Automatic Speech Segmentation in High Noise Condition

Rosen Ivanov

Department “Computer Systems and Technologies”, Technical University
4 “H.Dimitar”, 5300 Gabrovo, Bulgaria
rosen@tugab.bg

Abstract

The accurate segmentation of speech and end points detection in adverse condition is very important for building robust automatic speech recognition (ASR) systems. Segmentation of speech is not a trivial process - in high noise conditions it is very difficult to determine weak fricatives and nasals at end of the words. An efficient threshold (a priory defined) independent speech segmentation algorithm, robust to level of disturbance signals, is developed. The results show a significant improvement of robustness of proposed algorithm with respect to traditional algorithms.

Introduction

Current commercial systems for speech recognition are mainly used in the field of digital communication. Mobile access and extracting of information from global networks as Internet, as well as control through voice commands of devices connected to the network appear good condition for developing ASR systems which are able to keep their performance in the presence of various disturbance signals.

One of the methods for increasing the noise immunity of ASR systems is the implementation of training with a lot of speech material, obtained in different working conditions (multi-condition training). To model speech faster and accurately, it is necessary that training is implemented of the level of segments containing phonetically balanced information, rather than on the basis of asynchronously generated segments with a fixed length. Automatic segmentation has to be also used when the vocabulary size of the system is very large, in order to reduce the time for generation of the base with the training records.

During speech segmentation records inappropriate for the training process have to be removed. Those are records which mainly contain non-verbal sounds or transient noise (door slams, lip smacks, bread noise etc.) with high amplitude.

It is difficult to determine the beginning and the end of each word when weak fricatives (s, f, h), nasals (m, n) or trailing vowels at the end are present. In these cases, when SNR<5dB, the disturbance signals mask the speech signal and reliable segmentation is possible only when it is combined with noise reduction.

Basic Algorithms for Automatic Segmentation

Existing algorithms for speech segmentation and endpoints detection are based on features like: energy estimation, E; zero crossing rate, ZCR; spectral slope, periodicity, degree of likelihood of adjacent segments etc.

The main disadvantage of algorithms using E and ZCR is that a some threshold levels have to be specified in advance. This algorithms give satisfactory results only in the absence of high-amplitude noise.

With MLR (Maximum Likelihood Ratio) algorithm [1] an evaluation of the significant changes of the signal through the frequency content is used. It is assumed that within each segment speech features have Gaussian distribution. In this case the segmentation is reduced to comparison of the likelihood ratio logarithm of the segments with a thresholds defined in advance. The algorithm doesn’t give any satisfactory result when SNR<5dB.

Hirsch has developed an algorithm [2] based on analysis of histograms of the spectral magnitudes below a dynamically updated threshold. The segmentation is based on the smoothed along time noise level estimated as maximum of distribution in each spectral subband. The algorithm gives unsatisfactory result when transient noise present.

Histogram clustering algorithm [3] is similar to Hirsch histograms. In this case histograms of logarithmic short-time power are computed on relatively long speech segments. They are approximated by two normal distributions - one for speech and one for noise. The threshold value is defined by the point of intersection of the distributions. When level of noise is high, peaks are closer to each other and accurate segmentation is not possible.

Description of Algorithm

Proposed algorithm is based on the analysis of significant acoustic changes of speech signal in time-frequency domain. As with MLR algorithms, it is assumed that the input signal for the different frequency subbands has a distribution similar to Gaussian. In comparison with base MLR algorithm where analysis is realized by low-order IIR filters or Bark/Mel filters after FFT [4], the proposed algorithm is implemented by a wavelet transformation, WT. We use good time resolution for high frequencies and good frequency resolution for low frequencies (multi-resolutional analysis) of WT. Thus the accurate localization of spectral transitions and transient non-stationary noise is possible.

Daubechies 16-coefficient wavelet filter and 5 level of decomposition of signal are used:

level 1: 2000Hz - 4000Hz;
level 2: 1000Hz - 2000Hz;

level 3: 500Hz - 1000Hz;
level 4: 250Hz - 500Hz;
level 5: 125Hz - 250Hz.

Frequency range up to 125Hz is not used as it doesn't contain any information important for the segmentation.

If $\{\mathbf{x}_k^j\}$ is the vector obtained after DFT (Discrete WT) for segment *k*, level *j*, the signal distribution is

$$p\left(\mathbf{x}_k^j\right) = \prod_{i=1}^{N_j} \frac{1}{\sqrt{2\pi}\left(\sigma_k^j\right)^2}\, exp\left\{-\frac{1}{2}\left[\frac{x_k^j(i)}{\sigma_k^j}\right]^2\right\}, \text{ where}$$

N_j is the number of component in $\{\mathbf{x}_k^j\}$;

$\left(\sigma_k^j\right)^2$ is the variance of $\{\mathbf{x}_k^j\}$.

Parameters $\alpha_k^j = ln\left[p\left(\mathbf{x}_{k-1}^j\right) / p\left(\mathbf{x}_k^j\right)\right]$ give quantitative estimation for changes in the distribution of every two adjacent segments and they can be used for speech segmentation,

$$\alpha_k^j = \sum_{i=1}^{N_j}\left[ln\left(\frac{\sigma_k^j}{\sigma_{k-1}^j}\right)^2 + \frac{1}{2}\left(\frac{x_k^j(i)}{\sigma_k^j}\right)^2 - \frac{1}{2}\left(\frac{x_{k-1}^j(i)}{\sigma_{k-1}^j}\right)^2\right].$$

With L level of decomposition, taking into account that total distribution is a product of the distribution of different levels, it can be written:

$$\alpha_k \approx \sum_{j=1}^{L} \alpha_k^j .$$

For each record the segmentation is realised by analysis of $\{\alpha_k\}$. The changes in $\{\alpha_k\}$ correspond to acoustic changes in speech signal.

The algorithm is realised in following steps:

Step 1. After every 10 ms (sampling frequency 8kHz) a segment with a duration of 32 ms is formed.

Step 2. For each segment DWT is performed with 5 level of decomposition.

Step 3. For each level of wavelet decomposition, in order to accomplish additive noise reduction, we assume that the energy density function of the real signal is equal to the energy density of clean signal plus energy density of noise. Modification of spectral subtraction technique for every level of wavelet decomposition is applied,

$$\hat{\mathbf{x}}_k^j = \begin{cases} \left|\mathbf{x}_k^j\right|^2 - \alpha\left|\xi_k^j\right|^2, if \left|\mathbf{x}_k^j\right|^2 - \left|\xi_k^j\right|^2 > \beta\left|\mathbf{x}_k^j\right|^2, \\ \beta\left|\mathbf{x}_k^j\right|^2, otherwise. \end{cases}$$

where:

$\{\xi_k^j\}$ are wavelet coefficients of estimated noise signal in segment *k*, level *j*;

$\alpha = 1.2$ is an oversubtraction factor for the noise background;

$\beta = 0.01$ sets an energy flooring.

The estimation of noise level in wavelet coefficients can be driven from following recursive expression,

$$\xi_k^j = \frac{1}{M^j} \sum_{i=0}^{M^j-1}\left[\gamma \xi_{k-1}^j(i) + (1-\gamma)\, min\left(\mathbf{x}_k^j(i), \tilde{\mathbf{x}}_k^j\right)\right],$$

where:

$\tilde{\mathbf{x}}_k^j$ is mean value of wavelet coefficients' energy valleys at level *j*;

M^j is the mumber of wavelet coefficients at level *j*;

$\gamma = 0.986$.

Step 4. To reduce the "musical" and transient noise, filtering of signal energy for each decomposition level is realised. Taking into consideration that in practice noise is concentrated in particular frequency and time range, weighted median filtering (WMF) is used. When the filter response in the moment *n* is formed, the signal energy in the moments *n*-2, *n*-2, *n*, *n*+1, *n*+2 and *n*+3 are used,

$$\hat{y}_k^j(n) = med\left[w_1\lozenge\, y_k^j(n-2), w_2\lozenge\, y_k^j(n-1), ,\ldots, w_6\lozenge\, y_k^j(n+3)\right],$$

where operator $\lozenge$ means repetition, and $y_k^j = \hat{x}_k^j$. The initial value of weighted vector is $W = [2, 3, w3, 3, 2, 2]$. To set some significance to the current component of feature vector, the weighted coefficient *w*3 is computed depending on the value of standard deviation ,

$$a = round\left(y_k^j(n)/\sigma_k^j\right),\ a = a + w_3\, mod\, 2\ ,$$

$$w_3 = \begin{cases} a,\ a > 0, \\ 8 \end{cases}.$$

Median filtering allow reduction of the non-tonal energy spikes. Figure 1 shows a segment containing transient noise. Due to wavelet analysis, the beginning of the noise is localised exactly at levels 1 and 2. Due to weighted median filtering, the amplitude of non-tonal spikes is reduced under the level of standard deviation for the segment as shown in Fig. 2.

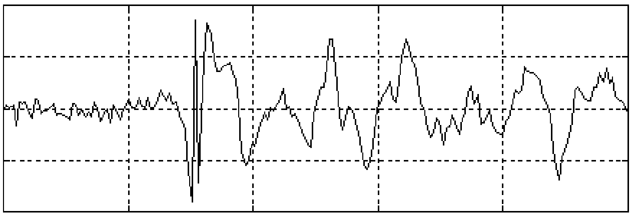

Figure 1: Transient noise (door slam)

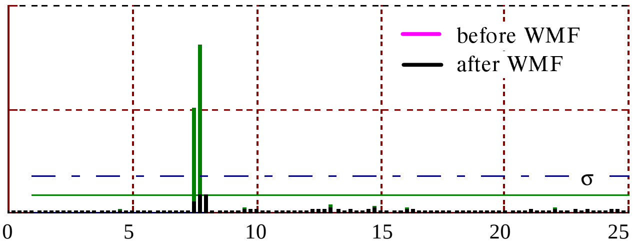

Figure 2: Noise reduction after WMF at level 1

Step 5. Go back to time domain,

$$\hat{x}_k^j = sign\left(x_k^j\right)\sqrt{\hat{y}_k^j}\ .$$

Step 6. For each segment the parameter α_k is obtained,

$$\alpha_k = \sum_{j=1}^{L} ln\left(\sigma_k^j / \sigma_{k-1}^j\right)^2 .$$

Step 7. Normalisation of vector $\{\alpha_k\}$. The vector must be normalised to be level independent. The statistical normalisation, based on mean value and standard deviation, is used instead of the absolute maximum value over frames. This normalisation avoids any errors occurred when very large with respect to all other values in $\{\alpha_k\}$ are present:

$$\hat{\alpha}_k = \left[\alpha_k - \mu(\alpha)\right] / \sigma(\alpha) .$$

Only the positive components of $\{\hat{\alpha}_k\}$ are taken as an end result, negative ones are make zero. A new vector $\{\tilde{\alpha}_k\}$ is generated. It's high-magnitude components correspond to the more significant acoustic changes in the speech signal.

Step 8. Making an end decision. Although the number of non-zero components in $\{\tilde{\alpha}_k\}$ is small, with high levels of noise some low-magnitude peaks appear which leads to the increase the number of segments. This problem can be partially solved in two ways:

- using a threshold value, *Th*;
- merge adjacent segments by clusterization.

With the proposed algorithm a threshold value which is function of noise level is used. It is obtained when the peaks and valleys in $\{\tilde{\alpha}_k\}$ have been found. The value of *Th* is the mean value of the magnitudes of all valleys, $Th = \frac{1}{L}\sum_{k \in V} \tilde{a}_k$, where *V* is the set of indexes corresponding to all energy valleys *L*. When segments are generated, only the peaks in $\{\tilde{\alpha}_k\}$ which have an magnitude higher than the threshold level *Th* are taken into account. With this initial segmentation, significant errors (incorrectly defined boundaries of phonemes or their omitting) are possible only when weak fricative and nasal sounds are present and when SNR<2dB. Because of the high level of the noise an increase of the number of segments are possible. The unification of acoustically similar segments is based on magnitude of peaks and the distance between them.

Since at a high level of noise the start and the end of each word are the most difficult to be obtained, the following correction of their position is implemented: the start is defined by index *i* of peak with a magnitude above *Th*, or by index *j* of the peak with a magnitude not lower than *Th*/2 and abs(*i*-*j*)<8 (80ms). In the same way the end of word is defined.

To find invalid records the following parameters are introduced:

- min_rec_th - minimal duration of record, 250ms;
- min_seg_len - minimal length segment, 30ms.

Records with a duration less than 'min_rec_th' are removed from the base and segments with a length less than 'min_seg_len' are merged with their adjacent segment.

The proposed algorithm can be used not only at the stage of training, but also in testing. Therefor an adaptive estimation of the mean and the standard deviation of $\{\alpha_k\}$ is introduced:

$$\mu(a_k) = \gamma\, \mu(a_{k-1}) + (1-\gamma)\, a_k \ ;$$
$$\sigma(a_k) = \gamma\, \sigma(a_{k-1}) + (1-\gamma)\, abs\left[\mu(a_k) - a_k\right], \text{ where } \gamma = .996.$$

Experimental Results

The proposed algorithm is tested at the stage of training of speech command recognition system with SNR from 25dB to 0dB. The base of records is generated as noise with necessary amplitude is added to the clean records. The disturbance signals are:

- moving train;
- moving car;
- babble noise (people chatting);
- transient noise (doors and windows closing, telephone rings);
- channel noise (radio, telephone).

Figure 3 shows the results after segmentation of the Bulgarian word 'sedem' (seven) at SNR 25dB, 10dB and 2dB (radio channel noise). The word begins with a weak fricative sound 's' and ends with a nasal sound 'm'. The values of vector $\{\tilde{\alpha}_k\}$ which is obtained for the word 'sedem' at SNR=2dB are shown in figure 4.

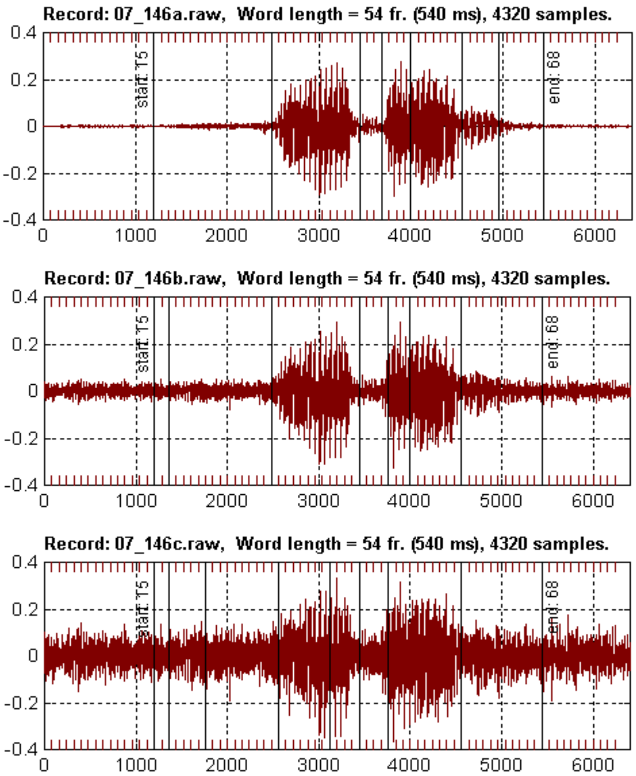

Figure 3: Segmentation of word 'sedem'
(before clusterisation); SNR=25dB, 10dB, 2dB

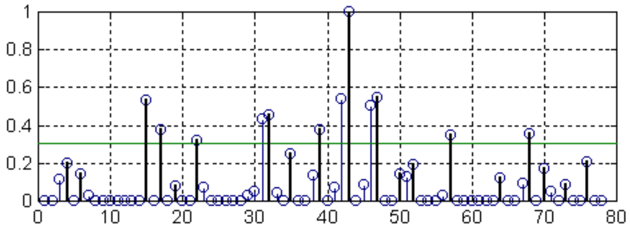

Figure 4: Values of vector $\{\tilde{\alpha}_k\}$ and *Th*, SNR=2dB

The operation of E-ZCR, base MLR and proposed algorithm (endpoint detection) are compared and the results shown in table 1. Classification marks (A=very good - absolute difference distance is up to 4 frames, B=good - from 5 to 9 frames, C=bad - from 10 to 15 frames, D=vary bad - above 15 frames) are obtained after the results of "manual" and automatic segmentation have been compared.

Algorithm	A	B	C	D
E-ZCR	55%	3.6%	35%	6.4%
MLR	58.6%	9.8%	25.4%	6.2%
proposed	72.3%	22%	5.1%	0.6%

Table 1. Results

Conclusion

The proposed algorithm can be used in quality evaluation of speech data base, as stage before multi-conditional training of ASR system. The algorithm is adaptive to noise level and it is independent on parameters defined in advance. It allows reliable segmentation when SNR>0dB.

When SNR<2dB, better results are obtained if vector $\{\alpha_k\}$ is formed not by summation of a_k^j, but by MAX operator of the normalised parameters. However this makes segmentation at low levels of noise worse.

References

Herrera A. et all, ″Speech Detection in high noise conditions″, ICSPAT, 1996, pp.1774-1778.

Hirsch H.G., ″Noise estimation techniques for robust speech recognition″, ICASSP, 1995, pp.153-156.

Korthauer A., ″Robust estimation of the SNR of noisy speech signals for the quality evaluation of speech data bases″, Workshop on Robust Methods for Speech Recognition in Adverse Condition, Tampere, 1999, pp.123-126.

Herrera A., et all, ″An acoustic isolated speech recognition approach using the KLT and VQ″, ICSPAT, 1997, pp.1739-1742.