SOUND SOURCE DETECTION USING MULTIPLE NOISE MODELS

Shoichi Matsunaga, Masahide Yamaguchi, Katsuya Yamauchi, Masaru Yamashita

Department of Computer and Information Sciences, Nagasaki University, Nagasaki, Japan

ABSTRACT

This paper describes a sound source detection approach based on elaborate noise-modeling techniques for audio indexing. For accurate detection, we devised two methods to generate multiple-noise models through clustering techniques. One method is based on frame-wise data similarity, and the other is based on noise source similarity. The former method employs K-means clustering and a smoothing technique to avoid inaccurate segmentation. The latter method involves noise modeling based on a tree data structure generated by the progressive merging of noise clusters. The classification experiments show that by using these proposed methods, audio sources can be detected with better accuracy than that achieved by a conventional method. When four noise models generated by the latter method were used, the noise detection performance increased by 3.9% for the periods in which the sound sources did not overlap. With regard to the experiments for an audio stream that included overlapped segments, the noise detection performance increased by 1.2% without a decrease in the speech detection performance.

***Index Terms*—** Clustering, Acoustic segmentation

1. INTRODUCTION

In recent years, the amount of new content requiring immediate access via the Internet has significantly increased. In comparison with searching for text content, effectively searching for desired multimedia content is very difficult. Useful descriptions of multimedia content (referred to as metadata) are required for effective searching. In general, such descriptions are manually transcribed from verbal and nonverbal information included in the content. Unfortunately, creating such descriptions is highly time-consuming and very expensive. Speech recognition and audio source detection are expected to be solutions to these problems. State-of-the-art speech recognition technology can recognize clear speech with high accuracy; however, noise or music from other sound sources, which commonly appear in most content, significantly hamper the recognition performance.

A number of studies have been conducted on multimedia content indexing for information retrieval [1-4]. In this regard, we have also studied an efficient strategy for indexing broadcast news [5]. The proposed broadcast-news indexing technique should be able to detect utterance boundaries with sufficient precision so as to enable better speech recognition [6] and locate significant acoustic points for topic segmentation.

Therefore, accurate audio source detection (segmentation) is required for audio content indexing. Voice activity detection (VAD), which delimits the beginning and end of the speech segments in audio content, has been widely studied for robust speech recognition [7], and this technology is applicable to content indexing. Noise detection is also an important research topic for indexing; however, it is very difficult to detect the various types of noise sources. This difficulty arises from the fact that the number of noise sources is large and that the acoustic features of each type of noise are very different. Furthermore, collecting a sufficient amount of data for each type of noise source for noise modeling is very difficult.

We have studied an audio source detection approach based on a stochastic method to detect speech, noise, music, and silence. Our approach uses not only conventional surface acoustic features such as signal energy and pitch frequency [8, 9] but also new features that are based on spectral correlation for more accurate detection [10]. These features measure spectral stability, white noise similarity, and spectral shape. The experiment with the broadcast news demonstrated that these feature parameters made it possible to capture the audio source more accurately. However, the experiment also showed that the detection performance for noise sources was still lower than that for the other sound sources. Devising a model to increase the detection performance for a noise source would require that the detection performance for other sound sources not be decreased.

Addressing the aforementioned issues, this paper proposes an audio source detection approach using multiple noise models for more accurate sound-source segmentation. We devised two methods for generating the multiple noise models. The first method was based on the similarity criterion among frame-wise noise data. In this method, K-means clustering was used to classify a noise-tagged data set into a number of noise clusters fixed *a priori*, and a smoothing technique was used to modify the clustering results. The clustering used a Mahalanobis metric to cope with the diversity of noise distribution, and the smoothing prevented the inaccurate segmentation of the data into very short noise periods. The second method was based on the noise-cluster similarity criterion. This method used a tree data structure made by the progressive merging of each type of noise data (hierarchical clustering), in which a Bhattacharyya metric was used. The results of our detection experiments that used Japanese broadcast news

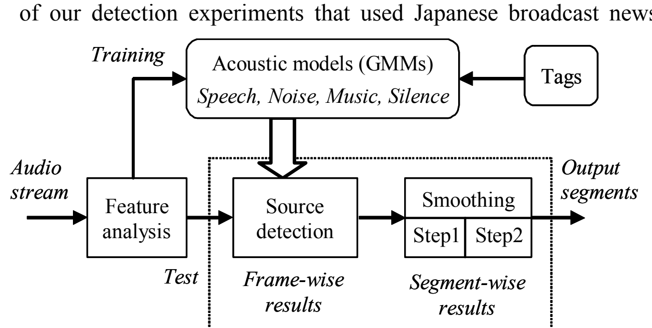

Figure 1. Sound source segmentation procedure

segments showed that both the proposed methods achieved higher noise recall rates than a conventional method using the single noise model. The latter method achieved the highest performance by increasing the noise detection rate considerably and the weighted average for the detection performances of all sound sources also increased.

2. SOUND SOURCE SEGMENTATION

2.1. Acoustic feature parameters

For acoustic source detection, we use seven acoustic features: the four conventional features of signal energy, pitch frequency, peak-frequency centroid, and peak-frequency bandwidth, and three parameters based on spectral cross-correlation: temporal stability, white noise similarity, and spectral shape [10].

2.2. Sound source segmentation procedure

In broadcast news preprocessing, the sound source segmentation procedure was carried out prior to speech recognition, and the sound content of the news was divided into four types of segments: speech, music, noise, and silence. Speech recognition and sound source segmentation results were used for indexing. The segmentation procedure is composed of an acoustic source detection part and a smoothing part (Figure 1). In the detection part, time-frame-wise detection results and the likelihood for each feature parameter in each frame are obtained. The likelihood is calculated by using the acoustic models of each sound source.

The smoothing part uses a merge method that consists of two processing steps. The first step (Step 1) takes into account the total likelihood of all feature parameters and each feature's likelihood for more elaborate smoothing. This technique is used because total likelihood is the most important factor in detecting boundaries; however, some sound source characteristics appear only with specific parameters, and the total likelihood frequently ignores these phenomena. The second step (Step 2) uses a conventional technique that smoothes the results of the first step using a longer window to avoid the inaccurate detection of very short periods. The details of segment smoothing are described elsewhere in the literature [10].

2.3. Audio source tagging

We prepared a tagged database for 28 broadcast news programs to evaluate the performance of audio source segmentation. These data comprised seven programs in each of the following categories—5, 10, 20, and 30 min segments. The data were manually tagged based on the beginning and ending times of speech, music, and noises. Some tagged periods overlapped, such as an interview in a crowd, an anchor's speech with background music, and so on. Non-tagged periods were the silent parts. The label definition of each audio source type was as follows:

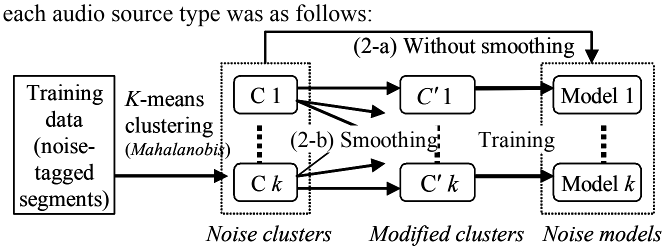

Figure 2. Modeling method based on noise data similarity (Method I)

- ***Speech*:** Speech of an anchor, reporter, interviewee, and all other transcribable utterances. The speech periods included short pauses.
- ***Music*:** Music and jingles.
- ***Noise*:** Sounds of crowds, cheers, traffic noise, and so on. Background voices, such as murmurs or shouts, were tagged as noise. Low noises, such as lip and paper noises, were also tagged as noise. Around thirty types of noise sources were separately tagged, and two or more types of noise-tags overlapped in many periods.

We investigated the amount ratios of non-overlapped (single-tagged) periods among different sources in our news content. Speech, noise, silence, and music segments accounted for 58.3%, 18.9%, 14.5%, and 8.3%, respectively. These results indicated that accurate detection is required for not only the speech segments but also the noise segments.

The sampling frequency of the audio data was 44.1 kHz. Every 23 ms, acoustic feature parameters were computed using a 46-ms Hamming window. The acoustic model for each sound source (speech, noise, music, or silence) was trained using single tagged periods and silence periods. Overlapped periods among the noise sources were only used for evaluation. The acoustic GMM (Gaussian mixture model) with a two-mixture Gaussian distribution was used for each sound source, and the mean, variance, and weighted value for each distribution were estimated using an EM (Expectation Maximization) algorithm.

3. ACOUSTIC NOISE MODELING

We developed two generation methods of the multiple noise models for more accurate detection: Methods I and II. Method I was based on noise frame-data similarity, and Method II was based on noise-source similarity. Method II required a more detailed noise tagging of the training data than Method I.

3.1. Modeling based on the similarity among noise data (Method I)

Figure 2 provides an overview of the model generation process. This method employed *K*-means clustering using all the frame-wise data in the noise segments. The frame-wise data set was partitioned into *K* clusters. The Mahalanobis metric was used to calculate the distance between each frame-wise data and the distribution of acoustic parameters of each subset. After the clustering was completed, a smoothing technique was applied to account for the time continuity of the noise sources. This technique used a time window to move the included cluster of unrealistic short-time data-sequences into the same cluster preceding or trailing the data sequence in the time alignment to which it belonged. This procedure was the same as Step 2 described in Section 2.2. The noise data subset in each obtained cluster was used as training data for each noise model.

3.2. Modeling based on the similarity among noise sources (Method II)

Figure 3 provides an overview of the model generation procedure. Our procedure consists of two processes. In the first process, hierarchical clustering was performed to generate a tree data structure (dendrogram) by using the acoustic features of sound segments from each noise source ((3-a) in Figure 3). In this clustering, the two closest clusters were selected by using the Bhattacharyya distance and were merged into a single cluster. The

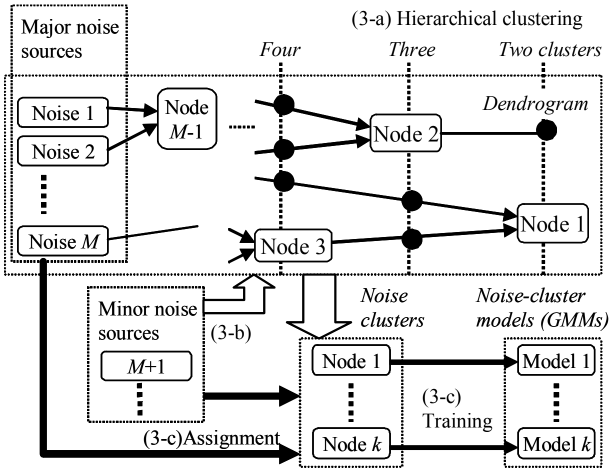

Figure 3. Modeling method based on noise source similarity (Method II)

Table 1. Audio source detection rates using a single model of noise [%] (Baseline)

Source	Recall	Precision	F-measure
Speech	95.3	91.7	93.5
Music	77.9	71.7	74.8
Silence	81.4	69.6	75.5
Noise	56.7	79.3	68.0
Average	84.5	84.5	84.5

Bhattacharyya metric calculated the distance by using the parameter distribution of each cluster. A single Gaussian model for each type of noise (cluster) was used to calculate the distance. The merging was progressively repeated until the desired number of noise clusters was attained. There were 32 types of noise sources in our database, and clustering was performed on 13 types of major noise sources that contained a larger number of samples. With regard to each noise cluster of other minor noise sources, the distance between each cluster (generated by merging) and each minor source was calculated. The noise cluster with the smallest distance was regarded as the belonging cluster of each minor noise source (3-b). In the second process, each noise-cluster model was generated using a data set of noise sources corresponding to each node of the dendrogram (3-c).

4. EVALUATION EXPERIMENTS

Our detection test was conducted using the 28 broadcast news programs. We performed a leave-one-out cross validation on these data. The overlapped periods among the different sound sources were not evaluated for the experiments in Sections 4.1, 4.2, and 4.3, where the evaluation data amounted to around 402k samples. In Section 4.4, both overlapped and non-overlapped segments (around 686k samples) were evaluated.

4.1. Detection accuracy of the baseline

A preliminary detection test was conducted to evaluate our baseline system using a single model of noise. The detection rate for each sound source is shown in Table 1. The average detection rate weighted with the data amount for each sound source is indicated as "Average." This table shows that the F-measure of noise detection is far lower than that of the other sources.

4.2. Experiments using *K*-means Clustering (Method I)

The detection rates using multiple noise models generated without smoothing ((2-a) in Figure 2) and with smoothing (2-b) are shown in Tables 2 and 3, respectively. By comparing the data of Table 1 with that of Table 3, it was observed that the detection using the two noise models generated by Method I increased the F-measures for both the noise and the average source detection by 2.7% and 0.8%, respectively. This result indicates that the proposed multiple noise modeling can increase the detection performance. However, the performance when the three noise models are used is lower than that of the baseline system. Devising a method to determine the number of noise models required to achieve sufficient detection performance is a problem for future investigation. Tables 2 and 3 show that smoothing always achieves a slight improvement.

4.3. Experiments using noise models based on source similarity (Method II)

Two to five noise-cluster models were generated in our experiments. While we performed a leave-one-out cross-validation, the same dendrogram was used for the training of the noise-cluster models. All the models for the other sources are the same as the baseline models. The detection results are shown in Table 4. The recall rate of noise detection increased with the increase in the number of noise-cluster models. On comparing the data in Tables 1 and 4, we observed that the proposed noise-cluster models constantly achieved higher F-measure values with regard to noise source detection and all source detection than the single noise model. The noise and average source detection rates using four noise-cluster models, which showed the highest performance, increased by 3.9% and 0.9%, respectively. This comparison also

Table 2. Audio source detection rates[%] (Method I, Mahalanobis metric, without smoothing)

No. of noise models	Source	Recall	Precision	F-measure
Two	Noise	63.4	77.1	70.3
	Average	85.4	85.0	85.2
Three	Noise	69.1	65.7	67.4
	Average	83.5	83.4	83.4

Table 3. Audio source detection rates [%] (Method I, Mahalanobis metric, with smoothing)

No. of noise models	Source	Recall	Precision	F-measure
Two	Speech	96.1	91.4	93.7
	Music	78.9	72.4	75.6
	Silence	72.7	78.9	75.8
	Noise	65.3	76.0	70.7
	Average	85.5	85.1	85.3
Three	Noise	70.1	65.7	67.9
	Average	83.6	83.6	83.6

showed that the F-measure values of both noise detection and music detection increased. Furthermore, there was no significant decrease in the F-measure values of speech and silence detection. These results indicate that the proposed noise-cluster modeling based on noise-source clustering is effective in audio source detection.

4.4. Detection experiments including overlapped segments

Finally, we conducted detection experiments on all data including the overlapped periods among two or three sound sources. The experiments were performed from two viewpoints: speech detection and noise detection. From the viewpoint of speech detection, only speech segments accounted for 31.4%, and overlapped segments of speech and other sources accounted for 40.2% in the evaluation data. The baseline method, Method I (two noise models, Mahalanobis metric, smoothing), and Method II (four noise models, Bhattacharyya metric) were evaluated. The speech detection performance is shown in Table 5. It is shown that the F-measure values using Methods I and II increased by slightly on employing the multiple noise models.

From the viewpoint of noise detection, experiments using the previous three methods were performed. Only noise segments accounted for 11.1%, and overlapped segments of noises and other sources accounted for 32.7%. The noise detection performance is also shown in Table 5. Our detection approach selected one sound source that achieved the highest likelihood. The number of anchors' or reporters' speech segments mixed with the background noises in our news content was large to the extent that the noise recall rates were low. Almost all these segments were detected as speech. However, using multiple noise models, the F-measure of noise detection was increased by 1.2% without a decrease in the performance of speech detection. These experiments also demonstrate the effectiveness of multiple noise models for audio source detection.

5. CONCLUSIONS

This paper described the sound source segmentation approach using multiple noise models for accurate detection. The noise models were generated by using clustering techniques applied to a noise-tagged acoustic data set. We devised two generation methods: one was based on frame-wise noise data similarity (Method I), and the other was based on noise cluster similarity (Method II). Classification experiments showed that by using these proposed methods, audio sources could be detected with greater accuracy than that achieved by a conventional method (baseline) using a single noise model. A detection approach using the latter method consistently achieved the highest performance among the conventional method and the two proposed methods.

Our proposed method (Method II) requires a significant amount of precise noise-tagging data, and determining the quantity of noise-tagged data required to achieve sufficient detection performance for indexing is a problem that requires future research. The detection of multiple sound sources for overlapped sound segments also needs to be tackled in the future.

Table 4. Audio source detection rates for each number of noise models [%] (Method II)

No. of models	Source	Recall	Precision	F-measure
Two	Noise	64.9	74.4	69.7
	Average	85.3	84.8	85.1
Three	Noise	67.2	75.7	71.4
	Average	85.7	85.2	85.4
Four	Speech	96.4	90.8	93.6
	Music	74.7	78.8	76.8
	Silence	71.0	79.7	75.3
	Noise	68.8	75.0	71.9
	Average	85.7	85.2	85.4
Five	Noise	74.7	67.5	71.1
	Average	84.8	85.0	84.9

6. REFERENCES

[1] Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen L. and Schwartz, R., "Speech and language technologies for audio indexing and retrieval," Proc. of the IEEE, vol.88, No.8, pp.1338-1353, 2000.

[2] Dharanipragada, S., Franz, M. and Roukos, S., "Audio indexing for broadcast news," Proc. TREC-7, pp.115-119, 1998.

[3] Johnson, S., Jourlin, P., Jones, K. and Woodland, P.C., "Audio indexing and retrieval of complete broadcast news shows," Proc. RIAO, pp.1163-1177, 2000

[4] Logan, B., Goddeau, D. and Thong, J., "Real-world audio indexing systems," Proc. of the ICASSP, 2005

[5] Ohtsuki, et al, "Automatic multimedia indexing," IEEE Signal Processing Magazine, Vol.23, No.2, pp.69-78, 2005

[6] Ramabhadran, B., et al, "Impact of audio segmentation and segment clustering on automated transcription accuracy of large spoken archives," Proc. Eurospeech, pp.2589-2592, 2003

[7] J. Ramirez, et al, "Efficient voice activity detection algorithm using long-term speech information," *Speech Communication,* Vol.42, pp.271-287, 2004

[8] Z. Liu, J.Huang, Y. Wang and T. Chen, "Audio feature extraction and analysis for scene classification," *Proc. IEEE Multimedia Signal Processing Workshop*, pp.343-348, 1997.

[9] T. Zhang and C. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. on Speech and Audio Processing*, Vol.9, No.4, pp.441-457, 2001.

[10] Matsunaga,S., et al, "Audio source segmentation using spectral correlation features for automatic indexing of broadcast news," *Proc. EUSIPCO*, pp.2103-2106, 2004.

Table 5. Speech and noise detection results for entire broadcast news test data [%]

Source (ratio)	Source	Recall	Precision	F-measure
Speech (71.6%)	Baseline	90.3	95.5	92.9
	Method I	91.0	95.3	93.1
	Method II	91.6	95.0	93.3
Noise (43.8%)	Baseline	20.5	80.6	50.5
	Method I	22.9	79.3	51.1
	Method II	24.4	79.0	51.7