ABSTRACT
This paper proposes a new segmentation procedure to detect audio source intervals for automatic indexing of broadcast news. The procedure is composed of an audio source detection part and a part that smooths the detected sequences. The detection part uses three new acoustic feature parameters that are based on spectral cross-correlation: spectral stability, white noise similarity, and sound spectral shape. These parameters make it possible to capture the audio sources more accurately than can be done with conventional parameters. The smoothing part has a new merging method that drops erroneous detection results of short duration. Audio source classification experiments are conducted on broadcast news segments. Performance is increased by 6.6% when the proposed parameters are used and by 3.1% when the proposed merging method is used, showing the usefulness of our approach. Experiments confirm the impact of this proposal on broadcast news indexing.

1. INTRODUCTION
The amount of accessible multimedia content has begun to increase more rapidly with the penetration of broadband networks. To look desired content up effectively, useful descriptions of the content, metadata, are necessary. Making such descriptions manually is, unfortunately, too time-consuming and too expensive. Speech recognition is expected to release us from this labor. State-of-art speech recognition technology can recognize clear speech with high accuracy, but noise or music from other sound sources, which is common in most content, decreases the performance significantly.

A number of studies have been made on multimedia content indexing for information retrieval [1-4]. In this regard, we have also been developing a system for indexing broadcast news [5]. This system uses speech recognition and topic segmentation techniques integrated with linguistic and acoustic processing results. The broadcast-news indexing it provides should detect utterance boundaries with sufficient precision so as to enable better speech recognition [6,7] and find significant acoustic points for topic segmentation.

This paper describes a sound-source segmentation procedure consisting of an acoustic detection part and a part that smooths the detected results. Conventional detection approaches use only surface acoustic features, such as signal energy and pitch frequency [8]. In addition to these features, our procedure uses three features which are based on spectral correlation for more accurate detection. These features measure spectral stability, white noise similarity, and spectral shape.

Short-duration analysis is usually required in the detection procedure, but the analysis sometimes results in misdetection. For instance, long vowels can be detected as music and fricatives detected as environmental noise. The likelihood of these midsections is high and they are difficult to identify. One conventional strategy to overcome this is merging detection results with longer duration windows to erase short misdetections. However, determining and setting the appropriate window size for every kind of audio source is difficult. To reduce the influence of these misdetections, we propose a new smoothing (segment detection) method that takes account of both the total likelihood of all features and each feature’s likelihood. Evaluation results of each part in the proposed segmentation procedure using Japanese broadcast news segments are shown. Experiments were also carried out using our indexing system to confirm the validity of the proposed segmentation procedure.

2. SOUND SOURCE SEGMENTATION
The sound source segmentation procedure is composed of an acoustic 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. Likelihood is calculated using Gaussian distribution functions. In the smoothing (segment detection) part, the duration of each sound source is decided by using a two-step merging procedure; likelihood-based smoothing and smoothing with long window.

Figure 1. Audio source segmentation flow
3. ACOUSTIC FEATURES FOR SEGMENTATION

For acoustic source detection, seven acoustic features are used: the four conventional features of signal energy, pitch frequency, peak-frequency centroid, and peak-frequency bandwidth, and three new parameters based on spectral correlation. These new features are temporal stability, whiteness (white noise similarity), and spectral shape.

3.1 Temporal Stability

The stability of spectral features is an important factor of source type detection. For example, spectral features of speech tend to change more dynamically than those of music. To measure stability, we calculate the correlations between adjacent power spectra \( S \) at time \( t \) and \( u (u = t + \Delta t) \). The generalized cross-correlation between \( S_t(f) \) and \( S_u(f) \) is defined as:

\[
R_{tu}(\tau) = \int_{-\infty}^{\infty} S_t(f)G_{tu}(f)e^{2\pi j \tau f} df
\]

where \( \tau \) denotes the complex conjugate, \( \psi \) is the frequency weighting filter, and \( G \) is the cross power spectrum. We use the variance of the correlation contour as a measure of temporal sound stability.

3.2 White Noise Similarities and Spectral Shape

The measurements of whiteness and sound spectral shape are also calculated using the generalized cross-correlation. As one of two consecutive power spectra is slid along the frequency domain, the correlation between the two power spectra \( S_t(f) \) and \( S_u(f + \Delta f) \) is calculated. Namely, \( G \) in Equation (2) is rewritten as follows.

\[
\hat{G}_{tu}(f, f + \Delta f) = S_t(f)S_u^*(f + \Delta f)
\]

where \( \Delta f \) is slide width in the frequency domain.

Figure 2 shows the concept of using cross-correlation to detect whiteness similarity. The power spectrum is shown on the left side and the correlation for white noise (Figure 2(a)) and colored sound (b) is shown on the right side. White noise tends to have a flat response along the frequency domain. In contrast, the contour of a processed voiced signal dissipates quickly and vibrates heavily. We extract contour shape by using weighted linear regression. The intercept and slope of the regression (dotted line in Figure 2(b)) represent the degree of whiteness and the difference between the high and low bands, respectively. If the intercept is a high value, the signal is similar to white noise. If it is a low value, the signal is regarded as colored sound.

4. SEGMENT SMOOTHING

The proposed smoothing (segment detection) part uses a new merge method consisting of two processing steps. The first step (Step-1) takes account of the total likelihood \( L_{\text{all}} \) of all feature parameters and each feature’s likelihood \( (L) \) of \( w \) is the weight value). The total likelihood is most important in detecting boundaries; however, some sound source characteristics appear only with specific parameters, and the total likelihood frequently ignores this phenomena. The Step-1 procedure is as follows (Figure 3):

1. Trace each acoustic feature parameter \( i \) \((1 \leq i \leq N)\) indicating the same detection result \( K_i \) from some time point \( t \), and acquire length \( D_i \).
2. Detect the longest segment length \( D_{\text{max}} \) and its audio source type \( K_{\text{max}} \) among all feature parameters.
3. For this segment candidate, from \( t \) to \( t + D_{\text{max}} \), calculate the total likelihood by all parameters and acquire detected audio source type \( K_{\text{all}} \) that gives the highest segmental likelihood \( L_{\text{all}} \).
4. If \( K_{\text{all}} \) and \( K_{\text{max}} \) are the same, then the audio source type of interval \( D_{\text{max}} \) is determined as \( K_{\text{all}} \). If they are not the same, then acquire the second longest \( D'_{\text{max}} \) and \( K'_{\text{max}} \).
5. While changing \( D'_{\text{max}} \) and \( K'_{\text{all}} \), repeat 3–4 until agreement is obtained between \( K_{\text{all}} \) and \( K_{\text{max}} \). The above procedure is carried out from the beginning \((t = 0)\) to the end of the content.

Step-2 is the conventional method; it uses a longer window, and so ignores short time misdetection.

5. EVALUATION EXPERIMENTS

5.1 Training and Evaluation Data

We prepared a tagged database for 30 broadcast news programs to evaluate the performance of audio source segmentation, speech recognition, and topic segmentation. These data were manually tagged on the basis of the beginning time and ending time of speech, music and noises. Some tagged periods overlapped, such as an interview in a crowd, anchor speech with background music etc.; non-tagged peri-
ods are silent parts. The label definition of each audio source type was as follows:

- [Speech]: Speech of anchor, reporters, interviewee, and all other transcribable utterances. The speech periods include short pauses.
- [Music]: Music and jingle.
- [Noise]: Sound of crowds, cheering, traffic noise, etc. Background voices, such as murmurs or shouts, are tagged as noise, but low noises, such as lip and paper noises, are not tagged as “noise”.

Transcriptions and topic tags were also entered into the database.

To train the acoustic feature models for each parameter, we used audio data of 10 broadcast news programs, the sampling frequency was 44.1 kHz. This training used single tagged periods and silence periods. The acoustic model with 2-mixture of Gaussian distribution was used for each parameter, and the mean, variance, and weight value for each distribution were estimated using an EM algorithm. Speech-tagged data was a one-hour voice-only content. Music data was a total of seven minutes of theme music, background music and jingles. Noise data was a total of eight minutes of a variety of noises. Silence data was a total of 13 minutes of attenuated sounds containing paper noise and anchor’s breath sound.

Evaluation data was provided by 20 other programs, i.e., five each of programs that were 5, 10, 20 and 30 minutes in length.

5.2 Evaluation of New Acoustic Parameters

We evaluated two types of analysis parameter sets. One set consisted of four conventional parameters: energy, pitch frequency, frequency centroid and bandwidth [8]. The other consisted of seven parameters, the first four plus three proposed parameters: stability, whiteness similarity, and spectral shape. This evaluation used the smoothing procedure.

Table 1 shows the evaluation results; overlapping segments with multiple sound sources were eliminated. The F-measure for each sound source was improved remarkably, and average F-measure weighted with data amount improved by 6.6%, showing the effectiveness of the proposed spectral-correlation-based parameters.

Table 2 shows two results of speech or non-speech segment detection for all the evaluation data including overlapping segments; one with no time margin and one with a margin of up to 100msec. Using these new features yielded a 94.9% F-measure of speech detection within a 100msec margin.

5.3 Evaluation of Smoothing Procedure Step-1

While Table 1 shows the results of the smoothing procedure composed of Step-1 and 2, detection experiments were carried out using only Step-2 to evaluate the Step-1. Experimental results using the seven parameters are shown in Table 3. Comparing Tables 1 and 3 indicates that Step-1 increases the average F-measure by 3.1%. This confirms the usefulness of considering each parameter’s detection result in the smoothing method.

Table 1. Acoustic signal detection rates (%)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Source</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four parameters</td>
<td>Speech</td>
<td>85.1</td>
<td>96.2</td>
<td>90.3</td>
</tr>
<tr>
<td></td>
<td>Music</td>
<td>68.7</td>
<td>62.6</td>
<td>65.5</td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td>55.2</td>
<td>52.9</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td>Silence</td>
<td>92.4</td>
<td>68.1</td>
<td>78.4</td>
</tr>
<tr>
<td>Average</td>
<td>Speech</td>
<td>95.6</td>
<td>95.0</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td>Music</td>
<td>77.5</td>
<td>87.7</td>
<td>82.3</td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td>66.4</td>
<td>72.5</td>
<td>69.3</td>
</tr>
<tr>
<td></td>
<td>Silence</td>
<td>87.3</td>
<td>80.7</td>
<td>83.9</td>
</tr>
<tr>
<td>Average</td>
<td>Speech</td>
<td>89.9</td>
<td>89.8</td>
<td>89.8</td>
</tr>
</tbody>
</table>

Table 2. Speech detection results for all test data (%)

<table>
<thead>
<tr>
<th>Margin</th>
<th>Source</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Speech</td>
<td>91.5</td>
<td>96.5</td>
<td>93.9</td>
</tr>
<tr>
<td></td>
<td>Non-speech</td>
<td>89.3</td>
<td>76.6</td>
<td>82.5</td>
</tr>
<tr>
<td>100msec</td>
<td>Speech</td>
<td>92.4</td>
<td>97.6</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>Non-speech</td>
<td>91.9</td>
<td>77.5</td>
<td>84.1</td>
</tr>
</tbody>
</table>

Table 3. Detection results without Step-1 (%)

<table>
<thead>
<tr>
<th>Source</th>
<th>Recall</th>
<th>Precision</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>94.0</td>
<td>90.4</td>
<td>92.2</td>
</tr>
<tr>
<td>Music</td>
<td>76.2</td>
<td>92.8</td>
<td>83.6</td>
</tr>
<tr>
<td>Noise</td>
<td>57.9</td>
<td>80.1</td>
<td>67.2</td>
</tr>
<tr>
<td>Silence</td>
<td>80.1</td>
<td>73.4</td>
<td>76.6</td>
</tr>
<tr>
<td>Average</td>
<td>86.7</td>
<td>86.7</td>
<td>86.7</td>
</tr>
</tbody>
</table>

6. EFFECT OF AUDIO SEGMENTATION FOR INDEXING BROADCAST NEWS

6.1 Indexing System [5]

The architecture of our indexing system with audio segmentation is shown in Figure 4. The system is composed of five modules: audio segmentation, speech recognition, topic segmentation, scene analysis, and information integration. The speech recognition module recognizes speech segments output by the audio segmentation module. Linguistic topic boundary candidates are detected by the topic segmentation...
6.2 Effect of Acoustic Segmentation on Indexing

Topic segmentation tests were carried out using 12 news programs. The speech recognition module used a 30k-word vocabulary trigram language model, which was trained using 600k sentences, and Gaussian-mixture triphone HMMs, which were trained using about 300 hours of speech. Speech recognition was carried out on speech segments detected by the audio segmentation module. The word error rate of all speech was 24.9%. In the topic segmentation module, the word conceptual vectors were trained using about 100k articles.

Table 4 shows the results of topic segmentation for four types of integration; AS-b means the results of segment boundaries for each type of sound sources detected in the audio segmentation module. TS(AS-a) means topic-boundary candidates detected among word sequences in speech segments as detected by the proposed audio source segmentation module. These were the recognition results of speech segments identified by the audio segmentation module. SC means the results of scene-change detection. Automatic topic segmentation results as taken from manually transcribed texts are also shown in this table. According to the table, the topic boundaries formed by topic segmentation (TS(AS-a)) are the best approach to detecting correct topic boundaries. This means that accurate speech segment detection is very important for indexing. The table also shows that the segment information (AS-b) for each sound source is useful.

7. CONCLUSION

This paper introduced new acoustic features and a new smoothing (segment detection) method for audio source segmentation. The proposed spectral correlation-based features improve detection performance remarkably. The smoothing method, which takes account of detection results provided by each parameter, also reduces the misdetection rate of the conventional method. Furthermore, the effect of the improved audio source segmentation approach was confirmed by experiments on broadcast news indexing.

Future work will be to improve noise detection performance through the use of multi-noise models.