A Way of Segmentation of Speech Recorded Simultaneously by Two Microphones

Petra Fischerová, Vlasta Radová
University of West Bohemia, Department of Cybernetics
Pilsen, Czech Republic

Abstract

This paper deals with a problem of segmentation of speech that was recorded by two microphones simultaneously. We suppose that we have a stereo record of an interview between two speakers. The interview is recorded in such a way that each speaker has his/her own microphone and the signal from each microphone is stored in one channel of a stereo signal. Although each microphone receives speech from the both speakers, the speech received by the microphone closer to the respective speaker is much cleaner and much more useful e.g. for speech recognition than the speech received by the farther microphone. Therefore our aim is to segment the stereo signal with respect to which speaker is just speaking. For a solution of this aim we propose two novel highly efficient methods that we call *CAMP* (Compare AMPlitude method) and *RESPEC* (REsiduum of SPECtrum).

1. Introduction

Two methods described in this paper deal with the type of microphone segmentation in which no prior knowledge of the identities of speakers is available. Everything what the algorithm needs for microphone segmentation is found in the audio signal, the algorithm does not need any other information. Successful solution of the task of microphone segmentation has a large number of possible applications, e.g. mainly as an improvement of speech recognition task. It can be also used for automatic transcription or indexing according to a speaker. A larger distance between the speaker and the microphone causes that the speech is influenced also by its environment. Nowadays it is an often task to re-write automatically reports from various meetings or interviews that are usually recorded by more than one microphone. In such a case the correct determination which of the microphones is the closest one to the just speaking person improves markedly recognition of speech. The microphone segmentation enables us to choose the closest microphone.

The problem of the microphone segmentation is illustrated in Fig. 1. In the ideal case each microphone contains clean speech from one speaker. Greater distances between each microphone and a particular speaker result in amplitude variations of the speech signal in the microphones and in influencing the speech by the surrounding environment.

After analyzing several interviews, we have found two important fundamental attributes that allowed us to solve the problem. In our data, there are two types of speech records:

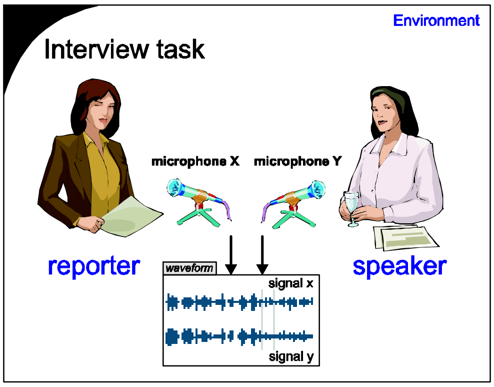

Fig. 1 Speech recorded simultaneously by two microphones.

- The speech in the microphone closer to the actual speaker has a dominant strengthening in comparison to the farther microphone (there is a markedly higher amplitude of the signal in the closer microphone).
- There is almost the same strengthening in both channels, but the speech in the farther microphone is influenced by the surrounding environment (the convolution of the speech and the environment is much stronger).

These two observations lead to a formation of two newly proposed methods.

The organization of the paper is as follows: First, in Sections 2 and 3 our new methods are explained. In Section 4 experiments are described and achieved results are presented. Finally, in Section 5 some conclusions are given.

2. CAMP Method

The *CAMP* method (*C*ompare *AMP*litude method) is a method that is suitable for the records in which there is a much greater amplitude in the microphone closer to the talking speaker than in the microphone that is farther from this speaker. The *CAMP* method is based on different intensities of the signals from particular microphones measured in the frequency domain.

2.1. Description of the method

- The record is split into small windows (the length of each window is 120 ms). A frequency spectrum is computed for each window by the application of the Fast Fourier Transform (FFT) [1]. The whole intensity of the spectrum is computed in the frequency band 120 – 1500 Hz which is the dominant band of the voiced speech, thereby unwanted noise at the lower and higher frequencies is eliminated. The window is shifted along the whole speech signal and the intensity is computed for all the windows.
- This process is done for both channels of the stereo record. The intensity of the first channel is marked as e_x and the intensity of the second channel as e_y. Now we can compute the difference between both intensities

$$residuum = e_x - e_y \tag{1.1}$$

  An example of the function of residuum is depicted in Fig. 2. The positive values of the residuum stand for the higher amplitude of the first channel, whereas the negative values of the residuum stand for the higher amplitude of the second channel. The vertical black lines in Fig. 2 represent the wanted positions of the changes of the residuum.

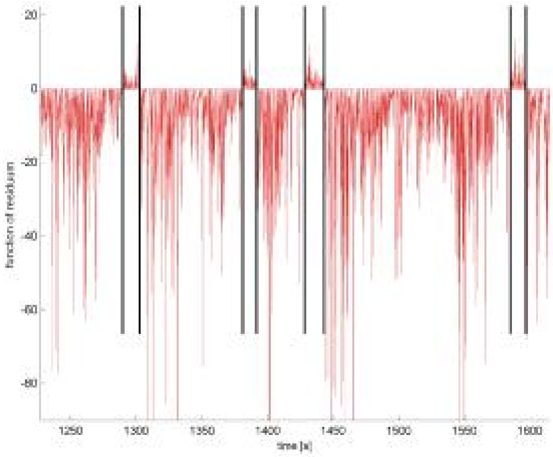

Fig. 2. An example of a function of residuum in the CAMP method.

- Each sample of the function of residuum obtained in the previous step is compared with a priory determined values SZONE and –SZONE according to the following rules:
  - If the value of the residuum is between the values *SZONE* and *–SZONE* (i.e. the value of the residuum is in so-called low energy band), then it is marked by the value –1. This situation occurs e.g. for silence or noise.
  - If the value of the residuum is higher than *SZONE*, then it is marked by the value 1 (which means the first channel).
  - If the value of the residuum is lower than *-SZONE*, then it is marked by the value 2 (which means the second channel).

  The value of the parameter *SZONE* was set to 1 in our experiments.
- The samples belonging according to the previous step into the low energy band are assigned to the first or to the second channel in this way: the values –1 are replaced by a rounded average of *n* previous values. After this step the function of residuum has only two stages, values 1 and 2. In our experiments the parameter *n* was set to 3.

- The short segments (in the length 1 or 2 samples) belonging to one channel but surrounded by another channel are replaced according to their neighbors.
  For example the function of residuum

  1 1 1 2 2 1 1 1 1 2 1 1 1 1 1 2 2 2 2 1 2 2 2 1 1 2 2 2 1 2 2 2

  is changed to

  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2.

- The positions of the changes between the microphones on the function of residuum are detected and converted into time values.

3. RESPEC Method

The *RESPEC* method is suitable for records where there is almost the same amplitude in both channels or one channel has a dominant amplitude all over the record. The method is based on the fact that the signal from the farther microphone is markedly influenced also by the surrounding environment. It means that the signal from the farther microphone is a convolution of the speech and the environment. The process of obtaining of the environment from the convolution is illustrated in Fig. 3.

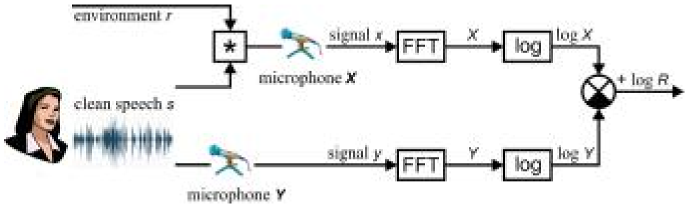

Fig.3. An illustration how the environment is obtained in the *RESPEC method*.

It holds that $x = s * r$ and $y = s$. We suppose that only the signal obtained from the farther microphone is influenced by the environment. In real situations it is not fully true because the signal from the closer microphone is also influenced by the environment. However this influence is not so significant and noticeable as in the farther microphone therefore we can ignore it. The procedure illustrated in Fig. 3 can be then mathematically described as

$$\left.\begin{aligned} x = s * r \xrightarrow{FFT} S \cdot R \xrightarrow{\log} \log S + \log R \\ y = s \xrightarrow{FFT} S \xrightarrow{\log} \log S \end{aligned}\right\} \Rightarrow \log S + \log R - \log S = + \log R \qquad (1.2)$$

3.1 Description of the method

- Similarly as in the *CAMP* method, we need information about the low energy band in the record. Therefore the beginning of the *RESPEC* method is the same as in the *CAMP* method up to the comparison of the samples of the function of residuum with the values *SZONE* and -*SZONE*. Then we create a vector of indexes of samples belonging to the low energy band in the whole record.
- A logarithm of the frequency spectrum for each window is computed for both channels and a residuum of spectrum is determined according to the formula

  $$specResiduum = \log F_x - \log F_y, \qquad (1.3)$$

  where F_x denotes the frequency spectrum of the first channel and F_y denotes the frequency spectrum of the second channel.
- The residuum of spectrum is represented by complex numbers. For our purpose, we use only the real part of the complex numbers. Another possibility is to use the absolute values of the complex numbers or the complex conjugate numbers, however results of our experiments showed that in such cases the results are much worse. The real parts of all *specResiduum* values in one window are added together. This is done for each window and as a result we obtain a

time function of spectrum residuum. An example of such a function is depicted in Fig. 4. The vertical black lines represent the wanted positions of the changes between microphones.

- The function of spectrum residuum is filtered by an MA (Moving Average) filter using 3 samples. The filter smoothes short peaks of the function of spectrum residuum.
- The smoothed function not containing samples belonging to the low energy band is used as a training set for two normal probability density functions. The *EM* algorithm [2] is used for the training of a two-mixtured GMM. Initialization values for the *EM* algorithm are computed by the *K*-means algorithm [3]. Each of the resultant normal density functions represents a model of one channel and is used for classification of individual samples of the smoothed function of spectrum residuum into two classes (two channels). The classification is performed using a Bayes classifier [2], [4]. However, we still do not know which class represents which channel.

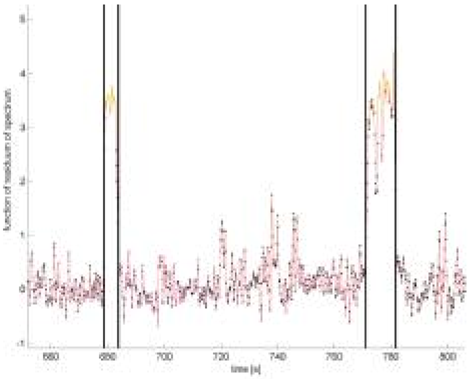

Fig.4. An example of a function of spectrum residuum in the *RESPEC* method.

- An identification of so-called evident samples of the spectrum residuum function follows. The evident sample is each sample which is located in the outer side from the means of the models. These samples cannot high probably belong to the other class when they are tilted into the opposite area.
- These samples are exactly sorted into the proper nearest class. The process of the identification of the evident samples is shown in Fig.5. We suppose that the values of the mean are m_1 and m_2, m_1 is higher than m_2. Each sample, whose value is higher than m_1, belongs evidently to the first class represented by the mean m_1. Each sample, whose value is lower than m_2, belongs evidently to the second class.

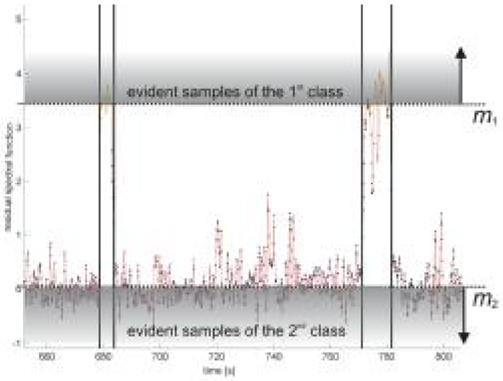

Fig.5. Classification of the evident samples

- The substitution function is created by two previous steps. The values of the substitution function of the samples, which belong to the first class, are marked by 1 and the values of the second class are marked by 2. The samples belonging to the low energy band are still marked by the value –1.
- The short segments (in the length of 1 or 2 samples) belonging to a channel that are surrounded by another channel are replaced according to their neighbors. It works as there are no values –1, they are ignored but their indexes are still kept in mind.
- The times of the changes of the microphones are determined now. The substitution function still includes the values 1, 2 and –1. For the segmentation, the values –1 are ignored but their indexes are still remembered. When the microphone change should be in a low energy band, then it is located into the middle of this interval of low energy. The process is illustrated in Fig. 6 where the third change is placed in the middle of the low energy band represented by the value –1.

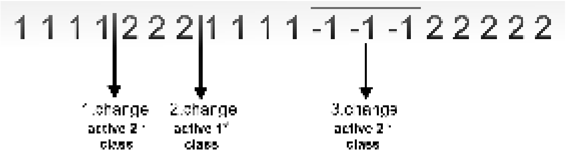

Fig.6. An illustration of the detection of microphone changes.

- Now we need to decide which class represents the first channel and which class corresponds to the second channel. The spectrum residuum function was computed as log F_x – log F_y in (1.3). The terms log F_x and log F_y can be substituted by the means of the trained models. The class that belongs to the model with the higher mean is the first channel (signal x). Thanks to this rule the found classes are assigned to the correct channels of the stereo record.

4. Experiments

Both methods described in this paper were tested in several experiments. The interview test set consisted of 16 records containing speech of a reporter and a reported person. The reported people spoke about their lives. The length of each record was about 30 minutes.

There were two types of interviews. The first group was characteristic by long parts where only the reported person spoke. The reporter spoke very rarely, usually only one or two words. Each record contained about 13 speaker changes on average.

The second group contained conversation based on questions or comments of the reporter. Some comments were longer, some were very short. Each conversation record contained about 100 speaker changes on average.

The percentage of correctly detected changes is summarized in Table 1.

Table 1. Results of the *CAMP* and the *RESPEC* methods.

	CAMP	**RESPEC**
	Accuracy [%]	Accuracy [%]
16227_03	100	93
17154_05	100	75
17249_04	67	100
13861_03	91	91
15060_04	100	82
15118_03	86	86
16373_04	99	98
21607_03	98	99
21633_05	96	97
21705_05	97	100
19687_04	98	100
22119_03	97	100
22131_03	100	94
23779_04	99	79
23915_03	100	94
23930_03	100	85

The *CAMP* method cannot be used for records in which the amplitude of the signals is almost the same or when the surrounding environment influences the speech in the farther microphone very strongly. In the majority of cases the amplitude is strong in the closer microphone and the *CAMP* method segments the record with high accuracy. The function of residuum is a robust function and the amplitude changes are clearly evident. Conversely the *RESPEC* method gives high-quality results for records, in which the speech in the farther microphone is strongly influenced by the surrounding environment caused by a great distance from the speaker. The method is able to compute results correctly also in such a case when the amplitude of the signal is lower than the amplitude in the channel corresponding to the farther microphone.

5. Conclusion

In this paper, we proposed two methods for segmentation of speech according to speakers recorded simultaneously by two microphones.

The *CAMP* method is a very fast, simple and very efficient method for the records with a greater amplitude in the microphone closer to the talking speaker than in the farther one.

The *RESPEC* method is suitable for records in which the amplitude of the signals is almost the same or even higher in the farther microphone or when the surrounding environment influences
the speech in the farther microphone very strongly. Although the *RESPEC* method comprises the *EM* algorithm for training the models of different channels, the *EM* algorithm is used only once and the time of processing is short.

In addition, both methods are capable to detect very quick changes of the microphone, i.e. they are capable to find also very short segments of the speech of particular speakers. Such a kind of speech is very common in interviews, however many other methods fail in such a case.

As the experiments show, the accuracy of both methods is very high. The right selection of the method is dependent on the type of the record, mainly on the way of its recording. Our following research will be focused on consolidation of both methods and on designing a global algorithm that will be able to decide which method is more suitable for a given record.

We can conclude that the proposed methods are robust and highly effective for microphone segmentation and can be easily applied in situations where more than two speakers are involved in conversation.

6. Acknowledgement

The work was supported by the Grant Agency of the Czech Republic, project no GAČR 102/05/0278.

References

1. *M. Cerna, A. F. Harvey*: The Fundamentals of FFT-Based Signal Analysis and Measurement, National Instruments, Application Note 041, 2000.

2. *T. F. Quatieri*: Discrete-Time Speech Signal Processing - Principles and Practice, Lincoln Laboratory, Prentice Hall PTR, 2002.

3. *S. Har-Peled, B. Sadri*: How Slow is the k-means Method?, Symposium on Computational Geometry 2006, Sedona, Arizona, USA, pp. 144–153, 2006.

4. *S. V. Vaseghi*: Advanced digital signal processing and noise reduction, John Wiley & Sons, LTD, Second Edition, New York, 2000.