Three-class Overlapped Speech Detection using a Convolutional Recurrent Neural Network

Jee-ween Jung, Hee-Soo Heo, Youngki Kwon, Joon Son Chung, Bong-Jin Lee

Naver Corporation, South Korea
jejeweon.jung@navercorp.com

Abstract

In this work, we propose an overlapped speech detection system trained as a three-class classifier. Unlike conventional systems that perform binary classification as to whether or not a frame contains overlapped speech, the proposed approach classifies into three classes: non-speech, single speaker speech, and overlapped speech. By training a network with the more detailed label definition, the model can learn a better notion on deciding the number of speakers included in a given frame. A convolutional recurrent neural network architecture is explored to benefit from both convolutional layer’s capability to model local patterns and recurrent layer’s ability to model sequential information. The proposed overlapped speech detection model establishes a state-of-the-art performance with a precision of 0.6648 and a recall of 0.3222 on the DIHARD II evaluation set, showing a 20% increase in recall along with higher precision. In addition, we also introduce a simple approach to utilize the proposed overlapped speech detection model for speaker diarization which ranked third place in the Track 1 of the DIHARD III challenge.

Index Terms: overlapped speech detection, convolutional recurrent neural network, speaker diarization, deep learning

1. Introduction

Overlapped speech detection (OSD) is a task that estimates on-sets and offsets of segments (i.e., a small part of an audio clip) within an audio clip (i.e., utterance, session, conversation as a whole) where more than one speaker is speaking simultaneously. OSD can be used for various applications, among which is speaker diarization where OSD is reported to play a critical role [1–9]. Without an OSD system, the performance of a speaker diarization system degrades, proportionally to the existing overlaps in a dataset, because the model conclusively misses the second speaker in the overlapped speech region.

Before the recent advances in deep neural networks (DNNs), hidden Markov models and Gaussian mixture models were widely used to build OSD systems [10, 11]. Martin et al. shows that the hidden Markov model can successfully perform OSD where each state is defined using a Gaussian mixture model. In recent works, OSD is treated as a sequence labeling task where convolutional neural networks (CNNs) or recurrent neural networks (RNNs) are trained as a binary classifier. However, when using DNNs for OSD, only two classes (overlapped and non-overlapped) are defined. We argue that the DNN can also benefit from defining more specific classes based on additional motivations addressed in Section 2.1. Second, we use a convolutional recurrent neural network (CRNN) architecture to merit from both architecture’s features. Although studies adopting CRNN show state-of-the-art performance in other sequence labeling tasks such as sound event detection [16], existing OSD models adopt variants of either a CNN or a RNN. We assume that OSD models can be also improved by adopting a CRNN architecture, benefiting from both CNN and RNN’s unique modeling capability. In the upcoming sections, we explain and experimentally show that both proposals increase the performance of an OSD model and establish state-of-the-art performance.

Exploiting the developed OSD model, we also introduce a simple algorithm for detecting simultaneously speaking speakers in a diarization task. It adds two sub-processes using an OSD model to the speaker diarization pipeline. First, we add a sub-process that further segments speech activity detection (SAD) module’s output where the onsets and offsets of overlapped segments are adjusted using the SAD result. Second, we provide an additional label to the overlapped segment using the second ranked speaker based on the clustering result of speaker diarization. This method works independently from the mainstream speaker diarization pipeline, making it simple to utilize yet effective, ranking third in the DIHARD III challenge Track 1 core evaluation [17].

2. Overlapped speech detection

Defined as a sequence labeling task, an OSD model inputs \( x = \{x_1, x_2, \ldots, x_t\} \in \mathbb{R}^t \) where \( t \) is the length of input sequence and \( f \) refers to the dimensionality of each element. Acoustic features (e.g., mel-spectrograms or mel frequency cepstral coefficients) are widely chosen as the input feature. Corresponding label \( y = \{y_1, y_2, \ldots, y_t\}, y_i \in \mathbb{R} \) is used to guide the training where \( 0 \) refers to not overlapped and \( 1 \) refers to overlapped (can be the other way). The output layer of the model is either one node, trained using a binary cross-entropy loss function, or two nodes, trained using a categorical cross-entropy loss function.

The overall framework of our proposed OSD model when conducting inference using an input audio clip is illustrated in Figure 1. Note that for conventional DNN-based OSD models, labels are either 0 or 1 as mentioned earlier, but for the proposed three-class OSD model, labels are either 0, 1, or 2 indicating non-speech, single speaker speech, and overlapped speech. We adopt mel-spectrograms as the input feature, using a 25ms window and 10ms shift size. The OSD model inputs 150 frames (≈ 1.5s) at a time with 50 frames shift. The node of the output layer that indicates overlapped speech is used. Except for the...
first and the last 100 frames, the sum of model output is divided by three accounting for the multiple predictions caused by shifting the input. To compensate for the decrease in the length of sequence affected by pooling layers, the values are duplicated. The main contributions of this paper, three-class classification and the CRNN architecture, are described in Sections 2.1 and 2.2 respectively.

2.1. OSD as a three-class classification

Conventional OSD systems using DNNs are trained to perform binary classification because the objective at the inference phase is to decide whether overlapped speech exists for each frame. It is a widely accepted straightforward approach, training the model to perform as one wants. However, by training the OSD model as a two-class classifier, both non-speech and overlapped speech, respectively. To process the entire audio clip longer than the defined sequence length of 1.5 seconds, every 150 frames are input to the model with 50 frames shift and model outputs are combined.

Table 1: Architecture of the proposed CRNN for three-class OSD. A mel-spectrogram of size (150, 128, 1) is input to the DNN where 150 refers to the length of sequence (1.5s), 128 is the number of mel bins, and 1 is the number of channels. Two numbers in the Conv bracket refers to filter length and stride size (BN: batch normalization, SE: squeeze-excitation). Average is applied to the mel bin dimension, before the GRU layer.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Input:(#seq,#mel,#filt)</th>
<th>Output shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>Conv(3,1)</td>
<td>×3 (25, 32, 128)</td>
</tr>
<tr>
<td></td>
<td>BN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conv(3,1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ReLU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AveragePool</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>Average</td>
<td>(25,128)</td>
</tr>
<tr>
<td>GRU</td>
<td>bi-GRU(256)×2</td>
<td>(512,)</td>
</tr>
<tr>
<td>Linear</td>
<td>FC(256)</td>
<td>(3,)</td>
</tr>
<tr>
<td></td>
<td>Dropout(0.5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LeakyReLU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linear(3)</td>
<td></td>
</tr>
</tbody>
</table>

viding not overlapped class into single speaker speech and non-speech classes, we argue that not only the problem occurred by grouping different categories may decrease, but also the DNN can leverage from additional information. It has commonalities in terms of motivation with the multitask learning [18] framework where more specific and concrete information is additionally provided to help learning. Also, precise class balancing using loss weights becomes possible. The proposed three-class scheme has no additional requirements when performing inference. One can simply use the node of the output layer that indicates overlapped speech (before or after applying softmax non-linearity) as the model’s score for the given input frame.

2.2. CRNN architecture

Most preceding studies that use DNN-based models for the OSD task either utilize a CNN or a RNN [5, 12–14]. However, we propose to use a CRNN architecture with three convolution blocks and two gated recurrent unit (GRU) layers. Three motivations led us to propose and use the CRNN architecture. First, CNNs and RNNs have different aspects that the OSD model can benefit from. CNNs are known for their capability of modeling local patterns and RNNs are widely used for their capability of modeling sequential data. We argue that by using a CRNN architecture, advantages of both CNNs and RNNs can be exploited. Second, decreasing the length of sequence via pooling layers included in CNN will facilitate the training of GRU layers. In addition, decreasing the sequence length will give a smoothing effect on the model’s output. For example, according to our configuration when 150 frames are input, the CRNN model will output 25 frames where each output frame represents six input frames. Third, a CRNN architecture shows state-of-the-art performance in sound event detection [16] which is also a sequence labeling task that predicts onsets and offsets of various predefined sound events. 

Table 1 denotes the proposed CRNN architecture used for OSD. Three convolution blocks are used where the length of a
sequence is divided by 6, followed by an average on the mel bin dimension. A squeeze-excitation [19] is applied to each block. Details on how the length of the sequence and the number of mel bins are decreased are addressed in Section 4.2. The output of CNN is input to two bidirectional GRU layers that output energies) as input to the DNN with a window size of 25ms and a 10ms shift. Pre-emphasis is applied before applying Fourier transforms and mean normalization is applied to the extracted log mel-spectrograms.

In order to utilize the proposed OSD model for speaker diarization, we introduce two simple steps that can be applied on top of the existing speaker diarization pipeline. First, we use the OSD model to further divide SAD results into single speaker segments and overlapped speech segments. Regions where both the OSD model and the SAD model detect as overlapped speech segment. Figure 2 illustrates three possible cases of the OSD model further segmenting the SAD segment. If the OSD onset is earlier than the SAD onset, SAD onset is used as the onset of an overlapped segment and OSD offset is used as modified SAD segment onset (Figure 2-(a)). If the OSD offset positions later than the SAD offset, SAD offset is used as an overlapped segment’s offset and SAD offset is replaced by OSD offset (Figure 2-(b)). When the OSD segment exists in the middle of an SAD segment, it is divided into three segments (Figure 2-(c)).

The other application of the OSD model for speaker diarization occurs after the clustering phase, in a similar manner with [14]. When performing clustering, there exists a speaker posterior matrix where speakers are aligned from the most likely to least likely for each segment. For overlapped segments, we simply allocate the second most likely speaker as an additional label. This approach is simple, easy to implement, and can be applied to the result of various clustering methods (e.g., spectral and agglomerative hierarchical clustering [22–24]) that can derive the second most likely speaker for each segment. However, note that the limitation of this approach is not being able to detect when three or more speakers are uttering simultaneously.

Aforementioned approach is applied to the Clova team’s speaker diarization model [25]. This model uses a reference SAD following the rules of Track 1 of the DIHARD III challenge. For speaker embedding extraction, a variant of ResNet34V2 in voxceleb_trainer is used. Speaker embeddings are then fed into two feature enhancement modules, followed by spectral clustering and agglomerative hierarchical clustering [22,23]. Further information regarding this speaker diarization system including feature enhancement techniques can be found in [21].

### 4. Experiments

#### 4.1. Datasets

We use four datasets to train the proposed OSD model: AMI Meeting corpus [26], DIHARD I development set [27], DIHARD II development set [28], and VoxConverse dataset [29]. The AMI corpus includes 100 hours of meeting conversations. DIHARD I and II development sets include 19 and 23 hours of data from diverse domains respectively. The VoxConverse dataset consists of approximately 20 hours of conversation from YouTube videos. For all datasets, multi-channel audio clips were modified to single-channel and the sample rate of all audio clips was adjusted to 16kHz. The DIHARD II evaluation set and the DIHARD III development set [17] are used to report the performance of the proposed OSD model. OSD model’s performance is reported using precision and recall. Speaker diarization result is reported in DER, the summation of false alarm, miss, and confusion.

#### 4.2. Configurations and details

We use a 128-dimensional log mel-spectrograms (i.e., filterbank energies) as input to the DNN with a window size of 25ms and 10ms shift. Pre-emphasis is applied before applying Fourier transforms and mean normalization is applied to the extracted log mel-spectrograms.

**Data augmentation.** A copy of all audio clips with a sampling rate of 16kHz is augmented by downsampling to 8kHz and upsampling back. We also apply data augmentation using reverberation and noise from RIRs and Musan datasets [30,31], following existing recipes [20,21]. Overlapped speech sample augmentation following the recipe introduced in Bullock et al. [14].

![Image](https://github.com/clovaai/voxceleb_trainer)
We propose an OSD model that is trained as a three-class classifier dividing the non-overlapped class into non-speech and single speaker speech classes. A CRNN architecture was also proposed to incorporate both CNN and RNN’s capability in detecting overlapped speech. Combining two proposals, our OSD model demonstrates state-of-the-art performance on the DIHARD II evaluation set. An approach of using a developed OSD model for speaker diarization was also addressed, where the system that followed this approach ranked third place in DIHARD III Track 1 challenge.
6. References


3090