Discrete Wavelet Transform for Automatic Speaker Recognition

Pavel Král
Dept. Informatics & Computer Science
University of West Bohemia
Plzeň, Czech Republic

***Abstract*—This paper deals with automatic speaker recognition. We consider here a context independent speaker recognition task with a closed set of speakers. We have shown in [1] a comparative study about the most frequently used parametriza-tion/classification methods for the Czech language. Wavelet Transform (WT) is a modern parametrization method successfully used for some signal processing tasks. WT often outperforms parametrizations based on Fourier Transform, due to its capability to represent the signal precisely, in both frequency and time domains. The main goal of this paper is thus to use and evaluate several Wavelet Transforms instead of the conventional parametrizations that were used previously as a parametrization method of automatic speaker recognition. All experiments are performed on two Czech speaker corpora that contain speech of ten and fifty Czech native speakers, respectively. Three discrete wavelet families with different number of coefficients have been used and evaluated: *Daubechies*, *Symlets* and *Coiflets* with two classifiers: Gaussian Mixture Model (GMM) and Multi-Layer Perceptron (MLP). We show that recognition accuracy of wavelet parametrizations is very good and sometimes outperform the best parametrizations that were presented in our previous work.**

I. Introduction

Automatic speaker recognition is the use of a computer to identify a person from his speech. Two main different tasks exist: speaker identification and speaker verification. Speaker identification is the use of a computer to decide who is currently speaking. Speaker verification is the use of a machine to prove that the speaking person is the claimed one or not. Information about the current speaker is useful for several applications: access control, automatic transcription of radio broadcasts (speaker segmentation), system adaptation to the voice of the current speaker, etc. Our work focuses on the access control system, where the audio speech signal will be the main information to authorize building entrance. Additional information (e.g. fingerprint, access card) will also be provided when audio information is ambiguous. In this paper, we focus on context independent[1] speaker recognition with a closed set of speakers.

In our previous studies, we have evaluated and compared in [1] the most promising conventional parametriza-tion methods and two classifiers in order to build an efficient Czech speaker recognition system. We have shown that all analyzed parametrizations/classifiers are comparable from recognition accuracy viewpoint and give good recognition scores.

[1]The content of utterances is general.

Wavelet Transform (WT) is a modern parametrization method successfully used for some signal processing tasks [2]. This method is also widely used in the field of image processing [3], [4]. It is often used instead of the Fourier Transform (FT) due to its capability to represent the signal in both frequency and time domains. Parametrizations based on the Fourier Transform are often used in speech processing task [5], [6], because speech signal varies very slowly and it can be thus considered as (quasi-)stationary in the short time interval. However this assumption is only a simplification of the reality and it is thus suitable to represent speech more precisely. Therefore, current efforts of researchers focus on the use of Wavelet Transform in several fields of automatic speech processing [7], [8], [9].

The main goal of this paper is to use and evaluate several Wavelet Transforms as parametrization method of automatic speaker recognition. The proposed parametrization will be evaluated on two Czech speech corpora and compared with our previous work presented in [1]. Note that to the best of our knowledge, there is no previous study that employs Wavelet Transform on automatic speaker identification in Czech language and there is also no comparative study about the different Wavelet families as a parametrization method for this task.

This paper is organized as follows. The next section presents a short review of automatic speaker recognition. Section III describes our use of Wavelet Transform for automatic speaker identification. Section IV presents our experimental setup and shows our results. Our speaker corpora are also described in this section. In the last section, we discuss the results and we propose some future research directions.

II. Related Work

The task of speaker identification is composed of two main steps: speech parametrization and speaker modeling. These steps are described below.

Several works successfully use, as shown in [10], Linear Prediction (LP) coefficients. LP coefficients are often non-linearly transformed in order to better represent the speech signal as in the Reflection Coefficients (RCs), Line Spectrum Pair (LSP) frequencies [11] or LP cepstrum [12]. Speaker characteristics may be also represented by prosodic features [13], such as fundamental frequency, energy, etc. Some work rather

use the Mel Frequency Cepstrum [14], [15] with high recognition accuracy. In recent years, some parametrization methods are based on the Wavelet Transform [16], [17], [18]. The main advantage of these approaches is the representation of speech signal in time-frequency domain which gives very good results when the signal is non-stationary.

Approaches of speaker modeling can be divided into three major groups: 1) template methods; 2) discriminative methods and 3) statistical methods. The first group includes for example Dynamic Time Warping (DTW) [19], Vector Quantization (VQ) [20] and Nearest Neighbours [21].

Discriminative methods are mainly represented by Neural Networks (NNs). In this case, a decision function between speakers is trained instead of individual speaker models. Different NNs topologies are used but the best results are mainly given by a Multilayer Perceptron (MLP) as shown in [22]. Neural networks usually need less parameters than the individual speaker models to achieve comparable results. However, the main drawback of NNs is the necessity to retrain the whole network when a new speaker appears. Another successful discriminative approach is Support Vector Machines (SVMs) [23].

Stochastic methods are the most popular and the most effective methods used in the speech processing domain (e.g automatic speech recognition, automatic speech understanding, etc.). In the speaker recognition task, these approaches consist in computing the probability of an observation given a speaker model. This observation is the value of a random variable, having a Probability Density Function (PDF) which depends on the speaker. The PDF function is estimated on a training corpus. During recognition, probabilistic scores are computed with every model and the model with the maximal probability is selected as the correct one. Hidden Markov Model (HMM) [24] and Gaussian Mixture Models (GMMs) [25] are very popular stochastic models used in the speaker recognition.

III. Wavelet Transform for Automatic Speaker Recognition

A. Signal Preprocessing

Speech signal is pre-processed in order to remove undesirable constituents and to be more robust for speaker identification. Two methods are used: *preemphasis* and *normalization*.

Speech is emphasized to reinforce spectral magnitudes of the high frequencies:

$$y(n) = x(n) - ax(n-1) \tag{1}$$

where $x(n)$ and $x(n-1)$ are the current one and the previous speech samples respectively, $a \in < 0.9; 1 >$ is a preemphasis coefficient and $y(n)$ is the result of preemphasis.

Normalization is performed as in [26] to minimize the differences in the speech intensity of the recordings by the following equation:

$$y(n) = \frac{x(n) - \mu}{\sigma} \tag{2}$$

where $x(n)$ is the original preemphasised sample, μ and σ are the mean and the standard deviation of the preemphasised signal and $y(n)$ is the resulting normalised sample.

B. Wavelet Transform

As the majority of other studies, we address only the Discrete Wavelet Transform (DWT). This chose has been made due to the high computational costs of the Continuous Wavelet Transform (CWT) and the discrete character of our task.

Individual wavelets are divided into several groups, so called *families*. We will consider the three discrete wavelets: *Daubechies*, *Symlets* and *Coiflets*. *Daubechies* wavelets are a family of orthogonal wavelets characterized by a maximal number of vanishing moments for some given support. There is a scaling function (father wavelet) which generates an orthogonal multiresolution analysis. *Symlets* are very close to the previous family. Hovewer, the main difference from the *Daubechies* is their symmetry. *Coiflets* are near symmetric, their wavelet functions have $\frac{N}{3}$ vanishing moments and scaling functions $\frac{N}{3} - 1$.

Wavelet Packet Decomposition (WPD) of the *fifth* level is calculated. Frequency spectrum of the each speaker is thus decomposed into 2^5 sub-signals, with 16 samples each. These values are used to compute the resulting feature vector as follows:

$$f_i = log_{10}(\frac{1}{N_i} \sum_{k=1}^{N_i - 1} |(w_i(k))^2 - w_i(k-1) * w_i(k+1)|) \tag{3}$$

where f_i is the value of the feature vector in the leave i, $w_i(k)$ is the coefficient k in the leave i, $w_i(k-1)$ and $w_i(k+1)$ are the previous and the next coefficients, respectively. The size of the feature vector is thus 32.

It is beneficial to choose an optimal Wavelet Transform method for speaker recognition. However, according to the literature [2], it is not possible to choose an optimal WT for a task given. Therefore, we must determine the best Wavelet Transform method for automatic speaker identification experimentally.

C. Classification

Let us call F the set of features for one sentence (created by some parametrization method) that have spoken by speaker S. We evaluate the performance of two classifiers: a Mutilayer Perceptron (MLP) that computes $P(S|F)$ and a Gaussian Mixture Model (GMM) that models $P(F|S)$. Both classifiers error rates are reported in the following experiments.

IV. Evaluation

A. Experimental Setup

The first experiment studies the recognition accuracy in function of the size of the training data. Our objective is to compute the minimal size of the training corpus in order to reach the desired recognition accuracy. This experiment has been motivated by the fact that the corpus preparation

is an expensive and time demanding task and it is thus not acceptable to create a large corpus without necessity.

The second experiment focuses on the relation between the size of the testing data and the resulting recognition rate. We would like to determinate the minimal length of the utterance to reach the desired accuracy. This experiment is very important to configure our speaker recognition system. These two experiments will be realized on a small clean speaker corpus.

The next experiment deals with speaker recognition on the larger speaker corpus. We would like to evaluate speaker recognition accuracy in "quasi" real condition (i.e. more speakers, some background noise, etc.).

All the previously described experiments are performed on the all wavelet parametrization methods and with the two classifiers. As reported previously, the three discrete *families* of Wavelets with different number of coefficients will be used and evaluated:

- *Daubechies*: eight different coefficients: $4, 6, .., 20$;
- *Symlets*: 14 different coefficients varies $8, 10, .., 20$;
- *Coiflets*: three different coefficients $6; 12; 18$.

These results are compared with four conventional parametrizations: Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficients (PLPC), Linear Prediction Reflection Coefficients (LPREFC) and Linear Prediction Cepstral Coefficients (LPCEPSTRA), that are described previously in [1].

B. Corpora

The first Czech corpus contains speech of ten Czech native speakers. It is composed of the speech of five women and five men. This corpus has been created in laboratory conditions in order to eliminate undesired effects (e.g. background noise, speaker overlapping, etc.). The detailed corpus structure is shown in Table I.

TABLE I
CZECH CORPUS THAT CONTAINS 10 SPEAKERS (5 FEMALES AND 5 MALES)

	Training	Testing
Number of speakers	10	
Size of the data	17 MB	9 MB
Number of recordings	521	262
Average number of recordings / speaker	52	26
Average length of recordings / speaker	9 min	5 min

The second corpus contains human-human dialogs of fifty Czech native speakers. It has been primarily created in order to build a dialog system for Czech Railways. Recorded speech is not as clean as in the previously described corpus, but it contains some low level stationary background noise. Table II shows the detailed structure of this corpus.

Both sets, the training and testing ones, are disjoint for both corpora.

TABLE II
CZECH CORPUS THAT CONTAINS 50 SPEAKERS (25 FEMALES AND 25 MALES)

	Training	Testing
Number of speakers	50	
Size of the data	10 MB	3.2 MB
Number of recordings	3190	1759
Average # of recordings / speaker	64	35
Average length of recordings / speaker	5.30 min	1.75 min

C. Experiments

All parametrizations use a window of 32ms length, and the size of the feature vector (see Section III) is 32. One GMM model with various number of Gaussian Mixtures is used. The number of mixtures varies from 1 to 256. Our MLP is composed of three layers: 32 inputs, one hidden layer and outputs correspond to the number of speakers (ten or fifty depend on the corpus). The optimal number of neurons in the hidden layer is set experimentally for each experiment. This value varies from 10 to 90. The GMM and MLP topologies with a similar number of training parameters are compared.

1) Study of the size of the training data: Figure 1 shows the speaker recognition accuracy in relation to the size of the training data. Ten Czech speakers from the previously described corpus are identified. The duration of the training data varies from 7.5 seconds to 9 minutes per speaker. The duration of the testing utterances is about five minutes and remains constant during the whole experiment. Results with a constant recognition accuracy of 100% are not reported in the figure. All twenty wavelets achieve very similar results. Therefore, only one representative wavelet, *Coifet* with 18 coefficients ($COIF18$[2]), is reported. This wavelet is compared only with MFCC parametrization due to the similar results of the other conventional ones. The performance of wavelets and conventional parametrizations are very close, while the results depend on the classifier. Recognition accuracy of the GMM model depends much more on the size of the training data than for the MLP one. GMM needs for correct training at least one minute of training data per speaker, while 30 seconds of training speech is sufficient for MLP parameters estimation. Furthermore, the reduction of GMM accuracy is much more significant than for the MLP model.

2) Study of the size of the testing data: Figure 2 shows the speaker recognition accuracy in relation to the length of the pronounced utterance. A similar set of speakers as in the previous experiment is used. The duration of the training data is 2.5 minutes per speaker and remains constant during the whole experiment, while the duration of the testing utterances varies in the interval of [0.5; 6] seconds. All previously reported number of coefficients have been evaluated for each

[2]We will used for wavelet specification the following notation: abbreviation of the wavelet family ($DAUB$ for *Daubechies*, $SYML$ for *Symlets* and $COIF$ for *Coiflets*) + number of coefficients (e.g. *Symlet* with 14 coefficients will be denoted as $SYML14$).

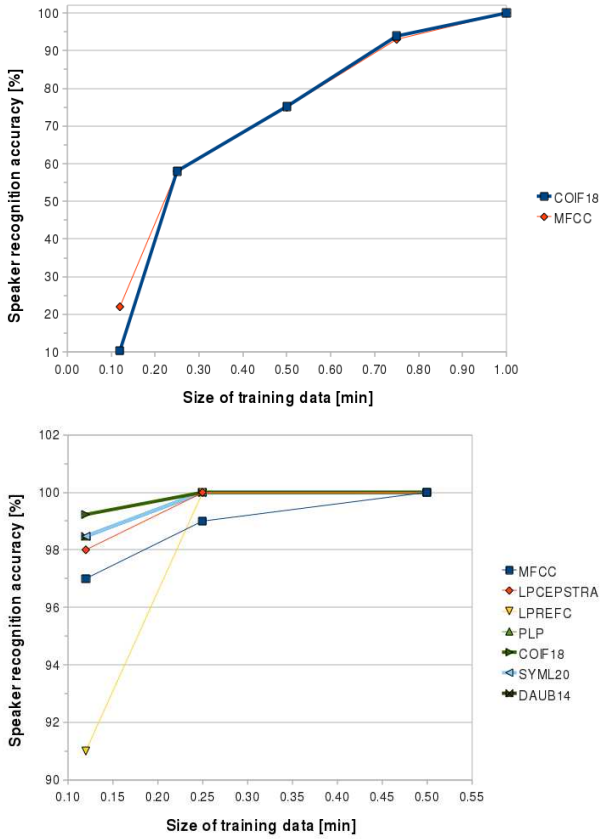

Fig. 1. Speaker recognition accuracy in relation to the size of the training data (GMM model on the top; MLP model on the bottom). The x-axis represents the size of the training data, while the y-axis shows the speaker recognition accuracy. The wavelet transform experiments are reported for the all experiments with strong lines, while the other ones with thin lines.

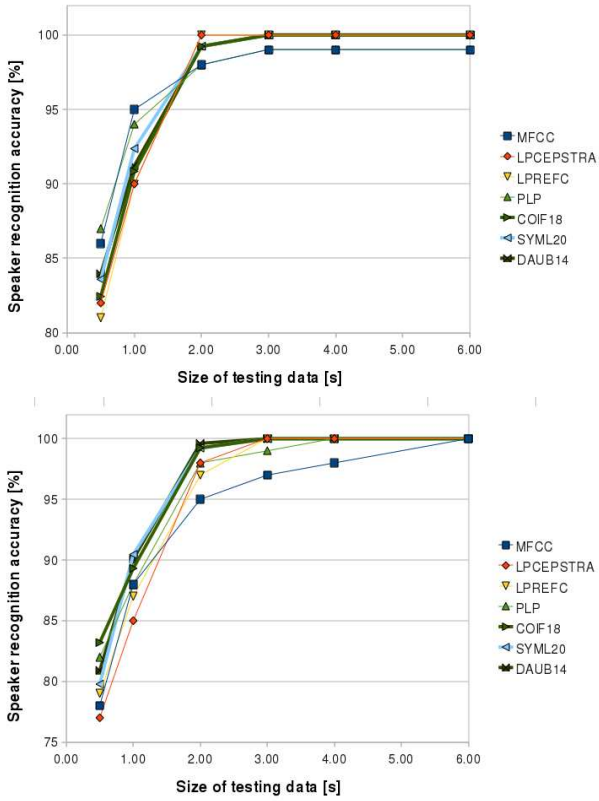

Fig. 2. Speaker recognition accuracy in relation to the size of the testing data (GMM model on the top; MLP model on the bottom). The x-axis represents the size of the testing data, while the y-axis shows the speaker recognition accuracy.

wavelet family, while only configurations with the best recognition accuracy is reported in the figure (one per family). The recognition accuracy of all parametrizations (wavelets and conventional) and both classifier are very close. We show that the minimal utterance length for the correct speaker recognition is about two seconds. We obtained 100% of accuracy for all the reported parametrizations (except MFCC and PLP) on the sentences of the duration of three seconds. We also show that all wavelets are comparable with LPCEPSTRA, the best conventional parametrization.

3) Experiments on the greater speaker corpus: Figure 3 shows the speaker recognition accuracy on the corpus that contains 50 speakers. The duration of the training data is as in the previous experiment about 2.5 minutes per speaker. The duration of the testing utterances is three second; duration with the 100% recognition accuracy of the previous experiment. The length of utterances remains constant during the whole experiment, while we will modify the number of model parameters (i.e. number of Gaussian in a GMM case and number of neurons in hidden layer for a MLP). The best recognition accuracy, about 99%, is reached with the combination of LPCEPSTRA (or LPREFC) parametrization and a GMM classifier. However, the best configuration of wavelets, $SYML20$ with MLP classifier, gives 98% which is still very good.

V. Conclusions

In this paper, three families of wavelet transform parametrizations, namely *Daubechies*, *Symlets* and *Coiflets*, have been evaluated and compared with four conventional parametrizations: MFCC, LPCEPSTRA, LPREFC and PLP on the automatic speaker recognition task with the two Czech corpora. Two classifiers, a GMM and a MLP have been used. Three experiments have been performed. In the first one, we studied the minimal training data size required for a correct estimation of the speaker models. We show that, from this point view, all parametrizations/classifiers are comparables. We also show that MLP requires less training data than GMM. It needs only 30 seconds of training data per speaker, while GMM needs at least one minute. The second experiment deals with the minimal duration of the test utterance for the correct recognition of the speaker. It has been demonstrated that recognition scores of the reported parametrizations/classifiers are very close. We further show that the minimal utterance length for the correct speaker recognition is about two seconds. In the last experiment, we show the performance of our speaker recognition system on the greater speaker corpus that

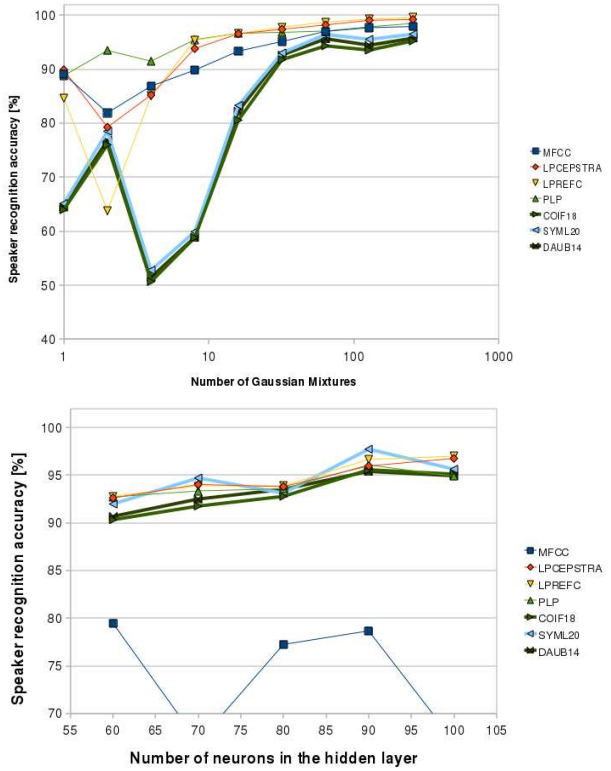

Fig. 3. Speaker recognition accuracy on the corpus with 50 speakers (GMM model on the top; MLP model on the bottom). The x-axis shows the number of model parameters (i.e. number of Gaussian in a GMM case and number of neurons in hidden layer for a MLP), while the y-axis shows the speaker recognition accuracy.

contains 50 speakers and some background noise. We show that results of the wavelet parametrizations are very closed to LPCEPSTRA, LPREFC, the best conventional parametrisations.

In this work, a closed set of speakers is considered. However, unknown speakers shall be also considered in real situation. Such a set of speakers is said to be open. We would like to modify our models in order to operate with an open set. Recognition accuracy of the reported experiments is very high. There are two main reasons: 1) low level of stationary background noise in the corpora; 2) small number of the speakers (only 50 in the greater corpus). Our second perspective thus consists in the evaluation of the parametrizations/classifiers on a larger corpus recorded in real conditions (e.g. with more non-stationary noise in the speech signal, etc.). We also would like to combine audio information with other modalities (e.g. fingerprint) in order to build a more efficient and secure access system.

ACKNOWLEDGMENT

This work has been partly supported by the Ministry of Education, Youth and Sports of Czech Republic grant (NPV II-2C06009). We would like to thank also to Mr. Michal Hrala for implementation and executions of experiments.

REFERENCES

[1] P. Král, K. Ježek, and P. Jedlička, "Evaluation of Automatic Speaker Recognition Approaches," in *WESPAC X 2009*, Beijing, China, 21-23 September 2009.

[2] S Mallat, *A Wavelet Tour of Signal Processing*, San Diego, Academic Press, 1999.

[3] I. Drori and Lischinski. D., "Fast Multiresolution Image Operations in The Wavelet Domain," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 3, pp. 395–412, 2003.

[4] O. Ryan, "Applications of the wavelet transform to image processing," in *Seminar on Wavelets*, Oslo, Norway, December 12 2004.

[5] M.D. Skowronski and J.G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition.," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1774–1780, 2004.

[6] Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, and Nikos Fakotakis, "Comparison of speech features on the speech recognition task," *Journal of Computer Science*, vol. 3, no. 8, pp. 608–616, 2007.

[7] B. Kotnik, Z. Kacic, and B. Horvat, "The usage of wavelet packet transform in automatic noisy speech recognition systems.," in *IEEE EUROCON 2003*, Slovenia, 2003, pp. 131–134.

[8] M Deviren, *Revisiting Speech Recognition Systems: Dynamic Bayesian Networks and New Computational Paradigms.*, Ph.D. thesis, Henri Poincaré University, Nancy, 2004.

[9] E. Didiot, I. Illina, D. Fohr, and O. Mella, "A wavelet-based parameterization for speech/music discrimination," *Computer Speech and Language*, vol. 24, no. 2, pp. 341 – 357, 2010.

[10] N. Z. Tishby, "On the application of mixture ar hidden markov models to text independent speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 39, no. 3, pp. 563570, 1991.

[11] G. Kang and L. Fransen, "Low bit rate speech encoder based on line-spectrum-frequency," Tech. Rep. 8857, NRL, 1985.

[12] A. L. Higgins and R. E. Wohlford, "A new method of text-independent speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, Tokio, Japan, 1986, pp. 869–872.

[13] I. Chmielewska, "Prosody-based text independent speaker identification method," in *From Sound to Sense*, Massachusetts Institute of Technology, June 2004, pp. 13–18.

[14] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," 1995.

[15] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining mfcc and phase information," in *Interspeech 2007*, Belgiun, Antwerp, August 2007.

[16] R. Sarikaya, B. L. Pellom, and J. H. L. Hansen, "Wavelet packet transform features with application to speaker identification.," in *IEEE Nordic signal processing symposium*, Denmark, 1998, pp. 81–84.

[17] Ching-Han Chen and Chia-Te Chu, "An high efficiency feature extraction based on wavelet transform for speaker recognition," in *Int. Computer Symposium*, Taipei, Taiwan, December 15-17 2004.

[18] S. Y. Lung, "Wavelet feature selection based neural networks with application to the text independent speaker identification," *Pattern Recognition*, vol. 39, pp. 1518–1521, 2006.

[19] A. Higgins *et al.*, "Speaker verification using randomized phrase promting," *Digitam Signal Processing*, vol. 1, no. 2, pp. 89–106, 1991.

[20] F. Soong, A. Rosenberg, L. Rabiner, and B-H. Juang, "A vector quantization approach to speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, USA, Florida, 1985, pp. 387–390.

[21] A. Higgins, L. Bahler, and J. Porter, "Voice identification using nearest neighbor distance measure," in *International Conference on Acoustics, Speech, and Signal Processing*, USA, Minneapolis, 1993, pp. 375–378.

[22] L. Rudasi and S. A. Zahorian, "Text-independent talker identification with neural networks," in *International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Ontario, Canada, 1991, pp. 389–392.

[23] H. Yang *et al.*, "Cluster adaptive training weights as features in svm-based speaker verification," in *Interspeech 2007*, Belgiun, Antwerp, August 2007.

[24] C. Che and Q. Lin, "Speaker recognition using hmm with experiments on the yoho database," in *EUROSPEECH'95*, Spain, Madrid, 1995, pp. 625– 628.

[25] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing 10*, pp. 19–41, 2000.

[26] X. Lou and K. A Loparo, "Bearing fault diagnosis on wavelet transform and fuzzy inference.," *Mechanical System and Signal Processing*, vol. 18, pp. 1077–1095, 2004.