BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **9**, No 1

Sofia • 2009

Further Results on Speaker Identification Using Robust Speech Detection and a Neural Network*

Atanas Ouzounov

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

Abstract: *A modification of the mean-delta parameter [7], intended for speech detection, is proposed in this paper. The effectiveness of this modified parameter and three other parameters – basic mean-delta parameter [7], multi-band spectral entropy [4] and frequency-filtering parameter [3] is experimentally studied in speaker identification task. The employed techniques are: for speech detection – each one of the mentioned above parameters as a feature and a single MLP as a classifier; for speaker identification – LPC cepstrum as a speaker identification feature and a common (for all speakers) MLP for speaker classification procedure. The training and testing has been done using noisy telephone speech data from BG-SrDat corpus [5]. The experiments have shown that the proposed modification of the mean-delta parameter improves the speech detection and yields better speaker recognition rate.*

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.

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

Text-independent speaker recognition experiments are carried out in the study. 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 a 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 analysis.

The present study is a continuation of the work described in [8] and it is focused on a modification of the Mean-Delta (MD) parameter [7]. The aim of the modification is to improve the speech detection performance without substantial increasing of the computational cost. This improved parameter is named as Modified Mean-Delta (MMD) parameter. In the paper the MD parameter described in [7] is named as Basic Mean-Delta (BMD) parameter.

In the study the performance of the MMD is compared with three other speech detection parameters – the BMD parameter, the Multi-Band Spectral Entropy (MBSE) parameter [4] and the Frequency-Filtering (FF) parameter [3].

The text-independent speaker recognition (closed set test) is realized using the Linear Predictive Coding (LPC) cepstrum as a feature and 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 analyzed parameters

2.1. The basic mean-delta parameter

The Mean-Delta (MD) parameter is proposed in [6] as a feature for trajectory-based speech detection. Its version intended for pattern recognition-based speech detection (i.e. BMD parameter) is described in [7, 8].

The BMD parameter is estimated using the delta spectral autocorrelation function of the power spectrum of speech signal. 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 $\Delta R_p(n,l)$ is

$$\Delta R_{\mathrm{p}}(n,l) = \frac{\sum_{q=-Q}^{Q} qR_{\mathrm{p}}(n,l+q)}{\sum_{q=-Q}^{Q} q^2}, \tag{1}$$

where $l = 0,...,L$; L is the number of correlation lags and $L = K/2 - 1$ (K is the FFT size); $n = 0,...,N-1$, N is the number of frames and $R_{\mathrm{p}}(n, l)$ is the biased spectral

autocorrelation function defined with the power spectrum [6]. The parameter Q determines the window width around the lag l and its effect over the accuracy of the approximation.

The BMD 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)

$$m_{\mathrm{d}}(j) = \max\left\{\left|\Delta R_{\mathrm{p}}(l)\right|\right\}_{l=L_j}^{l=L_{j+1}} \tag{2}$$

where $j = 1,\ldots,J$, J is the number of lags ranges and $\{L_1, L_2\}$ $\ldots\{L_j, L_{j+1}\}$ $\ldots\{L_{2J-1}, L_{2J}\}$ are pairs of boundary lags for each range.

The algorithm for the BMD feature vector estimation is summarized as follows (for each frame) [7]:

a) apply Hamming window to the analyzed signal;

b) compute the power spectrum of the windowed signal via FFT with size K;

a) compute the non-normalized biased spectral autocorrelation function with lags $L = K/4$;

d) compute the delta spectral autocorrelation function by equation (1);

e) take the absolute value of the delta spectral autocorrelation function;

b) divide the number of lags L into J non-overlapping lags ranges of equal size;

c) find the maximum values of $\left|\Delta R_{\mathrm{p}}(l)\right|$ in the lags ranges $\{L_1, L_2\}\ldots\{L_j, L_{j+1}\}\ldots\{L_{2J-1}, L_{2J}\}$ according to (2);

d) take the logarithm of the maximum values and obtain the BMD feature vector in the form $\{\log(m_{\mathrm{d}}(1)),\ldots,\log(m_{\mathrm{d}}(J))\}$;

i) mean normalization – the BMD feature vector for each frame is divided by the average BMD 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. The modified mean-delta parameter

The Modified Mean-Delta (MMD) parameter is obtained as a result of careful experimental research of the characteristics of the BMD parameter. The aim of this research is to improve the speech detection performance of the BMD feature without significantly increasing the computational cost.

The algorithm for MMD feature calculation is summarized as follows (for each frame). The steps a), b), c) and d) are the same as in the BMD algorithm (see 2.1):

a) apply Hamming window to the analyzed signal;

b) compute the power spectrum of the windowed signal via FFT;

c) compute the non-normalized biased spectral autocorrelation function with lags $L = K/4$;

d) compute the delta spectral autocorrelation function by equation (1);

e) use a sliding rectangular window across the autocorrelation lags with a wide of *Y* lags and a sliding step of *U* lags;

f) the MMD parameter $\mathrm{mm}_{\mathrm{d}}(v)$, where $v = 1,...,V$ is the number of sliding steps across the autocorrelation lags, is defined as

$$\mathrm{mm}_{\mathrm{d}}(v) = \left| \max\left\{ \Delta R_{\mathrm{p}}(l) \right\} \right|_{l=(v-1)*U}^{l=(v-1)*U+Y}; \tag{3}$$

g) take the logarithm of $\mathrm{mm}_{\mathrm{d}}(v)$ and obtain the MMD feature vector in the form $\{\log(\mathrm{mm}_{\mathrm{d}}(1)),\ldots,\log(\mathrm{mm}_{\mathrm{d}}(V))\}$;

h) mean normalization – the MMD feature vector for each frame is divided by the average MMD feature vector computed over all speech frames (over all used files).

In order to be able to compare the BMD and MMD features performance, the sliding window wide *Y* and sliding step *U* are selected in such a way so the number of sliding steps *V* across the autocorrelations lags to be equal to the number of lags regions *J* (see 2.1).

2.3. Multi-band spectral entropy [4]

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 $|X(n,k)|^2$ according to [4] is

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

where $k = 0,...,K/2$, K is the number of DFT-points and $n = 0,...,N-1$, N is the number of frames. The PMF in (4) is known as the full-band PMF.

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

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

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

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

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

2.4. Frequency-filtering parameter [3]

If $E(\omega)$ is the spectral envelope of analyzed speech signal and $S(\omega) = \log(E(\omega))$, then the derivative of $S(\omega)$ according to [3] is

$$\frac{d}{d\omega}S(\omega) = \frac{1}{E(\omega)}\frac{d}{d\omega}E(\omega)\,. \tag{7}$$

When the spectral energies are defined in a discrete frequency scale (e.g., obtained by filter banks), the derivative in (7) should be replaced by a difference. If the spectral slope is measured as the difference between the two samples surrounding the current one, then the log-spectral derivative results in the log-spectral difference or the Frequency Filtering (FF) parameter. This parameter is defined in [3] as follows

$$S_{\mathrm{FF}}(k) = S(k+1) - S(k-1)\,, \tag{8}$$

where k is the frequency sub-band index.

The FF parameter and linear discriminant analysis are used successfully in [9] to obtain robust speech detection.

3. Speech detection and speaker recognition modules

In the study two separate modules are designed. The first one is the speech detection module. This module is a particular two-class classification scheme utilizing MLP as classifier and mentioned above four parameters as features. The target sequences are formed using manually segmented speech data. For each one of mentioned above features a separate set of segmented speech data is obtained during the classification task. These sets are later used in the speaker recognition module. The speech detection module works in speaker independent mode.

The second module is the speaker recognition (i.e. speaker identification) module. It uses a MLP as classifiers and LPC-derived cepstral vectors as features. The data utilized in this module is segmented in speech and non-speech frames by speech detection module in advance. Only speech frames are used for speaker recognition.

3.1. Speech detection module

In this module a single MLP is used with structure 15-20-1. The network has 20 neurons in one hidden layer and a single output neuron. The activation functions of the neurons are a hyperbolic tangent function (in a hidden layer) and a sigmoidal function (in an output layer). The Rprop algorithm with the most typical parameters settings is applied according to the recommendation in [10]. The input vector size is set to 15. The target levels used are [0.1; 0.9] and the network is trained in a batch

mode. In a testing mode, in order to make the speech/non-speech decision, the output neuron level is thresholded at the mean of the output neuron values obtained over particular tested data (single file), i.e. this is the speech threshold (in [8] this threshold is set to 0.5).

The speech data for speech detection are sampled with a 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 speech preprocessing Hamming windowing and corresponded feature extraction are included – BMD feature, MMD feature, MBSE and FF feature. In this module the accepted feature vector size is 15. Therefore the number of sub-bands G=15 in the MBSE, the number of lags regions J=15 in the BMD, the number of sliding steps V=15 in the MMD, the number of Mel-scale triangular filters used to generate the FF parameter is 15 and the parameter Q=15 in (1).

The waveform of the noisy speech fragment with a length of 4 s and the corresponding trajectories of the output neuron level for analyzed features are shown in Fig. 1. These trajectories are obtained with already trained MLP and they are shown for illustrative purposes only.

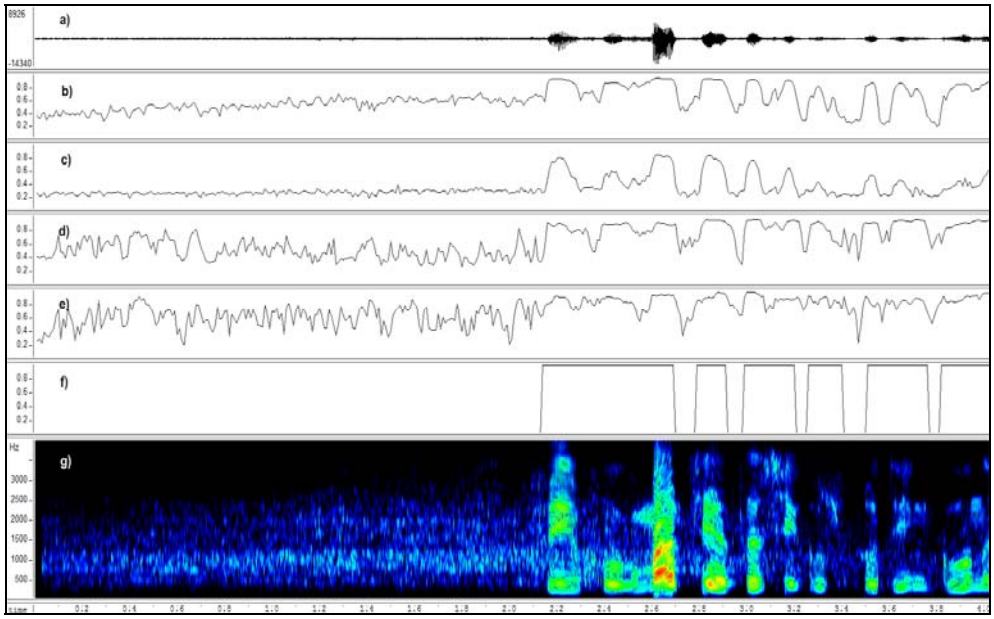

Fig. 1. Speech fragment and the corresponding output neuron level trajectory for analyzed features: (a) speech waveform, (b) BMD feature, (c) MMD feature, (d) MBSE feature, (e) FF feature, (f) manual segmentation and (g) spectrogram

3.2. Speaker recognition module

In this module a single neural network is used to perform the speaker classification task. The data in the experiments are selected from a small number of speakers (10)

and a 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 an activation function for all neurons. The Rprop algorithm with the most typical parameters settings is applied according to the recommendation in [10]. The target levels used are [–0.95; 0.95] and the network is trained in a batch mode. The structure of MLP is selected based on the heuristic considerations and advices given in [1, 2].

The speech data for speaker recognition are sampled with a frequency of 8 kHz at 16 bits, PCM format and a 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 Hamming windowing, a 14th order LPC-derived cepstral vector calculation and the cepstral mean subtraction technique are included.

4. Experiments

In the 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 comprise telephone speech collected under different noisy environments, i.e. the multi-style training is used.

The speech detection data are separated into three groups – for training, validation and testing. Each one of the first two groups includes roughly 40 000 frames selected from 10 speakers. The testing data are data which are used in a speaker recognition module (see in the text below). In further text the term ‘speech frames’ means the frames detected as speech by speech detection module.

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. The data for training and validation is formed by speech data sets. Each set consists of 1800 speech frames randomly collected from speech data obtained from a single telephone call. The training data for each speaker consists of 2 speech data sets (3600 speech frames from 2 different calls). The validation data consists of only one set per a speaker. In the testing mode supra segments-based technique is used. The length of a 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 the 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 the 200th epoch.

Since the neural network learning algorithms include random number based procedures, the speech data in the study are utilized by a MLP classifier in a multiple runs scheme [1]. This scheme is valid for both modules. The performed runs are 5 and 10, for speech detection and speaker recognition modules, respectively. Usually the recommended number of runs is not more than 20 [1].

In Table 1 the identification errors in percentage for each speaker and each feature are shown. These errors are calculated by averaging over the errors obtained in the 5×10 runs scheme, i.e. each speech detection run provides one set of segmented speech data (for each feature) that is later used in 10 speaker recognition runs.

Table 1. Identification errors in percentages

Speaker	Features			
	BMD	MMD	MBSE	FF
Spk No 1	57.04	22.72	18.08	38.40
Spk No 2	6.79	5.58	2.51	5.06
Spk No 3	0	0	0	0
Spk No 4	61.35	70.59	85.57	75.65
Spk No 5	16.87	20.17	31.26	41.06
Spk No 6	1.34	0.36	3.79	1.07
Spk No 7	3.33	1.36	0.9	0.81
Spk No 8	49.05	13.36	80.31	95.08
Spk No 9	55.65	46.68	33.95	21.18
Spk No 10	0.18	0	0	0.08
Average	**25.16**	**18.08**	**25.63**	**27.83**

5. 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.

The next two processing steps are important in the proposed modification of the MD parameter. First, instead of the non-overlapping lags ranges a sliding overlapping window is used. The width of this window is usually more than twice of the single lags range width. Thereby the final feature vector is additionally smoothed. And second, the mean normalization is done with the mean evaluated over all speech frames in a particular data set (typically it comprises a few files). In this way more reliable estimation of the mean vector is obtained. In the basic MD parameter algorithm the mean is evaluated only over speech frames from a single file (the files selected from the speech corpus have approximately the same length).

The trajectories (the raw speech detection module output, i.e., before using of the speech threshold) of all the features obtained on a noisy speech fragment with length of 4 s are shown in Fig. 1. It is evident that the noise has more effect on the trajectories of the MBSE and FF parameters than on the trajectories of both MD parameters – see the non-speech activity fragment in the first two seconds of data in Fig. 1 (d), (e) and (b), (c), respectively. Moreover, in the same fragment, the trajectory of the MMD parameter possesses less random variations than the BMD one (see Fig. 1 (c) and (b)).

It is worth to note that the noisy speech fragment shown in Fig. 1 is a part of the training data for the speaker noted as Spk No 8 in Table 1. The large variations

in MBSE and FF parameters trajectories (see Fig. 1 (d) and (e)) cause poor speech detection and this results in a very high error rate for that speaker (see Table 1, Spk No 8 – for MBSE and FF features).

Based on the results shown in Table 1 we conclude that the MMD feature provides the best speaker recognition rate among all the features. In comparison with the basic MD feature the modified MD feature improves the recognition rate for 7 of 10 speakers and makes worse for two of them. However for some speakers (i. e., see Table 1, Spk No 4 for all features) the error rate remains unacceptably high. The additional analysis of speech data for Spk No 4 revealed that the main cause for this low recognition rate is the high amount of distortions observed in these data.

Our forthcoming work will include some attempts to utilize the MMD feature in detection of the voiced part of noisy speech and to use the obtained detection results in speaker recognition tasks.

References

1. F l e x e r, A. Statistical Evaluation on Neural Network Experiments: Minimum Requirements and Current Practice. – In: Proc. of the 13th European Meeting on CSR, 1996, 1005-1008.
2. L e C u n, Y., L. B o t t o u, G. O r r, K.-R. M ü l l e r. Efficient Backprop, Neural Networks, Tricks of the Trade. Lecture Notes in Computer Science LNCS 1524, Springer Verlag, 1998.
3. M a c h o, D., C. N a d e u. Comparison of Spectral Derivative Parameters for Robust Speech Recognition. Eurospeech 2001, 205-208.
4. M i s r a, H., S. I k b a l, S. S i v a d a s, H. B o u r l a r d. Multi-Resolution Spectral Entropy Feature for Robust ASR. – In: Proc. of the ICASSP, 2005, 253-256.
5. O u z o u n o v, 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. O u z o u n o v, A. A Robust Feature for Speech Detection. – Cybernetics and Information Technologies, Vol. **4**, 2004, No 2, 3-14.
7. O u z o u n o v, A. Robust Features and Neural Network for Noisy Speech Detection. – Cybernetics and Information Technologies, Vol. **6**, 2006, No 3, 74-83.
8. O u z o u n o v, A. Speaker Identification using Robust Speech Detection and Neural Network. – Cybernetics and Information Technologies, Vol. **7**, 2007, No 3, 48-54.
9. P a d r e l l, J., D. M a c h o, C. N a d e u. Robust Speech Activity Detection using LDA Applied to FF. – In: Proc. of the ICASSP, 2005, I.557-I.560.
10. R i e d m i l l e r, M., H. B r a u n. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. – In: Proc. of the ICNN, 1993, 586-591.