END-TO-END OVERLAPPED SPEECH DETECTION AND SPEAKER COUNTING WITH RAW WAVEFORM

Wangyou Zhang[1], *Man Sun*[1], *Lan Wang*[2], *Yanmin Qian*[1†]

[1]MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

ABSTRACT

Overlapped speech processing has attracted more and more attention in recent years, and it is a key problem when processing multi-talker mixed speech under the cocktail party scenario. It is commonly observed that the performance of overlapped speech processing can be significantly improved if the number of speakers is given in advance. However, such prior knowledge is often unavailable in real-world conditions, so a robust overlapped speech detection and speaker counting system is demanded. Most existing works focus on combining different handcrafted features to tackle this task, which can be sub-optimal since there are no direct connections between the features and the task. In this work, we try to solve these two problems with an end-to-end manner. First, an end-to-end framework for overlapped speech detection and speaker counting is proposed, which extracts features from the raw waveform directly. Then a curriculum learning strategy is applied to make better use of the training data. The proposed methods are evaluated on multi-talker mixed speech generated from the LibriSpeech corpus. Experimental results show that our proposed methods outperform the model with handcrafted features on both tasks, achieving more than 2% and 4% absolute accuracy improvement on overlapped speech detection and speaker counting respectively.

Index Terms— end-to-end, raw waveform, overlapped speech detection, speaker counting, deep learning

1. INTRODUCTION

Although significant progress has been made in recent years on intelligent speech processing, the performance of current speech processing systems still degrades severely in the complex real-world conditions, especially under the cocktail party scenarios[1, 2]. The cocktail party problem defines a complicated scenario where multiple talkers speak simultaneously and other background noise is involved. Under such scenarios, the accurate overlapped speech detection and speaker counting can be very useful for later speech processing, such as speaker diarization [3, 4], speaker localization [5], speaker recognition [6, 7] and automatic speech recognition (ASR) [8, 9].

Recently, there have been several researches studying the methods for overlapped speech detection and competing speaker counting under the deep learning framework. [10] proposed to use a stacked convolutional network architecture for speaker counting, where each convolutional block consists of 3 pairs of convolutional layers and max pooling layers. [11] proposed a deep neural network (DNN) based multi-speaker localization system, whose output is encoded to represent the likelihood of a speaker being in each direction and can naturally detect the number of speakers by counting the peaks. In [12], a long-short term memory (LSTM) based network architecture is proposed to learn the presence of overlap in speech from the input spectrotemporal features. However, these works primarily focus on selecting and combining traditional handcrafted features, such as spectrogram, Mel-frequency cepstral coefficients (MFCCs), generalized cross correlation with phase transform (GCC-PHAT) and signal envelope, which might not be optimal for these two tasks.

In this work, we aim to learn a mapping from the observed mixed speech signal to the corresponding number of speakers directly. Our motivation is that, since neural networks have been proven capable of learning to extract appropriate task-specific features, we can also exploit this capability in the overlapped speech detection and speaker counting task, while the connections between existing handcrafted features and the two tasks are indirect and ambiguous. Therefore, in this paper, we utilize the convolutional neural network as a deep feature extractor, whose input is the raw waveform. Then the feature extractor is integrated into a larger convolutional architecture to train the entire model in an end-to-end manner.

To the best of our knowledge, this is the first work that performs end-to-end overlapped speech detection and speaker counting with raw waveform. Our experimental results have proven the effectiveness and robustness of the proposed

†Yanmin Qian is the corresponding author.

method in both tasks.

The remainder of the paper is organized as follows: Section 2 describes the overlapped speech detection and speaker counting tasks. Section 3 introduces the proposed end-to-end architecture with raw waveform as the input. The curriculum learning is designed to further improve system performance for this task and it is described in Section 4. Then Section 5 describes our experimental setups and used dataset, and demonstrates the results of both the baseline and the proposed methods. Finally, a conclusion is summarized in Section 6.

2. PROBLEM DESCRIPTION

Under the cocktail party scenario, the received speech signal often consists of overlapped speech from multiple speakers, which can be formulated as:

$$x(t) = \sum_{n=1}^{N} s_n(t) \tag{1}$$

where $s_n(t)$ denotes the speech from the n-th speaker and N is the total number of speakers.

For different speakers, the corresponding speech usually has different duration and onsets, which can be formulated as :

$$s_n(t) = \begin{cases} s'_n(t - \tau_{0n}), & \tau_{0n} \leq t \leq \tau_{0n} + T_n \\ 0, & \text{Otherwise} \end{cases} \tag{2}$$

where the subscript n represents the n-th speaker ($n = 1, 2, \cdots, N$), τ_{0n} denotes the beginning of the speech, T_n denotes the total duration, and $s'_n(t)$ is the onset-aligned speech.

Under the above assumption, the number of simultaneous speakers can change over time even within one segment of speech. In order to simplify the situation, we preprocess the input speech with voice activity detection (VAD), which ensures the presence of all speakers throughout each speech sample. And speech from different speakers is segmented into 500-ms fragments. Therefore, we only need to consider Equation (1) for both two tasks in the rest of this paper.

Based on this consideration, the speaker counting task is to estimate N from the mixed speech $x(t)$, while the overlapped speech detection task is to estimate whether N is larger than 1. Thus both two tasks can be formulated as an I-class classification problem. For overlapped speech detection, it is a binary classification problem with $I = 2$. For speaker counting, each class corresponds to a possible number of speakers, and we select $I = 4$ in our experiments. This selection is based on the investigation in [10], which reports that humans have difficulties in distinguishing more than four simultaneous speakers. Therefore, we can assume that handling at most four simultaneous speakers can satisfy the needs in most real-world applications.

3. MODELS FOR OVERLAPPED SPEECH DETECTION AND SPEAKER COUNTING

In this section, we first describe the baseline method using a stacked convolutional network architecture for overlapped speech detection and speaker counting, and then propose an end-to-end architecture to improve the performance on both two tasks. Since overlapped speech detection and speaker counting can be formulated as similar classification problems, as described in Section 2, we will adopt the same architecture for both two tasks in each method below.

3.1. Baseline CNN Model

The stacked convolutional network architecture proposed in [10] is adopted as the baseline model in our experiments. The model consists of three consecutive convolutional blocks and then a batch normalization layer, followed by three fully connected layers. Each convolutional block is composed of three pairs of convolutional layers and max pooling layers. Besides, the dropout is applied after the last convolutional block and after each fully connected layer.

As mentioned in Section 2, the duration of the input sample is 500 ms, which is moderately long and has been proven to result in relative high classification accuracy in the speaker counting task [10]. The input feature set of the network is a concatenation of three traditional handcrafted features, including the flattened spectrogram, signal envelope computed with Hilbert transform and histogram of speech signal, as described in Table II in [10]. The output is a 4-dimensional likelihood vector for speaker counting and a 2-dimensional vector for overlapped speech detection.

Although the baseline CNN architecture already produces good performance on the speaker counting task, it is still limited due to the handcrafted features, which might not be optimal for the task. In addition, the selection of features can vary under different scenarios, which requires careful design and more efforts. So we propose an end-to-end architecture to build the systems for overlapped speech detection and speaker counting directly, which will be described in the following section.

3.2. Proposed End-to-End Model

In order to extract the most related feature for speaker counting or overlapped speech detection, we introduced a CNN-based feature extraction module before the network in Section 3.1. It consists of a convolutional layer with 256 channels and a 64×1 kernel, followed by a max pooling layer with a 2×2 kernel. For a 500-ms input raw waveform, a 3968×256 feature is extracted as the input to the stacked convolutional model as described in Section 3.1, which can be illustrated in Fig. 1.

Different from the architecture in the previous section, a batch normalization layer, a rectified linear unit (ReLU)

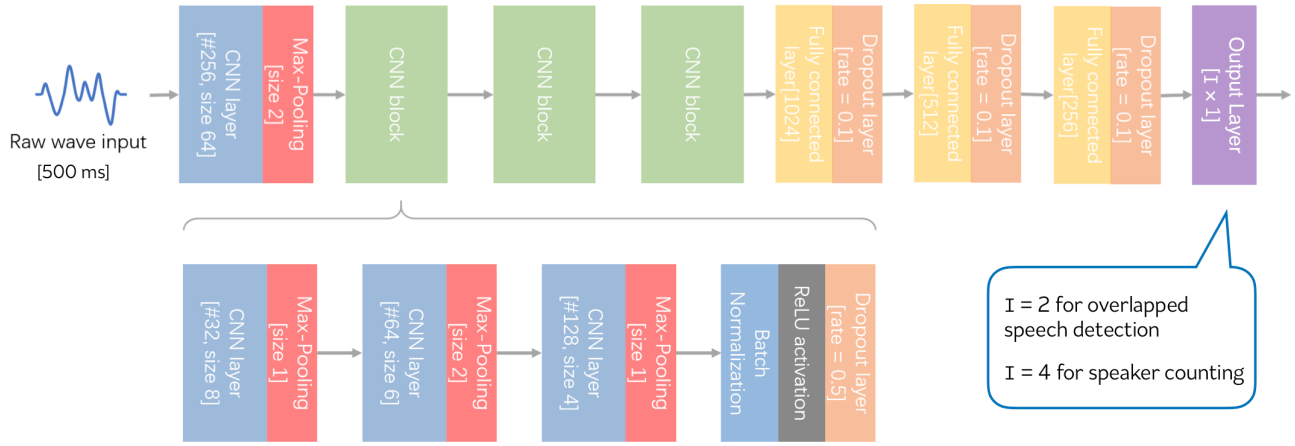


Fig. 1: Model Architecture of proposed End-to-End method

layer and a dropout layer are appended to the end of each convolutional block in our proposed model. This is based on the observation that the end-to-end model trained without constraints in the intermediate layers can face the problem that changes in the distribution of the input in previous layers are amplified layer-wisely, which increases the difficulty in adapting the model to different training samples. As pointed in [13], the batch normalization layer can achieve a stable distribution of activation values during training, thus enabling a more stable training process. Besides, since more parameters are introduced in the feature extraction module, the complexity of the model has increased, making it easier to overfit. Thus we add an extra dropout layer in each convolutional block to help reduce the possibility of overfitting.

The output layer in Fig. 1 is a linear layer followed by the SoftMax activation function, and the output dimensions for the two tasks are the same as those in Section 3.1.

In addition, a training trick is adopted to optimize the training process by reducing the learning rate by half when the accuracy on validation dataset stops increasing for 10 epochs. For CNN training, the cross entropy loss function is used, which is defined as

$$\mathcal{L} = -\sum_{n=1}^{N}\sum_{i=1}^{I} t_{n,i}\log(p_{n,i}) \tag{3}$$

where n denotes the n-th training sample, N is the number of samples in a batch, i denotes the i-th class, I is the number of classes, $p_{n,i}$ denotes the probability of the n-th sample belonging to the i-th class, and $t_{n,i}$ is defined as

$$t_{n,i} = \begin{cases} 1, & \text{if } c_n = i \\ 0, & \text{if } c_n \neq i \end{cases} \tag{4}$$

where c_n denotes the label of the n-th training sample.

4. CURRICULUM LEARNING FOR END-TO-END SPEAKER COUNTING

Since the overlapped speech detection task can be regarded as a subtask of speaker counting, we will focus on improving the latter task in this section.

In previous works, the classification model is trained directly with all the training data, disregarding the similarity and relation among different classes. However, some works [14, 15] have proven the order of training data can influence the training process, especially when the data are sorted from "easiest" to "hardest", which is called curriculum learning. The difficulty of a sample can vary for different tasks.

In this paper, we adopt the idea of curriculum learning to optimize the training process of our proposed model. Since the classification accuracy apparently decreases with a larger number of speakers, as illustrated in Section 5, we can define the difficulty of a sample according to its label, i.e. the number of involved speakers. However, the normal procedures in curriculum learning are not applied in our experiment, because the sorting criterion is directly related to the labels of the data, which will result in imbalanced data distribution in each minibatch and thus cause overfitting. To overcome this problem, we propose to train the model in three stages, as described in Algorithm 1. In each stage, the data are rearranged into I classes and an I-class classifier is trained based on the $(I-1)$-class classifier trained in the last stage[1], where I is 2, 3 and 4 for stage 1, 2 and 3 respectively.

5. EXPERIMENT

In this section, we first introduce the preparation process of the multi-speaker overlapped speech dataset. Then the baseline and our proposed models are evaluated on the generated dataset.

[1]For stage 1, the model is initialized randomly since no pre-trained model is available.

Algorithm 1: Curriculum learning for E2E speaker counting

for *stage :* $1 \rightarrow 3$ **do**
  Load the training dataset $\mathbf{X}_{\text{tr}}$ and validation dataset $\mathbf{X}_{\text{cv}}$;
  $I = stage + 1$;
  for $i : I \rightarrow 4$ **do**
    Relabel data in class i to class I;
  end
  Denote the relabelled datasets by $\mathbf{X}'_{\text{tr}}$ and $\mathbf{X}'_{\text{cv}}$;
  Shuffle $\mathbf{X}'_{\text{tr}}$ and $\mathbf{X}'_{\text{cv}}$, and divide them into minibatch sets respectively;
  if *stage* > 1 **then**
    Initialize an *I*-class classifier with the model trained in the last stage;
  else
    Initialize an *I*-class classifier randomly;
  end
  while *model is not converged* **do**
    for *each* $\mathcal{B}$ *in all minibatches* **do**
      Update the model with minibatch $\mathcal{B}$;
    end
  end
end

5.1. Experimental Setup

In our experiments, a well-annotated multi-speaker overlapped speech dataset is required, where the overlapped status and number of speakers in each segment of the speech should be labelled in order to ensure training convergence and accurate evaluation. However, currently there is no open-source dataset that matches our need, and we decide to artificially generate the multi-speaker mixed speech, as has been done in many previous works on overlapped speech detection [10, 16, 17].

To generate the multi-speaker data, we first randomly choose one to four single-speaker speech samples from LibriSpeech [18], an open-source ASR dataset consisting of recordings of 16kHz read English speech. Then these samples are preprocessed by voice activity detection (VAD)[2] and segmented to 500-ms fragments to ensure the presence of all speakers throughout each segment. Finally, these fragments from different speakers are mixed together directly to generate overlapped speech. The training and validation datasets are generated from LibriSpeech dev-clean subdataset, including a 5.4-hour corpus from 20 male and 20 female speakers. The evaluation dataset is generated from LibriSpeech test-clean subdataset, which is also a 5.4-hour corpus from another 20 male and 20 female speakers. Note that the speakers in the evaluation dataset are totally different from those in the training and validation dataset. The total number of samples in training, validation and evaluation datasets is 576000, 24000 and 100000 respectively, while the data containing different number of speakers are approximately of the same size within each dataset. The duration of each dataset is as follows: 80 hours for training and validation, and 14 hours for evaluation.

For the baseline model, the input is a 7197-dimensional mixed feature, which is the concatenation of flattened spectrogram, signal envelope and histogram of the speech signal. For our proposed model, the input is a 8000-dimensional raw waveform with the duration of 500 ms.

In both overlapped speech detection and speaker counting tasks, the learning rate is initially set to $\alpha = 0.0005$, and the Adam optimizer is used during training. All models are trained for 50 epochs with a batch size of 200, and the model with the highest accuracy on validation dataset is selected for evaluation.

5.2. Performance on Overlapped Speech Detection

First the performance of the models on overlapped speech detection is evaluated via classification accuracy and F1 score, and the results are presented in Table 1. We can observe that approximately 2% absolute accuracy improvement and 1.5% absolute F1 score improvement are achieved by our proposed end-to-end model, which indicates that our end-to-end architecture has the capability of extracting more related features for overlapped speech detection, thus improving the performance.

In addition, since the overlapped speech detection task can be regarded as a subtask of speaker counting, as described in Section 2, we also utilize the trained speaker counting model in Section 5.3 to initialize the overlapped speech detection model, and then fine tune the model under the new task. The results are presented in the last row in Table 1, which shows a further improvement can be obtained using the pre-trained method.

Table 1: Performance comparison of different methods for overlapped speech detection.

Model	**Accuracy** (%)	**F1 score**
baseline	94.98	0.9667
proposed	96.90	0.9794
+ pre-train	**97.60**	**0.9839**

5.3. Performance on Speaker Counting

In this section, we evaluate the performance of the models on speaker counting, which is described in Table 2. Note that the F1 score is the averaged F1 measure among different labels,

[2]VAD is performed using the open-source toolbox called VOICEBOX (www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html).

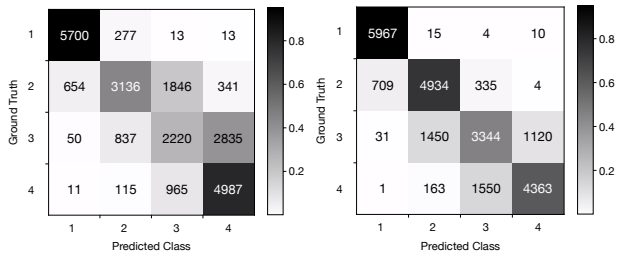
and for each label, the F1 measure is calculated as in binary classification.

As we can see, The classification accuracy of the baseline model is 72.42%, which is comparable to the result (70.5%) in [10] with similar model and experimental setup. And the proposed model outperforms the baseline model with over 4% absolute accuracy improvement, which confirms our previous assumption that more related features can better match the speaker counting task and improve the performance. Then we further utilize the pre-trained overlapped speech detection model to initialize this model before training, which also results in another improvement as shown at the bottom line in Table 2.

Table 2: Performance comparison of different methods for speaker counting.

Model	**Accuracy** (%)	**F1 score**
baseline	72.42	0.7218
proposed	76.31	0.7342
+ pre-train	**76.66**	**0.7609**

To better illustrate the effectiveness of our proposed model, we visualize the performance of both models in the form of a confusion matrix, which represents the distribution of the classified samples. As shown in Fig. 2, the baseline model with handcrafted input features shows less confidence for classification when a moderate number of speakers are involved. The number of misclassified samples is even larger than that of correctly classified samples in the third row. In contrast, our proposed model is more robust against different numbers of speakers and the distribution of the predicted samples are more concentrated on the correct class, which is consistent with the overall accuracy.



(a) Baseline (b) Proposed method

Fig. 2: Confusion matrices of different methods on the validation dataset[3].

5.4. Curriculum Learning

In this section, we investigate the effect of the training strategy on the performance of speaker counting. The curriculum learning strategy described in Section 4 is applied to the speaker counting model, with at most 20 epochs for the first two stages and 40 epochs for stage 3. In stage 1 and stage 2, the model with the highest accuracy on validation dataset is selected to initialize the model in the next stage, thus transferring the knowledge learned in a simpler task to a similar but harder one.

The performance of our proposed speaker counting model with different training strategies is presented in Table 3. We can observe that the curriculum learning strategy can improve both classification accuracy and F1 score based on our proposed model.

Table 3: Performance comparison of the proposed speaker counting model with or without curriculum learning on test dataset[4].

Training strategy	**Accuracy** (%)	**F1 score**
no curriculum	76.31	0.7342
curriculum	**76.64**	0.7557

5.5. Analysis of extracted features

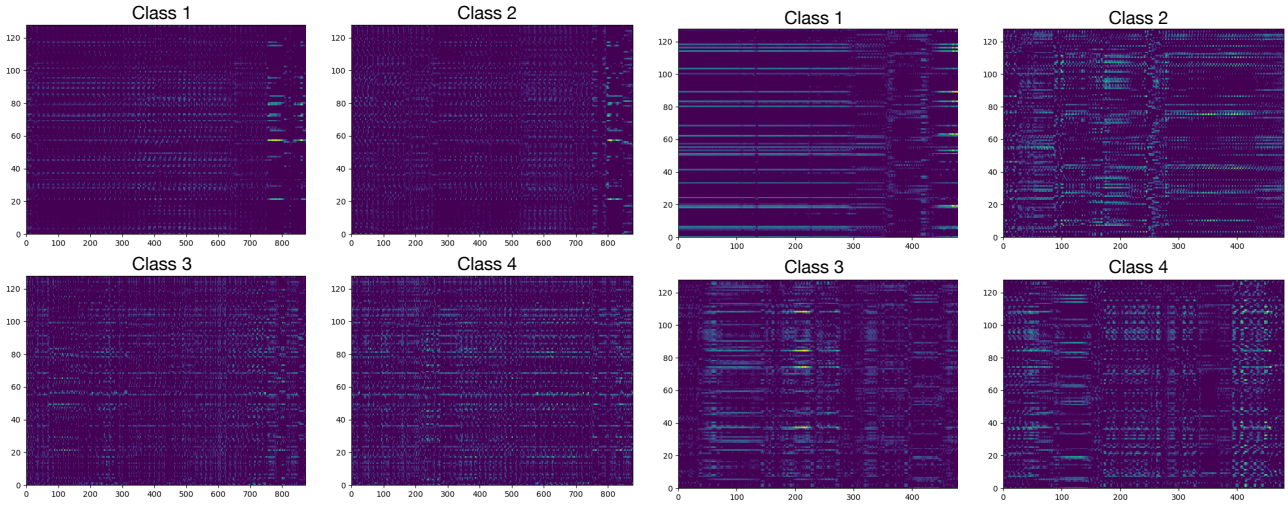
In order to understand the capability of extracting feature representations of our proposed model more intuitively, we randomly select input samples from four different classes and visualize the intermediate representations of these samples after the last convolutional block for both baseline and our proposed model, which are depicted in Fig. 3. Consider that the convolution and pooling operations are only performed along the time axis in our proposed model, each column in the representation corresponds to one time frame, while each row corresponds to one output channel in the CNN.

As we can see in Fig. 3(a), the difference among representations of samples from four classes is hardly observable, and the pattern of each class is ambiguous and unclear. In Fig. 3(b), however, the pattern of each class is more apparent and clearer, with more details in local areas. In addition, the patterns of different classes are more distinguishable, which is consistent with the classification performance in Section 5.3.

An interesting phenomenon is that there exist some discontinuous patterns along the horizontal (time) axis in the representations and more discontinuities are observed when a larger number of speakers are involved such as in Class 4. This is also common in sound localization, where different frames can be dominated by different speakers, resulting in different patterns. In addition, we can also observe that some continuous patterns may appear repeatedly along the time axis in one representation. Therefore, we can assume that our model is capable of learning the representations for different speakers, thus estimating the number of speakers.

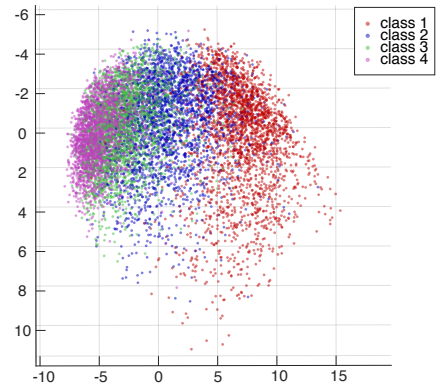
Furthermore, we perform principal component analysis (PCA) on both features of the test samples and compare the

[3]Class i ($i = 1, 2, 3, 4$) represents i speaker(s) in the speech.

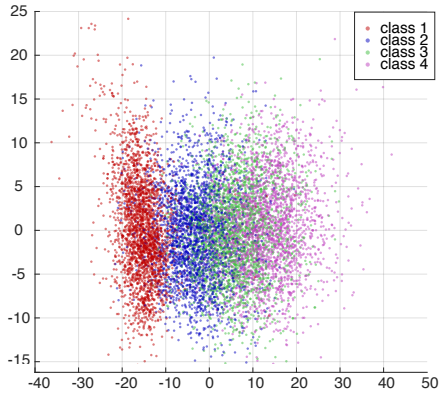


(a) Baseline model (b) Proposed model

Fig. 3: Representations of input data from different classes after the last convolutional block.



(a) Baseline model



(b) Proposed model

Fig. 4: Visualization results of the intermediate representations using PCA.

visualization results in Fig. 4, which also demonstrates that our proposed model can extract features with better separability among different classes.

With above comparison between the baseline and the proposed model, we further confirm that our proposed end-to-end model has stronger capability of extracting more related features for the speaker counting task, which is consistent with the results in Table 2.

6. CONCLUSION

In this work, we propose an end-to-end architecture with raw waveform input for both overlapped speech detection and speaker counting tasks. The proposed model is evaluated in the mixed speech generated from LibriSpeech and outperforms the baseline model with handcrafted input features in both two tasks. More than 2% and 4% absolute accuracy improvement is obtained in overlapped speech detection and speaker counting respectively. Besides, a curriculum learning strategy is applied to make better use of the training data, which also improves the performance. Future work will consider different acoustic scenarios with noise and reverberation as well as other curriculum learning strategies for these tasks.

7. ACKNOWLEDGEMENT

This work was supported by the China NSFC projects (No. 61603252 and No. U1736202) and joint project from Shenzhen Institutes of Advanced Technology in CAS. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

8. REFERENCES

[1] E Colin Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] Mark A Bee and Christophe Micheyl, "The cocktail party problem: what is it? how can it be solved? and why should animal behaviorists study it?," *Journal of comparative psychology*, vol. 122, no. 3, pp. 235, 2008.

[3] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4353–4356.

[4] Mireia Diez, Federico Landini, Lukáš Burget, Johan Rohdin, Anna Silnova, K Zmolıková, Ondřej Novotnỳ, Karel Veselỳ, Ondřej Glembek, Oldřich Plchot, et al., "BUT system for DIHARD speech diarization challenge 2018," in *Proc. Interspeech*, 2018, pp. 2798–2802.

[5] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, 2018.

[6] André G Adam, Sachin S Kajarekar, and Hynek Hermansky, "A new speaker change detection method for two-speaker segmentation," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002, vol. 4, pp. IV–3908.

[7] Navid Shokouhi and John HL Hansen, "Probabilistic linear discriminant analysis for robust speaker identification in co-channel speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Stuart N Wrigley, Guy J Brown, Vincent Wan, and Steve Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on speech and audio processing*, vol. 13, no. 1, pp. 84–91, 2004.

[9] Masayuki Suzuki, Gakuto Kurata, Tohru Nagano, and Ryuki Tachibana, "Speech recognition robust against speech overlapping in monaural recordings of telephone conversations," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5685–5689.

[10] Valentin Andrei, Horia Cucu, and Corneliu Burileanu, "Overlapped speech detection and competing speaker counting-humans vs. deep learning," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[11] Weipeng He, Petr Motlicek, and Jean-Marc Odobez, "Deep neural networks for multiple speaker detection and localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 74–79.

[12] Neeraj Sajjan, Shobhana Ganesh, Neeraj Sharma, Sriram Ganapathy, and Neville Ryant, "Leveraging LSTM models for overlap detection in multi-party meetings," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5249–5253.

[13] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[14] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[15] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[16] Fabian-Robert Stöter, Soumitro Chakrabarty, Bernd Edler, and Emanuël AP Habets, "Classification vs. regression in supervised learning for single channel speaker count estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 436–440.

[17] Fabian-Robert Stöter, Soumitro Chakrabarty, Bernd Edler, and Emanuël AP Habets, "CountNet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 268–282, 2018.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.