Detecting Utterance Scenes of a Specific Person

Kunihiko Sato
The University of Tokyo
Tokyo, Japan
kunihiko.k.r.r@gmail.com

Jun Rekimoto
The University of Tokyo
Sony Computer Science Laboratory
Tokyo, Japan
rekimoto@acm.org

ABSTRACT
We propose a system that detects the scene, where a specific speaker is speaking in the video, and displays the site as a heat map in the video's timeline. This system enables users to skip to the timeline they want to hear by detecting scenes in a drama, talk show, or discussion TV program, where a specific speaker is speaking. To detect a specific speaker's utterance, we develop a deep neural network (DNN) to extract only a specific speaker from the original sound source. We also implement the detection algorithm based on the output of the proposed DNN and the interface for displaying the detection result. We conduct two experiments on the proposed system. One is to confirm how much the amplitude of the other sounds can be suppressed and how much that of the specific person's utterance does not be suppressed by the proposed DNN. The second experiment is to confirm how accurately the proposed system can detect the utterance scene of a specific person.

Author Keywords
Scene detection; timeline; video; sound source separation; deep learning.

ACM Classification Keywords
H5.1. Information interfaces and presentation (e.g., HCI): Multimedia Information Systems; H5.2. Information interfaces and presentation (e.g., HCI): User Interface

INTRODUCTION
The demand for video streaming services, such as YouTube, Netflix, and Amazon Prime, is increasing as well as the amount of video contents on the Web. In this situation, in which too many videos have already been uploaded on the Web, the importance of supporting users to browse videos efficiently has also increased.

One method for efficient video browsing is fast-forwarding. Several researchers developed a content-aware fast-forwarding technique that dynamically changes playback speeds depending on the importance given to each video frame. This technique is enabled using key clips [1, 2], a skimming model [3], and the viewing histories of other people [4]. Direct manipulation techniques enable users to manipulate object positions in video frames to seek for specific video timelines [5, 6, 7, 8]. Video streaming services, such as YouTube, Netflix, and Amazon Prime, show a tiny picture of the video in relation to where the playhead is at in the timeline.

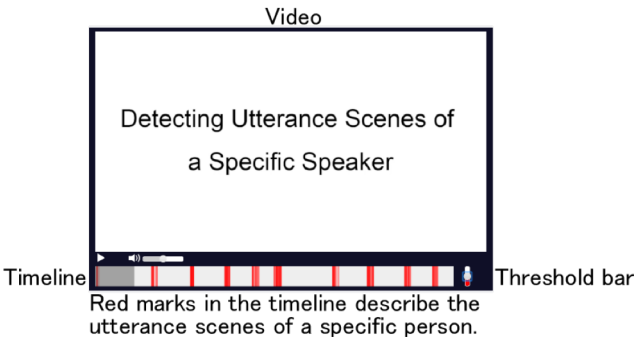

Figure 1. Proposed interface. The red marks in the timeline describes the utterance scenes of a specific person. The threshold bar can change the threshold of the scene detection algorithm.

Several studies on video navigations have used audio information. Conventional methods [9] using audio information summarize and classify videos based on silence, speech, and music. CinemaGazer [10] is an audio-based technique, which fast-forwards scenes without speech. This technique can only distinguish whether or not the scene includes speech, and cannot distinguish who speaks. As described, some studies supported video browsing using a sound class, but fewer audio-based methods have been used to seek specific video timelines than image or metadata-based methods.

We propose a system that detects the scene, where a specific speaker is speaking in the video, and displays the site as a heat map in the video's timeline, as shown in Figure 1. This system enables users to skip to the timeline they want to hear by detecting scenes in a drama, talk show, or discussion TV program, where a specific speaker is speaking. To detect a specific speaker's utterance, we develop a deep neural network (DNN) to extract only a specific speaker from the original sound source. Leveraging this sound source separation DNN, the system operates as follows: first, the system's DNN extracts the utterance of a specific person from the audio file of the target video and diminishes other sounds. As a result of DNN filtering, the

WII'18, March 11, Tokyo, Japan.

amplitude of the scene, in which the target person is speaking, does not become very small, while that of the other scenes becomes small. The system then calculates the difference between the amplitude of the original sound waveform and that of the filtered sound waveform. The system judges that scenes with the larger difference than a threshold are where the target person does not speak and those with the smaller difference are where the target person utters. The scenes, where the target person speaks, are displayed on the video timeline as a heat map based on the judgment result.

We conduct two experiments on the proposed system. One is to confirm how much the amplitude of the other sounds can be suppressed and how much that of the specific person's utterance does not be suppressed by the sound source separation DNN extracting only the specific person's utterance. The second experiment is to confirm how accurately the system can detect the utterance scene of a specific person.

Our contributions are summarized as follows.

- We propose a novel system that automatically detects the utterance scene of a specific person. We also confirm how accurately the system can detect the utterance scene of a specific person.
- We develop a sound source separation DNN that can extract only a specific person's utterance, and propose how to create a training dataset for the DNN. Many studies successfully tackled monaural sound source separation. However, these prior studies only confirmed the effects for separation between distinguished classes such as "speech and noise", or between multi-speakers. These studies did not clarify whether only a specific speaker can be separated when both diverse and various sounds are mixed in the sound source. We confirm how much the amplitude of the other sounds can be suppressed and how much that of the specific person's utterance does not be suppressed by the proposed DNN.

RELATED WORK

Browsing Support for Videos

Various techniques to support users in browsing videos are well studied. Fast-forwarding techniques, such as those in [11, 12], are useful in helping users watch videos in a reduced time. Several researchers also developed a content-aware fast-forwarding technique that dynamically changes playback speeds depending on the importance given to each video frame. Higuchi et al. [1] proposed a video fast-forwarding interface that helps users find important events from lengthy first-person videos continuously recorded with wearable cameras. The proposal of Pongnumkul et al. [2] makes it easy to find the scene change when sliding the video seek bar. Cheng et al. [3] proposed a video system to learn the user's favorite scene for fast-forwarding. Kim et al.'s method [4] shows the importance scene based on the viewing histories of other people. CinemaGazer [10] is an audio-based technique that fast-forwards scenes without speech.

Several techniques for indicating potential information in the video were also studied. These included spatio-temporal volume [13], positional information [14], and video synopsis [15, 16, 17]. Meanwhile, direct manipulation techniques enable users to manipulate object positions in video frames to seek for specific video timelines [5, 6, 7, 8]. Video lens allows users to interactively explore large collections of baseball videos and related metadata [18]. On-demand video streaming services, such as YouTube, Netflix, and Amazon Prime, show a tiny picture of the video in relation to where the playhead is at in the timeline.

Unlike the previous studies, ours focuses on providing an efficient method of allowing users to skip to the scenes, where a specific person that the user is searching for, is speaking.

Monaural Source Separation

Monaural sound source separation studies are closely related to the proposed method. We introduce these methods here and show their difference from the proposed method.

Wiener filtering is a classical method used for separating a specific sound source from a source waveform [19]. The Wiener filtering method heuristically determines parameters; hence, the parameters cannot be optimized for various sound sources [20].

In recent years, many studies attempted to separate monaural sound sources using deep learning. Previous deep network approaches [21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31] to separation showed promising performances in scenarios with sources belonging to a distinct signal class, such as "speech and noise" and "vocal and accompaniment." in addition, many researches attempted to separate multi-speakers using DNN [22, 32, 33, 34, 35, 36, 37]. These studies performed well in the speaker-dependent separation of two or three speakers. Deep clustering [29, 38, 39, 40] is a deep learning framework that can be used for a speaker-independent separation of two or more speakers, with no special constraint on vocabulary and grammar.

In spite of the advantages, these prior studies confirmed only the effects for separation between distinguished classes or between multi-speakers. The function required in the proposed approach is to isolate only the speech of a specific person from the sound source, including various noise and multiple speakers.

Speaker Recognition & Audio Event Detection

The speaker recognition technique seems effective in detecting the utterance section of a specific speaker. These techniques using phonemes [41, 42] perform well. However, speaker recognition methods are weak against noise. In addition, the shorter the input speech duration, the lesser the

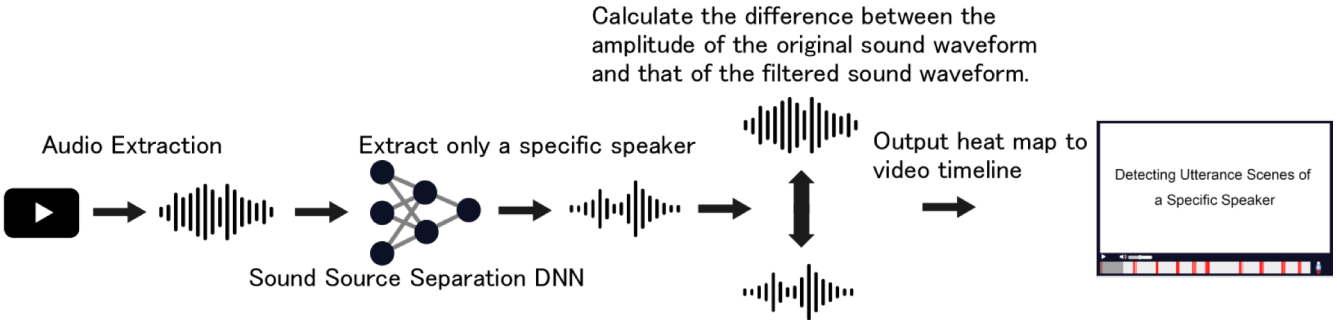

Figure 2. Proposed system's process. The system operates as follows: first, the system extracts the audio waveform from the target video. The system's DNN extracts the utterance of a specific person from the audio file of the target video and diminishes other sounds. The system calculates the difference between the amplitude of the original sound waveform and that of the filtered sound waveform. The system judges that scenes with a difference larger than a threshold are where the target person does not speak and those with a smaller difference are where the target person utters. The scene, where the target person speaks, is displayed on the video timeline as a heat map based on the judgment result.

speaker recognition precision. Ranjan et al. [43] reported that the equal error rate (false negative rate equals the false positive rate) becomes close to 40% when the input duration is 3 s. As described, this is not suitable for detecting the utterance scene of a specific speaker in videos because speaker recognition is vulnerable to noise and tiny duration input.

Jansen et al. [44] proposed the method for detecting recurring audio events in YouTube videos using a small portion of the manually annotated audio data set [45]. However, this method cannot distinguish who speaks while can distinguish between categories of sound, such as human voice and whistle.

IMPLEMENTATION

The proposed system detects the scene, where a specific speaker is speaking in the video, and displays the site as a heat map in the video's timeline. Figure 2 shows the system's process. The system first loads the sound of the target video once. Leveraging a DNN, the system then extracts only the specific speaker from the original sound source and diminishes the other sounds. In the sound waveform filtered by the DNN, the amplitude of the scene, where the target person is speaking, does not become too small, while that of the other scenes becomes small. The system calculates the difference between the amplitude value of the original sound waveform and that of the filtered sound waveform. The system then judges that scenes with the larger difference than a threshold are where the target person does not speak and those with the smaller difference are where the target person utters. The scenes, where the target person speaks, are displayed on the video timeline as a heat map based on the judgment result. The following subsections describe the implementation of the proposed sound source separation DNN, the detection and the interface.

	Class based	**Speaker separation**	**Proposed**
Relationship between separated sound sources	Speech-noise	Speaker-speaker	Specific speaker-others, including noise and other speakers

Table 1. Difference between the previous sound source separation and the proposed methods.

Sound Source Separation between a Specific Speaker and Other Sounds

We propose a DNN to detect the utterance of a specific person and separate this utterance from the other sounds. The difference of this DNN from the previous sound source separation methods is that the relationship between the separated sound sources is different as shown in Table 1. Many previous studies tackled the separation with different classes of sound sources, such as "sound and noise" and a fixed number of sound sources, such as "two or three speakers."

However, we assumed that the DNN models of the previous studies could be applied to our task if we change the training data. Therefore, we surveyed previous studies, and found that Rethage's method [31] was appropriate because it used a convolutional-based neural network, which allowed for parallel computation. Many previous methods [22, 23, 24, 25] employed recurrent neural networks

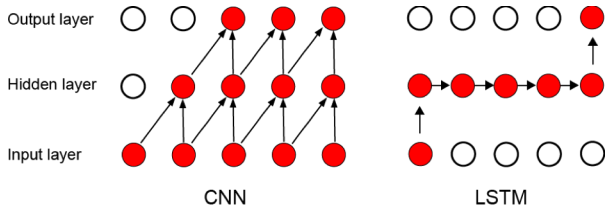

Figure 3. Diagrams showing the computational structure of typical CNN and LSTM architectures. Red signifies convolutions or matrix multiplications. The computation of LSTMs at each timestep is dependent on the results from the previous timestep. This why it is difficult to implement LSTMs using parallel processing.

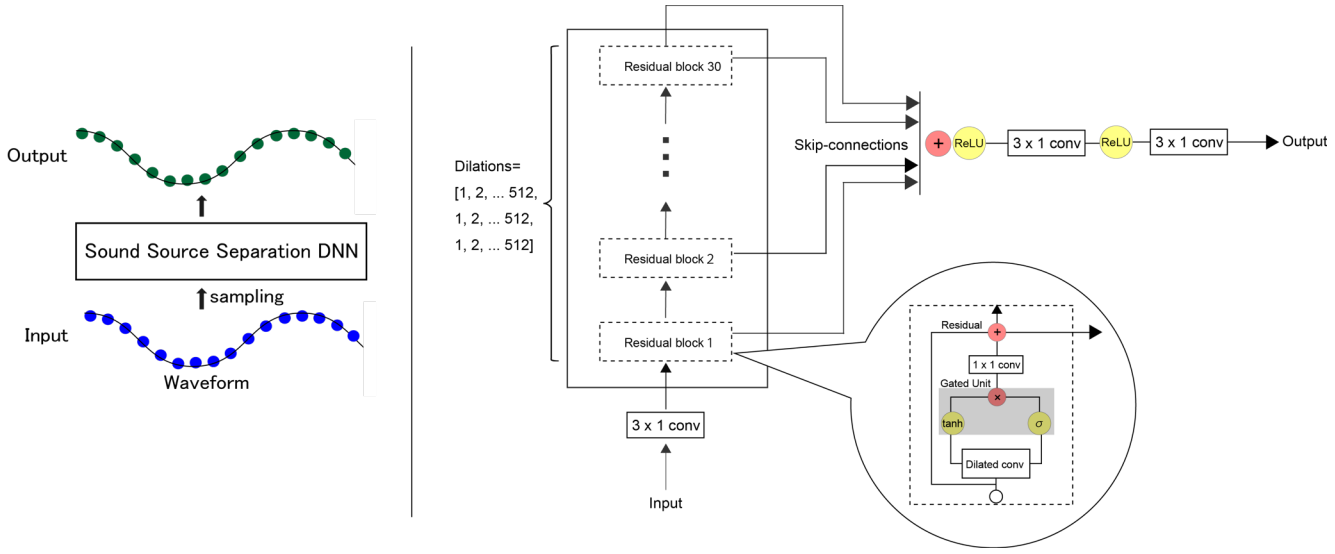

Figure 4. Left: Schematic diagram of the sound source separation DNN model. The waveform data is used as-is for input and output without using the features of the frequency domain. Right: Implementation details of the sound source separation DNN.

(RNNs), including long short-term memory (LSTM) networks, for source separation. As shown in Figure 3, the limitation of RNNs is that it is difficult for them to perform parallel computations because the computations at each timestep depend on the results from the previous timestep. Many videos on the web are several hours long; thus, the lack of parallel computations causes a significant problem of the processing time being linearly proportional to the video length. Furthermore, as the authors of deep clustering [38] reported, the most serious problem is that the LSTM performs poorly in the sound source separation of speakers, who are not in the training data.

To realize the proposed DNN, we devised a training dataset. As input data, we created the sound mixtures by merging the target speaker with the various environmental noises and the other speakers. We set clean speech of the target speaker as the ideal output value. By training the dataset, the proposed DNN was able to extract the speech of the target speaker and mute other sounds.

We implemented Rethage's DNN model as written in their article. Figure 4 shows the visualization of the implementation. The model is trained to extract a specific speaker by inputting and outputting the waveform data as-is. Their approach incorporated some techniques used in WaveNet [46], such as gated unit, skip connections, and residual blocks. The DNN model features 30 residual blocks. The dilation factor in each layer increases in the range 1, 2, ..., 256, 512 by powers of 2. This pattern is repeated thrice (three stacks). Prior to the first dilated convolution, the one-channel input is linearly projected to 128 channels by a standard 3 × 1 convolution to comply with the number of filters in each residual layer. The skip connections are 1 × 1 convolutions, which also feature 128 filters. A rectified linear unit (ReLU) is applied after summing all skip connections. The final two 3 × 1 convolutional layers are not dilated; contain 2048 and 256 filters, respectively; and are separated by a ReLU. The output layer linearly projects the feature map into a single-channel temporal signal using a 1 × 1 filter.

Detection

After the voice of a specific speaker is extracted by the sound source separation DNN, the algorithm for detecting the utterance scene of the speaker operates as follows: the algorithm segments the original and the filtered sound waveforms into certain window size, as shown in Figure 5. Then this algorithm calculates the difference between the amplitude value of both segments. This calculation aims to obtain the amplitude ratio of the original and filtered waveforms. The amplitude difference is obtained by the following equation:

$$diff\ [dB] = 20log_{10}\frac{A_{RMS\ (Original)}}{A_{RMS\ (Filtered)}}$$

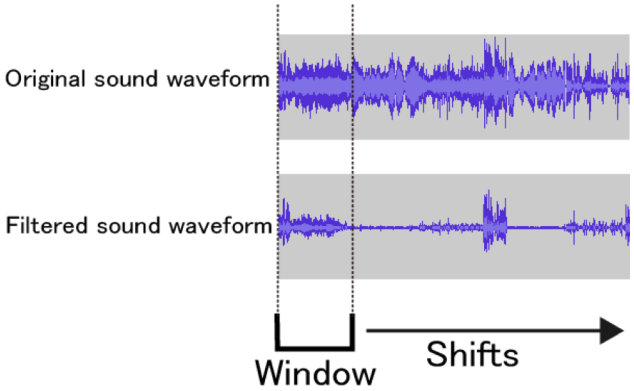

Figure 5. Visualization of segmenting the original and the filtered sound waveform into certain window size.

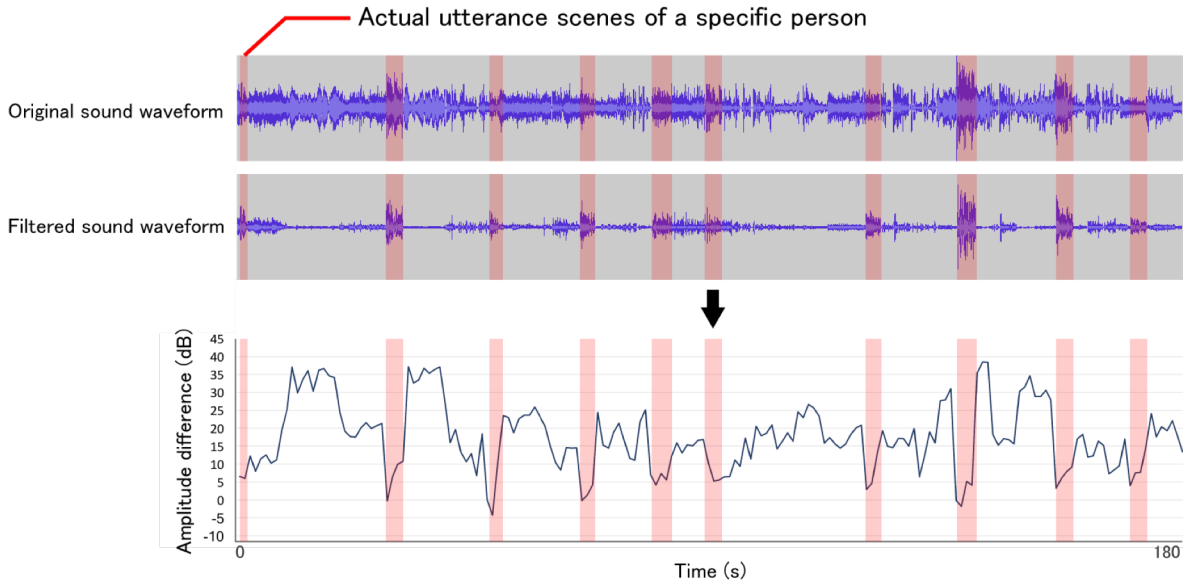

Figure 6. The line graph is plotted with the difference between the amplitude of the original sound waveform and that of the filtered sound waveform as the vertical axis, and time as the horizontal axis. The pale red marks represent actual utterance scenes of a specific person. The graph suggests that the amplitude difference in the utterance scenes of a specific person is smaller than that in the other scenes.

where $A_{RMS\,(Original)}$ represents root mean square of the amplitude of the original waveform segment and $A_{RMS\,(Filtered)}$ represents root mean square of that of the filtered waveform segment. The difference value (dB) indicates how much the amplitude of the original sound is attenuated after that is filtered by the proposed DNN. A small difference value means that the amplitude of the original sound is not much attenuated and a large difference value means that the amplitude is greatly attenuated. Leveraging the proposed DNN, the amplitude in the scenes, in which the target person is speaking, does not become very small (the difference is small), while that in the other scenes becomes small (the difference is large) as shown in Figure 6. Therefore, the algorithm can judge that the scenes with the larger difference than a threshold are where the target person does not speak, while those with the smaller difference are where the target person utters. After the judgement, the window shifts to the next segments. The abovementioned operation is repeated until the window reaches the end of each waveform.

The default value of the threshold is determined based on the average amplitude ratio of the original and filtered waveforms. This default value will be clarified by Experiment 1, which is described later.

Interface

After the speaking scenes of specific speakers are clarified, these scenes are displayed on the timeline as a heat map.

Figure 7. Left: The amount of red marks is decreased by lowering the bar. Right: The amount of red marks is increased by raising the bar.

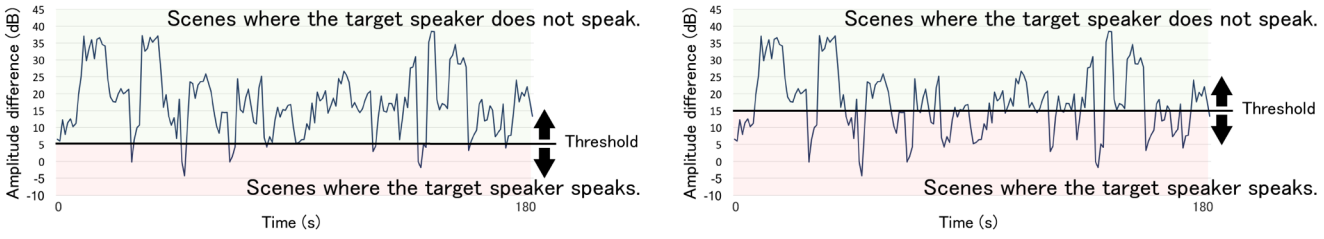

Figure 8. The figures visualize how the judgment for detecting the utterance scenes of the specific speaker changes when the threshold changes. These line graphs are the same as in Figure 6. When the threshold becomes lower, the number of scenes judged to be where the target person speaks, is decreased. When the threshold becomes higher, the number of scenes judged to be where the target speaks, is increased.

The red marks on the heat map represent the detected scenes. The user can jump to the scene uttered by the specific speaker by clicking the red mark position.

In addition, the user can change the threshold of the detection algorithm by operating the bar on the right side of the interface. Figure 7 shows the difference in the appearance of the heat map by operating the bar. Figure 8 shows how the judgment for detecting the utterance scenes of the specific speaker changes when the threshold changes. The amount of red marks in the timeline is decreased by lowering the bar because the threshold becomes lower. Only the scenes with a higher probability as the utterances of the specific speaker can be displayed. The amount of red marks is increased by raising the bar because the threshold becomes higher. The scenes with a low probability as a specific speaker's utterance may be included in the heat map, but this prevents the user from missing the scene of the speaker's utterance.

EXPERIMENT 1

This experiment is to confirm how much the amplitude of the other sounds can be suppressed and how much that of the specific person's utterance does not be suppressed by the sound source separation DNN extracting only the specific person's utterance. The ideal result is that the target speaker' utterance does not become very small but the other sounds become smaller. If the result is as described above, it can be said that the proposed DNN extracts only the utterance of the target speaker.

We let the sound source separation DNN model learn with the following setup. Then, we calculated how much of the decibel (dB) of the other sounds could be suppressed using the test dataset.

Setup

dataset

We created a training dataset of sound mixtures using noises from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [47], and utterances from TIMIT corpus [48] and CMU ARCTIC corpus [49]. Figure 9 describes the visualization of creating the training dataset. The target speaker of the detection was supplied by the CMU ARCTIC corpus. The subset of the CMU corpus we used features two native English speakers, including a man (ID: RMS) and a woman (ID: SLT). Note that it is common in speech research such as voice conversion that the target speakers are two. We randomly chose 593 sentences, which corresponds to 30 minutes, from each speaker for the training samples.

We mixed the training samples of each target speaker with the noise sounds provided by DEMAND. The subset of DEMAND that we used provided recordings in 17 different environmental conditions, such as in a park, a bus, or a cafe. Ten background noises were synthetically mixed with the target speech for training, while seven background noises

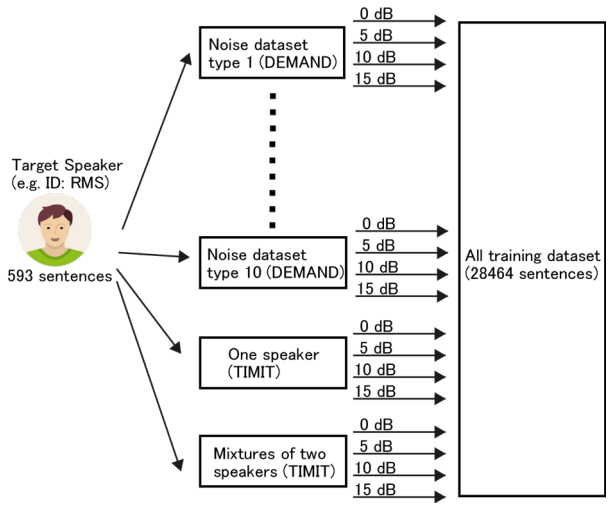

Figure 9. Visualization of creating training dataset. 593 sentences * 12 types of other sounds *4 type SNRs = 28464 sentences.

were used for testing. All training samples of each target speaker (593 sentences) were synthetically mixed with each ten noises type at each of the following single-to-noise ratios (SNRs): 0, 5, 10, and 15 dB. Note that the smaller the dB value, the bigger the noise value relative to the speech.

We also mixed the training samples of each target speaker with different speakers from the TIMIT corpus, which features 24 English speakers, including the following various dialects: New England, Northern, North Midland, South Midland, Southern, New York City, Western, and Army Brat. We synthetically mixed the all training samples of each target speaker with a TIMIT speaker at each SNRs (0, 5, 10, and 15dB). Additionally, we created new corpus of two-speaker mixtures using utterances from the TIMIT corpus. The mixtures were mixed with all training samples of each target speaker at each SNRs. As a result, the number of all training data per target speaker was 28464 sentences.

Learning

We let the sound source separation DNN learn with the

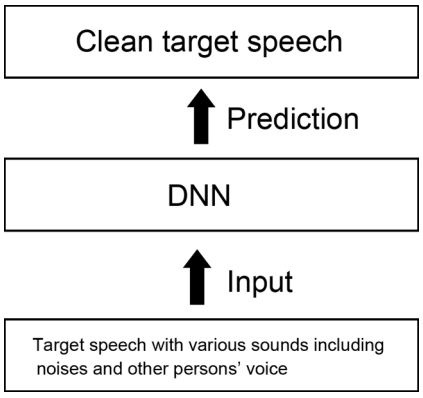

Figure 10: Let the DNN learn to output clean target speech from the target speech with various sound including noises and other persons' voice.

		Input source type				
		Noise only	**-10 dB**	**0 dB**	**10dB**	**Target only**
Average amplitude difference (dB)	**ID: RMS**	19.77 dB	8.75 dB	3.12 dB	0.64 dB	0.25 dB
	ID: SLT	22.99 dB	11.06 dB	3.20 dB	0.84 dB	0.45 dB

Table 2. Results of calculating the average difference between the output waveform and the input waveform. Top row represents the input source type: noise only, mixtures at -10dB, 0dB, 10dB, and target speech only. What the average amplitude difference is larger means that the input speeches were suppressed more. The result shows that the smaller the amplitude of the target speech included in the input source is, the larger the average amplitude difference becomes, and demonstrates the amplitude of target speech does not become very small while that of the other sounds becomes small.

above training dataset at 16 kHz, as shown in Figure 10. The loss function we used was the same as Rethage's [31]. The learning condition was as follows: a learning rate was 0.001, a batch size was 60, an early stopping epoch was 4 and the GPU we used was NVIDIA TITAN X Pascal.

Test

We randomly chose 100 sentences from the target speaker, which does not include the training dataset, for test samples. The test samples were synthetically mixed at each of the following SNRs: -10, 0 and 10dB, with the seven test-noise types from the DEMAND, one speaker, and two speaker mixtures from the TIMIT corpus. Furthermore, we used the noise only and target speaker only source, as the test dataset. We inputted 100 files of each source type (noise only, sound mixtures at -10, 0, 10 dB, and target only) into each learned DNN and calculated the average amplitude difference between the output waveform and the input waveform.

Result

Table 2 shows the results. What the average difference is larger means that the input speeches were suppressed more. The result demonstrates the amplitude of target speech does not become very small, while that of the other sounds becomes small. In addition, the result suggests that since the DNN decreases the amplitude of input waveform by about 20 dB at the maximum and about 0 dB at the minimum, it is appropriate to set the threshold during that interval.

EXPERIMENT 2

This experiment is to confirm how accurately the proposed system can detect the utterance scene of a specific person. We let the system perform the task of detecting the target speech included in the 10 minutes' sound.

Setup

The 10 minutes' sound was created by connecting DEMAND and TIMIT corpus which not in the training dataset. We chose the target speech randomly at 100 sentences and superimposed on that 10 minutes' sound. The SNRs of the target speech to 10 minutes' sound was chosen randomly from 0, 5, 10 and 15 dB. We used the sound source separation DNN learned in Experiment 1. The

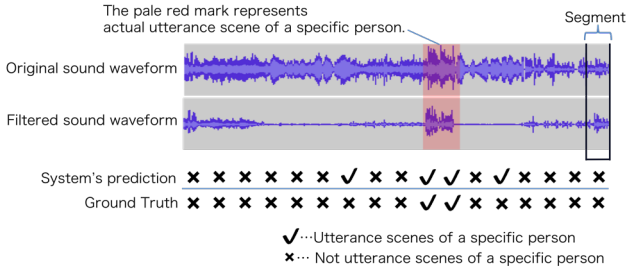

Figure 11. Visualization of predicting whether or not the scenes include the target speaker's utterance. The system performs prediction for each segment of the waveforms.

		True condition	
		Actual utterance scene of a specific person	**Not utterance scene of a specific person**
Predicted condition	**System predicts "utterance scene of a specific person"**	True Positive	False Positive
	System predicts "not utterance scene of a specific person"	False negative	True negative

Table 3. Contingency table of true positive, false positive, false negative and true negative.

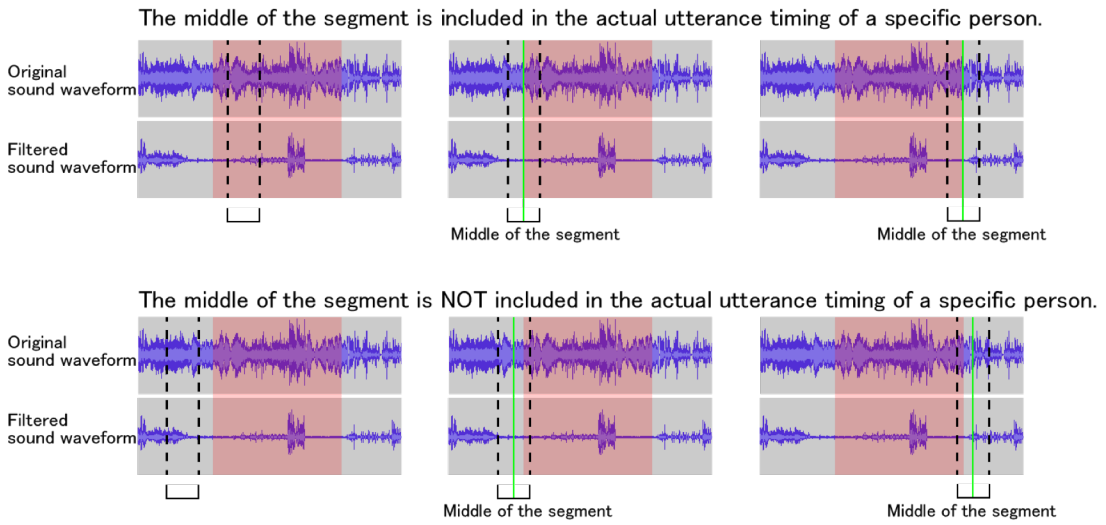

Figure 12. Upper: Case where the middle of the segment is included in the actual utterance timing of specific person. Lower: Case where the middle of the segment is not included in that timing. The green line represents the middle of the segment. The pale red marks represent actual utterance scenes of a specific person. When the middle of the segment is included in the actual utterance timing of a specific person, the true condition is "Actual utterance scene of a specific person".

window size of the detection was 0.1 s and the window's step length was also 0.1 s. We changed the threshold every 5 dB (-5, 0, 5, 10, 15, 20 dB) for confirming whether the result changes.

We used the following four events for test: True positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). Table 3 shows the definition of each event. Based on the four events, the following ratios were calculated: the accuracy and the precision. Accuracy and precision are formulated as follows:

$$Accuracy[\%] = (TP + TN) \,/\, (TP + FP + FN + TN)$$

$$Precision[\%] = TP \,/\, (TP + FP)$$

The system performs prediction for each segment of the waveforms as shown in Figure 11. When the middle of the segment is included in the actual utterance timing of a specific person, the true condition is "Actual utterance scene of a specific person" as shown in Figure 12.

Result

Table 4 shows the results. The result shows that the accuracy is 83% and the precision is 92% in the best case. The accuracy is higher when the threshold is around 10 to 15 dB and the precision is higher when the threshold is around 0 to 5 dB for each target speaker.

FUTURE WORK

User study

In this paper, we did the basic performance evaluation of the proposed system and did not do user study. We need to perform a user study and verify that the users can find the scenes they want to hear accurately and quickly.

We will need to refine the interface based on the user study. One alternative interface is to display the utterance scenes of a specific person as a graph in a video timeline. We will confirm how usability changes by changing the interface.

Improving accuracy

We need to explore a special DNN structure for extracting a

		Threshold					
		-5dB	**0dB**	**5dB**	**10dB**	**15dB**	**20dB**
Accuracy	**ID: RMS**	48%	59%	73%	79%	78%	72%
	ID: SLT	58%	67%	79%	83%	81%	74%
Precision	**ID: RMS**	83%	88%	89%	85%	78%	69%
	ID: SLT	88%	92%	91%	85%	81%	74%

Table 4. Result of the accuracy and precision for each target speaker

specific speaker more accurately. If we find this new structure, we could make the system improve the accuracy of the Experiment 2 task.

CONCLUSION

We propose a system that detects scenes, where a specific person speaks in the video, and displays them in the timeline. This system enables users to skip to the timeline they want to hear by detecting scenes in a drama, talk show, or discussion TV program, where a specific speaker is speaking.

We conducted two experiments on the proposed system. One was to confirm how much the amplitude of the other sounds can be suppressed and how much that of the specific person's utterance does not be suppressed by the sound source separation DNN extracting only the specific person's utterance. The result showed that the smaller the amplitude of the target speech included in the input source was, the larger the average amplitude difference between the input and output waveform became. That is, we got the result as expected.
The second experiment was to confirm how accurately the system can detect the utterance scene of a specific person. The result showed that the accuracy was 83% and the precision was 92% in the best case.

This system can be applied to voice services, like Podcast, Spotify, and SoundCloud. With the advent of smart speakers, such as Amazon Echo and Google home, audio contents are likely to increase along with the importance of searching timelines based on audio content.

REFERENCES

1. Keita Higuchi, Ryo Yonetani, and Yoichi Sato. 2017. EgoScanning: Quickly Scanning First-Person Videos with Egocentric Elastic Timelines. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17). ACM, New York, NY, USA, 6536-6546.
2. Suporn Pongnumkul, Jue Wang, Gonzalo Ramos, and Michael Cohen. 2010. Content-aware dynamic timeline for video browsing. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (UIST '10). ACM, New York, NY, USA, 139-142.
3. Kai-Yin Cheng, Sheng-Jie Luo, Bing-Yu Chen, and Hao-Hua Chu. 2009. SmartPlayer: user-centric video fast-forwarding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09). ACM, New York, NY, USA, 789-798.
4. Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology* (UIST '14). ACM, New York, NY, USA, 563-572.
5. Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video browsing by direct manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08). ACM, New York, NY, USA, 237-246.
6. Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2013. Direct manipulation video navigation in 3D. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13). ACM, New York, NY, USA, 1169-1172.
7. Thorsten Karrer, Malte Weiss, Eric Lee, and Jan Borchers. 2008. DRAGON: a direct manipulation interface for frame-accurate in-scene video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08). ACM, New York, NY, USA, 247-250.
8. Thorsten Karrer, Moritz Wittenhagen, and Jan Borchers. 2012. DragLocks: handling temporal ambiguities in direct manipulation video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, NY, USA, 623-626.
9. C. Saraceno and R. Leonardi, "Audio as a support to scene change detection and characterization of video sequences," *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, 1997, pp. 2597-2600 vol.4.
10. Kazutaka Kurihara. 2012. CinemaGazer: a system for watching videos at very high speed. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (AVI '12), Genny Tortora, Stefano Levialdi, and Maurizio Tucci (Eds.). ACM, New York, NY, USA, 108-115.
11. Abir Al-Hajri, Matthew Fong, Gregor Miller, and Sidney Fels. 2014. Fast forward with your VCR: visualizing single-video viewing statistics for navigation and sharing. In *Proceedings of Graphics Interface 2014* (GI '14). Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 123-128.
12. Neel Joshi, Wolf Kienzle, Mike Toelle, Matt Uyttendaele, and Michael F. Cohen. 2015. Real-time hyperlapse creation via optimal frame selection. *ACM Trans. Graph.* 34, 4, Article 63 (July 2015), 9 pages.
13. Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2012. Video summagator: an interface for video summarization and navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, NY, USA, 647-650.

14. Suporn Pongnumkul, Jue Wang, and Michael Cohen. 2008. Creating map-based storyboards for browsing tour videos. In *Proceedings of the 21st annual ACM symposium on User interface software and technology* (UIST '08). ACM, New York, NY, USA, 13-22.
15. Alex Rav-Acha, Yael Pritch, and Shmuel Peleg. In *In Proc. IEEE Conference on Computer Vision and Pattern Recognition* (CVPR' 06).
16. Yael Pritch, Alex Rav-Acha, Avital Gutman, and Shmuel Peleg. 2007. Webcam Synopsis: Peeking Around the World. *In In Proc. IEEE International Conference on Computer Vision* (ICCV'07).
17. Yael Pritch, Alex Rav-Acha, and Shmuel Peleg. 2008. Nonchronological Video Synopsis and Indexing. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 11 (November 2008), 1971-1984.
18. Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2014. Video lens: rapid playback and exploration of large video collections and associated metadata. In *Proceedings of the 27th annual ACM symposium on User interface software and technology* (UIST '14). ACM, New York, NY, USA, 541-550.
19. Pascal Scalart et al. Speech enhancement based on a priori signal to noise estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pp. 629–632, 1996.
20. Pal, Monisankha, et al. "Robustness of Voice Conversion Techniques Under Mismatched Conditions." *arXiv preprint arXiv:1612.07523* (2016).
21. Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pp. 436–440, 2013.
22. Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(12):2136–2147, 2015.
23. Y. Xu, J. Du, L. R. Dai and C. H. Lee, A Regression Approach to Speech Enhancement Based on Deep Neural Networks, in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7-19, Jan. 2015.
24. Anurag Kumar and Dinei Florencio. Speech enhancement in multiple-noise conditions using deep neural networks. *arXiv preprint arXiv:1605.02427*, 2016.
25. Jordi Pons, Jordi Janer, Thilo Rode, and Waldo Nogueira. Remixing music using source separation algorithms to improve the musical experience of cochlear implant users. *The Journal of the Acoustical Society of America*, 140(6):4338–4349, 2016.
26. Qian, Kaizhi, et al. "Speech enhancement using bayesian wavenet." *Proc. Interspeech 2017* (2017): 2013-2017.
27. Tu, Ming, and Xianxian Zhang. "Speech enhancement based on Deep Neural Networks with skip connections." *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017.
28. Pascual, Santiago, Antonio Bonafonte, and Joan Serrà. "SEGAN: Speech Enhancement Generative Adversarial Network." *arXiv preprint arXiv:1703.09452* (2017).
29. Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 61-65.
30. Fu, Szu-Wei, et al. "Raw Waveform-based Speech Enhancement by Fully Convolutional Networks." *arXiv preprint arXiv:1703.02205* (2017).
31. Rethage, Dario, Jordi Pons, and Xavier Serra. "A Wavenet for Speech Denoising." *arXiv preprint arXiv:1706.07162* (2017).
32. Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266. Springer, 2017.
33. Z. Q. Wang and D. Wang, "Recurrent deep stacking networks for supervised speech separation," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 71-75.
34. Chien, Jen-Tzung & Kuo, Kuan-Ting, "Variational Recurrent Neural Networks for Speech Separation", In *Interspeech,* pp. 1193-1197, 2017
35. K. Osako, Y. Mitsufuji, R. Singh and B. Raj, "Supervised monaural source separation based on autoencoders," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 11-15. doi: 10.1109/ICASSP.2017.7951788
36. Lee, Yuan-Shan, et al. "Fully complex deep neural network for phase-incorporating monaural source separation." *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017.
37. Wang, Yannan & Du, Jun & Dai, Li-Rong & Lee, Chin-Hui, "A Maximum Likelihood Approach to Deep Neural Network Based Nonlinear Spectral Mapping for Single-Channel Speech Separation", Interspeech, pp. 1178-1182, 2017

38. Hershey, John R., et al. Deep clustering: Discriminative embeddings for segmentation and separation. *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016.
39. Isik, Yusuf, et al. "Single-channel multi-speaker separation using deep clustering." *arXiv preprint arXiv:1607.02173* (2016).
40. Yu, Dong, et al. "Permutation invariant training of deep models for speaker-independent multi-talker speech separation." *arXiv preprint arXiv:1607.00325* (2016).
41. Y. Tian, L. He, M. Cai, W. Q. Zhang, and J. Liu, "Deep neural networks based speaker modeling at different levels of phonetic granularity," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2017, pp. 5440–5444.
42. Y. Lei, N. Scheffer, L. Ferrer and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 1695-1699.
43. S. Ranjan and J. H. L. Hansen, "Curriculum Learning Based Approaches for Noise Robust Speaker Recognition," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 197-210, Jan. 2018.
44. Jansen, Aren, et al. "Large-scale audio event discovery in one million youtube videos." *Proceedings of ICASSP*. 2017.
45. Gemmeke, Jort F., et al. "Audio Set: An ontology and human-labeled dataset for audio events." *IEEE ICASSP*. 2017.
46. Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).
47. Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. The diverse environments multichannel acoustic noise database: A database of multichannel environmental noise recordings. *The Journal of the Acoustical Society of America*, 133(5):3591–3591, 2013.
48. J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993.
49. J. Kominek and A. W. Black, "The CMU Arctic speech databases," in Fifth ISCA Workshop on Speech Synthesis, 2004.