Speaker Segmentation for Intelligent Responsive Space

Soonil Kwon

Korea Institute of Science and Technology, Intelligence & Interaction Research Center
P.O. BOX 131,Cheongryang,
Seoul 130-650, Korea
soonil@kist.re.kr

Abstract. Information drawn from conversational speech can be useful for enabling intelligent interactions between humans and computers. Speaker information can be obtained from speech signals by performing Speaker Segmentation. In this paper, a method for Speaker Segmentation is presented to address the challenge of identifying speakers even when utterances are very short (0.5sec). This method, involving the selective use of feature vectors, experimentally reduced the relative error rates by 27–42% for groups of 2 to 16 speakers as compared to the conventional approach for Speaker Segmentation. Thus, this new approach offers a way to significantly improve speech-data classification and retrieval systems.

Keywords: Speaker Segmentation, Speaker Recognition, Intelligent Responsive Space (IRS), Human Computer Interaction (HCI).

1 Introduction

Human Computer Interaction (HCI) technology has brought about a number of exciting and challenging applications. One of the applications showing significant promise is the development of Intelligent Responsive Space. Intelligent Responsive Space provides intelligent services and natural interfaces based on the information obtained by monitoring the behaviors, speech, and identities of interactive persons. Recently the development of intelligent interaction technology has focused on intelligent meeting technology as the physical object of Intelligent Responsive Space [1]. In intelligent meeting technology, Speaker Segmentation is very helpful for identifying the participants at meetings, transcribing the dialog at meetings, and providing related documents and the contents of meetings [2][3][4].

The conventional method for Speaker Segmentation is to use Gaussian Mixture Models (GMMs) of speech spectral features to identify a speaker with the minimum probability of error. In this case, segmentation performance depends highly on the amount of data available for recognizing speakers [5]. Experimental evidence has shown that utterances need to be long enough to ensure adequate speaker discrimination [6]. However, in some applications, it is necessary to operate with short speech segments. In spontaneous speech interactions such as telephone conversations and meetings, short utterances lasting an average of about 0.5 seconds

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 385–392, 2007.

(e.g., ``Yes", ``No", or ``Sure") occur frequently [7]. A smaller data set is usually more susceptible to segmentation errors since some feature vectors are more likely to skip the boundaries of short utterances. Therefore, a new Speaker Segmentation method was proposed for this study which excluded the consideration of the particular feature vectors which potentially cause segmentation errors.

The proposed method was experimentally evaluated. In actual conversations such as meetings and debates, the number of participants varies. Thus, in this study, four different groups consisting of 2, 4, 8 and 16 participants were tested. In addition, three lengths of utterances (0.5, 1, 2 seconds) were used for the experiments. Each utterance consisted of spontaneous speech from telephone conversations. Experimental results showed that the study method achieved consistently higher accuracy than the conventional method.

The rest of this paper is organized as follows: Section 2 explains the conventional Speaker Segmentation method; section 3 describes the proposed method; section 4 describes the experiments and discusses results; Conclusions and future plans are described in section 5.

2 Basics of Speaker Segmentation

Speaker Segmentation identifies the speaker of each speech segment. In other words, speech signals are indexed according to the speaker at each time unit. Speaker Segmentation can be regarded as the continuous and sequential execution of Speaker Identification. While speech recognition catches what a person is saying, Speaker Identification identifies the person who is talking.

Speaker Identification is essentially a kind of voice-pattern recognition problem. The system decides who the person is among a group of people. To do this, speech data for people in a group are collected as a training step. From this data, statistical speaker models are built. The model-based method uses a probabilistic formulation of feature space to measure the similarity between two vector sets. The Gaussian model is a basic parametric model. The Gaussian Mixture Model, a weighted sum of some Gaussian distributions, has been found to be effective for developing a speaker model. Model training is accomplished by using the Estimate Maximize (EM) algorithm. The next step in speaker identification is the task of comparing an unidentified utterance with the training models and making the identification. The goal of the identifying process is to choose the speaker model with the minimum probability of error [3][4][6][8].

To train speaker models and execute Speaker Identification and Speaker Segmentation, speech information needs to be analyzed by the short-time spectrum. Cepstral processing is useful to extract features from the speech signal. In addition, the filter bank method analyzes the speech signal through a bank of band-pass filters covering the available range of frequencies. The Mel Frequency Cepstral Coefficient (MFCC) can be obtained by transforming as follows:

- take the Fast Fourier Transform (FFT).
- take the magnitude.
- take the log: result is real valued and symmetric.

- warp the frequencies according to the mel scale.
- take the inverse Fast Fourier Transform.

The mel scale is based on non-linear human perception of sound frequency in which the higher frequency band is compressed since it is regarded as less important for the understanding of sounds than the lower frequency band.

The conventional method for Speaker Segmentation is similar to the conventional Speaker Identification method. The difference between the two methods is whether the procedure is continuously executed or not. To identify the speaker of each segment, speaker models are trained for each participant in a conversation. The set of data (input vectors) from each segment is then used to identify speakers via previously trained speaker models. Based on the Maximum Likelihood criterion, each segment of input data is sequentially mapped to the model of a certain speaker.

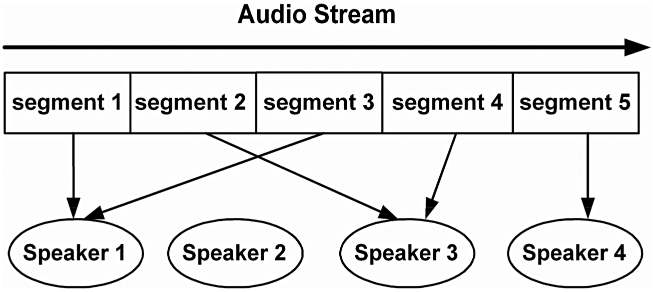

Fig. 1. Illustration of Speaker Segmentation

Fig. 1 shows that continuous speech data such as spoken dialogues and broadcast news can be divided into respective segments identifiable to certain speakers. A long segment is useful for improved Speaker Segmentation performance as it includes more information about the speakers and thus makes identification more accurate. However, it is apt to miss speaker changes that may occur within a segment. To solve this problem, a smaller segment size can be used. However, this requires a refined speaker change detection process to improve precision.

The length of the segment can be either variable or static. In variable length segmentation, the speech stream is divided into different lengths depending on several factors such as pauses and background changes. Pauses are an important consideration in speaker change analysis. In the middle of speaking, people usually breathe. Speaker changes are not likely to occur between these breathing points. The pause point is defined as a certain period within which the energy of a signal stays below the threshold. However, static segmentation assumes a fixed length. A static segmentation is attractive since it is computationally simple, but care has to be taken when choosing the length of segments. Too short a segment may not provide adequate data for analysis, while a longer segment may miss a speaker change point.

Speaker change detection is a very important step in accurate Speaker Segmentation. However, it is not easy to detect speaker changing points due to the lack of data, the variability of speech signal, and environmental noise. More sophisticated algorithms may overcome these difficulties. However, a good method for detecting speaker change has not yet been developed. To improve the speaker change detection algorithm, we can incorporate other features representing the speed and habit of speaking that have not been deeply explored yet. Multi-modal features, such as expression, emotion, gaze, and gesture, can also be useful in improving the performance of speaker change detection.

In this paper, speaker change detection is not considered because the focus is segmentation, and it is assumed that speakers do not change within a segment. For segmentation, the fixed-amount of data is extracted by a sliding window without overlapping.

3 New Method for Speaker Segmentation

A conventional method of Speaker Segmentation is to choose a speaker model with the minimum probability of error. Speaker models are usually overlapped. The reasons for overlapping models are environmental noise, pauses, and unpredictable similarity between the acoustical characteristics of speakers. These unwanted factors result in segmentation errors. Speaker Segmentation performance also depends highly on the amount of data available for identifying the voice patterns of speakers. However, in spontaneous speech interactions such as telephone conversations and meetings, short utterances are common. It is quite natural that a smaller data set is more susceptible to segmentation errors.

In order to reduce the impact of factors which induce segmentation errors, the proposed method split each speaker model into two: non-overlapped and overlapped models. Fig. 2 shows an illustration of one-dimensional speaker-model splitting in which there are two speakers: speaker 1 and speaker 2. Dotted lines represent conventional speaker models, and solid lines represent the new speaker models discussed in this paper. Speaker 1-a is the non-overlapped model of speaker 1, Speaker 1-b is the overlapped model of speaker 1, Speaker 2-a is the non-overlapped model of speaker 2, and Speaker 2-b is the overlapped model of speaker 2. A is the decision boundary of speaker 1 and speaker 2. Two conventional speaker models are split into 4 new speaker models to detect and eliminate undesirable factors.

For splitting speaker models, conventional speaker-models (GMMs) were first trained with sets of speaker-specific speech data (training vectors). Next, using the Maximum Likelihood criterion with the speaker models built in the previous step, we classified the training vectors for each speaker into two categories (non-overlap and overlap) since there could have been some vectors falsely identified if competing speaker models overlapped. In the last step of training, based on the reclassified training vectors, two models for each speaker were reconstructed: non-overlapped and overlapped speaker models [7].

For example, assume there are S single-speaker speech data sets. With feature vectors extracted from these data, we trained speaker models, Mi, where $i=1,\ldots,S$. Then we categorized feature vectors from each speaker data into non-overlapped and overlapped vectors using the Maximum Likelihood criterion as follows [7]:

- x_j: j-th input vector, $j=1,\ldots,N$.
- $I_j = \arg\max Pr(x_j \mid M_i)$, $i=1,\ldots,S$, $j=1,\ldots,N$.

- If I_j is a correct speaker index, $x_j \rightarrow P$ (a vector set of a non-overlapped category).
- Else $x_j \rightarrow Q$ (a vector set of a overlapped category).

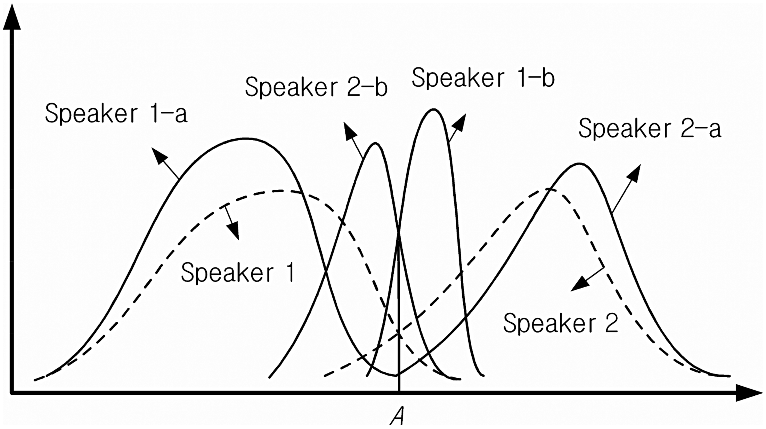

Fig. 2. Illustration of model splitting

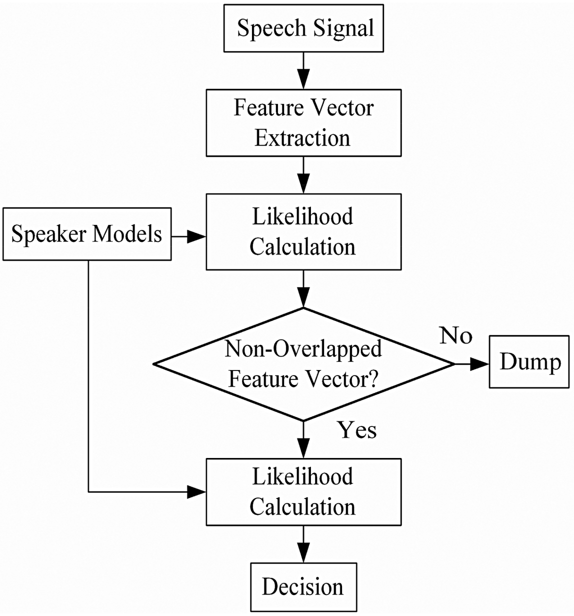

Fig. 3. Block diagram of Speaker Segmentation Procedure

After feature vector categorization, we reconstructed the speaker models. For each speaker *i*, we built two models, non-overlapped (MP_i) and overlapped (MQ_i), with the vectors of P and Q, respectively. Using the pairs of speaker models, we selected the feature vectors that would determine Speaker Segmentation.

As seen in Fig. 3, continuous speech data such as spoken dialogues and broadcast news need to be divided into respective segments of speakers. For segmentation, the fixed-amount of data was extracted by a sliding window with overlapping. The set of data (input vectors) from each segment was used to identify speakers with previously trained speaker models. Based on the Maximum Likelihood criterion, each input vector extracted from a speech signal was mapped to either the overlapped or non-overlapped model of a certain speaker. Next, taking only input vectors that were mapped to the non-overlapped model of any speaker, each segment was identified with a specific speaker. The point of this method was to exclude input vectors mapped to overlapped models for purposes of identification. In other words, the influence of common features inducing segmentation errors was reduced.

4 Experimental Results

In this experiment, Speaker Segmentation was executed with spontaneous speech data sets obtained from telephone conversations. The length of short utterances was 0.5, 1, and 2 seconds. We conducted experiments with 1 and 2 second utterances to compare with 0.5 second cases. Usually there are a varying number of people participating in conversations such as meetings and debates. Hence, in this experiment, data from 4 groups composed of varying numbers of participants (2, 4, 8, 16 persons) were examined. Twenty five different data sets were created for each group. Each speech data set was artificially made of 16 short utterances. For example, for an experiment with 0.5 second utterances of 4 participants, 25 test speech set, consisting of 16 utterances of 4 participants (8 second long), were used.

The performance was calculated based on the ratio of the number of correctly identified segments to the number of total segments as follows:

$$Error \quad rate = \frac{Number \quad of \quad correctly \quad identified \quad segments}{Number \quad of \quad total \quad segments} \, . \tag{1}$$

Experimental results showed that the new method tested in this research consistently achieved higher accuracy than the conventional method. It is interesting to observe that the new method outperformed the conventional GMM method for all the utterance lengths considered. In Fig. 4, the range of difference in the absolute error rate between the GMM (baseline) and the new method ranged from 3.2% to 8.4% absolute (27.4% to 42% relative error rate) with respect to the various number of participants (speakers) in the case of 0.5 sec utterances. The error rate of our method with 0.5 sec utterances (8.8%) was almost the same as the error rate of the conventional method (baseline) with 2 sec utterances (9.5%) in the case of a group with 4 participants. This means that the conventional method requires utterances approximately 4 times longer than the new method to achieve approximately the same level of accuracy.

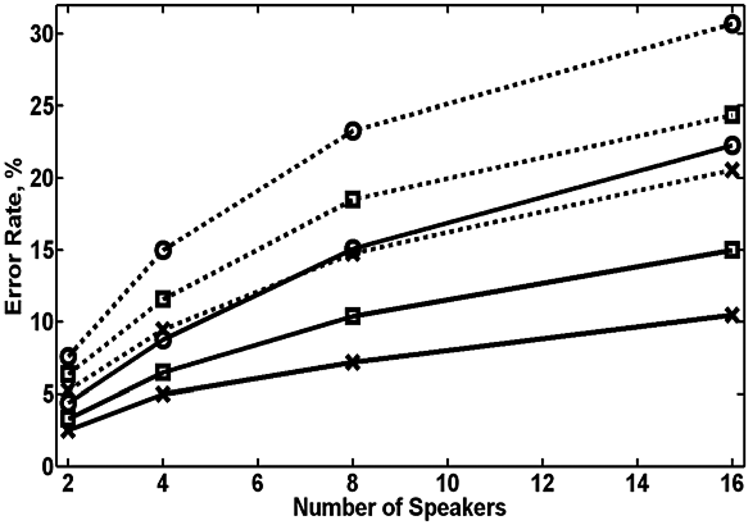

Fig. 4. Measured error rate of Speaker Segmentation with respect to the lengths of utterances, the number of speakers, and the methods of Speaker Segmentation (solid lines for the baseline method and dotted lines for the new method; **O** for 0.5 sec of utterances, □ for 1.0 sec, and **x** for 2.0 sec)

5 Conclusion

This paper examined a new Speaker Segmentation method designed to reduce identification errors. This method was useful for Speaker Segmentation by identifying speakers from short utterances. It also made it possible to detect the boundaries of short utterances. These results indicate that Speaker Segmentation can be applied to spontaneous speech in human-to-human, human-to-robot, and human-to-computer interactions. Future work should focus on the further refinement of identification methods in natural data streams, such as meetings and broadcast news.

References

1. Park, J.-H., Yeom, K.-W., Ha, S., Park, M.-W., Kim, L.: An overview of intelligent responsive space in tangible space initiative technology. In: Proc. Internt. Workshop on the Tangible Space Initiative (3rd), pp. 523–531 (2006)
2. Busso, C., Hernanz, S., Chu, C.-W., Kwon, S., Lee, C., Georgiou, P.G., Cohen, I., Narayanan, S.: Smart room: participant and speaker localization and identification. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, 2005, vol. 2, pp. 1117–1120 (2005)
3. Campbell, J.P.: Speaker recognition: A tutorial. Proc. IEEE 85, 1436–1462 (1997)
4. Kwon, S., Narayanan, S.: Unsupervised Speaker Indexing Using Generic Models. IEEE Trans. on Speech and Audio Processing 13(5) part 2, 1004–1013 (2005)

5. Nishida, M., Ariki, Y.: Speaker indexing for news articles, debates and drama in broadcasted TV programs. In: Proc. IEEE Internat. Conf. on Multimedia Computing and Systems, vol. 2, pp. 466-471 (1999)
6. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. on Speech Audio Processing 3(1), 334–337 (1995)
7. Kwon, S., Narayanan, S.: Robust speaker identification based on selective use of feature vectors. Pattern Recognition Letters 28, 85–89 (2007)
8. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signals, pp. 476–489. Prentice Hall, Englewood Cliffs (1978)