BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **7**, No 3

Sofia • 2007

# Speaker Identification Using Robust Speech Detection and Neural Network[1]

*Atanas Ouzounov*

*Institute of Information Technologies, 1113 Sofia*
*E-mail: atanas@iinf.bas.bg*

***Abstract:*** *An experimental study of the effectiveness of two speech detection parameters in the text-independent speaker identification task is presented in the paper. The first parameter is obtained by processing the spectral autocorrelation function derivative, while the second one is based on the multi-band spectral entropy. The techniques employed are: the two above mentioned parameters and a single MLP for speech detection, LPC cepstrum as a speaker identification feature and a common (for all speakers) MLP for speaker classification procedure. The training and testing have been done using noisy telephone speech data from BG-SrDat corpus. The experiments have shown that in comparison with the multi-band spectral entropy, the use of the spectral autocorrelation function derivative in speech detection results in a lower speaker recognition error.*

***Keywords:*** *Speech detection, multilayer perceptron, speaker recognition.*

## 1. Introduction

The speech detector is one of the key components in speaker recognition systems designed to operate in noisy real-world environments. The recognition error in such systems is due to many causes, one of which is the inaccurate speech fragments detection. The speech fragments usually provide data for speaker model estimation. The non-speech ones are discarded or are used for noise parameters estimation with the purpose of reducing the noise effect on the recognition performance.

[1] This research is supported in part by the Contract BY-TH-202/2006 with the Ministry of Education and Sciences in Bulgaria.

In the study text-independent speaker recognition experiments are carried out. During these experiments the speech part of the analyzed signal is separated using a speech detection module. This module is a particular two-class classification scheme utilizing MultiLayer Perceptron (MLP) as classifier and selected parameters as features. The idea in the paper is to study the effect of the speech detection features on the speaker recognition rate. In the work, the raw speech detection (without speech enhancement and hangover mechanisms) is under study.

The speech detection features used in the study are the Mean-Delta (MD) feature [6, 7, 8] and the Multi-Band Spectral Entropy (MBSE) feature [4]. The previous speech detection experiments with both features and corresponded performance analysis using the ROC-graphs [1] are described in detail in [8].

The text-independent speaker recognition (closed set test) is realized using the Linear Predictive Coding (LPC) cepstrum as feature and a common (for all speakers) MLP as a classifier. The training and testing is carried out using a limited amount of noisy telephone speech data from BG-SrDat corpus [5].

## 2. The robust features

### 2.1. The Mean-Delta feature

The MD parameter is proposed in [6] as a feature for trajectory-based speech detection. Some its modifications intended for pattern recognition-based speech detection are described in [7, 8]. In this study, the modification proposed in [8] will be used.

The MD parameter is estimated using the delta spectral autocorrelation function of the power spectrum of speech signal. Let $x(i)$ is a discrete signal, where $i = 0, \ldots, I-1$, $I$ being the number of samples and the spectrum $X(k)$ of $x(i)$ is obtained by the Discrete Fourier Transform (DFT), where $k = 0,...,K/2$, where $K$ is the number of points in the DFT.

The biased spectral autocorrelation function $R_{\mathrm{p}}(l)$ is defined with the power spectrum $\left|X(k)\right|^2$ as [6]

$$R_{\mathrm{p}}(l) = \sum_{k=0}^{K/2-1-l} \left|X(k)\right|^2 . \left|X(k+l)\right|^2 , \tag{1}$$

where $l = 0,...,L$, $L$ is the number of correlation lags and $L = K/2-1$.

In order to remove the tilt in the spectral autocorrelation function and enhance its peaks, in [6] is proposed a parameter obtained in a way similar to the delta cepstrum evaluation. It is named as Delta Spectral AutoCorrelation Function (DSACF). This parameter is computed as an orthogonal polynomial fit of the first-order derivative (in correlation domain) of the spectral autocorrelation function.

For a particular frame, the DSACF is computed utilizing only the frame's spectral autocorrelation lags. For the $n$th frame, the DSACF $\Delta R_{\mathrm{p}}(n,l)$ is

$$\Delta R_{\mathrm{p}}(n,l) = \frac{\sum_{q=-Q}^{Q} qR_{\mathrm{p}}(n,l+q)}{\sum_{q=-Q}^{Q} q^2}, \tag{2}$$

where $l = 0,...,L$; $n = 0,...,N-1$, $N$ is the number of frames.

The parameter $Q$ defines the window width around the lag $l$ and it effects over the accuracy of the approximation. For the purpose of this study, it is chosen to be between 10 and 15 lags (based on the preliminary experiments).

To design the frame feature vector we find the maximum values of $\left|\Delta R_{\mathrm{p}}(n)\right|$ in different non-overlapping ranges of lags. The MD feature vector for $n$th frame is formed as $\{m_{\mathrm{d}}(1),\ldots,m_{\mathrm{d}}(J)\}$. Its components are defined as follows (for simplicity, the frame index is omitted)

$$m_d(j) = \max\left\{\left|\Delta R_{\mathrm{p}}(l)\right|\right\}_{l=L_j}^{l=L_{j+1}} \tag{3}$$

where $j = 1,\ldots,J$, $J$ is the number of ranges and

$$\{L_1, L_2\},...,\{L_j, L_{j+1}\},...,\{L_{2J-1}, L_{2J}\}$$

are pairs of the boundary lags for each range.

The algorithm for the MD feature vector estimation is summarized as follows (for each frame) [8]:

- apply Hamming window to the analyzed signal;
- compute the power spectrum of the windowed signal via FFT;
- compute the non-normalized biased spectral autocorrelation function by equation (1) with lags $L = K/4$;
- compute the delta spectral autocorrelation function by equation (2);
- take the absolute value of the delta spectral autocorrelation function;
- divide the number of lags $L$ into $J$ non-overlapping lags ranges of equal size;
- find the maximum values of $\left|\Delta R_{\mathrm{p}}(l)\right|$ in the lags ranges $\{L_1, L_2\},...,\{L_j, L_{j+1}\},...,\{L_{2J-1}, L_{2J}\}$ according to (3);
- take the logarithm of the maximum values and obtain the MD feature vector in the form $\{\log(m_{\mathrm{d}}(1)),\ldots,\log(m_{\mathrm{d}}(J))\}$.

The last step in the MD feature vector estimation is the mean normalization. It is done by dividing the MD feature vector for each frame by the average MD feature vector computed over all frames. If the speech data consists of different speech records (files), the mean normalization should be applied for each file separately.

2.2. Multi-band spectral entropy

The spectral entropy for the *n*th frame is estimated in the following steps [4]. First, the Probability Mass Function (PMF) $P(|X(n,k)|^2)$ for the full-band power spectrum $|X(n,k)|^2$ is computed as

$$P(|X(n,k)|^2) = \frac{|X(n,k)|^2}{\sum_{l=0}^{K/2} |X(n,l)|^2}, \quad (4)$$

where $k = 0,...,K/2$, $K$ is the number of DFT-points and $n = 0,...,N-1$, $N$ is the number of frames.

Second, the spectral entropy $H(n)$ for *n*th frame is computed as follows

$$H(n) = -\sum_{k=0}^{K/2} P(|X(n,k)|^2).\log_2(P(|X(n,k)|^2)). \quad (5)$$

The entropy in (5) is named as full-band spectral entropy [4]. To capture a local variation in the spectrum, the idea of multi-band spectral entropy is introduced in [4]. The core of this idea is to divide the full-band PMF into sub-bands and then the spectral entropy to be computed for each sub-band using full-band PMF. In this case, one entropy value is obtained for each sub-band.

According to [4] the Multi-Band Spectral Entropy (MBSE) feature vector for the *n*th frame is formed as $\{H_{\mathrm{MBSE}}(n,1),\ldots,H_{\mathrm{MBSE}}(n,G)\}$ and its components are computed as

$$H_{\mathrm{MBSE}}(n,g) = -\sum_{k=B_g}^{B_{g+1}} P(|X(n,k)|^2)\log_2(P(|X(n,k)|^2)), \quad (6)$$

where $P(|X(n,k)|^2)$ is the full-band PMF in (5); $g = 1,\ldots,G$, $G$ is the number of sub-bands and $\{B_1,B_2\},...,\{B_g,B_{g+1}\},...,\{B_{2G-1},B_{2G}\}$ are pairs of boundary spectral bins for each sub-band.

## 3. Speech detection

The speech detection module is a particular two-class classification scheme utilizing MLP as classifier and selected parameters as features. The MLP with a structure 15-20-1 is selected. The network has 20 neurons in one hidden layer and a single output neuron. The activation functions of the neurons are hyperbolic tangent function (in hidden layer) and sigmoidal function (in output layer). The Rprop algorithm with most typical parameters settings is applied according to recommendation in [9]. The input vector size is set to 15. The used target levels are [0.1; 0.9] and the network is trained in batch mode. In testing mode, in order to make the speech/non-speech decision, the output neuron level is thresholded at 0.5 (speech threshold).

In speech detection module the number of sub-bands in the entropy estimation is $G$=15. The number of lags regions is the same, i.e., $J$ = 15 and $Q$ = 15 in (2) [8].

The speech detection training always precedes the training procedure for the speaker recognition and the training data for these two procedures are not correlated. The training of speech detection module is speaker independent and it is done only once. For more details about speech detection features, training and testing procedures, see [8].

## 4. Speech processing and speaker classification

In the study one large neural network is used to perform the speaker classification task. The data in experiments are selected from small number of speakers (10) and a single MLP with structure 14-100-10 is used. The network has 100 neurons in one hidden layer and 10 output neurons (number of speakers). The input vector size is set to 14. The hyperbolic tangent function is selected as activation function for all neurons. The Rprop algorithm with most typical parameters settings is applied according to recommendation in [9]. The used target levels are [–0.95; 0.95] and the network is trained in batch mode. The structure of MLP is selected based on heuristic considerations and advices given in [2, 3].

The speech data are sampled with frequency of 8 kHz at 16 bits, PCM format and mono mode. The analyzed frequency range is up to 4000 Hz. No additional filtering is applied. The analysis parameters are frame length-30 ms and frame shift-10 ms. In the speech preprocessing are included hamming windowing and a 14th order LPC-derived cepstral vector calculation.

## 5. Experiments

In the speaker recognition experiments are utilized speech samples selected from updated version of the BG-SrDat corpus [5]. The BG-SrDat is a corpus in Bulgarian language collected over noisy analog telephone channels and designed for speaker recognition.

The selected data for speaker recognition included speech material from 10 speakers (male). This data is divided into three groups - for training, testing and validation.

In further text the term ‘speech frames’ means the frames detected as speech by speech detection module.

The data for training and validation is formed by speech data sets. Each set consists of 2000 speech frames randomly collected from speech data obtained from single telephone call. The training data for each speaker consists of 2 speech data set (4000 speech frames from 2 different calls). The validation data consists of only one set per speaker. In testing mode supra segments-based technique is used. The length of supra segment is 200 speech frames and the shift is 100 speech frames. The speaker identification is performed for each supra segment separately. The recognized class is the class with maximum value in the average MLP outputs vector obtained over frames belonging to the particular supra segment. The MLP

training is stopped, when based on the validation test a global minimum in the output mean square error is found or this error is not changed significantly up to 200th epoch.

Since the neural network learning algorithms include random number based procedures, the speech data in the study are utilized by MLP classifier in a multiple runs scheme [2]. In the experiments, 10 runs are performed (typically, runs are not more then 20 [2]). In Table 1 are shown the identification errors in percentage for each speaker of the 10 speakers and for both features. These errors are calculated by averaging over the errors obtained in the 10 runs scheme. In testing data, the number of supra segments are 260 for MD feature and 235 for MBSE.

Table 1. Identification errors in percentages

| № | SPEAKER | FEATURES | |
|---|---|---|---|
| | | MD | MBSE |
| 1 | Spk1 | 38.69 | 45.02 |
| 2 | Spk2 | 4.78 | 3.44 |
| 3 | Spk3 | 7.03 | 0.30 |
| 4 | Spk4 | 5.75 | 31.11 |
| 5 | Spk5 | 12.38 | 79.79 |
| 6 | Spk6 | 2.59 | 98.21 |
| 7 | Spk7 | 13.61 | 0.0 |
| 8 | Spk8 | 66.00 | 3.41 |
| 9 | Spk9 | 74.54 | 89.58 |
| 10 | Spk10 | 0.0 | 0.0 |
| 11 | **Average** | **22.53** | **35.08** |

## 6. Discussion and conclusions

In the experiments with noisy speech data, we study the raw speech detection effect on the speaker recognition rate. The raw speech detection does not utilize any additional techniques to improve speech/non-speech decision. It is often used for development of speech detection algorithms because the lack of improvement techniques helps to identify easily which feature is more effective.

Based on the results shown in Table 1 we conclude that the MD feature provides better speaker recognition rate then the MBSE one. It can see in the table that the identification error for speaker Spk6 is 98.21% when using the MBSE feature and it is only 2.59% when using the MD one. The results for Spk8 are in reverse order. Partial analysis of speech data by spectrograms shows that in signal fragments with nasalized sounds, the raw MBSE trajectory (i.e., speech detection module output before using the threshold) falls often below the speech threshold. In such cases these frames are classified as non-speech ones. Moreover, the random variations in the raw MBSE trajectory due to the telephone noise are more significant then these ones in the raw MD trajectory, which results in more detection errors.

In fact, we are not sure what cause the high error for some speakers. It can be due to wrong speech detection or to bad generalization in the MLP classifier or due to both reasons. It is advisable to do a further research in order to find the error cause.

The forthcoming work will include improvements in the MD feature and attempts to use it in detection of the voiced part of noisy speech data in speaker recognition tasks.

## References


1. Fawcett, T. An Introduction to ROC analysis. – Pattern Recognition Letters, Vol. **27**, 2006, No 8, 861-874.
2. Flexer, A. Statistical Evaluation on Neural Network Experiments: Minimum Requirements and Current Practice. – In: Proc. of the 13th European Meeting on CSR, 1996, 1005-1008.
3. LeCun, Y., Léon Bottou, Genevieve B. Orr, Klaus-Robert Müller. Efficient Backprop, Neural Networks, Tricks of the Trade. Lecture Notes in Computer Science LNCS 1524. Springer Verlag, 1998.
4. Misra, H., S. Ikbal, S. Sivadas, H. Bourlard. Multi-resolution Spectral Entropy Feature for Robust ASR. – In Proc. of ICASSP, 2005, 253-256.
5. Ouzounov, A. BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition Over Telephone Channels. – Cybernetics and Information Technologies, Vol. **3**, 2003, No 2, 101-109.
6. Ouzounov, A. A Robust Feature for Speech Detection. – Cybernetics and Information Technologies, Vol. **4**, 2004, No 2, 3-14.
7. Ouzounov, A. Noisy Speech Detection Using Robust Features and Neural Network. – In: Proc. of the International Conference on Automatics & Informatics 2006, 143-146.
8. Ouzounov, A. Robust Features and Neural Network for Noisy Speech Detection. – Cybernetics and Information Technologies, Vol. **6**, 2006, No 3, 74-83.
9. Riedmiller M., H. Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. – In: Proc. of the ICNN, 1993, 586-591.