Speaker Verification Using Line Spectrum Frequency, Formant, and Support Vector Machine

Shi-Huang Chen[1], Yu-Ren Luo[2], and Rodrigo Capobianco Guido[3]
[1, 2] *Department of Computer Science & Information Engineering, Shu-Te University Kaohsiung County 824, Taiwan.*
Email: shchen@mail.stu.edu.tw; s96639118@mail.student.stu.edu.tw
[2] *Institute of Physics of São Carlos - University of São Paulo, São Carlos, São Paulo 13566-590, Brazil.*
Email: guido@ifsc.usp.br

Abstract

Speaker verification is desirable widely in many speech related applications, such as automatic telephone banking and biometric security system. This paper proposes an application of line spectrum frequency (LSF), formant, and support vector machine (SVM) to develop an algorithm of text-dependent speaker verification system. First, LSF and formant are extracted from the voiced password provided by the user. Then the proposed algorithm will make use of SVM to train the speaker characteristics from these speaker features and finally generate a claimed speaker model to discriminate between the speaker and other impostors. Experiments were conducted on the real speech signals and shown the performance of the proposed algorithm yields an equal error rate (EER) of 2.12% with 8-order LSFs and formant information. In addition, both of the false acceptance rate (FAR) and the false rejection rate (FRR) are also improved remarkably.

1. Introduction

Generally, the speaker recognition system can be classified into two subcategories: speaker identification and speaker verification. Even though these two subcategories use similar strategies, their applications are quite dissimilar. The purpose of speaker identification is to solve the problem of "who is the speaker?" while the speaker verification is proposed to answer the question of "is the speaker who they claim to be?" [1]. Therefore, the problem of speaker identification belongs to multiple-choice question and the speaker verification is a true-false question. Both of these two techniques, i.e., speaker identification and speaker verification, are desirable widely in many speech related applications, such as automatic telephone banking and biometric security system [1]. Meanwhile, depended on the differences of recognition target, the systems of speaker identification and speaker verification could fall into two types: text-dependent and text-independent. The former one requires that the speaker should provide keywords or sentences of the same text for both training and recognition, while the latter one dose not depend on the specific text being spoken [2]. For security consideration, this paper will focus on the problem of the text-dependent speaker verification system.

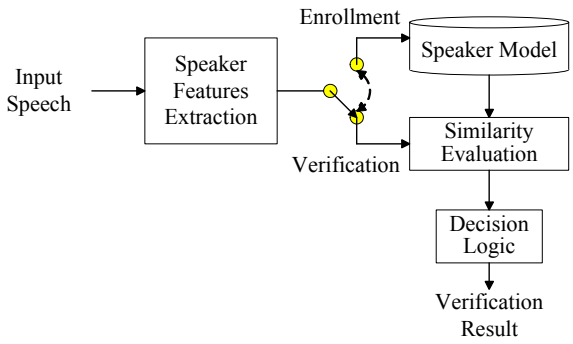

Fig. 1. The typical speaker verification system.

It follows from [1-2, 8] that a typical speaker verification system consists of two tasks: enrollment and verification as shown in Fig. 1. Enrollment is the task to construct a speaker model. This step will capture the speaker features, e.g., spectral or temporal parameters. Then these parameters of speaker features are used to build a model that could authenticate the speaker during the verification phase. In the speaker verification task, the speaker features of the input speech from test subject will be extracted and matched against the speaker model. A likelihood ratio will evaluate the similarity between the model and the measured observations. The general approach is based on a threshold set for the acoustic likelihood ratio to decide the test speaker is accepted or rejected.

Conventional speaker verification systems use dynamic time warping (DTW), hidden Markov models (HMM), or neural network (NN) to perform the likelihood ratio test [2]. The technique of DTW does not

978-0-7695-3890-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ISM.2009.132

need explicit model, but it requires pruning to limit storage computing. The systems which make use of HMM or NN have the well performance. However, their enrollment training is computationally expensive.

In addition, the choice of speaker features for speaker verification is of another primary concern. The ideal speaker feature set should have higher inter-class variance and lower intra-class variability. In addition, the selected speaker features should be independent of each other as in order to minimize redundancy. Based on the above discussion, the goal of this paper is to develop a new approach to the text-dependent speaker verification using the line spectrum frequency (LSF), formant, and support vector machines (SVM). Previous researches have shown that LSF is more robust in the noisy environment and is a good measure to discriminate between different speakers [3]. On the other hand, SVM is a two-class classifier based on the principles of structural risk minimization. Therefore SVM has well generalization ability when compared to HMM and NN based classifier [4]. Furthermore, since speaker verification is basically a binary decision, SVM seems to be a promising candidate to perform this task.

In this paper, LSF plus formant are used as speaker features, and SVM is used as likelihood ratio evaluation to perform speaker verification. In the beginning, the user has to provide a voiced password and the corresponding LSF as well as formant will be extracted from this spoken password. Then the proposed text-dependent speaker verification system will make use of SVM to train the speaker features from these LSF and formant information. It will finally generate a speaker model to discriminate between the speaker and other impostors. Using the speech signals recorded in real environment, experimental results shown the performance of the proposed speaker verification algorithm yields an equal error rate (EER) of 2.12% with 8-order LSFs and formant information. Furthermore, both of the false acceptance rate (FAR) and the false rejection rate (FRR) are also improved remarkably.

The remainder of this paper is organized as follows. The introductions to speaker feature extraction and SVM are briefly reviewed in Sections 2 and 3, respectively. Section 4 will describe the training procedure of the proposed text-dependent speaker verification system. Section 5 illustrates the experimental results. Finally, conclusions are given in the last Section.

2. Speaker Feature Extraction

2.1. Pre-processing of speech signal

The input speech signals are sampled at 8000 Hz with 16-bit resolution. Each speech signal is divided into frames. The frame length is 256 samples (32ms) with 128 samples (50%) overlap between adjacent frames. Each frame is processed via a pre-emphasizing filter that is defined as

$$s_n^{'} = s_n - 0.96 \times s_{n-1},\ n = 1, \cdots, 255 \tag{1}$$

where s_n is the n-th sample of the frame s and $s_0^{'} = s_0$. Then, the pre-emphasized frame is Hamming-windowed by

$$s_i^h = s_i^{'} * h_i, i = 0, \cdots, 255 \tag{2}$$

where $h_i = 0.54 - 0.46 \times \cos(2\pi\, i / 255)$.

2.2. Line spectrum frequency

Line spectrum frequency (LSF) or called line spectrum pair (LSP) introduced in [5] have been regarded as a method to represent linear predictive coding (LPC) parameters and are widely used in speech coding systems. In LSP computation, a linear predictive analysis filter $A(z)$ is decomposed into both a symmetric and an antisymmetric real polynomials. These two polynomials are called the LSP polynomials.

Let p be the order of a given LPC, the minimum phase LPC polynomial is expressed by [6] as follows:

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p} = 1 + \sum_{i=1}^{p} a_i z^{-i} \tag{3}$$

where $a_1, a_2, \cdots, a_p$ are the LPC coefficients. Then, LPC polynomial can be decomposed into a symmetric $P(z)$ and an antisymmetric $Q(z)$ polynomials. The definition of $P(z)$ and $Q(z)$ are given as

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}), \tag{4}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}). \tag{5}$$

One can observe that the polynomial $A(z)$ can be easily reconstructed via $P(z)$ and $Q(z)$ by

$$A(z) = \frac{1}{2}[P(z) + Q(z)] \tag{6}$$

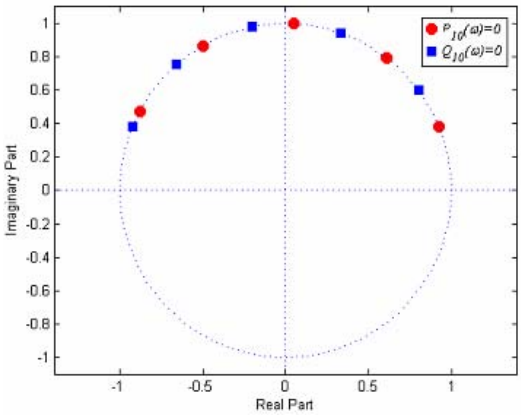

Fig. 2. Interlaced LSFs on the unit circle of z-plane.

The LSFs are the frequencies of the zeros of $P(z)$ and $Q(z)$. It is worthy to notice that the LSFs satisfy an interlacing property on the unit circle of z-plane, which holds for all minimum phase polynomials. That is

$$0 < \omega_1^{(P)} < \omega_2^{(Q)} \cdots < \omega_{p-1}^{(P)} < \omega_P^{(Q)} < \pi \quad (7)$$

where ω_i^K is the i-th LSF obtained from K (P or Q) polynomial. Fig. 2 demonstrates this behavior. In this paper, 8-, 10-, and 12-order LSFs are used.

2.3. Formant

Formant is the meaningful frequency components of human speech and is not directly related to pitch. By the definition [7], formant can be determined from the spectral peaks of the speech spectrum $|P(f)|$. In this paper, $|P(f)|$ is given as

$$|P(f)| = \log|F(s_i^h)|, i = 0, \cdots, 127 \quad (8)$$

where $|F(x)|$ is the magnitude of Fourier transform of x. The formant with the lowest frequency is called F1, the second F2, the third F3, and so on. Because most often the two first formants are enough to disambiguate the speech, this paper uses F1 and F2 as formant information.

3. Support Vector Machine

An SVM is a two-class classifier constructed from sums of a known kernel function $K(\cdot, \cdot)$ to define a hyperplane.

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \quad (9)$$

where $y_i \in \{1, -1\}$ are the target values, $\sum_{i=1}^{N} \alpha_i y_i = 0$, and $\alpha_i > 0$. The vector $\mathbf{x}_i \subseteq R^n$ are support vectors and obtained from the training. This hyperplane will separate given points into two predefined classes. Suppose a training set $S = \{(x_1, y_1), \cdots, (x_l, y_l)\}_{i=1}^{l} \subseteq (X \times Y)^l$ and a kernel function $K(x_i, x_j) = <\phi(x_i), \phi(x_j)>$ on $X \times X$ is given, where $<\cdot,\cdot>$ denotes the inner product and ϕ maps the input space X to another high dimensional feature space F. With suitably chosen ϕ, the given nonlinearly separable samples S may be linearly separated in F, as shown in Fig. 3. An improved SVM called soft-margin SVM can tolerate minor misclassifications [4] and use in this paper.

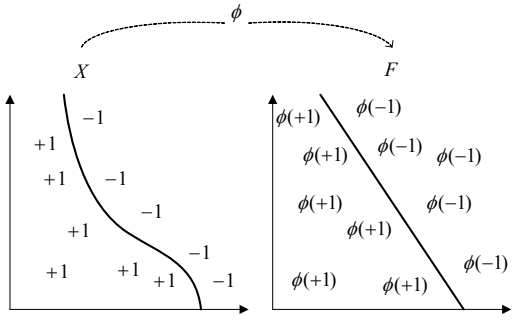

Fig. 3. A feature map simplifies the classification task.

Many hyperplanes can achieve the above separation purpose but the SVM used in this paper is to find the one that maximizes the margin (the minimal distance from the hyperplane to each point). The soft-margin SVM, which includes slack variable $\xi_i \geq 0$, is proposed to solve non-separable problems. Fig. 4 shows the slack variables, where ξ_i is defined as

$$\xi_i = \max(0, \gamma - y_i(< w, x_i > + b)). \quad (10)$$

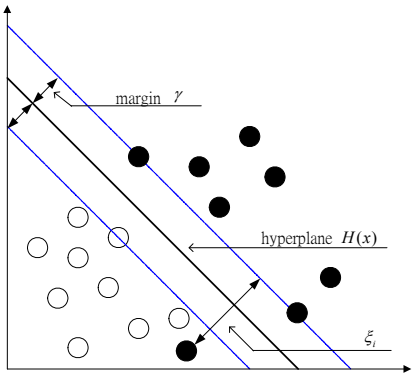

Fig. 4. The margin and the slack variable for a classification problem.

In (10), the parameter ξ_i can measure the amount by which the training set fails to have margin γ, and take into account any misclassification of the training data. Consequently, the training process tolerates some points misclassified and is suitable in most classification cases. The kernel function of exponential radial basis function (ERBF) that is defined as

$$K(x, \bar{x}) = \exp(-|x - \bar{x}| / 2\sigma^2) \quad (11)$$

where parameter σ^2 is the variance of the Gaussian function.

4. The Training Procedure of the Proposed Speaker Verification System

Fig. 5 illustrates the block diagram of the proposed text-dependent speaker verification system. Before performing speaker verification, one has to build a claimed speaker model and an imposter model via SVM training. The training procedure is described as follows. Assume that n_T is the number of the obtained LSF and formant vectors. The training set T is then defined to be the $n_T \times (p+2)$ array with row vectors being these p-order LSF and F1, F2 formants vectors. In this paper, p could be 8, 10, or 12. The next section will discuss the performances of these three settings of p. Let $T(i, j)$denote the (i, j)-position of T. Use this array T to construct another $n_T \times (p+2)$ array T' whose (i, j) position $T'(i, j)$ is defined to be $T'(i, j) = T(i, j) - \mu_j$,

where $\mu_j = \sum_i T(i,j)/n_T$ is the mean of column j. Next, one normalizes T' by computing $T^N(i,j) = T'(i,j)/m_j$, where m_j is the maximum of the absolute value of elements in column j. Thus, each LSF and formant feature will have similar weights after the normalization process.

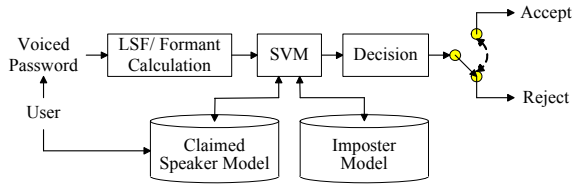

Fig. 5. The block diagram of the proposed text-dependent speaker verification system.

To train a model for a specific speaker, this paper utilizes a SVM classification method called one-against-all strategy. That is the speaker data are trained to an SVM target value of +1 whereas the imposter data are trained to an SVM target value of –1. Finally SVM will find a nonlinear hyperplane that can separate speaker and imposter features. The parameters σ^2 in (11) will be determined after training. The determination criterion of σ^2 is based on the separation capability of speaker and imposter. In this paper, σ^2 is set to be 50.

5. Experimental Results

The experimental results of the proposed text-dependent speaker verification system are achieved by using 20 male and 10 female speakers. Among these 30 speakers, the 26 speakers (18 male and 8 female) are used in training and the 4 speakers which are not contained in training set are used as imposters. All of the test speech signals are recorded in real environment and are sampled at 8000 Hz with 16-bit resolution. Each test speech signal consists of 4~6 words.

Speaker verification performance will be reported using the false acceptance rate (FAR), the false rejection rate (FRR), and the equal error rate (EER).

The definitions of FAR and FRR are given as follows:

$$\text{FAR} = \frac{\#\,\text{aceepted imposter claims}}{\#\,\text{imposter accesses}} \times 100\% \qquad (12)$$

$$\text{FRR} = \frac{\#\,\text{rejected genuine claims}}{\#\,\text{genuine accesses}} \times 100\% \qquad (13)$$

Once the receiver operating characteristic (ROC) curve of FAR vs. FRR is obtained, one can determine the EER, which FAR and the FRR at this point is the same for both of them.

This paper compared the results obtained on the SVM based speaker verification system with F1, F2 formants , and three settings of LSF order, namely p= 8, 10, and 12. An impostor model was trained on all the LSF and formant in the impostor data set while the speaker model was built using the corresponding speaker data set. During speaker verification task, a likelihood ratio was computed between the speaker model and the impostor model. The likelihood ratio was defined as:

$$LR = \log P(x \mid \text{speaker model}) - \log P(x \mid \text{impostor model}) \qquad (14)$$

where x is the input test LSF and formants vector. Table 1 shows a summary of the experimental results of the proposed text-dependent speaker verification system with various combinations of speaker features. One can find that with the combination of 8 order LSF and F1, F2 formants, the proposed speaker verification system has lowest EER, *i.e.*, the best verification rate. An EER of 2.12% is achieved using the proposed system. The ROC plots of FRR and FAR with LSF order = 8, 10, and 12 are shown in Figs. 6, 7, and 8, respectively.

It also follows Table 1 that the LSF dominates the performance of the proposed text-dependent speaker verification system. Without using F1 and F2 formant information, the proposed system can yield an EER of 2.78% with 8 orders LSF.

Table 1. Comparison of the proposed SVM based text-dependent speaker verification system with various combinations of speaker features.

Speaker features	EER
SVM + F1+F2	8.92
SVM + 8 LSFs	2.78
SVM + 10 LSFs	3.70
SVM + 12 LSFs	7.41
SVM + F1+F2+8 LSFs	2.12
SVM + F1+F2+10 LSFs	3.21
SVM + F1+F2+12 LSFs	6.89

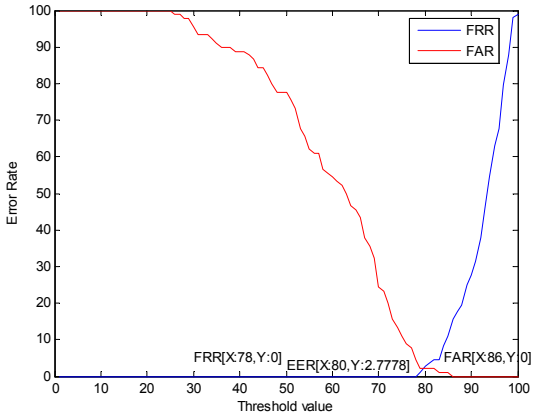

Fig. 6. A ROC plot of FRR and FAR with 8 orders LSF.

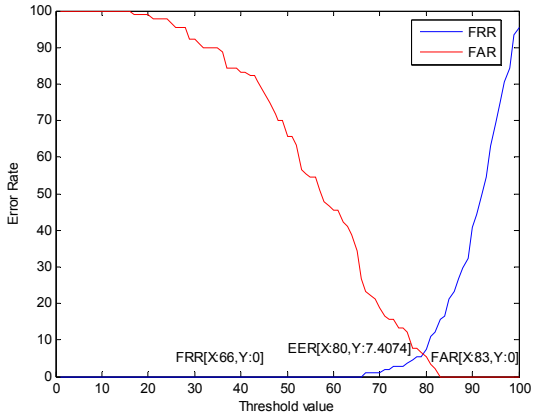

Fig. 7. A ROC plot of FRR and FAR with 10 orders LSF.

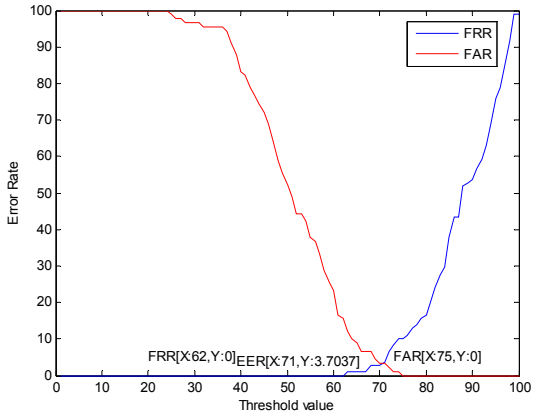

Fig. 8. A ROC plot of FRR and FAR with 12 orders LSF.

6. Conclusions

In this paper, line spectrum frequency (LSF), formant, and support vector machine (SVM) are applied to the task of text-dependent speaker verification system. First, the LSF and formant information will be extracted from the voiced password provided by user. Then the proposed algorithm will make use of SVM to train the speaker characteristics model from the speaker features and generate a claimed speaker model to discriminate between the speaker and other impostors. Experiments were conducted on real speech signals and shown the performance of the proposed algorithm yields an equal error rate (EER) of 2.12% with 8 orders LSF and the first and second formants. In addition, both of the false acceptance rate (FAR) and the false rejection rate (FRR) are also improved remarkably.

7. Acknowledgement

This work was supported by the National Science Council (NSC), Taiwan, R.O.C., under Grant NSC 97-2221-E-366-010-MY3.

8. References

[1] Peter Day and Asoke K. Nandi, "Robust Text-Independent Speaker Verification Using Genetic Programming," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 1, pp. 285-295, 2007.

[2] Minho Jin, Frank K. Soong, and Chang D. Yoo, "A Syllable Lattice Approach to Speaker Verification," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2476-2484, 2007.

[3] H. Cordeiro, C.M. Ribeiro, "Speaker Characterization with MLSFs," IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, pp. 1-4, June 2006.

[4] C.C. Lin, S.H. Chen, T. K. Truong, and Yukon Chang, "Audio Classification and Categorization Based on Wavelets and Support Vector Machine," *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 5, pp. 644-651, Sept. 2005.

[5] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," *J. Acoust. Soc. Am.*, 57, 535(A), 1975.

[6] W. C. Chu, *Speech Coding Algorithm: Foundation and Evolution of Standardized Coders*, Wiley-Interscience, 2003.

[7] G. Fant, Acoustic Theory of Speech Production, Mouton & Co, The Hague, Netherlands, 1960.

[8] A.E. Rosenberg, "Automatic speaker verification: A review," *IEEE Proceedings*, Vol. 64, pp. 475-487, 1976.