Robust Features for Speech Detection – A Comparative Study

Atanas Ouzounov

Abstract: *In the paper are presented the results from an experimental comparative study of four robust features intended for speech detection. These features are the Mean-Delta (MD) feature [6], the Spectral Entropy (SE) [3], the Spectral Entropy with Normalized frame Spectrum (SENS) [7] and the Relative Spectral Entropy (RSE) [1]. For noisy speech, the trajectory's variations of the features are compared by visual evaluation on their graphical representations. Noisy speech samples from two databases (the SpEAR database [2] and the BG-SRDat corpus [5]) are selected and used in the study. In all tests, the performance of the MD feature is similar or better than that of the other features.*

Key words: *Speech Detection, Voice Activity Detection, Spectral Entropy.*

INTRODUCTION

The location of speech embedded in various non-speech events has many names, of which some are speech detection, endpoints detection, voice (speech) activity detection, and speech/non-speech segmentation [4].

The speech detection algorithms can be divided into two general categories. The first one includes the algorithms that analyze the time variations (trajectories) of selected parameters and utilize a set of thresholds in order to produce a speech/non-speech decision for a particular segment. The second category is comprised of algorithms based on a pattern recognition technique. In this case, during the training mode the reference models for two classes (i.e., speech and non-speech) are created based on selected speech features [3, 4].

The feature selection for speech detection tasks is composed usually of two stages. The first stage is a preliminary selection. It is based on a visual evaluation on the graphically represented parameters. This selection is a feasible task only in cases when the parameters possess reasonable graphical representation. The latter stage is the final feature selection and usually a recognition scheme is applied. The developed speech detection algorithm is embedded as a component of a complete speech or speaker recognition system. The effectiveness of different speech detection features is estimated experimentally based on their indirect influence on the recognition performance [3, 4, 7].

In the paper, four robust features indented for trajectory-based speech detection are experimentally studied. These features are spectral entropy [3], the spectral entropy with normalized frame spectrum [7], the relative spectral entropy [1] and the mean-delta feature [6]. For different noisy speech signals, the trajectory's variations of the features are compared by visual evaluation on their graphical representations.

ROBUST FEATURES

The Mean-Delta feature

The Mean-Delta (MD) feature is proposed in [6] and it is defined as the mean of the absolute values of the delta spectral autocorrelation function of the power spectrum of speech signal. Let $x(i)$ is a discrete speech signal, where $i = 0,...,I-1$, I is the number of samples and the spectrum $X(k)$ of $x(i)$ is obtained by the Discrete Fourier Transform (DFT), where $k = 0,...,K/2$, K is the number of points in the DFT.

The spectral autocorrelation function $R_p(l)$ is defined with the power spectrum as [6]

$$R_p(l) = \sum_{k=0}^{K/2-1-l} \left|X(k)\right|^2 \cdot \left|X(k+l)\right|^2 , \qquad (1)$$

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

This research is supported in part by the National Science Fund of the Bulgarian Ministry of Education and Science, Contract No. И-1302/2003

The Delta Spectral AutoCorrelation Function (DSACF) is the first-order derivative of the spectral autocorrelation function obtained by a polynomial approximation in a manner similar to the delta cepstrum evaluation [6]. For particular frame it is computed using only frame's spectral autocorrelation lags (intra-frame processing).

For the *n*th frame the DSACF $\Delta R_P(n,l)$ is computed as

$$\Delta R_P(n,l) = \frac{\sum_{q=-Q}^{Q} qR_P(n,l+q)}{\sum_{q=-Q}^{Q} q^2}, \quad (2)$$

where $l = 0,...,L$; *Q* is typically between 2 and 5, i.e. regions from 5 to 11 lags are analyzed in the autocorrelation domain, and $n = 0,...,N-1$, *N* is the number of frames.

For *n*th frame the MD feature $m_d(n)$ is computed as follows

$$m_d(n) = \frac{1}{\Delta L}\sum_{l=L_1}^{L_2} \left|\Delta R_p(n,l)\right|, \quad (3)$$

where $\Delta R_P(n,l)$ is the DSACF in (2) for lag l, L_1 and L_2 are the boundary lags and $\Delta L = L_2 - L_1 + 1$. For more details about the MD feature, see the full article [6].

The spectral entropy

The Spectral Entropy (SE) for the *n*th frame is estimated in the following steps [3]. First, the probability density function $P(|X(n,k)|^2)$ for the spectrum $|X(n,k)|^2$ is computed as

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

where $k = 0,...,K/2$ and $n = 0,...,N-1$. The speech spectrum is limited to the frequency range from 250 Hz to 3750 Hz and some heuristic rules are added, namely if frequency component *k* is below 250 Hz and above 3750 Hz then $|X(n,k)| = 0$; if $P(|X(n,k)|^2) \geq 0.9$ then $P(|X(n,k)|^2) = 0$. After the above constrains are applied the spectral entropy $H_c(n)$ for *n*th frame is computed as follows

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

The negative SE $H_c^-(n)$ is defined as $H_c^-(n) = -H_c(n)$. It is more convenient in the trajectory-based speech detection algorithms to be used the negative SE, especially when this entropy will be combined or be compared with the energy-based features.

The spectral entropy with normalized frame spectrum

The entropy measure of the magnitude spectrum is proposed in [7] to be used as speech detection feature. It is known that the entropy curve of the speech regions with colored noise is very similar to the entropy curve of the non-speech regions. To make the speech detection with entropy feature under colored noise conditions more reliable, in [7] is proposed to divide the spectrum of each frame by the average spectrum computed over all frames of the analyzed speech data (i.e. to normalize the frame spectrum). If $|X(n,k)|$ is the magnitude spectrum for the *n*th speech frame, where $n = 0,...,N-1$; $k = 0,...,K/2$ and *K* is the number of points in the DFT and *N* is the number of frames, so the normalized spectrum $|\hat{X}(n,k)|$ is computed as follows

$$\left|\hat{X}(n,k)\right| = \frac{\left|X(n,k)\right|}{\frac{1}{N}\sum_{n=0}^{N-1}\left|X(n,k)\right|}. \quad (6)$$

The probability density function $P(\left|\hat{X}(n,k)\right|^2)$ for the spectrum $\left|\hat{X}(n,k)\right|$ is estimated by normalizing the frequency components

$$P(\left|\hat{X}(n,k)\right|^2) = \frac{\left|\hat{X}(n,k)\right|^2}{\sum_{k=0}^{K/2}\left|\hat{X}(n,k)\right|^2}, \quad (7)$$

and the Spectral Entropy with Normalized frame Spectrum (SENS) $H_w(n)$ for *n*th frame is computed as

$$H_w(n) = -\sum_{k=0}^{K/2} P(\left|\hat{X}(n,k)\right|^2).\log(P(\left|\hat{X}(n,k)\right|^2)). \quad (8)$$

The negative SENS $H_w^-(n)$ is defined as $H_w^-(n) = -H_w(n)$.

The relative spectral entropy

The Relative Spectral Entropy (RSE) is proposed in [1] as a speech detection feature. The RSE with respect to the mean spectrum is very useful in situations where the speech signal is contaminated with constant voicing source, e.g. the fan noise, etc. The mean spectrum for the particular frame is computed over its neighbouring frames. These frames form a shifting (along the speech data) supra-segment which length is typically a few hundred milliseconds.

Let the probability density function $P(\left|X(n,k)\right|^2)$ for the spectrum $\left|X(n,k)\right|^2$ is computed as follows

$$P(\left|X(n,k)\right|^2) = \frac{\left|X(n,k)\right|^2}{\sum_{k=0}^{K/2}\left|X(n,k)\right|^2}, \quad (9)$$

where $n = 0,...,N-1$; $k = 0,...,K/2$, *K* is the number of points in the DFT and *N* is the number of frames. The mean spectrum $\left|Y(n,k)\right|^2$ for the *n*th speech frame is computed as follows

$$\left|Y(n,k)\right|^2 = \frac{1}{M+1}\sum_{m=n-M/2}^{n+M/2}\left|X(m,k)\right|^2, \quad (10)$$

where M is the number of frames belong to the shifting supra-segment.
In that case, the RSE $H_r(n)$ for the *n*th frame is defined as [1]

$$H_r(n) = -\sum_{k=0}^{K/2} P(\left|X(n,k)\right|^2).\log\frac{P(\left|X(n,k)\right|^2)}{\left|Y(n,k)\right|^2}. \quad (11)$$

The negative RSE $H_r^-(n)$ is defined as $H_r^-(n) = -H_r(n)$.

EXPERIMENTS

We carried out series of experiments that can be divided into two groups. The aim of these experiments is to perform graphically and to evaluate visually the trajectories of analyzed features. During the first group of experiments, we used selected noise-corrupted speech samples from the SpEAR database [2]. In the second group of experiments, we utilized noisy telephone speech sample from the BG-SRDat corpus [5].

In order to make a correct comparison between different features we have to compute all of them in the same frequency range. We selected the range accepted in [3], i.e. from 250 Hz to 3750 Hz. In all experiments, the obtained trajectories are normalized in

the range from zero to one to allow direct comparison between them. The frame length is 30 ms, the frame shift is 10 ms and the FFT-points are 1024. The mean spectrums in (6) and (10) are estimated over entire analyzed speech phrase. All contours are smoothed by 3-points moving-average filter.

Hereafter in the text, the attribute 'negative' will be omitted before all entropy measures for more convenience.

Experiment No.1

We selected one record from Lombard section and one from noisy speech recordings section in the SpEAR database. All records have clear reference and corresponded noisy speech samples with different Signal-to-Noise Ratio (SNR). All selected wave files are with sampling frequency of 16 kHz at 16 bits, PCM format and mono mode [2].

The record from the Lombard section contains speech corrupted with factory noise. The SNR of clean reference is 27.28 dB and for noisy recording SNR = - 9.96 dB. The record from the noisy section contains speech corrupted with bursting noise with SNR = 0.16 dB. The noise is computer generated using a white Gaussian random number generator.

In Figures 1 and 2 are shown the noisy speech examples from SpEAR database and the corresponded trajectories of the SE, SENS, RSE and MD feature. The factory noise speech record is shown in Figure 1, while the bursting noise speech record is shown in Figure 2.

Experiment No.2

The BG-SRDat is a corpus in Bulgarian language recorded over analog telephone channels and intended for speaker recognition. The speech data included in the BG-SRDat are sampled with frequency of 8 kHz at 16 bits, PCM format and mono mode [5].

We selected one record, which is typical of the BG-SRDat. It distinguishes for the presence of high-level pulse noise. In Figure 3 are shown the noisy speech example from the BG-SRDat corpus and the corresponded trajectories of the analyzed features.

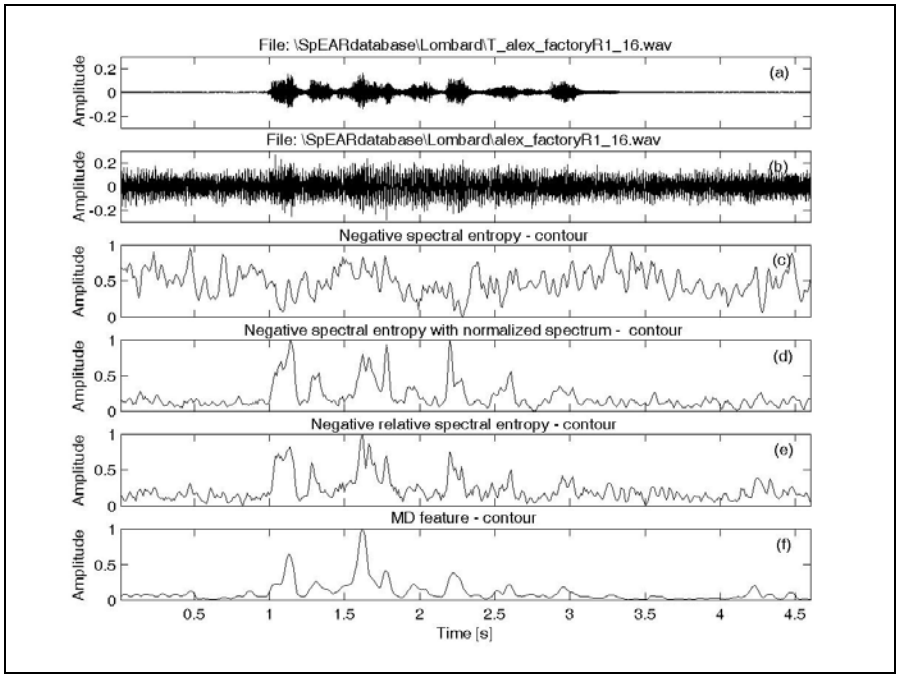

Fig.1. Examples from the SpEAR database: (a) clean speech sample, (b) noisy version of (a) with factory noise, (c) SE contour for noisy speech in (b), (d) SENS contour for noisy speech in (b), (e) RSE contour for noisy speech in (b) and (f) MD feature contour for noisy speech in (b).

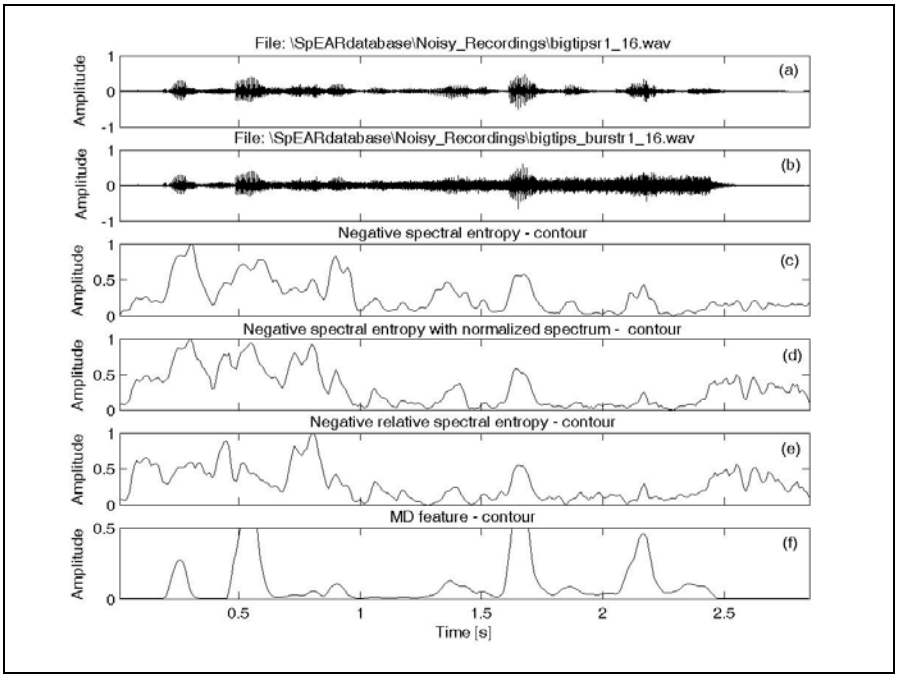

Fig.2. Examples from the SpEAR database: (a) clean speech sample, (b) noisy version of (a) with bursting noise, (c) SE contour for noisy speech in (b), (d) SENS contour for noisy speech in (b), (e) RSE contour for noisy speech in (b) and (f) MD feature contour (with 2x zoom on Y-axis) for noisy speech in (b).

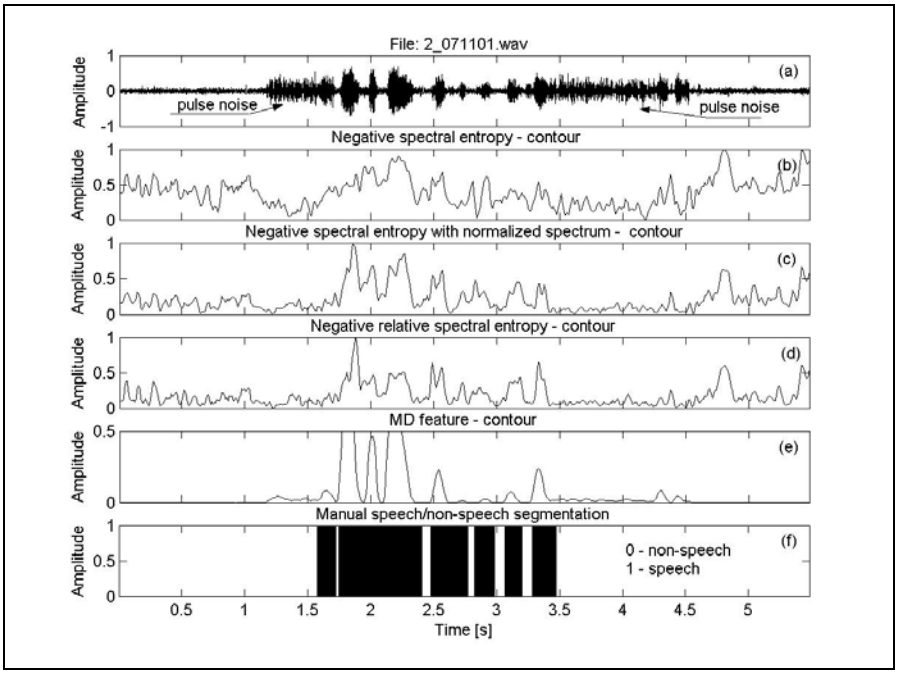

Fig.3. An example from the BG-SRDat corpus: (a) noisy speech sample, (b) SE contour (c) SENS contour, (d) RSE contour, (e) MD feature contour (with 2x zoom on Y-axis) and (f) manual speech detection contour.

DISCUSSION

In the experiments with noisy speech data, we compared the trajectories variations of the MD feature and the group of the spectral entropy-based features. We decided to use the feature value (contour level) as measure for the presence of speech. This is a so-called 'energy-type' approach for speech detection. It is based on the assumption that the low levels in feature's contour correspond to the non-speech frames or frames with consonants and the high levels ones – mainly to the voiced or semi-voiced frames.

The results from experiment No.1 are shown in Fig.1 and Fig.2. The factory noise used in the test is a kind of colored noise. In this case, the SE contour provides the worst results. As can be seen in Figure 1 (c) it is very difficult to make reliable decision (based

only on the contour level) about the place of the speech and non-speech parts in the analyzed data. Conversely, the trajectories of the SENS, RSE and the MD feature allow finding out more easily the speech and non-speech fragments – see Figure 1 (d), (e) and (f). The results shown in Figure 2 are obtained with bursting noise. In this case, the entropy-based contours are not suitable for trajectory-based speech detection. Again, the MD feature performs itself very well.

The results from experiment No.2 are shown in Fig. 3. Again, the SE provides the worst result, while the MD feature allows easy to find speech and non-speech fragments based only on contour levels.

CONCLUSIONS AND FUTURE WORK

In the study, we analyzed four features intended for speech detection. We carried out visual evaluation on their graphical representations for noisy speech samples selected from two databases. Based on experimental results we made the following conclusions:

-the behaviour of all features depends on the type of noise - this dependence is more significant for the spectral entropy-based features;

-the spectral entropy-based features produce the slightly better results than the MD feature only for white noise test (these results are not presented here due to the lack of space);

-the worst results for all tests are obtained for SE feature (except for the white noise test);

-the SENS and the RSE are very promising features but they are more influenced by non-stationary noises (as bursting and pulse noises) than the MD feature;

-in the MD feature trajectory can be noticed an extra smoothing, especially for the low-energy speech sounds (see Fig. 3 (e) – between time axis ticks 2.5 s and 3 s).

Our further work will include the development of an integrated feature-based speech detection algorithm (e.g., a combination of the MD feature and one of the entropy-based features – RSE or SENS). We will evaluate this algorithm in the context of speaker recognition system, in order to estimate the efficiency of this new feature as a component of a complete system.

REFERENCES

[1] Basu S., A Linked-HMM Model for Robust Voicing and Speech Detection, ICASSP 2003, vol.1, pp.I-816-I-819.

[2] Center for Spoken Language Understanding, Speech Enhancement and Assessment Resource (SpEAR) Database, Oregon Graduate Institute of Science and Technology, *http://cslu.ece.ogi.edu/nsel/data/SpEAR_database.html*

[3] Liang-sheng Huang and Chung-ho Yang, A Novel Approach to Robust Speech Endpoint Detection in Car Environment, ICASSP'2000, pp.1751-1754.

[4] Li Q., J. Zheng, A. Tsai, Q. Zhou, Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition, *IEEE Transaction on SAP*, vol.10, No.3, March 2002, pp.146-157.

[5] Ouzounov A., BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels, *Cybernetics and Information Technologies*, vol. 3, No.2, 2003, pp.101-109, *http://www.iit.bas.bg/staff_en/BG_SRDat.pdf*

[6] Ouzounov A., Robust Feature for Speech Detection, *Cybernetics and Information Technologies*, vol.4, No.2, 2004, pp.3-14,

http://www.iit.bas.bg/staff_en/SpeechDetectionFeature.pdf

[7] Renevey Ph. and A. Drygajlo, Entropy Based Voice Activity Detection in Very Noisy Conditions, EUROSPEECH'01, 2001, pp.1883-1886.

ABOUT THE AUTHOR

Assist. Prof. Atanas Ouzounov, Institute of Information Technologies, Bulgarian Academy of Sciences, Sofia, Phone + 359 28 706493, E-mail: atanas@iinf.bas.bg.