Speaker Change Detection Using Ensemble Prediction in Conversations

Qian-Bei Hong
Department of Electrical Engineering
Southern Taiwan University of Science and Technology
qbhong75@gmail.com

Chung-Hsien Wu
Department of Computer Science and Information Engineering
National Cheng Kung University
chunghsienwu@gmail.com

Yeou-Jiunn Chen
Department of Electrical Engineering
Southern Taiwan University of Science and Technology
chenyj@stust.edu.tw

Abstract

Speaker change detection (SCD) is an important technique for separating sentences from different speakers in a conversation. This study proposes a novel approach that utilizes ensemble prediction for SCD. First, a contour predictor is created to predict the frame-based speaker change probability contour for a speech segment. Next, the highest SCD probability for each frame among the sequential segments is selected as the contour value, where each specific frame appears in sequential segments with different timestamps. Finally, the speaker change boundaries are determined using a threshold. In the experiments, the proposed SCD model was evaluated on the 2000 NIST Speaker Recognition Evaluation corpus and the AMI meeting corpus. The NIST corpus was used to train the baseline and proposed models for SCD evaluation. The proposed ensemble prediction achieved the best performance and improved the precision of speaker change boundaries. In addition, the AMI corpus was used to evaluate the effects of out-of-domain prediction. The experiment shows that even though domain mismatch greatly affected the SCD performance, the proposed ensemble prediction, which considers the prediction probabilities of a change boundary from sequential segments achieved improved detection results.

Keywords: Speaker change detection, ensemble prediction, sequential segments

1 Introduction

Speaker change detection (SCD) is an important technique for separating sentences from different speakers in a conversation to understand the relationships between contexts. SCD has been widely used in the automatic speech recognition (ASR) tasks as in (S. Kumar et al., 2024; J. Wu et al., 2023; L. Sarı et al., 2019). It is adopted to split the signal into several speaker-specific speech segments before ASR to avoid transcription error of sequential words presented from different speakers. On the other hand, SCD is an important step for the speaker diarization task (X. Anguera et al., 2012; K. VijayKumar and R.R. Rao, 2023; W. Xia et al., 2022), in which the conversation signal needs to use uniform segmentation to extract the speaker embeddings and cluster them according to speaker features. If a speech segment contains the utterance from more than one speaker, then using SCD to discard the multi-speaker segments before extracting speaker embedding can effectively improve the performance of speaker clustering.

Traditional SCD methods used distance metric such as Bayesian information criterion (BIC) (S. Chen and P.S. Gopalakrishnan, 1998), Kullback-Leibler (KL) divergence (J. E. Rougui et al., 2006), etc. for speaker change detection. However, these traditional techniques often perform poorly due to

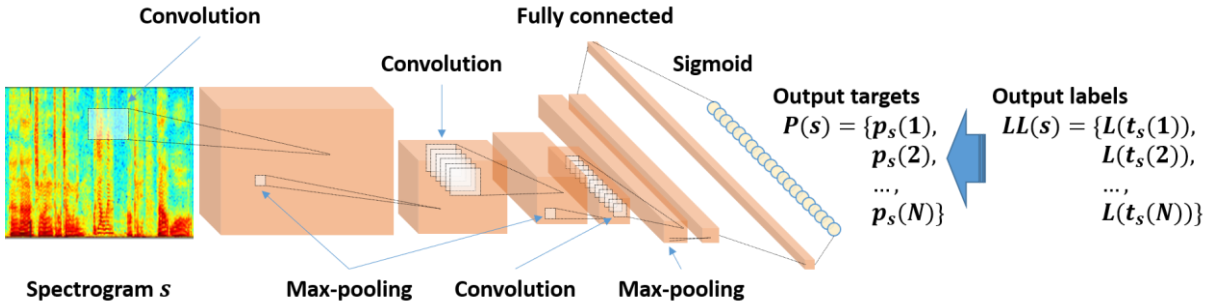

Figure 1. The contour prediction architecture for SCD task.

the challenges associated with processing high-dimensional features in complex multi-speaker conversations. In the past few years, deep neural network (DNN)-based model has been successfully applied to the SCD task (L. Mateju et al., 2019; V. Gupta, 2015; R. Wang et al., 2017), and the performance was better than the traditional methods. (M. Hrúz and Z. Zajíc, 2017) proposed a convolutional neural network (CNN) model with fuzzy label to predict the speaker change boundaries, which has been applied in the diarization systems (Z. Zhou et al., 2018; Z. Zajíc et al., 2018). Fuzzy labelling considers the distance between the detected boundary and the real boundary. The fuzzy labels are used for model training and can effectively improve the detection accuracy. However, in most studies, each speaker change probability is only determined once by the model. Accordingly, the predicted probability is easily affected by different energy distributions and degrades the accuracy of speaker change detection. Thus, this study tries to utilize the ensemble prediction approach to enhance the SCD performance based on the combination of different weights and prediction results.

Ensemble is to integrate different prediction results for performance improvement. The common approach is integrated from multiple models (J. Yi et al., 2017), in which the major point is to predict the output from multiple models with different weights. However, using multiple models for prediction will increase the time cost. Therefore, this study proposes a novel ensemble approach based on multi-label prediction, which requires only one model for prediction.

In summary, this paper uses multi-label prediction to predict the probabilities of each speaker change boundary. The main contributions are as follows: (1) Each potential speaker change boundary can be predicted with multiple values from different interconnection weights, and these probabilities of the same speaker change boundary are integrated to obtain the final prediction results. (2) Using multi-label prediction for segment-based spectrogram can increase the flexibility of frame-level prediction.

2 Speaker Change Contour Prediction

In (M. Hrúz and Z. Zajíc, 2017), a CNN-based SCD model was proposed for speaker diarization, in which the spectrogram of the speech signal is used to predict the speaker change probability of each frame. According to fuzzy labelling definition, the predicted probability is based on the distance between the time point of the current frame and the time point of the real speaker change boundary as follows.

$$L(t) = max\left(0, -\frac{\tau \cdot min_i(|t - b_i|)}{\lambda} + 1\right) \quad (1)$$

where t is the time point of the frame, b_i is the time point of the i-th speaker change boundary closest to t, τ is the shift time between two adjacent frames and $\lambda = 0.6$ is the tolerance. As shown in equation (1), the prediction probability is determined by considering the closest speaker change boundary and the probability of each time point. Thus, the predicted results are easily affected by other magnitudes around the speaker change boundary and decreases the performance of SCD.

2.1 Multi-label Prediction for SCD

In this study, a multi-label prediction approach to SCD is proposed. Figure 1 shows the procedure for multi-label SCD. The multi-label prediction is defined as frame-level prediction for an input

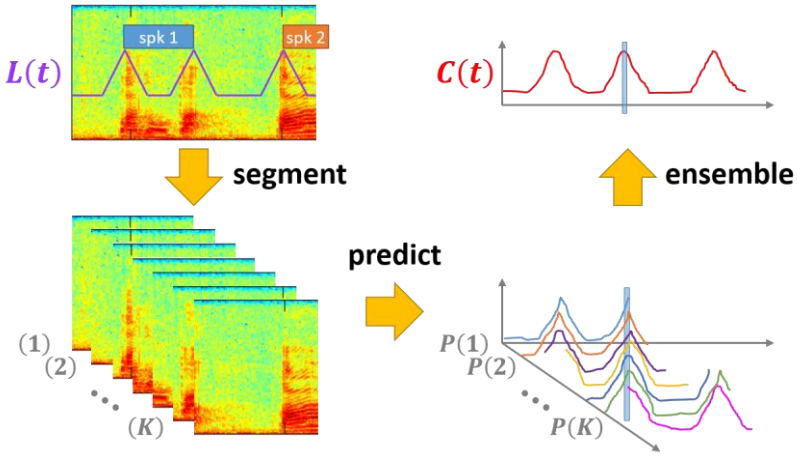

Figure 2. The procedure for the ensemble of prediction probabilities from sequential segments.

Algorithm 1. Boundary merging for SCD.

Input: Contour vector $C \in \mathbb{R}^T$ and change threshold θ
Output: Change time points $Q \in \mathbb{R}^J$

Initial $Q = \emptyset$
Let $\delta = 0.5$ be the onset of a boundary in second.
Let $\tau = 0.01$ be the shift time (in seconds) between two adjacent frames.
$\rho = round(\delta/\tau)$
for $t = 1$ to T **do**
  if $C(t) > \theta$ **then**
    $u = 0, v = 0$
    for $r = t - \rho$ to $t + \rho$ **do**
      if $C(r) > u$ **then**
        $u = C(r), v = r$
      end
    end
    if $v > 0$ and v not in Q **then**
      $Q = \{Q, v\}$
    end
  end
end
return Q

speech segment. The label function is defined as follows.

$$LL(s) = \{L(t) : t_s(1) \le t \le t_s(N)\} \quad (2)$$

where N is the number of frames in the speech segment, $t_s(1)$ is the starting time point (first frame) of the s-th segment and $t_s(N)$ is the end time point (last frame) of the s-th segment. We can see that a segment-based spectrogram trained through multi-labels obtains frame-level prediction results, and it only needs to be determined once. In this case, multi-label prediction not only keeps the determined level of fuzzy label, but also increases the prediction precision.

2.2 Ensemble of Predicted Results

After all results of the speech frames in the whole speech signal are predicted, if each time point is included in K segments, the speaker change probability for each time point is determined K times. Therefore, the speaker change contour could be estimated by the ensemble of the K predicted results. Figure 2 shows the ensemble process using multiple prediction to estimate a speaker change contour in the conversation. As the speaker change probability at time point t is predicted K times, the max function is used to estimate the contour value.

$$C(t) = \max_{s \in \Theta_t} \left(p_s(t - t_s(1) + 1) \right) \quad (3)$$

where Θ_t is the set of the speech segments in which the time range includes time point t and $p_s(\cdot)$ is the predicted probability of speaker change for the t-th frame.

2.3 Boundary Merging

The speaker change contour can be used to determine when the speaker changes. The simplest way is to determine the speaker change boundaries by a threshold. Assuming that the threshold $\theta = 0.5$, according to the fuzzy labelling, more than one time points of the same speaker change boundary will be alarmed, which means that many incorrect boundaries are found. Thus, we need to merge the boundaries when these boundaries are too close. In

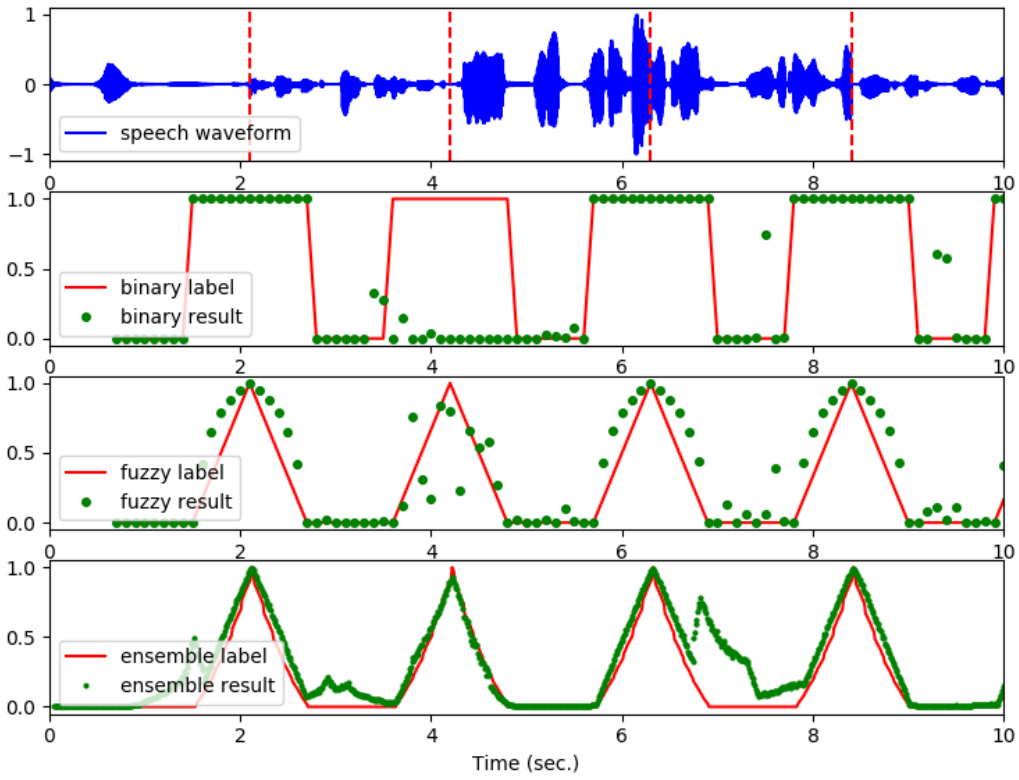

Figure 3. Comparison of the speaker change probability contours obtained from different approaches.

this study, the boundary merging method is shown in Algorithm 1.

3 Experimental Results

3.1 Corpora

This study used two corpora to evaluate the SCD performance and analyzed the effects of domain mismatch.

- **2000 NIST Speaker Recognition Evaluation (LDC2001S97):** This corpus is widely used in speaker diarization experiments. Disk 1 provides audios recorded from 546 females and Disk 2 provides audios recorded from 457 males. We split all audios in these two disks into a sequence of segments of 2.1 second with a time shift of 0.1 second, and we randomly selected and concatenated the segments to generate synthetic conversations for training and testing. This study randomly selected a total of five hours of segments to generate a synthetic conversation for SCD model training and used the randomly selected segments of one hour to generate a synthetic conversation for in-domain data analysis.
- **AMI Meeting Corpus:** The corpus contains 100 hours of meeting recordings. This study utilized the boundary labels of 16 meeting recordings (A. O. T. Hogg et al., 2019) to evaluate the SCD performance in real applications and the effects of out-of-domain data.

3.2 Experimental Setup

The baseline SCD models and the proposed SCD model were trained based on the same backbone architecture as in (M. Hrúz and Z. Zajíc, 2017). The input features were 13-dimentional Mel-frequency cepstral coefficients (MFCCs) with first and second order derivatives. The spectrogram of a segment (consisting of 137 frames) was extracted from a 25ms window with a stride of 10ms, and a

	Acc. (%)	False Alarm (%)	Hit (%)
Binary	86.29	74.41	40.57
Fuzzy	91.14	**17.13**	86.22
Ensemble (proposed)	**92.97**	17.37	**86.29**

Table 1. Results of SCD on the NIST corpus.

Recording Name	Binary		Fuzzy		Ensemble (proposed)	
	Hit (%)	Std. (sec.)	Hit (%)	Std. (sec.)	Hit (%)	Std. (sec.)
ES2004a	72.80	0.2892	62.40	0.2500	75.20	0.2706
ES2004b	63.64	0.2696	57.14	0.2731	76.19	0.2908
ES2004c	76.26	0.2868	65.66	0.2856	80.30	0.3008
ES2004d	60.94	0.2637	68.67	0.2828	72.96	0.2994
IS1009a	72.09	0.2524	81.40	0.3212	76.74	0.2914
IS1009b	76.37	0.2731	75.27	0.2852	73.63	0.2926
IS1009c	69.75	0.2688	69.14	0.2968	74.07	0.2959
IS1009d	81.36	0.2794	66.10	0.2778	79.66	0.2794
EN2002a	66.90	0.2834	68.99	0.2727	73.87	0.2931
EN2002b	62.23	0.2913	59.86	0.2766	70.07	0.2870
EN2002c	60.31	0.2801	59.79	0.2832	67.24	0.3054
EN2002d	69.93	0.2859	65.54	0.2692	75.00	0.3073
TS3003a	85.71	0.2202	47.62	0.2799	80.95	0.3112
TS3003b	84.33	0.2566	63.13	0.3140	73.73	0.3055
TS3003c	78.41	0.2478	60.98	0.3109	77.27	0.3070
TS3003d	75.63	0.2493	65.55	0.3038	78.15	0.3025
Average	72.29	**0.2686**	69.83	0.2864	**75.31**	0.2962

Table 2. Results of SCD on the AMI meeting corpus (domain mismatch).

shift of 0.1 seconds was applied to obtain a sequence of segment-based spectrogram of the input speech signal. The two baseline models ($N = 1$) were trained by the binary labels (M. Hrúz and M. Kunešová, 2016) and fuzzy labels (M. Hrúz and Z. Zajíc, 2017). The speaker change probability is predicted based on the spectrogram of each input frame. The proposed model ($N = 137$) outputted a multi-label prediction.

3.3 Evaluation on NIST Corpus

Because the SCD models were trained by the NIST corpus, the SCD performance can be regarded as in-domain performance. In this experiment, the test signal was obtained by concatenating the speech signals, each with 2.1 seconds from a specific speaker, to form a synthetic conversation, and every ground-truth speaker change applied a collar of 0.2 seconds to tolerate the prediction error.

Figure 3 shows the prediction score contours for speaker change, and the red vertical lines in the first subfigure represents the ground-truth boundaries. We can see that the binary model cannot successfully detect the speaker change boundary, and the fuzzy model does not precisely detect the speaker change boundary and will increase the error of SCD. However, using ensemble prediction can not only increase the time precision of change boundaries, but also improve the detection performance. In Table 1, the experimental results show the performance of SCD in the one-hour synthetic conversations, which contain a total of 1,714 boundaries in the conversations. In the boundary detection task (TRUE or FALSE, decided by the threshold θ), the proposed ensemble prediction achieved the best accuracy of 92.97%, while the binary prediction method achieved an accuracy of 86.29%, which had a very high false alarm rate and a low hit rate. This is because the time point of maximum probability for speaker change is not close to the ground-truth boundary. Therefore, if the test data is in-domain, the ensemble prediction not only achieved the best accuracy of boundary classification, but also obtained the precise time points of SCD.

3.4 Evaluation on AMI Corpus

The AMI corpus has been widely used for SCD system evaluation as in (Z. Fan et al., 2022; M. Kunešová and Z. Zajíc, 2023; S. Kumar et al., 2024), and we used this corpus to evaluate the effects of out-of-domain data. (A. O. T. Hogg et al., 2019) utilized pitch tracking to detect the speaker change boundaries in the AMI corpus. In this study, we used the ground-truth boundaries provided in (A. O. T. Hogg et al., 2019) to evaluate our proposed model. Table 2 shows the SCD performance in 16 AMI recordings. As the SCD models were trained by NIST corpus, using AMI

corpus as the testing data will cause the domain mismatch problem. Thus, this experiment used a collar of 0.5 seconds as the tolerance of the prediction error and decreased the threshold θ to increase the hit rate. At the same conditions, we can see that even there are domain mismatch problem in the experiments, our proposed ensemble prediction also achieved the best hit rate. It is worth noting that the hit rate of fuzzy prediction is worse than binary prediction. The reason may be that the fuzzy model with high complexity is difficult to keep accurate prediction for the case of domain mismatch. As the ensemble prediction is based on fuzzy labelling, the performance of ensemble-based SCD will be affected more than binary prediction. In the experiments, the ensemble prediction obtained the average standard deviation of 0.296 seconds for detection error which was slightly worse than the baseline models, but the ensemble prediction achieved an average hit rate of 75.31%, significantly outperforming the binary prediction (72.29%) and fuzzy prediction (69.83%). Therefore, even though the domain mismatch greatly affects the SCD performance, the proposed SCD model based on ensemble prediction can further consider the predicted probabilities of a change boundary in the sequential segments and integrate them to obtain the final prediction results.

4 Conclusions

This paper proposes an ensemble prediction approach to SCD. First, a CNN-based SCD model is trained for multi-label prediction. This method can effectively predict the frame-level results from a segment-level spectrogram. Next, considering each boundary can be predicted more than once through the sequential segments, the maximum probability of speaker change of the same boundary was regarded as the final prediction probability. In the experiments, we found that if the data is in-domain, the proposed ensemble prediction achieved the best hit rate and obtained precise time points of SCD. If the data is out-of-domain, the ensemble prediction can further consider the prediction probabilities of a change boundary in the sequential segments to improve the performance of SCD.

References

S. Kumar, S. Madikeri, I. Nigmatulina, E. Villatoro-Tello, P. Motlicek, K. Pandia, S. P. Dubagunta, A. Ganapathiraju. 2024. Multitask speech recognition and speaker change detection for unknown number of speakers. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12592–12596.

J. Wu, Z. Chen, M. Hu, X. Xiao, and J. Li. 2023. Speaker change detection for transformer transducer ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

L. Sarı, S. Thomas, M. Hasegawa-Johnson and M. Picheny. 2019. Pre-training of speaker embeddings for low-latency speaker change detection in broadcast news. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6286–6290.

X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. 2012. Speaker diarization: a review of recent research. *Journal of the IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356-370.

K. VijayKumar and R.R. Rao. 2023. Optimized speaker change detection approach for speaker segmentation towards speaker diarization based on deep learning. *Journal of the Data & Knowledge Engineering*, 144:102121.

W. Xia, H. Lu, Q. Wang, A. Tripathi, Y. Huang, I. L. Moreno, and H. Sak. 2022. Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8077–8081.

S. Chen and P.S. Gopalakrishnan. 1998. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*, pages 127–132.

J. E. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez. 2006. Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages V–V.

L. Mateju, P. Cerva, and J. Zdansky. 2019. An approach to online speaker change point detection using DNNs and WFSTs. In *Proceedings of the INTERSPEECH*, pages 649–653.

V. Gupta. 2015. Speaker change point detection using deep neural nets. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4420–4424.

R. Wang, M. Gu, L. Li, M. Xu, and T. F. Zheng. 2017. Speaker segmentation using deep speaker vectors for fast speaker change scenarios. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5420–5424.

M. Hrúz and Z. Zajíc. 2017. Convolutional neural network for speaker change detection in telephone speaker diarization system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4945–4949.

Z. Zhou, Y. Zhang, and Z. Duan. 2018. Joint speaker diarization and recognition using convolutional and recurrent neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2496–2500.

Z. Zajíc, M. Kunešová, J. Zelinka, and M. Hrúz. 2018. Zcu-ntis speaker diarization system for the dihard 2018 challenge. In *Proceedings of the INTERSPEECH*, pages 2788–2792.

J. Yi, J. Tao, Z. Wen, and Y. Li. 2017. Distilling knowledge from an ensemble of models for punctuation prediction. In *Proceedings of the INTERSPEECH*, pages 2779–2783.

A. O. T. Hogg, C. Evers, and P. A. Naylor. 2019. Speaker change detection using fundamental frequency with application to multi-talker segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5826–5830.

M. Hrúz and M. Kunešová. 2016. Convolutional neural network in the task of speaker change detection. In *Proceedings of the International Conference on Speech and Computer*, pages 191–198.

Z. Fan, L. Dong, M. Cai, Z. Ma, and B. Xu. 2022. Sequence-level speaker change detection with difference-based continuous integrate-and-fire. *Journal of IEEE Signal Processing Letters*, 29:1551-1554.

M. Kunešová and Z. Zajíc. 2023. Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.