EVALUATION OF AUTOMATIC SPEAKER RECOGNITION APPROACHES

Pavel Král and Václav Matoušek

University of West Bohemia in Plzeň (Pilsen), Czech Republic
pkral | matousek@kiv.zcu.cz

Abstract: This paper deals with automatic speech recognition in Czech. We focus here on context independent speaker recognition with a closed set of speakers. To the best of our knowledge, there is no comparative study about different speaker recognition approaches on the Czech language. The main goal of this paper is thus to evaluate and compare several parametrization/classification methods in order to build an efficient Czech speaker recognition system. All experiments are performed on a Czech speaker corpus that contains approximately half one hour of speech from ten Czech native speakers. Four parameterizations, which are mentioned in other studies as particularly successful for the speaker recognition task, are compared: MEL Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficients (PLPC), Linear Prediction Reflection Coefficients (LPREFC) and Linear Prediction Cepstral Coefficients (LPCEPSTRA). Two classifiers are compared: Hidden Markov Models (HMMs) and Multi-Layer Perceptron (MLP). In this work, we further study the impact of varying sizes of training corpus and test sentence on the recognition accuracy for different parametrizations and classifiers. For instance, we experimentally found that the recognition is still very accurate for test utterances as short as two seconds. The best recognition accuracy is obtained with LPCEPSTRA/ LPREFC parametrizations and HMM classifier.

1 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 consists in using 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 emissions (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 be also provided when audio information is ambiguous. In this paper, we focus on context independent speaker recognition with a closed set of speakers.

To the best of our knowledge, there is no previous study that compares several different speaker recognition approaches on the Czech language. The main goal of this paper is thus to evaluate and compare several parametrizations methods and classification models in order to build an efficient speaker recognition system. Four parameterizations, which are mentioned in other studies as particularly successful for speaker recognition in other European languages, are here compared: MEL Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficients (PLPC), Linear Prediction Reflection Coefficients (LPREFC) and Linear Prediction Cepstral Coefficients (LPCEPSTRA). Two classifiers are also compared: Hidden Markov Models (HMMs) and Multi-Layer Perceptron (MLP).

The next section presents a short review of automatic speaker recognition approaches. A short description of the most important parametrizations and models is also given. Section 3

presents our experimental setup and shows our results. Our speaker corpus is also described in this section. In the last section, we discuss the results and we propose some future research directions.

2 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 [1], Linear Prediction (LP) coefficients. Linear prediction is based on the fact that the speech signal varies slowly in time and it is thus possible to model the current signal value by the *n* previous ones. 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 [2] or LP cepstrum [3]. Speaker characteristics may be also represented by prosodic features [4], such as fundamental frequency, energy, etc. The most recent works rather use the Mel Frequency Cepstrum [5, 6] with high recognition accuracy.
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) [7, 8], Vector Quantization (VQ) [9] and Nearest Neighbours [10].
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 Multilayer Perceptrons (MLPs) as shown in [11]. Neural networks need usually 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) [12].
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 a value of a random variable, which Probability Density Function (PDF) 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. The most popular stochastic model used in the speaker recognition is Hidden Markov Model (HMM) [5, 13, 14]. For non-stochastic variables, we use the Gaussian Mixture Model (GMM) [15].

3 Evaluation

3.1 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 a 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 a desired accuracy. This experiment is very important to configure our speaker recognition system.

The last experiment focuses on the recognition of two similar voices that belong to twin brothers. It is quite difficult to distinguish their two voices by humans. The human recognition rate is a bit low (about 50 % on the telephone speech).
All the previously described experiments are performed on the four parametrization methods and with the above mentioned two classifiers.

3.2 Corpus

The Czech corpus contains eleven Czech native speakers. It is composed of the speech of five women and six men (two twins). Every record is manually labeled with its corresponding speaker labels. This corpus has been created in laboratory condition in order to eliminate undesired effects (e.g. background noise, speaker overlapping, etc.). The detailed corpus structure is shown in Table 1.

Table 1: Czech corpus size

	Training		**Testing**	
Speaker number	Recording	# Length [min]	Recording	# Length [min]
1	100	9.4	31	5.1
2	46	9.3	25	4.9
3	41	9.4	28	5.1
4	40	8.9	17	5.1
5	28	9.1	16	5.0
6	32	9.5	20	4.8
7	35	9.0	29	5.0
8	86	8.9	41	4.9
9	65	9.0	27	5.4
10	48	9.2	28	4.8
11	50	9.1	26	5.0
Total	**571**	**135**	**288**	**92**

The number of recordings differs between the speakers because of their different duration. However, the length of the recorded speech is for every speaker almost equal (about 9 minutes for training and about 5 minutes for testing). Both sets, the training and testing ones, are disjoint.

3.3 Experiments

All parametrizations use a Hamming window of 32 ms length, and the size of the feature vector is 32. One state HMM 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 10 outputs (correspond to the number of speakers). The optimal number of neurons in the hidden layer is set experimentally for each experiment. This value varies from 10 to 22. The HMM and MLP topologies with a similar number of training parameters are compared. The HTK [16] toolkit is used for implementation of the HMMs and the LNKnet [17] for the implementation of the MLP.

3.3.1 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 constant recognition accuracy of 100 % are not reported on the figure. The HMM recognition scores are almost equal for all four parametrizations. Therefore, only MFCC is reported in the left figure. Recognition accuracy of the HMM model (on the left) depends much more on the size of the training data than for the MLP one (right). HMM 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 HMM accuracy is much more significant (up to 20 %) than for the MLP model.

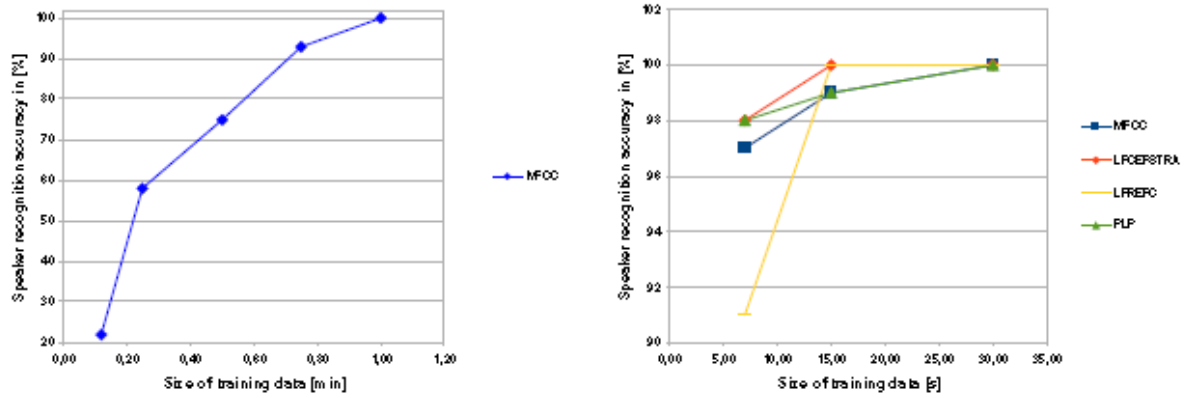

Figure 1: Speaker recognition accuracy in relation to the size of the training data (HMM model on the left; MLP model on the right). The x-axis represents the size of the training data, while the y-axis shows the speaker recognition accuracy

3.3.2 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.

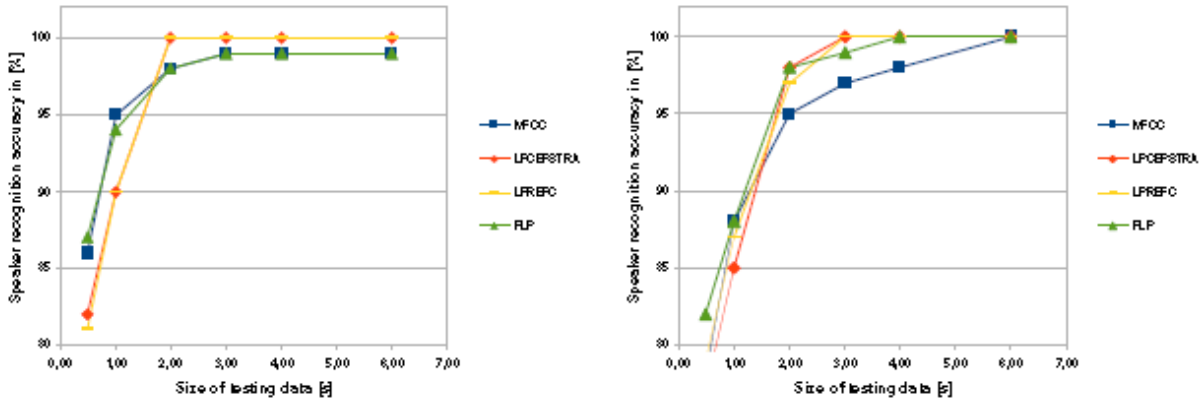

Figure 2: Speaker recognition accuracy in relation to the length of the testing utterance (HMM model on the left; MLP model on the right). The x-axis represents the size of the training data, while the y-axis shows the speaker recognition accuracy

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. Figure 2 shows that the recognition accuracy of all four parametrizations and both classifier are almost similar. We show that the minimal utterance length for the correct speaker recognition is about two seconds. We obtained 100 % of accuracy for LPCEPSTRA/ LPREFC parametrizations and the HMM classifier and 98 % of accuracy for LPCEPSTRA/ PLP parametrizations and the MLP classifier. Furthermore, we show that the HMM is a better

classifier than MLP. From the parametrization point of view, LPCEPSTRA and LPREFC are more accurate than MFCC and PLP for the HMM model, while in the MLP case the three parametrizations (LPCEPSTRA, LPREFC and PLP) are almost similar, only the MFCC parametrization gives worse results.

3.3.3 Automatic recognition of similar voices of two brothers

This experiment concerns only two speakers, brothers with subjectively similar voices. The obtained recognition accuracy is closed to 100 % for all four parametrizations and both classifiers with at least 2.5 minutes of the training data and with the testing utterances of a minimal duration of 2 seconds.

4 Conclusions and Perspectives

In this paper, four parametrizations, namely MFCC, LPCEPSTRA, LPREFC and PLP, and two classifiers, HMM and MLP have been evaluated and compared on the automatic speaker recognition task on the Czech corpus. 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 HMM. It needs only 30 seconds of training data per speaker, while HMM 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 all reported parametrizations/classifiers are almost comparables. We further show that the minimal utterance length for the correct speaker recognition is about two seconds. Furthermore, we show that the HMM is quite a better classifier than the MLP in this task. In the last experiment, we show that it is possible to automatically recognize speakers with subjectively similar voices with a high accuracy.
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) no noise in the corpus; 2) small number of the speakers. Our second perspective thus consists in the evaluation of the parametrizations/classifiers on a larger corpus recorded in real conditions (e.g., with noise in the speech signal). In addition, we studied all parametrizations/classifiers independently. Another extension of this work thus consists in combining these classifiers in order to improve the final result. 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.

5 Acknowledgement

This work has been partly supported by the grant of the Ministry of Education, Youth and Sports of Czech Republic No. NPV II-2C06009.

References

[1] Tishby, N. Z.: „On the application of mixture AR hidden Markov models to text independent speaker recognition.“ In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 39, No. 3, pp. 563-570, 1991.
[2] Kang, G. and Fransen, L.: „Low bit rate speech encoder based on line-spectrum frequency.“ *Tech. Rep. 8857*, NRL, 1985.

[3] Higgins, A. L. and Wohlford, R. E.: „A new method of text-independent speaker recognition." In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Tokio, Japan, pp. 869-872, 1986.
[4] Chmielewska, I.: "Prosody-based text independent speaker identification method." In: *From Sound to Sense*, Massachusetts Institute of Technology, pp. 13-18, June 2004.
[5] Reynolds, D.: "Speaker identification and verification using Gaussian mixture speaker models." In: *Speech Communication*, vol. 17, pp. 91-108, 1995.
[6] Nakagawa, S., Asakawa, K., and Wang, L.: "Speaker recognition by combining MFCC and phase information." In: *Proceedings of Interspeech 2007*, Belgium, Antwerp, August 2007.
[7] Doddington, G. R.: "Speaker recognition-identifying people by their voices." In: *IEEE Proceedings*, vol. 73, no. 11, pp. 1651-1664, 1985.
[8] Higgins, A. et al.: "Speaker verification using randomized phrase prompting." In: *Digital Signal Processing*, vol. 1, no. 2, pp. 89-106, 1991.
[9] Soong, F., Rosenberg, A., Rabiner, L., and Juang, B. H.: "A vector quantization approach to speaker recognition." In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, USA, Florida, pp. 387-390, 1985.
[10] Higgins, A., Bahler, L., and Porter, J.: "Voice identification using nearest neighbor distance measure." In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, USA, Minneapolis, pp. 375-378, 1993.
[11] Rudasi, L. and Zahorian, S. A.: "Text-independent talker identification with neural network." In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Ontario, Canada, pp. 389-392, 1991.
[12] Yang, H. et al., "Cluster adaptive training weights as features in SVM-based speaker verification." In: *Proceedings of Interspeech 2007*, Belgium, Antwerp, August 2007.
[13] Che, C. and Lin, Q.: "Speaker recognition using HMM with experiments on the YOHO Diabase." In: *Proceedings of the International Conference Eurospeech 95*, Spain, Madrid, pp. 625-628, 1995.
[14] Reynolds, D. and Carlson, B.: "Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers." In: *Proceedings of the International Conference Eurospeech 95*, Spain, Madrid, pp. 647-650, 1995.
[15] Douglas, A., Reynolds, D., Quatieri, T. F. and Dunn, R. B.: "Speaker verification using adapted Gaussian mixture models." In: *Digital Signal Processing 10*, pp. 19-41, 2000.
[16] Young, S. et al.: "The HTK Book," Cambridge University, Engineering department, December 2006.
[17] Kukolich, L., Lippman, R.: "LNKnet user's guide." Lincoln laboratory, Massachusetts Institute of Technology, February 2004.