An Extraction of Speech Data from Audio Stream Using Unsupervised Pre-Segmentation

K. Biatov

Fraunhofer Institute for Media Communication,
Sankt Augustin, Germany
biatov@imk.fraunhofer.de

Abstract

In this paper we investigate an extraction of speech data from audio stream. Our method includes unsupervised optimal self-segmentation of the audio stream into small, homogeneous segments. The homogeneity is defined on a base of the average amplitude and a zero-crossing in a frame. A measure of the homogeneity is entropy. In our approach we calculate a relative ratio between the average amplitudes of the neighboring homogeneous segments. For a speech signal this ratio is less than a threshold defined on a short pure speech signal. As a discriminative feature we use a percent of the homogeneous segments within 1 sec interval having high relative amplitude ratio. In the process of the classification each 1 sec is labeled incrementally as a speech or a non-speech segment. The discrimination technique shows high performance for more than six-hour data that include different types of audio.

1. Introduction

In the last decade many papers describing speech/music, speech/singing, speech/environmental sounds discrimination and audio data segmentation [1], [2], [3], [4] were published.
For the classification were used both statistical characteristics and structure patterns of the audio data that describe the structural features for spectral and temporal representation [4]. Below is the list of main features used for speech and other sound discrimination.

Zero-crossing and variance of zero-crossing.
Spectral centroid and variance of spectral centroid.
Rolloff point.
RMS (root mean square).
Entropy and dynamism.
Micro and macro modulation.
Rhythm.
Harmonic coefficients.
MFCC coefficients.
High zero-crossing rate ratio.
Low short-time energy ratio.
Spectral flux and variance of spectral flux.
Band periodicity.
Speech/silence ratio.

For the classification, different models were also used including GMM (Gaussian Mixture Model), KNN (K-Nearest Neighbors), HMM, ANN (Artificial Neural Networks), SVM, PCA (Principal Component Analysis).

In [3] a simple and effective way for speech/music discrimination using speech/silence ratio and variation of zero-crossing rate was described. Speech signals have more higher silence ratio than music. The most important task is how to calculate silence ratio and what is a measure of the silence ratio to separate speech and non-speech.
In [3] the method for the silence ratio calculation was suggested. At first authors find the amplitude threshold below which a sample is considered being silent. A silent period is found when the number of consecutive silent samples is larger than time threshold. The silence ratio is the number of silent samples divided by the total number of samples. For such kind of classification three thresholds were used: the amplitude threshold, the time threshold and the silence ratio threshold. All these thresholds are obtained experimentally from training audio database.

In this paper we suggest another method to calculate the silence ratio. In the contrast to the mentioned approach, we exploit more structural patterns than statistical characteristics of audio features. Our method doesn’t require three thresholds obtained from the training data. It uses two thresholds motivated by the specific of human speech.
We suggest doing non-linear self-segmentation of the audio signal into homogeneous segments with the high and the low amplitudes that fit the speech audio signal. We use two features for this segmentation – average amplitude and average zero-crossing rate in a frame. A measure of the homogeneity is an entropy and the result of the segmentation is a sequence of small segments having the high or the low amplitude. The detailed description of the unsupervised segmentation is presented below.

2. Unsupervised segmentation of audio data

In [5] was described algorithm for quasi-linear reduction of the speech signal. We apply this idea for optimal segmentation of the audio signals. Suppose that $L = (l_1, l_2, \ldots, l_i)$ is a current sequence of frames. We define the segmentation as a sequence of numbers (segment boundaries) s_j, *j = 1: m*, where $s_m = i$. The number of the segments is the result of the segmentation. 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. The minimal duration of the segment is 50 ms and the maximal duration is 1 sec. According to the syllabic rate 3-20 Hz it is natural to consider that in speech signals each homogeneous segment has the duration, no less than 50 ms. We also define the measure of homogeneity of the segment. Our measure of homogeneity is based on the entropy. An average amplitude and a zero-crossing rate in 10 ms frame are used as the features. We suppose that these two features have Gaussian distribution. For the segment that is started at the moment i_s and is finished at the moment i_e we calculate a mean and a variance of the average amplitude and the zero-crossing rate in each frame of this segment. A multivariate Gaussian density with diagonal covariance matrix is used to calculate the probability of these features in the frame. The entropy as the measure of homogeneity of the segment is calculated as:

$$H(s, e) = -\sum_{i=s}^{e} p_i \log p_i , \qquad (1)$$

p_i is the probability value in the frame *i.* In this approach, the smaller the entropy is, the more homogeneous the corresponding segment is. 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 will be minimal. So, we are looking for such segments boundaries that the entropy (homogeneity) will be minimal. We use the dynamic programming to find optimal boundaries. Formally, the criterion of optimization is formulated as:

$$G(q, s_1, s_2 \ldots, s_q) = \min \sum_{t=1}^{q-1} H(s_t, s_{t+1}) \qquad (2)$$

Current optimal value of G_i is calculated as

$$z^* = \mathit{argmin}(G_{i-z} + H(i-z,i)), \qquad (3)$$

$$LMIN < z < LMAX$$
$$G_i = G(i-z^*) + H(i-z^*,i)$$

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

$$s_{j-1} = s_j - z(j), \; j = 2,\ldots, q. \qquad (4)$$

Used algorithm is the algorithm of self-segmentation. It does not require a definition of the number of the segments. The dynamic programming segmentation gives automatically the number of the segments. The values *m* and *q* are defined automatically during the process of backtracking. The maximal values of *m* and *q* will be in the cases when the size of each segment is equal to LMIN (50 ms).
Now each homogeneous segment can be described by the mean and the variance of used features. There are the average amplitude and the zero-crossing rate, calculated within this segment. The durations of optimal segments for the speech data are not equal. The example of optimal segment boundaries for the speech signal is presented in Figure 1.

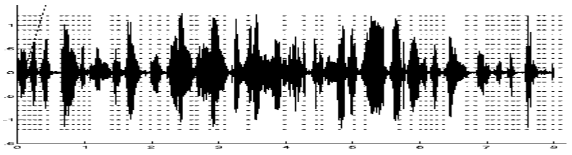


Figure 1: The results of unsupervised segmentation of the speech data.

The experiments with unsupervised segmentation show that in the most cases we can find the boundaries that correspond to rapid changes of the amplitude in the speech signal. Below we describe how the features that correspond to the optimal segments are used for speech/non-speech classification.

3. Feature extraction

For each optimal segment that includes at least 50 ms, we calculate the average amplitude, the variance of the amplitude, the average zero-crossing rate and the variance of zero-crossing rate. We use these four primary features to get the new structural features. Within 1 sec we calculate ratios between corresponding features for each of the two neighboring segments. When the ratio is larger than 1 instead of *R* we use reverse value of *R..* We calculate all ratios successively in each 1 sec interval of audio signal. These ratios describe the dynamic of the average amplitudes, zero-crossings and their variances between neighboring segments. We suppose that these ratios have Gaussian distribution and for each of these new features we calculate the mean and the variance.
In Figure 2 the similarity matrix between Gaussian densities of ratios and corresponding to it audio signal are presented. The similarity is calculated for each pairs of 1 sec intervals of the audio signal. For the similarity the Kullback Leibler (KL) distance between the N-dimensional normal densities is used. The KL distance is not symmetric and we construct resulting distance from two KL distance as was described in [6]. This audio signal mostly includes speech data and also includes music, music with the speech, artificial sounds, singing and silence.

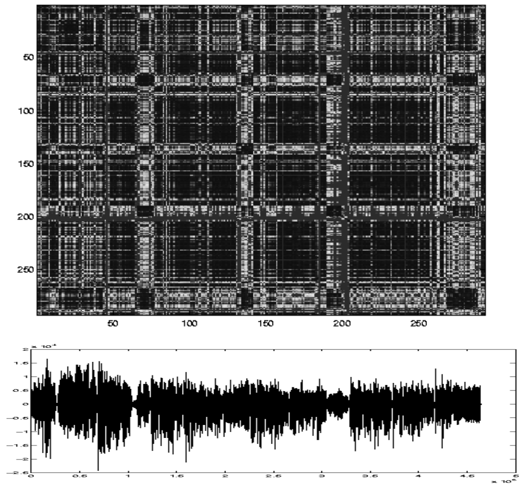
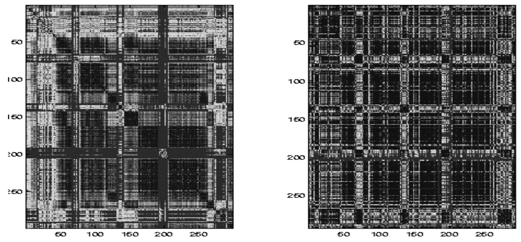


Figure 2. An example of the similarity matrix and corresponding audio signal.

The dark regions correspond to the high similarity and the bright regions correspond to the low similarity. The presented similarity matrix demonstrates high separation between different audio segments, in particular between music, speech, singing and artificial sounds.

For comparison purpose the similarity matrix using MFCC coefficient was computed. For the same audio signal for each 1 sec the mean and the variance for 12 MFCC coefficients were calculated. This similarity matrix was computed in the similar way as the matrix for the features based on the ratios.

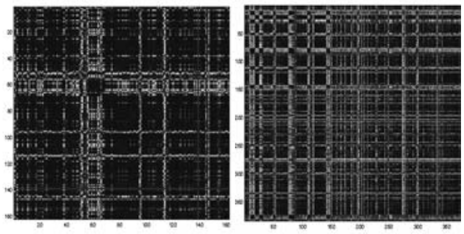
In Figure 3 is presented the comparison between similarity matrix calculated using suggested features and the similarity matrix calculated using 12 MFCC coefficients.



(a) (b)

Figure 3. Comparison between similarity matrix computed using 12 MFCC coefficients (a) and similarity matrix computed using ratios (b).

Black squares on the diagonal correspond to the high self-similarity. The larger black squares are the larger are the regions of the similar data. The segments boundaries of the signal correspond to the boundaries of the black squares on the diagonal. The black squares on the periphery of the diagonal correspond to the similar non-neighboring segments. The sharp changes in the color of the squares corresponding to the different kinds of data. It shows the ability of the features to separate the changes in the content. The matrix of the similarity based on the ratios has good comparability with the matrix of the similarity based on MFCC coefficients. We use only four features and sharpness of the matrix in particular in the boundaries of the speech and non-speech data shows that the suggested features fit for speech and non-speech separation. In Figure 4 are presented the similarity matrixes for jazz and for parliamentary speech. Jazz includes short segment of the vocal. The bright regions on the periphery of diagonal on the picture presented jazz data show high separation between vocal and jazz. On the same picture the black squares on the diagonal show good self-similarity both for vocal and for jazz. In the left part of Figure 4 a short audio fragment of the parliamentary speech is presented. This fragment includes regions of applause and silence. The applause and the silence don't have high similarity with the speech. They are presented by the bright regions on the periphery of the main diagonal. On the periphery of the main diagonal are many black squares. This shows high similarity between non-neighboring speech segments. Mostly on the diagonal are black squares with the clear boundaries. This shows high self-similarity of the different kind of audio data.



(a) (b)

Figure 4. The examples of the similarity matrixes for different kinds of signals: (a) jazz and vocal, (b) parliamentary speech.

In the next section we describe how the features based on ratios are used for speech non-speech classification.

4. Classification

We do classification between speech and non-speech data that include different kinds of music (instrumental, jazz, and orchestra), singing with music background, environmental sounds, sound effects and a silence. At the first step each 1 sec interval is segmented into homogeneous segments, using described segmentation algorithm. The result of segmentation is a sequence of small homogeneous segments. Then for each two neighbouring segments the ratio of the average amplitudes is calculated. When this ratio is more than 1 it is replaced by the reverse value. We calculate the number of the cases when this ratio is less than a threshold. The threshold is selected using short speech signal. This threshold describes the rapid changes of the amplitude for neighbouring segments in speech data. We have found that for wide class of speech data this threshold is 0.4. Then the number of cases when ratio is less than the threshold is normalized using the number of segments in 1 sec interval. When the normalized number is more than the second threshold we make decision that this 1 sec interval is speech, otherwise it is a non-speech. The selection of the second threshold is based on the features specific to the speech. Normally for the speech data this value is no less than 0.2. This mean that in 1 sec interval the speech has at least in 20% cases rapid amplitude changes between the neighbouring segments.

As an alternative technique we examinate multivariate Gaussian classification. For each class of the data, speech and non-speech data, we estimate means and covariance in a supervised training phase. The duration of the training data is approximately 5 minute for both classes. The estimated parameters are used to classify the incoming new audio data. We estimate parameters for two kind of the features. The first kind of the features is four features, based on the ratios described in the section 3.

The second kind of the features is 39 MFCC that are state-of-art features for speech recognition, classification and segmentation. We exploit these 39 MFCC features for the baseline test using multivariate Gaussian classification.

5. Experiments

We have conducted series of experiments for speech/non-speech discrimination. In the Table 1 are presented the descriptions of the audio data. In the Table 2 are presented the results of the classification using described rule, using suggested 4 features and using as a baseline 39 MFCC features. The accuracy is evaluated with the tolerance 1 sec.

Table 1. Description of the audio data.

Audio data description	Duration in sec.	Type of data
Daily radio news, mixed speech/non-speech, speech with the music background	**speech-6214** **non-speech -572**	**1**
Parliamentary speeches, mixed speech/non-speech, speech with the applause background	**speech-7250** **non-speech-640**	**2**
Instrumental music from CD, homogeneous data	**2106**	**3**
Orchestra from CD, homogeneous data	**1800**	**4**
Jazz and jazz vocal from CD, homogeneous data	**3792**	**5**
Classical singing with music from CD, homogeneous data	**1341**	**6**

Table 2. The results of the testing (1 sec tolerance).

Type of audio data	Accuracy (rule based)	Accuracy (features based on ratios)	Accuracy (39 MFCC features)
1	**82.1%**	**92.1%**	**91.6%**
2	**85.9%**	**90.5%**	**80.8%**
3	**96.3%**	**97.5%**	**97.9%**
4	**97.4%**	**97.4%**	**95.6%**
5	**87.9%**	**82.6%**	**97.3%**
6	**94.3%**	**93.6%**	**96.7%**

6. Smoothing

For post-processing smoothing we use the results of the unsupervised segmentation via Bayes Information Criterion (BIC) [7]. BIC is used for dissimilarity measurement between two adjacent windows and is based on parametric statistical models that correspond to their windows. After BIC segmentation within each segment we apply the voting rule to find the most dominant label. The most dominant label becomes as the label of the whole segment. The results of the smoothing are presented in the Table 3.

Table 3. The results of the smoothing.

Type of audio data	Accuracy (rule based)	Accuracy (features based on ratios)	Accuracy (39 mfcc base features)
2	**97.6%**	**97.6%**	**97.6%**

7. Conclusions

The conducted experiments show that suggested technique is effective for speech/non-speech classification. New technique can be used separately and can also be combined with existing approaches. It is really hard to compare our results with the published before because the used audio data are different. The comparison with the results published in [1], [2], [3] shows that our results are comparable with the published before, but are achieved by using comparatively more simple technique. Our results also overperform the results of the prototyping system [3]. The suggested technique doesn’t require large training data. It works with the different types of the speech and the non-speech audio data.

8. References

[1] Scheirer, E. and Slaney, M., “Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator,” in *Proc. ICASSP*, Munich, Germany, vol.2, pp. 1331-1334, 1997.

[2] Carey, M., Parris, E. and Lloyd-Thomas, H., “A Comparison of Features for Speech, Music Discrimination,” in *Proc. ICASSP*, Phoenix, USA, pp. 149-152, 1999.

[3] Lu, G. and Hankinson, T., “An Investigation of Automatic Audio Classification and Segmentation,” in *Proc. ICSLP 2000*, Beijing, China, 776-781, 2000.

[4] R. Cai, L. Lu, H.-J. Zhang and L.-H. Cai, “Using Structure Patterns of Temporal and Spectral Feature in Audio Similarity Measure,” in *Proc. ACM Multimedia 2003*, Berkeley, USA, Nov. 2003, pp. 219-222.

[5] Ludovik E., ”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.

[6] Foote J. and M. Cooper M., “Media Segmentation using Self-Similarity Decomposition”, *Proceedings of SPIE* Volume 5021. Storage and Retrieval for Media Databases 2003. Jan. 2003, vol. 5021, NO. 5021, 506 pages.

[7] Chen S. and Gopalkrishman P., “Speaker, Environment and Channel Change Detection and Clustering via the Bayes Information Criterion,” in *Proc. Broadcast News Transcription and Understanding Workshop*, pp. 127-132. 1998.