The Method of an Audio Data Classification and Segmentation

Konstantin Biatov
Fraunhofer Institute for Media Communication
Schloss Birlinghoven , 53754,
Sankt Augustin, Germany
Phone:+49(0)2241-14-1908
biatov@imk.fraunhofer.de

ABSTRACT
In this paper we investigate on-line zero-crossing based audio stream segmentation and classification into speech and other segments. We consider such segments as applause, noise of the auditorium, and silence. We demonstrate that the features extracted from zero-crossing are stable and valid to be used for speech and other signal discrimination and classification and don't require large amount of data for the training. We describe the optimal segmentation of unlimited audio signals in the flight of classification using results of the frames classification based on multivariate Gaussian classifier. We demonstrate that using optimal segmentation is better than using traditional sliding window technique.

Keywords
Audio stream analysis, classification and optimal segmentation.

1. INTRODUCTION
Currently an explosive growth of multimedia data (audio and video) is observed worldwide. The multimedia data include broadcast news, movies, TV programs and many other types of data. One of the greatest interests with respect to these data is to make indexing and information retrieval. The recognition of a large vocabulary continuous speech gives possibilities to transcribe unrestricted audio data. The difficulty is that very often audio data include different kind of non-speech segments such as music, song, noise of auditorium, applause and etc. A quick and reliable segmentation of audio data into homogeneous speech and non-speech segments increase the transcription accuracy and the speed of preparation data for information retrieval. In some application of information retrieval will be interested to find location of non-speech segments in the audio stream. In this paper we suggest and investigate real-time zero-crossing based optimal audio stream segmentation and classification into speech and other segments.

2. FEATURE EXTRACTION
There are several published researches of using zero-crossing for speech processing [1], [2], [3], [4], [5]. In [1] are described possibilities of the spectral analysis based on zero-crossing. In [2] is demonstrated the spectral analyser based on zero-crossing. The formant frequency estimation on the base of zero-crossing is presented in [3]. Experiments on speech recognition when frequency information of the signal is obtained from zero-crossing intervals are described in [4]. We consider that the signal is normalized, i.e. a mean is removed. We define zero-crossing by the time location t_i , $i = 1,2,3\ldots$, as a time when the signal changes the sign or is equal to zero. The successive zero-crossing intervals are defined as $n_i = t_i - t_{i-1}$. As it was mentioned in [3] the successive zero-crossing intervals of sinusoidal signal exhibit high consistency and are inversely related to the frequency. The resulting signal from the sum of more than one sinusoidal component may not satisfy this principle. We'll consider the representation of the audio signal as a sequence of successive zero-crossing intervals and maximal signal amplitudes for such intervals which keep a positive sign and minimal amplitude for the intervals which keep a negative sign. This representation is described in [7] and is the result of resolving the approximation task. This original audio signal is represented as a sequence of pairs (n_i, A_i), $i=1,2,3,\ldots n$, that each pair indexed by odd number i has positive amplitude and indexed by even number i has negative amplitude. The restored signal in each zero-crossing interval is represented as sinusoidal and doesn't depend on the real form of the audio signal. The audio signal restored from this sequence (n_i, A_i) performs good quality. This restored audio signal has the same pitch as the original signal. For the restoring of the original signal we use the following formulas described in [7]. We define as x_n^i a discrete element of the signal from the interval i (where the signal keeps sign):

$$x_n^i = 2*\cos\left(\frac{\pi}{n_i}\right) * x_{n-1}^i - x_{i-2}^n \quad (1)$$

$$x_0^i = (-1)^i * A_i * \sin\left(\frac{\pi}{2n_i}\right) \quad (2)$$

$$x_1^i = (-1)^{i+1} * A_i * \sin\left(\frac{\pi}{2n_i}\right) \quad (3)$$

One interesting characteristic of such signal presentation is that we can generate all possible amplitude modulated signals using

clipped signal (all A_i are 1) and the transformations described in (1), (2), (3). This idea could be applied for search frequences of the amplitude modualtions in the original signal that is important in speech non-speech segmentation task. In Figure 1 we demonstrate one fragment of the original speech and corresponding to it the fragment of restored speech.

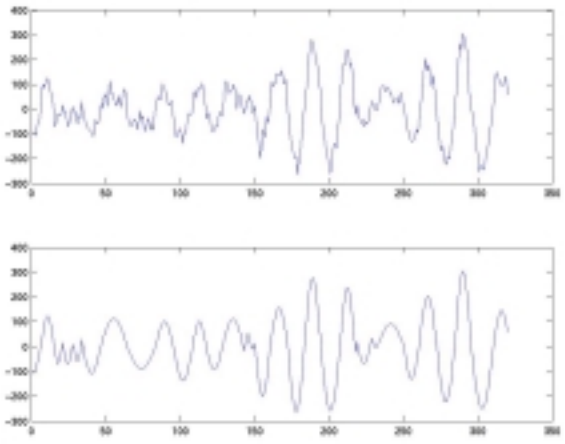

Figure. 1. Comparison of the original signal and recovered signal

The restored signal is not a clipped signal. It keeps exactly the same pitch as the original signal. The quality of the restored signal is good for the discrimination of the speech from other signals and for classification of these signals and the later classifiction stage is carried out on the restored signal.

We make the following procedure for the signal analysis. From each 20 ms frame with 10 ms overlapping we extract the interval between successive zero-crossings. Then the extracted periods are converted into the frequency as $f = 1/x$ where x is the interval between successive zero-crossings. For each interval between successive zero-crossings we extract the amplitude which is maximal for the interval with positive sign of the signal and is minimal for the interval with a negative sign of the signal. For each interval we calculate energy E :

$$\boldsymbol{E = ABS\,(\frac{A_i * n_i}{2})} \qquad (4)$$

Then using 24 frequency bands according to the Bark scale we group energies which correspond to one band together and then normalize all groups using the sum of all energies in the frame. As the result for each 20 ms frame, we have the distribution of energies in accordance with the frequency. Such description keeps important features of the speech signal and is valid to be used for different discrimination tasks. We use this new description to generate new features for the audio signal classification. We consider that following features will be useful. Some of them are well known, some are new and specific for this approach.

1. Average energy in a frame.

2. Percentage of the intervals n_i in a frame with the value less that 50% of average value.

3. Gravity center:

$$CG = \frac{\sum_{i=0}^{23} E_i f_i}{\sum_{i=o}^{23} E_i} \qquad (5)$$

4. The frequency (from Bark scale) below which is 25% of the energy:

$$\sum_{j=0}^{f} E_j > 0.25 \left(\sum_{j=0}^{23} E_j \right) \qquad (6)$$

5. The frequency (from Bark scale) below which is 55% of the energy.

6. The frequency (from Bark scale) below which is 85% of the energy.

7. As it was mentioned the signal is described by the sequence of (n_i, A_i). In accordance with the algorithm all odd amplitudes are positive, all others are negative. We define a derivative of the same sign amplitude as:

$$D_i = 2*abs(\frac{A_{i+2} - A_i}{n_{i+2} + 2*n_{i+1} + n_i}) \qquad (7)$$

As a feature we use average value of D_i for one frame. We consider that this feature is correlated with the amplitude modulation of the initial signal.

8. Homogeneity of the frame:

$$H = \sum_{j=0}^{23} n f_j * (\frac{n f_j}{n_i}) * \log(\frac{n f_j}{n_i}) \qquad (8)$$

where n_i is a number of zero-crossing intervals in a frame and $n f_j$ is a number of intervals that correspond to Bark frequency f_j. Maximal homogeneity is equal to 0, for example for periodical normalized signal *sin(x).*

9. Maximal zero-crossing interval in frame:

$$M = \max_i (n_i) \qquad (9)$$

10. Auto-correlation of the feature vector that describes the frame:

$$R(\tau) = \sum_{j=0}^{23-\tau} E_j * E_{j+\tau} \qquad (10)$$

11. Correlation between the vectors that describes the neighboring frames.

$$R(\tau) = \sum_{j=0}^{23-\tau} E_j^i * E_{j+\tau}^{i+1} \qquad (11)$$

12. Euclidean distance between the vectors that describe the neighboring frames *i* and *i+1*:

$$D_e = \sum_{j=0}^{23} (E_j^i - E_j^{i+1})^2 \qquad (12)$$

For the discrimination and classification we use a multivariate Gaussian classifier. We consider these features as independent and use diagonal covariance matrix. As the result of classification we have labeling of each 20 ms frame by the labels corresponding to the type of considered audio event : a speech, a silence, a noise, an applause.

3. OPTIMAL SEGMENTATION

3.1 Optimal Segmentation Algorithm

In [8] was described algorithm for quasi-linear reduction of the speech signal. We apply this idea for optimal segmentation of the audio signal. As was mentioned in the process of classification we generate the labels for each frame. At the same time with the classification we do optimal segmentation and smoothing the results of classification. Suppose that $L = (l_1, l_2, \ldots, l_i)$ is a current sequence of classified frames. We define the segmentation as a sequence of numbers (segments boundaries) s_j, $j = 1: m$, where $s_m = i$. We also define that $(s_{t+1} - s_t) > LMIN$, where *LMIN* is a minimal duration of the segment, and that $(s_{t+1} - s_t) < LMAX$, where *LMAX* is a maximal duration of the segment. It is natural to consider that in the real audio signal each homogeneous segment has a duration, for example, not less than 0.25 sec. We also would like to define the measure of homogeneity of a segment. Our measure of homogeneity is based on the entropy. If the segment started in the moment i_s and finished in the moment i_e includes *N* labels and each label l_i is repeated m_i times, the homogeneity of this segment will be:

$$H(s, e) = \sum_{i=s}^{e} m_i * (\frac{m_i}{N}) * \log(\frac{m_i}{N}) \qquad (13)$$

When all labels in one segment are equal, the entropy for this segment is equal to 0. In other cases, the value of the entropy is less than 0.

The task of the optimal segmentation is to find such boundaries of the segments (in accordance with the duration limitation) so that the criterion of the optimization (homogeneity) will be maximal. We use the dynamic programming to find optimal boundaries. Formally, the criterion of optimization is formulated as:

$$G(q, s_1, s_2 \ldots, s_q) = max \sum_{t=1}^{q-1} H(s_t, s_{t+1}) \qquad (14)$$

Current optimal value of G_i is calculated as

$$z^* = argmax(G_{i-z} + H(i-z, i)),$$

$$LMIN < z < LMAX$$

$$G_i = G(i-z^*) + H(i-z^*, i) \qquad (15)$$

Suppose that *q(i)* is a current number of segments. The value of *q(i)* can be calculated as *q(i) = q(i-z*) + 1*. Suppose that z(i) is a current optimal length of the segment. For *z(i)* we can write that *z(i) = z**. The optimal boundaries can be defined as

$$s_{j-1} = s_j - z(j), j = 2, \ldots, q \qquad (16)$$

For each homogeneous segment we apply a voting rule to select one the most probable label. At the final stage, the found homogeneous segments are united with the neighboring segments having the same homogeneity label . The optimal segmentation also was extended into a case when the limitation for the duration is defined individually for each class of a segment. For example, we can define that a minimal duration for a speech segment is 1 sec, for silence 0.25, for noise of the auditorium 0.25 sec. and for applause 0.5 sec.

The described optimal segmentation is better than, for example, a smoothing window which smoothes labels only in the fixed size window and doesn't take into account other labels outside this window.

3.2 Experimental Results

We conducted a series of experiments for the speech/non-speech discrimination and non-speech signals classification. We have considered such non-speech data as applause, noise of the auditorium, and silence. We use audio signals sampled at 16 kHz. For training and testing were used audio data from parliamentary speeches. These data include speech signals, applause sounds, noise of the auditorium and silence. The audio data were collected from two different acoustic sources. For training limited amount of audio data were used. In the Table 1 is presented the size of the training data.

Table 1. The size of the training data

	Applause	**Speech**	**Noise**	**Silence**
Size in frames	986	533	9709	520

In the Table 2 are presented some preliminary results of the classification. The accuracy is evaluated using the frame level.

Table 2. The results of the testing (frame level)

Recongnized as	**Noise**	**Speech**	**Applause**	**Silence**
Noise	50508 87.8%	6828 11.9%	48	119 0.2%
Speech	1947 9.4%	15724 76.6%	1080 5.2%	1752 8.5%
Applause	821 7%	812 6.9%	9951 85.2%	91 0.7%
Silence	552 5.3%	111 0.1%	10	9626 93.4%

Some experiments were conducted to compare suggested optimal segmentation with the smoothing by sliding window. In the experiments the value of minimal duration LMIN was 0.25 second

and the value of maximal duration LMAX was 5 second. The size of sliding window was 0.5 second. For the segmentation test was used 5 minutes audio segment from the parliamentary speeches. In this audio data were presented all type of audio events in approximately equal proportion. The results of segmentation were evaluated with the tolerance 0.2 second. The results are presented in the Table 3.

Table 3. The results of comparison the local optimal segmentation and the segmentation using sliding window

	Number of segments (correct boundaries)	**Total number of segments**
Optimal segmentation	100 (**75.7**%)	132
Sliding window	90 (**71.4**%)	126

We have compared the results of recognition of the applause and the speech segments using optimal segmentation and sliding window. The minimal duration for the speech and the applause segments using optimal segmentation was 0.25 sec. The size of sliding window was 0.5 sec. The results of comparison are presented in the Table 4. These preliminary results show that using optimal segmentatation gives improvement in comparison with traditional sliding window technique.

Table 4. The results of recognition using local optimal segmentation and using sliding window

	Applause segments recognition	**Speech segments recognition**
Optimal segmentation	**95.1%**	**96.7%**
Sliding window	**92.%**	**95.5%**

4. CONCLUSIONS

The conducted experiments show that optimal segmentation gives better results. The described methodology is implemented in the system that do classification and local optimal segmentation of audio data and work in the flight with the unlimited audio signal.

5. ACKNOWLEDGMENTS

This work was funded by the German Federal Ministry for Research and Education and done in the frame of the official project AGMA.

6. REFERENCES

[1] B.Kedem: Spectral analysis and discrimination by zero-crossing, Proc. of IEEE, Vol. 74, No.11, pp.1477-1493, 1986.

[2] S.Kay: A zero-crossing based spectrum analyzer, IEEE Transaction on Acoustic, Speech and Signal Processing, Vol. 34, No.1, pp.96-104, 1986.

[3] T.Sreenivas, R.Niederjohn: Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise, IEEE Transaction on Signal Processing, Vol. 40, No.2, pp.282-293, 1992.

[4] D.S.Kim, S.Y.Lee, R.M.Kil: Auditory processing of speech signal for robust speech recognition in real-world noise environment, IEEE Transaction on Speech and Audio Processing, Vol. 7, No.1, pp.59-69, 1999.

[5] R.De Mori, L.Moisa, R.Gemello, F.Mana, D.Albesano: Augmenting standard speech recognition features with energy gravity centres, Computer Speech and Language 15, pp.341-354, 2001.

[6] J.Saunders: Real-time Discrimination of Broadcast Speech/Music, Proc. ICASSP, pp.993-996, 1996.

[7] T.Vintsiuk, A.Kulias, A.Dys: A economic presentation of acoustic signal in computer, Proc. ARSO-12, Institute of Cybernetics of AS USSR, Kiev, pp.74-75, 1982.

[8] E.K.Ludovik: The algorithm for optimal quasi-linear reduction of the speech signals, Proc. ARSO-12, Institute of Cybernetics of AS USSR, Kiev, pp.114-116, 1982.

[9] K.Biatov, M.Larson and S.Eickeler: Zero-Crossing-based Temporal Segmentation and Classification of Audio Signals, Proc. of the Sixth All-Ukrainian International Conference on Signal/Image Processing and Pattern Recognition UkrObraz'2002, Kiev, pp.71-74, 2002.

[10] K.Biatov: Classification and Optimal Segmentation of Acoustic Signals into Homogeneous Segments, Proc. of Speech Processing Workshop, Otto-von-Guericke University, Magdeburg, 2003.