한국음향학회지 제40권 제5호 pp. 466～472 (2021)
The Journal of the Acoustical Society of Korea Vol.40, No.5 (2021)
https://doi.org/10.7776/ASK.2021.40.5.466

pISSN : 1225-4428
eISSN : 2287-3775

Proposal of speaker change detection system considering speaker overlap

화자 겹침을 고려한 화자 전환 검출 시스템 제안

Jisu Park,[1] Young-Sun Yun,[1†] Shin Cha,[1] and Jeon Gue Park[2]
(박지수,[1] 윤영선,[1†] 차신,[1] 박전규[2])

[1]Department of Information and Communication Engineering, Hannam University, [2]ETRI

(Received July 9, 2021; accepted August 26, 2021)

ABSTRACT: Speaker Change Detection (SCD) refers to finding the moment when the main speaker changes from one person to the next in a speech conversation. In speaker change detection, difficulties arise due to overlapping speakers, inaccuracy in the information labeling, and data imbalance. To solve these problems, TIMIT corpus widely used in speech recognition have been concatenated artificially to obtain a sufficient amount of training data, and the detection of changing speaker has performed after identifying overlapping speakers. In this paper, we propose an speaker change detection system that considers the speaker overlapping. We evaluated and verified the performance using various approaches. As a result, a detection system similar to the X-Vector structure was proposed to remove the speaker overlapping region, while the Bi-LSTM method was selected to model the speaker change system. The experimental results show a relative performance improvement of 4.6 % and 13.8 % respectively, compared to the baseline system. Additionally, we determined that a robust speaker change detection system can be built by conducting related studies based on the experimental results, taking into consideration text and speaker information.

Keywords: Speaker overlap detection, Speaker representation, Speaker change detection, Deep neural networks

PACS numbers: 43.71.Bp, 43.72.Fx

초　　록: 화자 전환 검출은 대화 중에 발성 화자가 다른 사람으로 바뀌는 시점을 검출하는 것을 의미한다. 이 과정에서 화자 중복, 화자 정보 표기의 부정확성, 데이터 불균형 등으로 화자가 바뀌는 순간을 검출하는 데 어려움이 발생한다. 본 논문에서는 이러한 문제를 해결하기 위해 음성 인식에 널리 사용되는 TIMIT 데이터를 가공하여 충분한 양의 훈련 데이터를 얻었으며, 화자가 겹치는지를 파악한 후에 화자 전환 여부를 판단하였다. 본 논문에서는 화자 겹침을 고려한 화자 전환 검출 시스템을 구축하기 위하여 다양한 접근법을 사용하여 성능을 평가하고 검증했다. 그 결과 화자 겹침 영역을 제거하기 위해 X-Vector 구조와 유사한 형태의 검출 시스템과 화자 전환 검출 시스템을 모델링하기 위한 Bi-LSTM 모델을 제안하였다. 실험 결과 기준 시스템보다 상대적으로 각각 4.6 %, 13.8 % 성능 향상을 확인하였다. 또한, 실험 결과를 기반으로 텍스트 정보와 화자 정보 등을 고려한다면 좀 더 강인한 화자 전환 검출 시스템을 구축할 수 있을 것으로 판단한다.

핵심용어: 화자 겹침 검출, 화자표현, 화자 전환 검출, 심층 신경망

Ⅰ. Introduction

Speaker Change Detection (SCD) is a process of finding boundaries between speaker turns in conversation. A typical speaker change detection method divides the speech signal into short time segments. Subsequently, the similarity of the signals in two adjacent segments is calculated to determine whether there is a change in the

†**Corresponding author:** Young-Sun Yun (ysyun@hnu.kr)
Department of Information and Communication Engineering, Hannam University, 70 Hannam-ro Daedeok-gu, Daejeon 34430, Republic of Korea
(Tel: 82-42-629-7569, Fax: 82-42-629-7843)

speaker or not. However, if multiple speakers talk simultaneously or a speaker interrupts while another speaker is talking, speaker interference occurs. When speaker intrusion occurs, a speaker overlapping is found. Speaker overlapping make it difficult to detect a change in the speakers. It has been reported that there are four types of speech conversation as follows.[1]

A. Short response or feedback that does not interrupt the conversation

B. A person starts talking while another person is speaking because he/she could not accurately predict when the other person was going to stop talking

C. Some people start talking at the same time after a long period of silence

D. When a person forcibly interrupts another person's talking and tells his/her own story

A recent study on speaker change detection has created an independent module,[2] transformed the acoustic features and artificial neural network model in various ways to detect speaker overlapping. To overcome the insufficiency of data and their imbalance on related studies, the TIMIT corpus was used to generate speaker overlapping datasets, where the single speaker or speaker overlapping region were distinguished through various acoustic characteristics and artificial neural networks.[3] Moreover, the overlap region was randomly selected when artificially creating the speaker overlap data using the TIMIT corpus. Based on various acoustic characteristics and a system with a convolutional neural network, the effect of overlapping regions on speaker change detection was studied and presented.[4] Another study evaluated the performance of the TED Talks corpus based on the speaker's gender. The study artificially generated data using the TED Talks corpus and used a convolutional neural network, and the research results were reported.[5]

In this paper, we proposed a system built by separating the detection of speaker overlapping region and the detection of speaker change region. Additionally, we suggested various artificial neural network models to effectively detect changes in the speaker.

This paper is organized as follows. Section 2 discusses previous studies for solving difficulties in speaker change detection. Section 3 describes the proposed speaker change detection system, and Section 4 analyzes the experimental results of the proposed system. Finally, Section 5 presents the conclusion of the paper and direction of future studies.

II. Related Work

In a speaker change detection system, many difficulties are encountered while detecting changes in the speaker owing to issues such as speaker overlap, recognition of unregistered speakers, and preregistration of speakers. Various studies are being conducted to solve these problems.

Research on the detection of speaker overlap typically distinguishes between single speaker, overlapping speaker, and other speaker waves and processes them separately. It has been reported that it is advantageous to disregard the short overlapping regions for effective use of real speech data.[6] To minimize inaccurate labeling of the speaker and the speech overlap information, a method was presented to mark the before and after 50 ms regions including a real boundary as "speaker changed regions". Alternatively, this information (change probability) could be displayed either linearly or non linearly in both directions from the specified speaker change boundary.[7] One study has modeled to linearly reduce the probabilities by 200 ms in both directions, before and after the speaker change, by maximizing the probability on the speaker change boundary.[8] On the other hand,[5] represented a non linearly reduced probability by 600 ms in both directions. In the ETAPE corpus, the speaker change region account for approximately 0.4 % of the entire corpus.[7] It was also observed that the overlapping regions in a conversation less than 5 %.[9] Many studies generally extend the speaker overlapping region to resolve the data imbalance by augmentation. For example, the neighborhood frames of the boundary where the speaker changed are marked as "speaker change" and they are extended evenly by 200 ms.[10] There was also a study that divided the data into speaker overlapping regions and non-overlapping regions, and each region consisted of similar size.[4]

In this study, we designed both a dataset and a speaker change detection system that considers the speaker overlap by analyzing the advantages and disadvantages of the previous studies. We also propose a variety of approaches to determine the optimization model for the artificial neural networks used in the system.

III. Proposed Speaker Change Detection System

The speaker change detection system proposed in this paper is configured as shown in Fig. 1. When the speech data entered, overlapped speech are removed by the proposed speaker overlap detection system. For the data having non overlapping speech but containing pauses or silence, WebRTC[11] was used to remove pauses or silences.

In this way, the data were preprocessed to be readily used for speaker change detection system. From the data composed of only the speech information of multiple speakers, a segment including the speaker change was detected through an artificial neural network. Additionally, the frame including speaker change was expressed as a probability value rather than a binary value. Fig. 2. shows examples of whether the speaker changed or not.

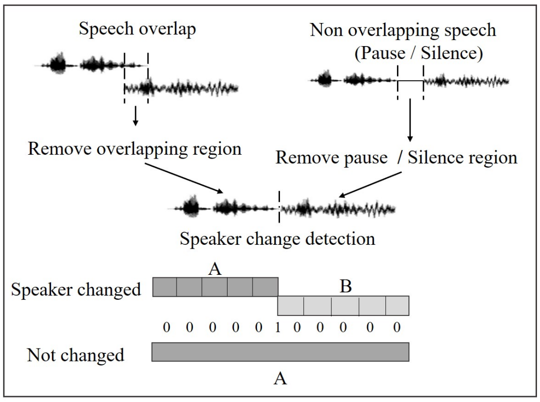
Speech overlap
Non overlapping speech (Pause / Silence)
Remove overlapping region
Remove pause / Silence region
Speaker change detection
Speaker changed
A
B
0 0 0 0 0 1 0 0 0 0 0
Not changed
A

Fig. 1. Speaker change detection system overview.

3.1 Artificial Speaker Change Data

To detect the point when the speaker changes completely during a conversation between two speakers, the speaker overlapping scenario was considered. The conversation was divided into the case where the speech between the two speakers overlaps and the case with a region where a speaker does not speak (silence) or pauses while speaking.

Case (1) in Fig. 2 shows an example of a typical case where the speaker changes with overlapping. Case (2) shows that a speaker interrupts while the other speaker is speaking, but the other speaker continues to speak. Therefore, the speaker does not change in this case. Case (3) indicates that there is silence or pauses while speaking. In case (4), the two speakers take turns to speak continuously, and their speech does not overlap. Cases (1) and (2) are data types of speaker overlap, and the overlapping speech region was removed using an artificial neural network. Cases (3) and (4) are data types where the speakers do not overlap, but the data contain regions where the speaker stops speaking or pauses while talking. For (3) and (4), only the Voice Activity Detection (VAD) region was used by utilizing WebRTC. Obtaining a sufficient amount of high quality data is a crucial factor that can determine the performance of an artificial neural network based system.

The system proposed in this paper is a speaker change detection system based on the classification model that utilizes an artificial neural network. Therefore, it is essential to have a sufficient number of speakers, an appropriate distribution of speaker changes, and accurate information labeling in order to obtain good performance.

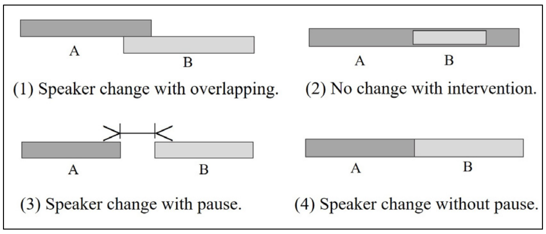
A
B
(1) Speaker change with overlapping.
A
B
(2) No change with intervention.
A
B
(3) Speaker change with pause.
A
B
(4) Speaker change without pause.

Fig. 2. Examples of speaker change.

However, it is difficult to obtain a speaker change detection dataset. It is also difficult to obtain a sufficient amount of data to train artificial neural networks. Accordingly, TIMIT dataset was used to artificially generate speaker changes and overlaps.

3.2 Speaker Overlap Detection Neural Network Model

A total of four artificial neural network were used and compared for detecting with speaker overlapping. These four models include the Multi-Layer Perceptron (MLP), MLP (*d*-vector), Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM), and Pseudo X-Vector. Their architectures are shown in Table 1.

The MLP model is the baseline, and it uses filter bank energies to perform a simple classification in units of frames. The MLP (*d*-vector) model receives a speaker embedding vector known as the *d*-vector in the form of a segment from the baseline model. The other models, excluding the MLP (*d*-vector) model, receive 11 frames as an input, which is similar to the length of one segment. The model using CNN and LSTM was built with a relatively simple structure. It was modeled to facilitate the extraction of spatial features of data by changing the kernel size. The Pseudo X-Vector structure was created to model the adjacent frames' features well, similar to the X-Vector[12] structure, in which the output of the final hidden layer of the artificial neural network represent the speaker information. The Pseudo X-Vector model uses a convolutional neural network to change the stride and dilation. Thus, it was possible to efficiently model the entire speech region using the sampling method without increasing data, which is similar to the 1D CNN model used to detect speaker changes.

Table 1. Speaker overlap detection neural network models.

MLP (baseline)	MLP (*d*-vector)	CNN + LSTM	Pseudo X-Vector
Linear (1024)	Linear (2560)	Conv1D (4, 384)	Conv1D (512, 9)
Linear (512)	Linear (1024)	Max Pooling 1D (1, 384)	Conv1D (512, 5)
Linear (256)	Linear (512)	LSTM (384)	Conv1D (512, 1)
Linear (64)	Linear (256)	LSTM (384)	Conv1D (256, 1)
Linear (1)	Linear (1)	Linear (128)	Conv1D (256, 1)
		Linear (64)	Linear (256)
		Linear (1)	Linear (128)
			Linear (1)

3.3 Speaker Change Detection Neural Network Model

A *d*-vector, fixed-length speaker embedding vector representing speaker characteristics, was used as the input for the speaker change detection model. The forward-backward difference based *d*-vector representation and the Universal Background Model (UBM) based speaker similarity methods, which had the best results among the combinations of methods proposed in a previous paper[13] were selected for the *d*-vector feature method. The forward-backward difference means the difference of the *d*-vectors extracted in forward and backward directions on the Bidirectional LSTM (Bi-LSTM) model. The UBM fine tunes the speaker vector so that it is not dependent on a specific speaker. The average feature of all speakers was used to improve discrimination when comparing individual speakers.

Four artificial neural network models MLP, 2D CNN, 1D CNN, and Bi-LSTM were compared and analyzed to detect speaker changes in a conversation. Their architectures are shown in Table 2. The 2D CNN model was constructed with a structure similar to that of the CNN + LSTM model, one of the models used to detect speaker overlap. The loss of spatial information was minimized in the 1D CNN model through dilation convolution, similar to the X-Vector model, which is a feature vector specialized in detecting speaker overlapping frames, to maintain the spatial features according to the 1D CNN model's features

Table 2. Speaker change detection neural network models.

MLP (baseline)	2D CNN	1D CNN	Bi-LSTM
Linear (1024)	Conv 2D (1, 6, 3, 1)	Conv 1D (512, 3, 1)	LSTM (756, 3)
Linear (1024, 512)	Conv 2D (6, 12, 3, 1)	Conv 1D (512, 512, 3, 1)	Linear (1512, 1024)
Linear (512, 256)	Linear (6096, 1024)	Linear (512, 256)	Linear (1024, 512)
Linear (256, 1)	Linear (1024, 384)	Linear (256, 1)	Linear (512, 64)
	Linear (384, 1)		Linear (64, 1)

that reflect temporal variation characteristics. The Bi-LSTM model is built with the same architecture, but takes the *d*-vector, the output of Bi-LSTM model, as input data.

IV. Experiments

4.1 Data Preparation

The TIMIT corpus, widely used in speech recognition, was used to evaluate the speaker change detection system. To simplify the problem, the speech data were generated artificially by separating the data for detecting overlap speakers and the data for detecting speaker changes. The TIMIT corpus provides reading data for 10 phonetically rich sentences from 630 different speakers. The data of 462 speakers were classified as the training dataset, and the data of 168 speakers were classified as the test dataset. The speech data for detecting changes in the speaker, which include speaker overlap, were generated by additionally setting overlapping speaker regions of arbitrary length between 500 ms and 2 s in the corpus. For speaker change detection system, the dataset was configured by combining 10 arbitrary sentence speeches for 10 arbitrary speakers so that there would be one point of speaker change. For the training data, 10 different sentences were selected for each sentence of 462 speakers. The selected data were combined for one speaker who was selected out of 461 speakers. In the same way, 16,800 test data were generated.[14]

4.2 Experimental environments

The experiment was conducted to evaluate and analyze the artificial neural network used, based on whether overlapping speakers exist. Precision and recall were used to measure the performance of the proposed model. In addition, the results were verified using the F1 score, which calculates the harmonic mean of the precision and recall. The F1 score can accurately evaluate the performance of the model when the data label is unbalanced, and it can be expressed as a single number. Therefore, the F1 score was chosen and used for the final comparison of the results. All experiments are performed on Ubuntu 20.04 with the Pytorch 1.7 framework. Speaker overlap detection experiments first developed as Tensorflow based Keras and converted to the Pytorch framework.[15] On the other hand, a speaker change detection system was developed from scratch in the Pytorch framework. For the training of the speaker overlap and change detection system, we adopted the binary cross entropy for loss function, 0.001 learning rate, and the Adam optimizer. Separately, the *d*-vector used a generalized end-to-end loss function based on cosine similarity.

4.3 Speaker Overlap Detection

The baseline MLP model using the frames had an F1 score of 86.9 % which was the second highest score next to that of the Pseudo X-Vector model as shown in Table 3. The MLP model using the *d*-vector as the input had an F1 score of 80.5 %. The performance degradation is because the frame-based MLP and the *d*-vector-based MLP have different feature characteristics. The *d*-vector is a speaker embedding vector that represents the speaker characteristics for a given set of frames. When speakers overlap in the same section, the frame feature can represent different characteristics for each speaker, but the d-vector feature shows only representative speaker characteristics. The model constructed by combining the CNN and LSTM, similar to the 2D CNN mode, had an F1 score of 84.9 % which is similar to the MLP (baseline) model.

The Pseudo X-Vector model was compared with the

Table 3. Speaker overlap detection results.

NN Model	Feature Type	Precision (%)	Recall (%)	F1 Score (%)
MLP (baseline)	frame	86.2	87.5	86.9
MLP	*d*-vector	85.5	76.0	80.5
CNN + LSTM	frame	93.8	77.6	84.9
Pseudo X-Vector	frame	94.0	88.0	**90.9**

Table 4. Speaker change detection results.

NN Model	Precision (%)	Recall (%)	F1 Score (%)
MLP (baseline)	69.8	69.6	69.8
2D CNN	73.1	74.5	73.8
1D CNN	71.9	83.1	77.1
Bi-LSTM	83.5	75.8	**79.5**

CNN + LSTM model because it has a structure similar to that of the 1D CNN model. The Pseudo X-Vector model performs redundancy processing in the time series, so it was determined to be suitable for extracting speech characteristics for overlapping speakers. The Pseudo X-Vector model had an F1 score of 90.9 %. Thus, it was confirmed that this model produces the best result. Based on the detection of overlapping speaker proposed in this paper, it was confirmed that the Pseudo X-Vector model's result was relatively 4.6 % higher than that of the baseline model. The overall results are shown in Table 3.

4.4 Speaker Change Detection

The results of the speaker change detection experiment, which was conducted using the *d*-vector,[13] showed that the baseline MLP model had an F1 score of 69.8 %. The 2D CNN model, which has a similar structure as the CNN + LSTM, had an F1 score of 73.8 %. The 1D CNN model was made similar to the Pseudo X-Vector model, which minimizes information loss by maintaining spatial features using dilation convolution. The 1D CNN model achieved an F1 score of 77.1 %. The Bi-LSTM model had an F1 score of 79.5 %.

The results are summarized in Table 4, and all NN models detect whether the speaker changes based on the *d*-vector. The *d*-vector is obtained by the difference between the forward and backward LSTM vectors (each LSTM vector has 256 dimensions) based on the UBM similarity metric so that it does not depend on any particular speaker. As a result, it was confirmed that the Bi-LSTM model obtained higher results than other models, and showed relatively 13.8 % higher results than the basic model.

V. Conclusions and Future work

In this paper, we proposed a speaker change detection system that considers overlap speaker. The types of speaker change detection data were analyzed to solve the difficulties arising from overlapping speakers, inaccuracy in labeling the information, and data imbalance in speaker change detection. In addition, datasets for speaker change detection were artificially generated. Moreover, we proposed various approaches for detecting changes in the speaker and introduced the optimized model through the experiments. The results showed that, for the overlapping speaker detection, the Pseudo X-Vector model showed the highest F1 score of 90.9 %. For speaker change detection, the Bi-LSTM model exhibited the highest F1 score of 79.5 %. These results confirmed the applicability of the proposed approaches through various experiments. Based on them, future research will be able to improve the performance of the speaker change detection system by integrating acoustic characteristics, speaker information, and text information.

Acknowledgement

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-01376, Development of the multi-speaker conversational speech recognition technology).

References

1. A. G. Adam, S. S. Kajarekar, and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," Proc. ICASSP. 3908-3911 (2002).
2. L. Bullock, H. Bredin, and L. P. Garcia Perera, "Overlap aware diarization: Resegmentation using neural end-to-end overlapped speech detection," Proc. ICASSP. 7114-7118 (2020).
3. N. Sajjan, S. Ganesh, N. Sharma, S. Ganapathy, and N. Ryant, "Leveraging lstm models for overlap detection in multi party meetings," Proc. ICASSP. 5249-5253 (2018).
4. V. Andrei, H. Cucu, and C. Burileanu. "Detecting overlapped speech on short time frames using deep learning," Proc. Interspeech, 1198-1202 (2017).
5. E. Kazimirova, A. Belyaev, "Automatic detection of multi speaker fragments with high time resolution," Proc. ICASSP. 1338-1392 (2018).
6. Z. Ge, A. N. Iyer, S. Cheluvaraja, and A. Ganapathiraju, "Speaker change detection using features through a neural network speaker classier," Proc. IEEE SAI Intelligent Systems Conference, 1111-1116 (2017).
7. R. Yin, H. Bredin, and C. Barras, "Speaker change detection in broadcast tv using bidirectional long short term memory networks," Proc. Interspeech, 3827-3831 (2017).
8. M. Kunesova, M. Hruz, Z. Zajc, and V. Radova, "Detection of overlapping speech for the purposes of speaker diarization," Proc. ICSC. 247-257 (2019).
9. S. C. Levinson, "Turn-taking in human communication - Origins and implications for language processing," Trends in Cognitive Sciences, **20**, 6-14 (2016).
10. H. Bredin, "TristouNet: Triplet loss for speaker turn embedding," Proc. Interspeech, 5430-5434 (2017).
11. *WebRTC Homepage*, http://webrtc.org, (Last viewed November 21, 2020).
12. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," Proc. ICASSP. 5329-5333 (2018).
13. J. Park, S. Cha, S. Eun, J. G. Park, and Y.-S. Yun, "Data augmentation and d-vector representation methods for speaker change detection," Proc. ICRACS. 67-71 (2020).
14. V. Zue, S. Sene, and S. Glass, "Speech database development at MIT: TIMIT and beyond," Speech communication, **9**, 351-356 (1990).
15. H. Kim, J. Park, S. Cha, K. A Son, Y.-S. Yun, and J. G. Park, "Framework switching of speaker overlap detection system" (in Korean), J. SW Assessment and Valuation, **17**, 101-113 (2021).

Profile

Jisu Park (박지수)

She graduated with a BSc in Department of Medical Information Technology Engineering from Konyang University in 2017. She took the MSc degree in Information and Communication Engineering from Hannam University in 2019 and is currently working on PhD. program. Her research interests includes speech recognition, speaker change detection, speaker recognition and artificial intelligence.

Young-Sun Yun (윤영선)

He received the BSc, the MSc and PhD degrees in computer science in 1990, 1992, and 2001, respectively, from the Korea Advanced Institute of Science and Technology, Daejeon, Republic Of Korea. Since 2001, he has been a professor at Hannam University. From Apr. 2006 to Feb. 2007, he was a visiting researcher of ETRI, Daejeon, Korea. From Aug. 2012 to Jul. 2013, he stayed as a visiting scholar at University of Washington, USA. His research interests includes speech recognition, speaker verification, diarization and related speech processing.

Shin Cha (차신)

He received the PhD degree in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Korea in 1995. He has been with LG Electronics as a principle engineer (1986~2000), and with IA corporation as a vice president (2000~ 2015). He is currently a professor at Hannam University, Daejeon, Korea. His research interest includes risk modeling and analysis, software functional safety, speaker verification and diarization.

Jeon Gue Park (박전규)

He received his PhD degree in information and communication engineering from Paichai University, Rep. of Korea, in 2010. He had worked for ETRI, Rep. of Korea as a senior researcher since 1991, L&H Korea, Rep. of Korea as a director since 2000, and Donga Seetech Inc. Rep. of Korea as a director and CTO since 2002. During 2001~2002 he stayed as a visiting scholar at Carnegie Mellon University, USA. He rejoined ETRI since 2004 and is currently leading Integrated Intelligence Research Section as a Research Fellow. His current research interests include artificial intelligence, computer–assisted language learning, spoken dialogue system, and cognitive systems.