker verification method. The proposed method consists of CELP coding block and speaker verification block (see Figure 1). For example, a mobile terminal with the function of speaker verification is realized by adding only the speaker verification block in Figure 1 to a current mobile terminal. Moreover, the speaker verification block consists of enrollment process and verification process. In the following subsections, we will explain the above-mentioned blocks.

2.1 CELP coding block

In the CELP coding block, speech signal is encoded by a CELP coder and the encoded parameters are extracted. In general, CELP is an AbS (Analysis by Synthesis) coding scheme with an excited sound source which is prepared in a codebook. In the proposed method, we use a CS-ACELP (Conjugate-Structure Algebraic-Code-Excited Linear-Prediction)[3] coder, a kind of CELP coder which is standardized by ITU-T as ITU-T Recommendation G.729 (see Figure 2). This coder operates on speech frames of 10 ms corresponding to 80

0-7803-8653-1/04/$20.00

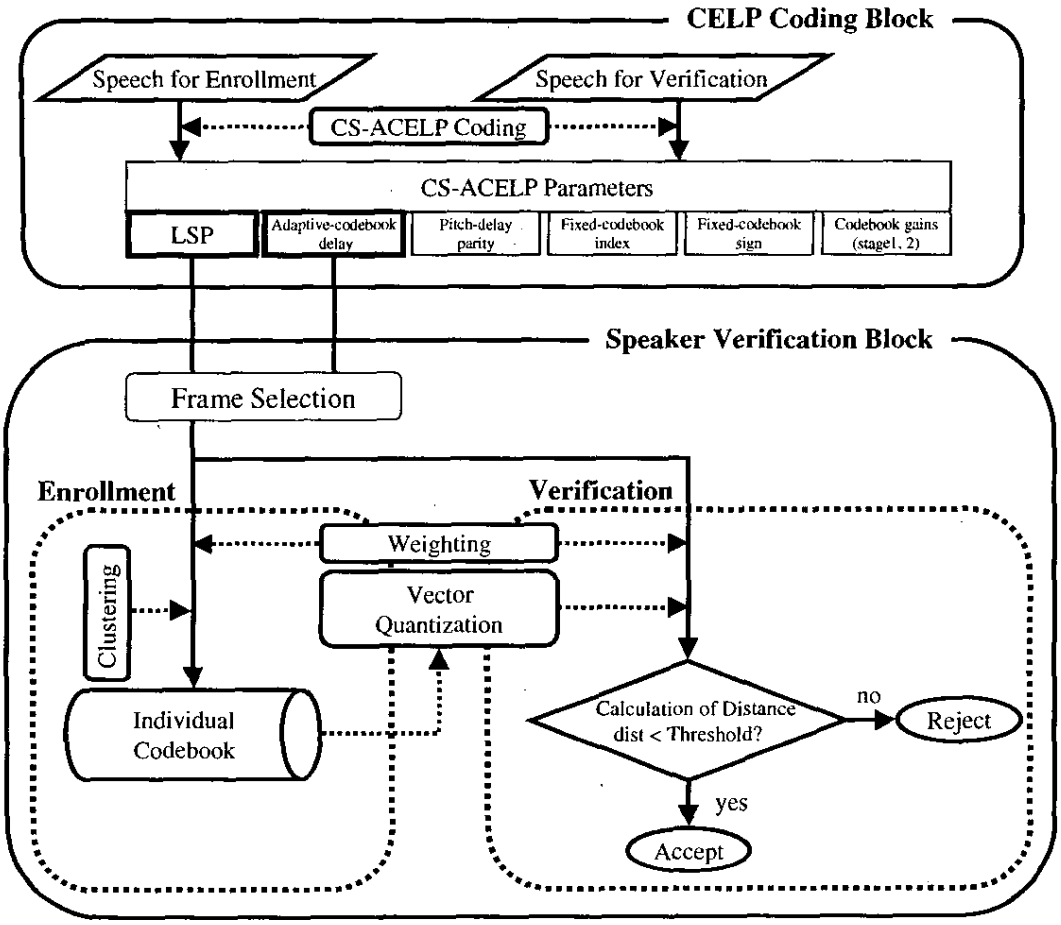

Figure 1: Enrollment and verification process of speaker verification

samples at a sampling rate of 8000 samples per second. For every 10 ms frame, the speech signal is analyzed to extract the parameters of the CELP model. Among those parameters, we focus on a parameter called LSP (Line Spectrum Pair), which corresponds to a parameter for articulation of speech. It is well known that a speaker's individual characteristics is shown in articulation. Therefore, it is expected that LSP contains a speaker's personal features which are unique to each speaker and useful for speaker verification. Considering above points, we use LSP as a principle parameter of the proposed speaker verification method.

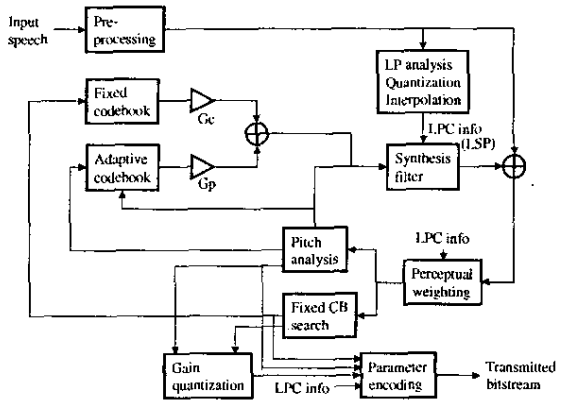

Figure 2: Principle of CS-ACELP coding[3]

2.2 Speaker verification block

The speaker verification block contains two subprocesses, enrollment process and verification process. Each subprocess is preceded by another process called frame selection, which is explained below. In the enrollment process, we extract LSP (hereafter, enrollment LSP) for enrollment and define a codebook (hereafter, an individual codebook) which is produced by clustering the enrollment LSP as a speaker's personal features. On the other hand, in the verification process, we verify a speaker based on the quantization error in the process of vector quantization of LSP (hereafter, verification LSP) by referring to the individual codebook.

2.2.1 Frame selection

In this process, speech frames which are effective for speaker verification are extracted. To realize noise-robust speaker verification, it is important to select stable frames in which the value of LSP is not affected by noises. In our preparatory experiment, it is clear that there exist some frames in which the value of LSP varies little regardless of the existence of noises. Moreover, the following should be taken into account: (1) The fluctuation of LSP is small in a frame where pitch is stable. (2) Pitch is more stable in voiced sounds than in unvoiced sounds. (3) Pitch is more stable in active voice than in non-active voice (silence). (4) In some cases, pitch can be stable in non-active voice. Taking the above suggestions into account, we propose a frame selection method by the combination of two processes, pitch stability decision and voice activity decision (see Figure 3).

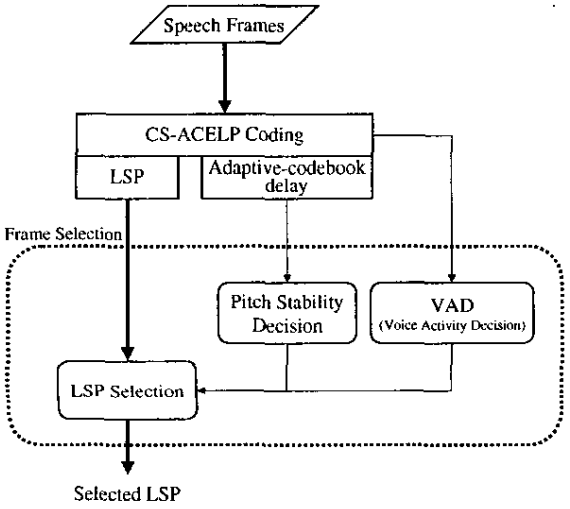

Figure 3: Outline of the frame selection

In the process of pitch stability decision, it is decided whether the pitch in the target frame is stable or not based on 'adaptive-codebook delay', a parameter of the CS-ACELP coding as follows:

1. A pitch sequence is extracted from a set of values of the adaptive-codebook delay. The sequence consists of 10 points in which the fifth point P, the center of the sequence, corresponds to the target frame.
2. The extracted pitch sequence is linearly approximated by LSM (Least Square Method) and D, the corresponding MSE (Mean Square Error), is calculated.
3. When the following condition is satisfied, it is decided that the target frame is stable; $P < P_{th}$ and $D < D_{th}$, where P_{th} and D_{th} are the thresholds of P and D, respectively.

On the other hand, in the process of voice activity decision, it is decided whether the target frame corresponds to active voice or non-active voice based on the VAD (Voice Activity Decision) algorithm[4]. The VAD algorithm is used to reduce the transmission rate during silence periods of speech and standardized by ITU-T as ITU-T Recommendation G.729 (Annex B). In the VAD algorithm, voice activity decision is made by using the following measures; a spectral distortion, an energy difference, a low-band energy difference, and a zero-crossing difference (see Figure 4).

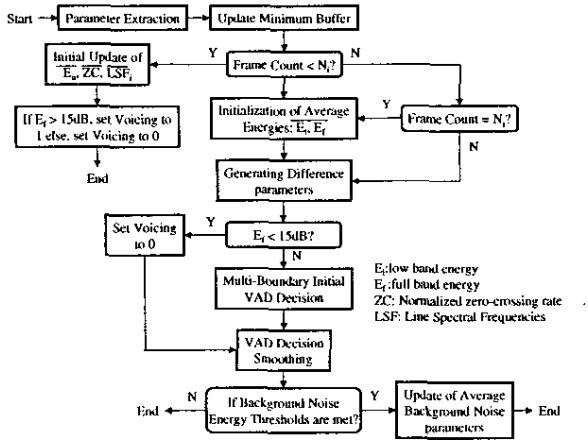

Figure 4: VAD algorithm[4]

In the process of frame selection, the LSP in the target frame is selected and utilized in the subsequent processes when the following conditions are satisfied:

1. It is decided in the pitch stability decision that the pitch is stable in the target frame.
2. It is decided in the voice activity decision that the target frame corresponds to active voice.

2.2.2 Enrollment process

First, a weighting process for emphasizing the characteristics of the enrollment LSP is carried out with a view to emphasize each speaker's personal features. In the proposed method, a value of a speaker's i-th order LSP at n-th frame is defined as $LSP_n(i)$. Similarly, an average value of the speaker's i-th order LSP over all N frames is defined as $\overline{LSP(i)}$. In the weighting process, $LSP_n(i)$ is emphasized into $LSP_n(i)'$ by multiplying a constant value k as follows:

$$LSP_n(i)' = LSP_n(i) + k \cdot \overline{LSP(i)} \quad (1)$$

$$\overline{LSP(i)} = \frac{1}{N} \sum_{n=1}^{N} LSP_n(i) \quad (2)$$

Second, an individual codebook is produced by applying a clustering algorithm to the weighted enrollment LSP. In the proposed method, we use the LBG algorithm[5] as the clustering algorithm. The individual codebook can be stored in a user's smart card or her/his mobile terminal. In the case of using a smart card, for example, the card is inserted into her/his mobile terminal and will be activated only when she/he is verified to be an authorized user by the proposed method.

2.2.3 Verification process

First, a speaker's speech is encoded by the CS-ACELP coder and the produced LSP (verification LSP) is used for verification. Second, the Euclidean distance between the weighted verification LSP and the vectors in the individual codebook is calculated. In the case that the calculated distance 'dist' is smaller than the preset threshold value, the speaker is accepted as an authorized user.

3 Reliability test

In this section, the reliability of the proposed method is shown using simulation results. In the reliability test, we used a Japanese speech database constructed by ATR (Advanced Telecommunication Research Institute)[6]. Moreover, we used a noise database constructed by JEIDA (Japan Electronic Industry Development Association)[7] to evaluate the reliability under the condition that speech signal is contaminated by some noises. A detailed description of the noise data extracted from the above database and used in the test is shown in Table 1. From these databases, we made a set of noise-mixed speech data, in which SN ratio was set to some kind of levels. Parameters for the reliability test are shown in Table 2. These parameters were chosen, based on the results of a preparatory experiment, so as to achieve stable extraction of personal features. It should be noted that the speech data for enrollment are different from those for verification.

name	description
[car]	Background noise in a running car
[crowd]	Background noise at a crowded plaza
[station]	Background noise at a crowded station
[street]	Background noise at a congested crossing

Table 1: Description of noise data

speech data	ATR Japanese speech database (20 males and 20 females)
noise data	JEIDA noise database ([car], [crowd], [station], [street])
SN ratio	5dB, 10dB, 15dB, 20dB
sampling frequency	8 kHz
cut-off frequency	3.1 kHz
speech length	enrollment : 15 seconds verification : 10 seconds
individual codebook	10 dimensions, 16 levels
frame selection	$P_{th} = 120, D_{th} = 20$
weighting	$k = 20$

Table 2: Parameters for the reliability test

Figure 5 and 6 show the FRR (False Rejection Rate) and the FAR (False Acceptance Rate) for noise-free speech data. The EER (Equal Error Rate) is 1.2% in Figure 5 and 3.6% in Figure 6. These results suggest that the proposed method is superior to our previous one, which are caused by suitable selection of LSP by the frame selection process. Also, it is clear that the proposed method is effective for text-independent speaker verification.

On the other hand, Table 3 shows the EER under the condition that the speech signal was contaminated by some noises. Comparing the case of using the frame selection and the case of not using the frame selection, the former is superior to the latter from the viewpoint of verification accuracy. In the former case, the EER is relatively small in the noisy conditions. It is obvious that the proposed frame selection is effective for enhance the robustness against background noises. These results suggest that the proposed method verifies a speaker under a realistic condition in which speech and noise coexist.

EER [%] with frame selection				
SNR[dB]	[car]	[crowd]	[station]	[street]
5	1.7	8.6	22.0	14.7
10	1.4	3.4	12.5	5.9
15	1.3	1.8	5.0	2.7
20	1.1	1.2	2.4	1.7
noise free	1.2			
EER [%] without frame selection				
SNR[dB]	[car]	[crowd]	[station]	[street]
5	31.8	14.8	21.8	23.6
10	23.8	9.7	11.9	18.1
15	18.1	7.9	7.2	15.0
20	11.4	6.7	4.8	9.0
noise free	3.6			

Table 3: Verification results under noisy conditions

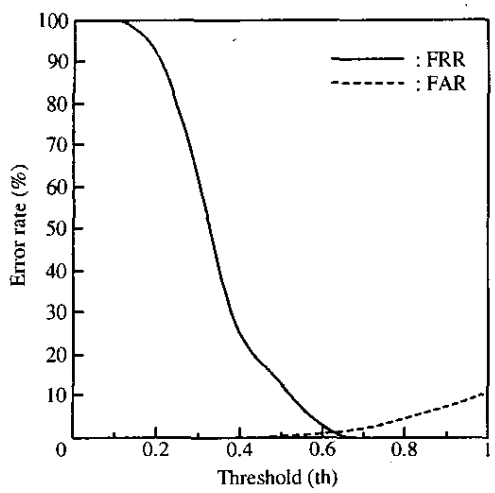

Figure 5: Verification results with frame selection

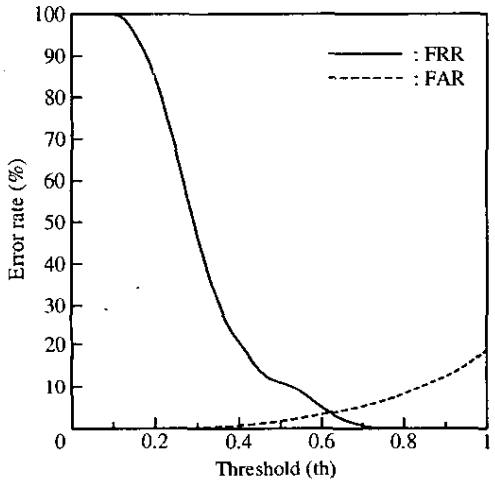

Figure 6: Verification results without frame selection

4 Conclusion

In this paper, we proposed a text-independent speaker verification method using CELP parameters. We also evaluated the performance of the proposed method under noisy conditions. From the simulation results, it is clear that the proposed method is effective for speaker verification using encoded speech information such as speaker verification in mobile communication systems. Our further research may involve the determination of appropriate parameters for speaker verification, and the evaluation of verification accuracy using an increased number of speakers and various kind of noises.

Acknowledgments

This work was supported in part by Grant-in-Aid for Young Scientists (B) (No.15760275) from MEXT (Ministry of Education, Culture, Sports, Science and Technology), Japan.

References

[1] A.Jain et al., "BIOMETRICS —Personal Identification in Networked Society", Kluwer Academic Publishers, 1999.

[2] Y.Yamazaki et al., "A speaker verification method using CELP parameters", Proc. of 6-th International Conference on Knowledge-based Intelligent Information & Engineering Systems (KES'2002), Vol.2, pp.1202-1206, 2002.

[3] ITU-T, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)", ITU-T Recommendation G.729, 1996.

[4] ITU-T, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP), Annex B: A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70", ITU-T Recommendation G.729 (Annex B), 1996.

[5] Y.Linde et al., "An algorithm for vector quantizer design", IEEE Trans. Comm. COM-28, 1, pp.84-95, 1980.

[6] H.Kuwabara et al., "Construction of ATR Japanese speech database as a research tool (Appendix I)", TR-I-0086, ATR Interpreting Telephony Research Laboratories, 1989.

[7] JEIDA, "JEIDA noise database", JEIDA (Japan Electronic Industry Development Association).