DLF-based Speech Segment Detection and Its Application to Audio Noise Removal for Video Conferences

Kazuto Sasaki (member)†, Takahiro Ogawa (member)†, Sho Takahashi (member)†, Miki Haseyama (member)†

Abstract A new decision-level fusion (DLF)-based speech segment detection method and its application to audio noise removal for video conferences are presented in this paper. The proposed method calculates visual and audio features from video sequences and audio signals, respectively, obtained in video conferences. Features extracted from mouth regions of participants and attribution degrees of speech class are used as visual and audio features, respectively, and Support Vector Machine (SVM)-based classification is performed by using each kind of feature. The SVM classifier performs two-class classification of speech and non-speech segments to realize speech segment detection. From the detection results obtained from the visual and audio features, DLF based on Supervised Learning from Multiple Experts is performed to successfully obtain the final detection results with focus on the accuracy of each detection result. Then, from audio signals in the non-speech segments detected by our method, we can extract noise information to realize accurate audio noise removal in the speech segments.

Key words: speech segment detection, decision-level fusion, multimodal scheme, noise removal.

1. Introduction

In recent years, audiovisual video conference services have been used by many users due to the widespread use of mobile devices. Users can use such services anywhere, but audiovisual sequences tend to be degraded by background noises. Audio noises severely degrade the quality of services in video conferences, and their removal is an important issue in the field of multimedia processing.

A number of audio noise removal methods have been proposed, and their performance has also been improved by many researchers. If the target application of audio noise removal is video conferences, there are the following two useful characteristics.

1) Since there are conversations among participants, non-speech segments are also included.
2) Not only audio signals but also video sequences can be utilized.

From the first characteristic, we can obtain noise information from non-speech segments in audio signals. Since the other parts in audio signals correspond to speech segments, their noise removal becomes feasible by using the obtained noise information. Next, from the second characteristic, speech segments can be found by using not only audio signals but also video sequences, and improvement of speech segment detection using these two kinds of sequences can be expected. Detection of speech segments is important for subsequent noise removal, and it can be improved by collaboratively using audiovisual sequences obtained in video conferences. In this paper, we focus on the detection of speech segments in video conferences for audio noise removal.

Speech segment detection has been intensively studied by many researchers. Traditionally, such studies were conducted on the basis of audio signal analysis. Recently, with the rapid growth of multimodal schemes, many methods for audiovisual analysis have been proposed, and speech segment detection has been improved. However, there have been few studies showing that speech segment detection can definitely be improved by integration of different modalities, i.e., audiovisual sequences. One possible problem, for example, is that if the levels of accuracy of different modalities are different, simple fusion may cause degradation of speech segment detection.

In this paper, we propose a new decision-level fusion (DLF)-based speech segment detection method for audio noise removal in video conferences. In the proposed method, we first extract mouth regions of participants in video sequences to detect speech segments from their movements based on Support Vector Machine (SVM). Speech segment detection based on SVM using attribution degrees of four audio classes obtained by our previously reported audio classification method is also performed. From the results of speech segment detection using these two methods, we perform DLF based on Supervised Learning from Multiple Experts, which enables successful integration by focusing on the accuracy of each detection result. The proposed method can learn the relationship between audiovisual
features and the classification result obtained by each kind of modality with consideration of their classification performances. Since this DLF-based approach can consider how correctly each kind of modality can perform the classification for target audiovisual features, improvement in performance is guaranteed. Consequently, by using audio signals in the remaining non-speech segments, we can extract their pure noise information to realize audio noise removal of the speech segments.

The biggest contribution of the proposed method is multimodal audiovisual speech segment detection using the DLF scheme. On the other hand, several ensemble learning methods including Bagging \(13)\), Boosting \(14)\), Random Forest \(15)\) and Bayesian Network \(16)\) have been proposed. Among them, the proposed method adopts the DLF scheme proposed by Raykar et al. \(12)\) with consideration of the following points:

- Since it is difficult to prepare a number of training samples for each participant and each noise environment, the fusion scheme integrating several classification results should be successfully trained from a limited number of training samples.

- For several SVM-based classification results for which classification performances are different, the fusion scheme should provide a classifier that can consider the relationship between their performances and the features input to the classifier.

Since the DLF scheme based on Supervised Learning from Multiple Experts proposed by Raykar et al. satisfies the above points, this scheme is adopted in the proposed method. In the field of medical imaging, this method was originally proposed for realizing a computer-aided diagnosis (CAD) system that can integrate examination results obtained by radiologists whose examination accuracies are different. This method performs DLF with consideration of the features and the examination accuracy of each radiologist. Therefore, by regarding the radiologists as multiple SVM-based classifiers, this DLF scheme is suitable for our target problem in this paper.

The method presented in this paper is an improved version of our previously reported methods \(17,18)\). Specifically, the proposed method shown in this paper improves the single feature-based classification method using SVM by using improved features. Furthermore, a new DLF-based approach is introduced in the proposed method for obtaining the final detection results.

This paper is organized as follows. First, in Section 2, we present the multimodal speech segment detection method using audiovisual sequences obtained in video conferences. Experimental results are shown to verify the performance of the proposed speech segment detection method in Section 3. In this section, we also apply audio noise removal to the results obtained by the speech segment detection method to verify its effectiveness. Finally, concluding remarks are presented in Section 4.

2. DLF-based Speech Segment Detection

The DLF-based speech segment detection method is presented in this section. In the proposed method, we try to realize the detection of speech segments by classifying each segment in a target video including audio signals into two classes, i.e., speech segments and non-speech segments. Specifically, the proposed method consists of two phases. In the first phase, we construct two kinds of classifiers that perform the classification of speech segments and non-speech segments, the two classifiers being obtained from visual features extracted from video sequences and audio features extracted from audio signals, respectively. Two sets of classification results are obtained by using these classifiers. In the second phase, we use DLF to integrate the classification results with consideration of their classification accuracy. The details of each phase are shown in the following subsections, and a block diagram of the proposed method is shown in Fig. 1.

2.1 Audiovisual Speech Segment Detection

The first phase of the proposed method is presented in this subsection. The visual and audio features used in the proposed method are first explained, and then the single feature-based speech segment detection method used in the first phase is presented.
(1) Extraction of Visual Features

For each frame \((X \times Y \text{ pixels})\) of video sequences, the proposed method extracts visual features from the mouth region of a participant. First, the participant’s face is detected by using Adaboost\(^{19}\), and the lower part of the face is clipped as the mouth region \((X_m \times Y_m \text{ pixels})\) as shown in Fig. 2. Next, from the obtained mouth region, the HOG features that are robust to local geometric and brightness changes are extracted\(^{20}\). Specifically, the proposed method uses blocks of \(\frac{X_m}{3} \times \frac{Y_m}{3} \text{ pixels}\) and cells of \(\frac{X_m}{9} \times \frac{Y_m}{9} \text{ pixels}\), i.e., each block is composed of \(3 \times 3\) cells. These blocks are obtained at intervals of \(\frac{X_m}{6} \times \frac{Y_m}{6} \text{ pixels}\) (same as the cell size), and 16 blocks are totally extracted from the mouth region. Furthermore, for each block, a 9-bin gradient histogram is calculated from each cell. Therefore, since \(3 \times 3 \times 9 = 81\) features are obtained from each block, \(16 \times 81 = 1296\)-dimensional HOG features can be obtained.

In addition, the following two kinds of color histogram-based features are calculated from the obtained mouth region. First, a color histogram is calculated from the obtained mouth region, and the number of bins, i.e., the dimension of the features, is set to 512, where 8 bins are used for each RGB channel. Second, we calculate a 512-dimensional histogram intersection vector that is obtained between the above color histogram vector and a synthesized color histogram vector of non-speech segments, the synthesized color histogram vector of non-speech segments being obtained by calculating a vector median from color histogram vectors of non-speech segments that are detected by the audio features. This detection scheme based on audio features can be found in the remaining explanations of this subsection. Generally, since the performance for detection of speech segments based on audio features is better than that based on visual features, we use color histogram vectors of the non-speech segments detected by using the audio features for obtaining the synthesized color histogram vector of non-speech segments. Then the obtained histogram intersection vector becomes similar to background subtraction features.

In this way, we can finally obtain \(2320(= 1296 + 512 + 512)\)-dimensional visual features. The proposed method uses HOG and the color histogram, i.e., static visual features, for respectively representing lip shapes and their color distributions in order to determine whether the mouth is open or not. Furthermore, in order to detect mouth movements, histogram intersection features, i.e., non-static visual features, are calculated as the difference between the histogram features of each frame and those of non-speech segments. We use only the color histogram-based features to represent mouth movements for reducing computation complexity and do not calculate the HOG-based ones. Nevertheless, since it is important to adopt better spatio-temporal features such as Dynamic Texture\(^{21}\) or CHLAC (Cubic Higher-order Local Auto Correlation)\(^{22}\) for improving the detection performance using visual features, their improvement is necessary in our future work.

(2) Extraction of Audio Features

First, in the proposed method, audio signal segmentation is performed at a fixed interval, which is the same as the interval between two consecutive frames in video sequences. Then audio features are extracted from each audio segment by using our previously reported audio signal classification method\(^{11}\). From each audio segment, we calculate the attribution degree of this segment belonging to one of four audio classes: silence, speech, music and noise. Specifically, our previously reported method\(^{11}\) first extracts features including volume, zero-crossing rate, pitch, frequency centroid, frequency bandwidth, sub-band energy ratio, non-silence ratio and zero-ratio and performs their dimensionality reduction based on principal component analysis. Next, pre-clustering of target audio signals is performed on the basis of Mahalanobis generalized distance to obtain their attribution possibility for each cluster. Finally, audio signal classification using a Fuzzy algorithm\(^{23}\) is performed by using the obtained attribution possibilities, and then the attribution degrees of the four audio classes (silence, speech, music and noise) are calculated. Therefore, in the proposed method, these four attribution degrees are used as four-dimensional audio features of each audio segment.

(3) Single Feature-based Classification

We present the single feature-based classification method for speech segments and non-speech segments in the first phase. By the above-described procedures, the proposed
method can obtain visual and audio features. Then feature selection based on the Max-Relevance and Min-Redundancy (mRMR) algorithm is applied to the visual features to effectively obtain a visual feature vector for classification. On the other hand, the four-dimensional audio features are directly used to obtain an audio feature vector. In the proposed method, we calculate the raw low-level visual features. However, all of the features are not always effective for the classification of speech and non-speech segments. Thus, in the proposed method, we adopt the mRMR algorithm, which is one of the most representative feature selection algorithms, to remove such unnecessary visual features. On the other hand, the audio features are soft indexing results obtained by using our previously reported audio signal classification method. Therefore, since the audio features become higher-level features compared to the visual features, we do not apply the mRMR algorithm to the audio features.

In the proposed method, we separately construct two kinds of classifiers, i.e., classifiers using visual and audio feature vectors. Given a training dataset consisting of feature vectors \( \tilde{x}_i \in \mathbb{R}^d \) (\( i = 1, 2, \ldots, N \)) and their corresponding labels \( y_i \in \{1, -1\} \), the feature vectors \( \tilde{x}_i \) are mapped into a high-dimensional feature space to obtain \( \psi(\tilde{x}_i) \in \mathbb{R}^{d'} \), where we use a nonlinear map whose kernel function is the Gaussian kernel. Note that if the proposed method cannot detect the participant’s face, we cannot calculate the visual features, and SVM-based classification results are not also obtained. In such cases, the proposed method directly outputs classification results based on audio features as the final results without performing the following DLF procedures, and the final classification performance becomes the same as that by only using the audio features.

\begin{equation}
    f_{\text{svm}}(\tilde{x}) = v^T \psi(\tilde{x}) + \beta,
\end{equation}

where \( \tilde{x} \) is an input vector whose class label is unknown, i.e., speech segment or non-speech segment of \( \tilde{x} \) is unknown. The class label of \( \tilde{x} \) is determined according to \( y_{\text{svm}} = \text{sign}(f_{\text{svm}}(\tilde{x})) \). If \( y_{\text{svm}} \) is positive, \( \tilde{x} \) belongs to the positive class whose label is 1. Otherwise, it belongs to the negative class of \(-1\). In this way, SVM-based classification becomes feasible for the new input vector \( \tilde{x} \). The proposed method adopts SVM since it is one of the most benchmarking classifiers that have high generalization abilities for a limited number of training samples.

The proposed method separately uses the visual feature vector and the audio feature vector extracted from each segment in the target video as \( \tilde{x} \) and inputs them into the constructed SVM classifiers as shown in Eq. (1). Note that in the classification based on the visual feature vector, there exist some impulsive miss-detections. Thus, we prepare a sliding window size of 30 segments centered at a non-speech segment, and if this window includes more than 15 speech segments, we compensate the target non-speech segment to a speech segment. In this way, two classification results are obtained for each segment.

### 2.2 Decision-level Fusion of Multiple Results

The second phase of the proposed method, i.e., DLF of multiple results obtained by the procedures described in the previous subsection, is presented in this subsection. In the proposed method, we integrate the two sets of classification results based on Supervised Learning from Multiple Experts to obtain the final detection results for speech segments.

Since the DLF used in our method comes from the research field of computer-aided diagnosis (CAD), it merges multiple classification results from human annotators, e.g., radiologists. In the proposed method, we regard the two classifiers based on visual and audio features as two annotators. In order to integrate the two sets of classification results, we focus on the classification accuracy of each annotator and assign higher weights to classification results of annotators that have a higher level of classification accuracy. The details are shown below.

Note that if the proposed method cannot detect the participant’s face, we cannot calculate the visual features, and SVM-based classification results are not also obtained. In such cases, the proposed method directly outputs classification results based on audio features as the final results without performing the following DLF procedures, and the final classification performance becomes the same as that by only using the audio features.

\begin{enumerate}
  \item Each annotator’s classification accuracy and classification model
  
  We explain the classification accuracy of each annotator and the classification model defined in our method. Let \( y^j \in \{0, 1\} \) be the label assigned to the feature vector \( x \) by annotator \( j \in J \), where \( J = \{v, a\} \) is a set of annotators, and “v” and “a” correspond to Visual and Audio, respectively. Furthermore, \( y^j = 0 \) and 1 correspond to non-speech and speech segments, respectively.

  The details of \( x \) are shown in the following paragraph. Given the actual label \( y \in \{0, 1\} \), i.e., ground truth, the levels of classification accuracy of the two annotators, \( P_{se} \) (sensitivity) and \( P_{sp} \) (specificity), are respectively defined as follows:
\end{enumerate}
\[ P_{se}^i := \Pr[y^i = 1 | y^i = 1], \quad P_{sp}^i := \Pr[y^i = 0 | y^i = 0]. \tag{2} \]
\[ f_w(x) = w^\top x, \tag{4} \]

where \( w \) is a weight. In the second phase of our method, the feature vector \( x \in \mathbb{R}^r \) is a two-dimensional vector whose elements are the SVM outputs obtained in Eq. (1) in the classification based on visual and audio features, i.e., \( r = 2 \). Several methods that focus on the SVM outputs have been proposed \(^{27} \). In addition, it has been reported that the SVM outputs can be used as important features for the DLF, we adopt them for \( x \) in Eq. (4). The final classification result \( \hat{y} \) is obtained as follows:

\[ \hat{y} = \begin{cases} 1 & f_w(x) \geq T_h \\ 0 & \text{otherwise,} \end{cases} \tag{5} \]

where \( T_h \) is a predetermined threshold. As mentioned above, \( \hat{y} = 1 \) and \( 0 \) represent speech and non-speech segments, respectively, i.e., they correspond to the class labels \(+1\) and \(-1\) of SVM, respectively, shown in the previous subsection.

(2) Training Stage

Given the training data \( D \) consisting of \( N \) feature vectors with the classification results by the two annotators \( D = \{ x_i, y^s_i, y^a_i \}_{i=1}^N \) and their corresponding actual labels \( y_i \) \((i = 1, 2, \ldots, N)\). In order to determine the discriminating function \( f_w(x) \), we have to obtain \( w \) from the training data by using each annotator’s classification accuracy defined in Eqs. (2) and (3). The details are shown below.

From the training data \( D \), the likelihood of \( w \) is defined as

\[ \Pr[D|w] = \prod_{i=1}^N \Pr[y^s_i, y^a_i | x_i, w]. \tag{6} \]

By using the values of sensitivity \( P_{se} = \{P_{se}^j | j \in J\} \) obtained from each annotator and those of specificity \( P_{sp} = \{P_{sp}^j | j \in J\} \), the above equation is rewritten as follows:

\[ \Pr[D|w] = \prod_{i=1}^N \left[ \Pr[y^s_i, y^a_i | y_i = 1, P_{se}] \Pr[y_i = 1 | x_i, w] \\
+ \Pr[y^s_i, y^a_i | y_i = 0, P_{sp}] \Pr[y_i = 0 | x_i, w] \right]. \tag{7} \]

If it is assumed that each annotator \( j \in J \) is independent of each other, \( \Pr[y^s_i, y^a_i | y_i = 1, P_{se}] \) can be rewritten as follows:

\[ \Pr[y^s_i, y^a_i | y_i = 1, P_{se}] = \prod_{j \in J} \Pr[y^s_i | y_i = 1, P_{se}^j] = \prod_{j \in J} [P_{se}^j | 1 - P_{se}^j]^{y_i}. \tag{8} \]

In the same manner, \( \Pr[y^s_i, y^a_i | y_i = 0, P_{sp}] \) can be rewritten as follows:

\[ \Pr[y^s_i, y^a_i | y_i = 0, P_{sp}] = \prod_{j \in J} \Pr[y^a_i | y_i = 0, P_{sp}^j] = \prod_{j \in J} [P_{sp}^j | 1 - P_{sp}^j]^{y_i}. \tag{9} \]

Then the likelihood in Eq. (7) is rewritten as

\[ \Pr[D|w] = \prod_{i=1}^N \left[ a_i p_i + b_i (1 - p_i) \right]. \tag{10} \]

Note that

\[ p_i = \Pr[y_i = 1 | x_i, w] = \sigma(w^\top x_i) = \frac{1}{1 + \exp(-w^\top x_i)}, \tag{11} \]

and

\[ a_i = \prod_{j \in J} [P_{se}^j | 1 - P_{se}^j]^{y_i}, \tag{12} \]

\[ b_i = \prod_{j \in J} [P_{sp}^j | 1 - P_{sp}^j]^{y_i}. \tag{13} \]

The maximum-likelihood estimator is found by maximizing the following log-likelihood:

\[ \hat{w}_{ML} = \arg \max_w \left\{ \ln \Pr[D|w] \right\}. \tag{14} \]

Note that the DLF method proposed by Raykar et al.,\(^{12} \) is an unsupervised learning approach. On the other hand, since the ground truth is available in the training stage of the proposed method, we perform reformulation of the target problem. Specifically, Raykar et al. proposed an unsupervised DLF method that can perform integration of several classification results under the condition that their ground truth is unknown. Then the above equations shown in this subsection can be derived even if the ground truth \((y \text{ and } y_i)\) is unknown. Although the method proposed by Raykar et al. tries to estimate the probabilistic estimate of \( y_i \) \((i = 1, 2, \ldots, N)\), their labels have been provided as the training data in advance in our method. Therefore, the proposed method uses the labels \( y_i \) directly and performs the supervised learning by using \( y_i \) as the ground truth. Let \( y = [y_1, \ldots, y_N] \) be the set of the actual labels, and the complete data log-likelihood can be written as

\[ \ln \Pr[D, y|w] = \sum_{i=1}^N [y_i \ln p_i a_i + (1 - y_i) \ln (1 - p_i) b_i]. \tag{15} \]
The vector $w$ is estimated by maximizing this likelihood. Specifically, by solving
\[
\frac{\partial}{\partial w} \ln \mathbb{P}(D, y|w) = 0,
\]
we obtain the estimated weight vector $w$ as
\[
w \leftarrow w - \eta \mathbf{H}^{-1} g.
\]
In Eq. (17), $g$ is a gradient vector, $\mathbf{H}$ is a Hessian matrix and $\eta$ is a step length. The gradient vector $g$ and the Hessian matrix $\mathbf{H}$ are respectively obtained as follows:
\[
g = \sum_{i=1}^{N} [y_i - \sigma(w^\top x_i)] x_i,
\]
\[
\mathbf{H} = -\sum_{i=1}^{N} [\sigma(w^\top x_i)][1 - \sigma(w^\top x_i)] x_i x_i^\top,
\]
where $\sigma(w^\top x_i)$ is shown in Eq. (11). In the proposed method, we perform the iteration shown in Eq. (17) until $\|\Delta w_i\|^2 < 0.001$, where $\Delta w_i = w_i - w_{i-1}$, and $w_i$ is $w$ obtained in the $i$th iteration.

3. Testing Stage

Given the test data $[x, y^v, y^s]$ whose label $y$ is unknown, the final classification result can be obtained as follows. We first calculate
\[
p = \frac{1}{1 + \exp(-w^\top x)}.
\]
Furthermore,
\[
a = \prod_{j \in F} [P_{se}^j y^v | 1 - P_{se}^j]^{1 - y^v},
\]
\[
b = \prod_{j \in F} [1 - P_{sp}^j y^v | 1 - P_{sp}^j]^{y^v},
\]
where $P_{se}^j$ and $P_{sp}^j$ of annotator $j$ are calculated from training data, and $y^v$ is the classification result of the test data by annotator $j$. Therefore, we obtain the final classification result considering each annotator’s accuracy. Then $\mu$ is computed by using $p$, $a$ and $b$ as follows:
\[
\mu = \mathbb{P}[y = 1|y^v, y^s, x, w] \\
\approx \mathbb{P}[y^v, y^s|y = 1, w] \mathbb{P}[y = 1|x, w] \\
= \frac{ap}{ap + b(1 - p)}.
\]
The value of $\mu$ corresponds to a soft probabilistic estimate of the actual ground truth. Finally, we obtain the final classification result as
\[
\hat{y} = \begin{cases} 1 & \mu \geq T_\mu \\ 0 & \text{otherwise} \end{cases}
\]
instead of using Eq. (5), where $T_\mu$ is a pre-defined threshold. In this way, we can realize speech segment detection from the above classification results. Finally, we also perform sliding window-based compensation to avoid impulsive miss-detection in the same manner as the visual feature-based classification.

3. Experimental Results

In this section, we show some experimental results in order to verify the performance of the proposed method. In the experiment, we used 10 videos, each video being about 30 seconds, $1280 \times 720$ pixels, 30 fps, and 44.1 kHz. Therefore, the time length of each segment became $\frac{1}{30}$ second. Among these 10 videos, five videos were obtained from conversations by males and the remaining five videos were obtained from conversations by females. Each video contained approximately 15-second speech segments, i.e., the remaining approximately 15 seconds corresponded to non-speech segments.

In this section, we first show results of the speech segment detection in 3.1. Next, results of noise removal realized by using the speech segment detection results obtained in 3.1 are shown in 3.2.

3.1 Results of Speech Segment Detection

In this subsection, we verify the performance of the proposed speech segment detection. As described above, we used 10 videos and employed the leave-one-out scheme for performance verification. In this experiment, we performed a comparison with unimodal methods using only video sequences or audio signals, i.e., we simply used the two sets of detection results obtained in the single feature-based classification. We also compared the performance of our method with that of an early fusion (feature-level fusion)-based method and that of a late fusion-based method using majority voting, the details of which are shown later. It is reasonable to adopt these comparisons since our main contribution is the DLF-based approach. Note that in the proposed method, we employed the Gaussian kernel for SVM in all methods, with its kernel parameter being automatically determined by a grid search. For evaluating the accuracy of our speech segment detection, we used Recall, Precision and F-measure defined as follows:
\[
\text{Recall} = \frac{N_c}{N_m},
\]
\[
\text{Precision} = \frac{N_c}{N_a},
\]
\[
\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}},
\]
where $N_c$ is the number of segments detected correctly as...
speech segments, $N_m$ is the number of true segments judged manually as speech segments, and $N_a$ is the number of segments detected by the method automatically as speech segments.

In this experiment, we added three kinds of noises, white noises, those of an air-conditioner and those of a computation server, to the audio signals of the videos with changes in their SNR. From these videos including noisy audio signals, we performed speech segment detection and calculated Recall, Precision and F-measure from the detection results. An example of the detection results is shown in Fig. 3. Furthermore, the relationship between SNR of noisy audio signals and detection performance is shown in Fig. 4. As mentioned above, the following methods were adopted in this experiment to compare the performance with that of the proposed method:

(i) Method using only video sequences
(ii) Method using only audio signals
(iii) Early fusion-based method
(iv) Late fusion-based method using majority voting

The first and second methods (i) and (ii) were implemented by only using Section 2.1. The third method uses the integrated vector concatenating both of the visual and audio features and inputs it to SVM to obtain the final detection result. We regard this feature-level fusion as early fusion. The fourth method performs majority voting. Note that when the SVM classification results of visual and audio features are different, we compare probabilities calculated from the SVM outputs and output the result for which the probability is larger.

Since Recall and Precision have a trade-off relationship, we focus on F-measure for comparing the proposed method with the other methods. From the obtained results shown in Figs. 3 and 4, the proposed method tends to output better results than those obtained by the other methods. Therefore, it can be seen that the proposed DLF-based method is effective for realizing successful speech segment detection. Since the proposed method enables integration of the two sets of detection results with consideration of the accuracy of each modality, performance improvement can be guaranteed.

In Fig. 4, F-measure of the proposed method tends to become worse than that of comparative method (ii) when the target noise is white noise. In the case of white noise, the performance of comparative method (ii) using only audio signals becomes higher than the performance in the case of noise of an air-conditioner and that in the case of a computation server. Although the proposed method enables DLF that can consider the classification performance of each modality, its model is the linear classifier whose input is the two-dimensional feature vector consisting of the SVM outputs as shown in Eq. (4). Thus, since its representation performance is limited, our DLF-based approach cannot outperform comparative method (ii), the performance of which is sufficiently high. For solving this problem, it is necessary to introduce the kernel method in the classification model of the DLF or add new features to $x$ in Eq. (4), and these improvements are our future work.

Finally, we show computation times of the proposed method. The computation time of each procedure taken for processing each segment for which the time length is 30 seconds is shown in Table 1. From the obtained results, we can see that high-speed calculation is realized in the single feature-based classification using SVM and DLF. On the other hand, computation of the visual features is time-
3.2 Results of Noise Removal

Application of the proposed speech segment detection method to noise removal is shown in this subsection. From the detection results described in the previous subsection, we tried to obtain noise information from the non-speech segments to realize noise removal in the audio signals. In the proposed method, we used a simple Spectral Subtraction method. Spectral Subtraction is a method for recovering the noise-free audio signal from a signal with added noise by performing an inverse Fourier transform of the difference between the power spectrum of the noise and the power spectrum of the noisy signal. The noise detected in the non-speech segments can then be removed from the speech segments, and the original signal can be estimated. Due to the limitation of pages, we simply give an overview of the above method. Details of the mathematical formulations can be found in the original paper.

We extracted noise information from the non-speech seg-
ments, which were obtained by performing speech segment detection shown in the previous subsection, and removed the noise from the target audio signal by using the Spectral Subtraction method. Thus, recall and precision shown in the previous subsection are respectively related to removal of noises from speech segments and extraction of noise information from non-speech segments. In order to perform successful noise reduction, we have to detect speech segments with high recall rates, i.e., short false positives are not important in the following noise reduction. Thus, as shown in an example of Fig. 3(c), the short false positives caused in comparative method (ii) will not affect the performance of the following noise reduction. On the other hand, since false negatives in the speech segments severely affect the following noise reduction, it can be seen that the proposed method, the recall rates of which are the highest with preservation of the high F-measure as shown in Fig. 4, is effective from practical viewpoints. This means that the proposed speech segment detection method is suitable for noise reduction.

Figure 5 shows the results of noise reduction obtained by the proposed method and the comparative methods shown in the previous subsection. Specifically, this figure shows the relationship between SNR of the noisy speech signal and SNR of the recovered speech signal. In this experiment, we calculated SNR from only the true speech segments.

From the obtained results, we can see that successful noise removal using the noise information in the non-speech segments obtained by the proposed method becomes feasible compared to the other comparative methods. These results are closely related to the speech segment detection results shown in the previous subsection. Therefore, the results of this experiment indicate that it is important for noise removal to accurately detect speech segments and non-speech segments.

4. Conclusions

In this paper, we have presented a new DLF-based speech segment detection method and its application to audio noise removal for video conferences. The proposed method tries to realize accurate speech segment detection using audio-visual features extracted from videos including both video sequences and audio signals. Therefore, by using these two kinds of signals, we separately obtain detection results and perform their DLF. In this approach, we introduce Supervised Learning from Multiple Experts into the DLF and achieve successful detection that can consider the level of accuracy of the detection results obtained from these two kinds of signals. Then, by applying the detection results to noise removal, the proposed method also realizes more successful estimation of the original signals than do methods only using video sequences or audio signals. Experimental results show that the proposed method realizes more accurate speech segment detection and noise removal.

5. ACKNOWLEDGEMENT

This work was partly supported by Grant-in-Aid for Scientific Research (B) 25280036, Japan Society for the Promotion of Science (JSPS).

References

5) D. Ying, X. Lu, J. Li, Y. Yan, J. Dang and F. Soong, “Noise estimation using a


