OVERLAP-AWARE DIARIZATION: RESEGMENTATION USING NEURAL END-TO-END OVERLAPPED SPEECH DETECTION

Latané Bullock [1]   *Hervé Bredin* [2]   *Leibny Paola Garcia-Perera* [3]

[1] Rice University, Houston, USA

[2] LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Orsay, France

[3] Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, USA

ABSTRACT

We address the problem of effectively handling overlapping speech in a diarization system. First, we detail a neural Long Short-Term Memory- based architecture for overlap detection. Secondly, detected overlap regions are exploited in conjunction with a frame-level speaker posterior matrix to make two-speaker assignments for overlapped frames in the resegmentation step. The overlap detection module achieves state-of-the-art performance on the AMI, DIHARD, and ETAPE corpora. We apply overlap-aware resegmentation on AMI, resulting in a 20% relative DER reduction over the baseline system. While this approach is by no means an end-all solution to overlap-aware diarization, it reveals promising directions for handling overlap.

Index Terms— speaker diarization, overlapped speech detection, resegmentation

1. INTRODUCTION

Speaker diarization answers the question, 'Who spoke when?' in an audio recording. In favorable conditions, modern diarization systems are able to achieve error rates nearly on par with those of humans. However, even the best diarization systems struggle to identify who was speaking in adverse scenarios. An audio file can be *adverse* in terms of the number of speakers (and, in particular, the amount of overlapping speech), the age of the speakers in the recording, the proximity of the microphone to the speakers, or any combination of these. There is a need for robust speaker diarization to process child-centered and other naturalistic recordings, massive amounts of online audio and video, and clinical interviews, to name a just a few.

There have been several studies on overlap detection and its impact on diarization. One of the first is [1], which investi-

gates how overlap detection could help diarization. Later, the authors in [2] detect overlap with a three-state Hidden Markov Model, and subsequently sum over frame-level posteriors for all of the frames within a segment to make second-speaker assignments. [3] propose a 'two-pass' system to first detect overlap, then use it to purify speaker models and make assignments. [4] used information external to the overlapped speech - namely the surrounding silence - in its detection. [5] introduce neural networks to the overlap detection problem. Their main findings are that LSTM-based detection provides comparable results to HMM, and LSTM+HMM is better than HMM. Later in [6], a convolutional neural network (CNN) architecture was used for detection. [7] defend the use of artificially mixed data for training in order to combat the imbalance of overlapped and monospeaker regions. Most recently, [8] report CNN-based overlap detection accuracy and evaluate the resulting potential change in diarization error rate (DER), but assume access to perfect two-label assignment.

Some other studies, such as DIHARD I and DIHARD II [9, 10], clearly show that handling overlap is crucial and remains an open problem. In this research, we investigate the use of overlap information to improve diarization performance. Our two-stage process combines detecting overlap in the audio with recurrent neural networks, and hypothesizing two speaker labels in regions with overlap.

2. OVERLAPPED SPEECH DETECTION

Overlapped speech detection is the task of detecting regions where at least two speakers are speaking at the same time. Detecting regions of overlapped speech is most effectively solved with a temporal approach, where we take into account the sequential nature of speech.

2.1. Principle

We address overlapped speech detection as a sequence labeling task where the input is the sequence of feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ and the expected output is the corresponding sequence of labels $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$ with $y_t = 0$ if there is zero or one speaker at time step t and $y_t = 1$
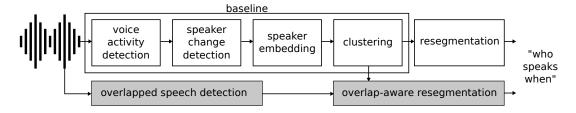
Fig. 1. The proposed pipeline for speaker diarization. The baseline incorporates end-to-end neural voice activity detection, speaker change detection, and speaker embeddings, with clustering performed via affinity propagation [11]. It is available in `pyannote.audio` toolkit [12]. The grey boxes highlight the our contributions: neural (LSTM-based) detection of overlapping speech and a simple frame-level resegmentation module designed to account for the overlapping speech.
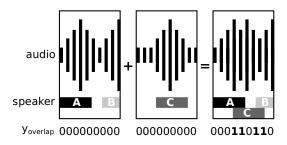

Fig. 2. To increase the number of positive training samples for overlapped speech detection, artificial audio chunks are created by summing two random audio chunks.

if there are two speakers or more.

Because processing long audio files of variable lengths is neither practical nor efficient, we rely on shorter fixed-length sub-sequences. At training time, fixed-length sub-sequences are drawn randomly from the training set to form mini-batches, increasing training sample variability (data augmentation) and training time (shorter sequences). To address the class imbalance problem, half of the training sub-sequences are artificially made from the weighted sum of two random sub-sequences, as depicted in Figure 2.

At test time, audio files are processed using overlapping sliding windows of the same length as used in training. For each time step t, this results in several overlapping sequences of prediction scores, which are averaged to obtain the final score of each class. Finally, time steps with prediction scores greater than a tunable threshold θ_{OSD} are marked as overlapped speech.

2.2. Implementation details

Models are based on the architecture depicted in Figure 3. They are trained on 2s audio chunks, either with handcrafted MFCC features (19 coefficients extracted every 10ms on 25ms windows, with first- and second-order derivatives) or with trainable SincNet features (using the configuration of the original paper [13]). The rest of the network includes two stacked bi-directional Long Short-Term Memory (LSTM) recurrent layers (each with 128 units in both forward and backward directions), two feed-forward layers (128 units, *tanh* activation) and a final classification layer (2 units, *softmax* activation), fed into binary cross-entropy loss.

3. OVERLAP-AWARE RESEGMENTATION

While most diarization systems hypothesize a single speaker in all voiced regions, a robust overlapping speech detector opens up the possibility of hypothesizing an additional speaker in overlapping regions.

3.1. Principle

Depicted in Figure 4, our proposed approach relies heavily on the i-vector-based Variational Bayes Hidden Markov Model (VB-HMM) introduced for speaker diarization in [14], and applied to resegmentation in [15]. We use the output of the speaker diarization baseline as the binary initialization of the per-frame speaker posterior matrix: Q_{st} is initialized to 1 if speaker s is responsible for the speech at the voiced frame t, and 0 otherwise. After VB-HMM resegmentation, the previously one-hot hard assignments of speakers to frames in Q become soft probabilities. The most likely speaker is assigned to frames detected as speech by the voice activity detector. A second most likely speaker is only assigned for frames detected as overlapped speech.

3.2. Implementation details

We first perform resegmentation using [14]'s VB-HMM module. Feature vectors for the module are length-60 MFCCs with deltas and double deltas, extracted in 10ms steps with a 25ms window. The (400-dimensional) i-vector extractor and diagonal (1024-component) Universal Background Model are trained on the training portion of AMI Headset mix. We use a single VB inference iteration (default 10) and adjust the loop probability parameter to 0.95 (default 0.9). Otherwise, we keep the default parameters of the VB-HMM diarization module.
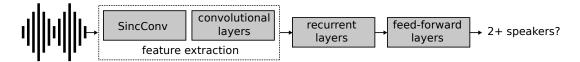
Fig. 3. Architecture of the neural network used for end-to-end overlapped speech detection. We also report detection results where the trainable feature extraction part is replaced by handcrafted MFCC features.
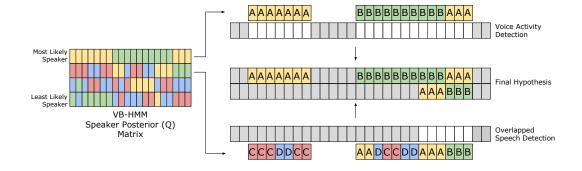

Fig. 4. Illustration of proposed method for assigning secondary speakers in overlap regions. The speaker posterior matrix from VB-HMM resegmentation (on the left) serves as the source of speaker label hypotheses. The most likely speaker sequence is masked with the voice activity detection (upper right), while the second most likely speaker sequence is masked by the overlap detection output (lower right). The final diarization hypothesis is the union of the two.

4. EXPERIMENTS

Overlapped speech detection models were trained, tuned, and tested on three different datasets whose statistics are summarized in Table 1:

- AMI (Headset mix) [16] is a subset of the AMI corpus that consists of summed recordings of spontaneous speech of mainly four speakers;
- DIHARD II [9] contains single channel wide-band audio from 11 challenging domains that range from very clean (near-field recordings of read audiobooks) to noisy, far-field recordings;
- ETAPE (TV subset) [17] consists of TV content in French (news, talk shows, debates).

The proposed overlap-aware resegmentation module has only been tested on AMI Headset mix. 81% of the total speech in voiced regions is single-speaker and 15% of the time two-speaker, leaving approximately 4% of the time to three or more speakers. This implies that the two-speaker situation accounts for about 75% of the overlap regions – justifying our initial focus on this case.

Code, configuration files, and pre-trained models for reproducing the speaker diarization baseline and overlapped speech detection results are available in the `pyannote.audio` repository [12]. Code for VB-HMM resegmentation is provided by Brno University of Technology[1], and all assignment

---

[1]https://speech.fit.vutbr.cz/software

code can be found in the JSALT 2019 Speaker Detection team repository[2]. The `pyannote.metrics` toolkit [19] is used to evaluate overlapped speech detection in terms of precision and recall, and resegmentation in terms of diarization error rate (DER). DER is the portion of the recording that is labelled incorrectly, with three possible types of errors: false alarm, missed detection, and speaker confusion.

4.1. Overlapped speech detection

As reported in Table 2, the end-to-end variant consistently outperforms the one based on handcrafted features for all datasets, setting a new state-of-the-art performance on all three datasets[3] – though we could not find any previously published overlapped speech detection results for DIHARD. When tuned for high (90%) precision, the proposed approach gets a very low recall of 1.5% on DIHARD, making it almost useless for the subsequent overlap assignment step.

4.2. Overlap-aware resegmentation

The impact of our second contribution on the performance of the diarization pipeline is reported in Table 3. Overall, our proposed overlap-aware resegmentation approach brings a significant 20% relative (or 5.9% absolute) improvement in terms of diarization error rate (from 29.7% down to 23.8%).

---

[2]https://github.com/jsalt2019-diadet/jsalt2019-diadet

[3]Thanks to Claude Barras for providing the overlapped speech detection output corresponding to system L_1 in Table 2 of [20], and Marie Kunešová for providing the overlapped speech detection output corresponding to system *"AMI test (all subsets) + dereverberation"* in Table 2 of [8].

Dataset		Train			Development			Evaluation		
AMI (Headset mix)	[16]	70h	85%	19%	14h	84%	20%	**14h**	**82%**	**19%**
DIHARD II	[18]	15h	75%	9%	8h	77%	11%	22h	74%	9%
ETAPE (TV)	[17]	14h	94%	6%	4h	93%	5%	4h	92%	7%

Table 1. Datasets statistics. For each subset, we report the total audio duration (in hours), the amount of speech (as percentage of audio duration), and the amount of overlapped speech (as percentage of speech duration). For instance, AMI evaluation set amounts to 14h of audio, 82% of which is speech (11.5h), among which 19% is overlapped speech (2.2h). Note that DIHARD does not come with a training set so the official development set was divided into two thirds for training and one third for development.

	AMI		DIHARD		ETAPE	
	Precision	Recall	Precision	Recall	Precision	Recall
Baseline	75.8 80.5 [8]	44.6 50.2 [8]			60.3 [20]	52.7 [20]
Proposed (MFCC)	91.9 90.0	48.4 52.5	58.0 73.8	17.6 14.0	67.1 55.0	57.3 55.3
Proposed (waveform)	86.8 90.0	65.8 63.8	64.5 75.3	26.7 24.4	69.6 60.0	61.7 63.6

Table 2. Evaluation of overlapped speech detection models, in terms of precision (%) and recall (%). Results on the development set are reported using small font size. We report two variants: the first one is based on handcrafted features (MFCCs) and the other one is an end-to-end model processing the waveform directly. *Baseline* corresponds to the best result we could find in the literature as of October 2019.

	DER%	FA%	Miss%	Conf%
Baseline	29.7	**3.0**	20.8	5.8
+ VB resegmentation	28.9	**3.0**	20.9	**5.0**
+ overlap assignment	**23.8**	3.6	**13.0**	7.2
+ oracle detection	22.2	3.1	6.0	13.2
+ oracle assignment	11.8	0.6	11.2	0.0

Table 3. AMI Headset mix diarization, false alarm, missed detection, and speaker confusion error rates after VB-HMM resegmentation and overlap assignment. The proposed assignment technique using oracle overlap detection and using oracle assignment are also reported. Oracle assignment refers to ideal both primary and secondary speaker labels.

A detailed analysis shows that the VB-HMM resegmentation step reduces confusion error by less than 1% while leaving – by design – false alarm and miss detection rates unchanged. Tuned for high precision, overlapped speech detection reduces missed detection by 38% relative (or 7.9% absolute), at the expense of a small increase in false alarm rate (from 3.0% to 3.6%). The secondary speaker assignment does increase speaker confusion by more than 2% (out of the 7.9% of correctly detected overlapped speech regions). Overall, the combination of our two contributions (overlapped speech detection and assignment) leads to a new state of the art on AMI Headset Mix, by a large margin.

Switching to oracle overlapped speech detection only brings a minor performance boost (from DER=23.8% down to 22.2%). This indicates that most future improvements will likely come from a better speaker assignment – which is confirmed by the oracle assignment experiment.

5. CONCLUSION

In this paper we have highlighted two contributions to the speaker diarization task. The first is an neural architecture for overlap detection, and the second is a simple yet effective resegmentation module that assigns two speakers in frames detected as overlapping speech.

The overlap detector predicts overlapping speech regions with state-of-the-art accuracy on AMI and ETAPE, and sets the baseline for future experimentation on DIHARD II. Our proposed solution for overlap-aware resegmentation was tested on AMI and beats state-of-the-art systems in DER due to a drastic decrease in missed detection error. However, further work is needed to more accurately assign secondary speakers, as evidenced by the large increase in speaker confusion error with oracle detection. Additionally, testing on other datasets will be necessary to establish the method as a robust approach.

Our hope is that the present study will encourage more research on both overlap detection and its practical uses in diarization systems. Two of the primary remaining questions are: how to increase accuracy of secondary speaker assignment, and how to extend assignment to more than two speakers. A system inspired by [21] could integrate the detection and assignment steps to improve secondary speakers hypotheses. [22] recently proposed a neural architecture to count the number of concurrent speakers in a signal, which could enable three or more speaker assignment.

6. REFERENCES

[1] Scott Otterson and Mari Ostendorf, "Efficient use of overlap information in speaker diarization," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 683–686.

[2] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. ICASSP 2008*. IEEE, 2008, pp. 4353–4356.

[3] Marijn Huijbregts, David A. van Leeuwen, and Franciska de Jong, "Speech overlap detection in a two-pass speaker diarization system," in *Proc. Interspeech 2009*, 2009.

[4] Sree Harsha Yella and Fabio Valente, "Speaker diarization of overlapping speech based on silence distribution in meeting recordings," in *Proc. Interspeech 2012*, 2012.

[5] Jürgen T Geiger, Florian Eyben, Björn Schuller, and Gerhard Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in *Proc. Interspeech 2013*, 2013.

[6] Valentin Andrei, Horia Cucu, and Corneliu Burileanu, "Detecting Overlapped Speech on Short Timeframes Using Deep Learning," in *Proc. Interspeech 2017*, 2017.

[7] Gerhard Hagerer, Vedhas Pandit, Florian Eyben, and Björn Schuller, "Enhancing lstm rnn-based speech overlap detection by artificially mixed data," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*. Audio Engineering Society, 2017.

[8] Marie Kunešová, Marek Hrúz, Zbyněk Zajíc, and Vlasta Radová, "Detection of overlapping speech for the purposes of speaker diarization," in *Speech and Computer*, Albert Ali Salah, Alexey Karpov, and Rodmonga Potapova, Eds., Cham, 2019, pp. 247–257, Springer International Publishing.

[9] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "Second dihard challenge evaluation plan," *Linguistic Data Consortium, Tech. Rep*, 2019.

[10] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "First dihard challenge evaluation plan," 2018.

[11] Ruiqing Yin, Hervé Bredin, and Claude Barras, "Neural Speech Turn Segmentation and Affinity Propagation for Speaker Diarization," in *Proc. Interspeech 2018*, 2018, pp. 1393–1397.

[12] pyannote.audio contributors, "pyannote.audio: Neural Building Blocks for Speaker Diarization," Submitted to ICASSP 2020.

[13] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *Proc. SLT 2018*, 2018.

[14] Mireia Diez, Lukas Burget, and Pavel Matejka, "Speaker Diarization based on Bayesian HMM with Eigenvoice Priors," in *Odyssey 2018 The Speaker and Language Recognition Workshop*. June 2018, pp. 147–154, ISCA.

[15] Gregory Sell and Daniel Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4794–4798.

[16] Jean Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, 2007.

[17] Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert, "The ETAPE Corpus for the Evaluation of Speech-based TV Content Processing in the French Language," in *Proc. LREC 2012*, 2012.

[18] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *Proc. Interspeech 2019*, 2019, pp. 978–982.

[19] Hervé Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Proc. Interspeech 2017*, Stockholm, Sweden, August 2017.

[20] D. Charlet, C. Barras, and J. Linard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7707–7711.

[21] Shaojin Ding, Quan Wang, Shuo-yiin Chang, Li Wan, and Ignacio Lopez Moreno, "Personal vad: Speaker-conditioned voice activity detection," *arXiv preprint arXiv:1908.04284*, 2019.

[22] Fabian-Robert Stöter, Soumitro Chakrabarty, Bernd Andreas Edler, and Emanuël A. P. Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 268–282, 2019.