Abstract

Speaker diarization for real-life scenarios is an extremely challenging problem. Widely used clustering-based diarization approaches perform rather poorly in such conditions, mainly due to the limited ability to handle overlapping speech. We propose a novel Target-Speaker Voice Activity Detection (TS-VAD) approach, which directly predicts an activity of each speaker on each time frame. TS-VAD model takes conventional speech features (e.g., MFCC) along with i-vectors for each speaker as inputs. A set of binary classification output layers produces activities of each speaker. I-vectors can be estimated iteratively, starting with a strong clustering-based diarization.

We also extend the TS-VAD approach to the multichannel microphone case using a simple attention mechanism on top of the current speaker's features (i-vectors, d-vectors, x-vectors etc.) followed by clustering (k-means, agglomerative hierarchical, spectral, etc.) according to some similarity metrics (Probabilistic Linear Discriminant Analysis (PLDA) [10, 8], score, cosine, etc.). These stages can also be followed by re-segmentation (such as GMM [11], Variational Bayes [12] or LSTM-based [13]) and overlapping speech segments post-processing.

Currently, high diarization accuracy is achieved for many benchmarks, such as CallHome. However, the development of a diarization system for complex acoustic environments is still an unsolved task. This was a motivation for the DIHARD Challenges [14, 15] focused on the development of systems for highly overlapping speech. One of the most promising methods of such kind is the end-to-end neural diarization (EEND) [20], which performs diarization in a single stage and outputs frame-level activity probabilities for each speaker independently. Another direction that we found to be promising consists of using pre-computed features of a speaker of interest to draw the system’s attention to only their speech. This direction is represented by such approaches as Target-Speaker ASR [21], Speaker Beam [22, 23] and Voice Filter [24] aimed at the target-speaker speech extraction, etc. Moreover, the TS-ASR approach may be used for simultaneous speech recognition and diarization [25]. One more representative of this direction is the Personal VAD [26] approach allowing to detect speech boundaries for only that speaker whose acoustic “profile” is fed into the system.

Inspired by the the ideas mentioned above, we combined their benefits in our own diarization approach for the CHiME-6 referred to as Target-Speaker VAD (TS-VAD). In the course of the development, our model evolved from the simple one, very similar to the Personal VAD, to the sophisticated one, which processes multi-channel recordings and outputs independent speaker activity streams like the EEND model. The main difficulty for the successful TS-VAD application is that, unlike TS-ASR or Personal VAD scenarios, we do not have any pre-computed speakers’ features and have to compute them directly from severely distorted and highly overlapping speech. Nevertheless, we managed to find such a way of applying TS-VAD which reduced DER to the values of 33% and 36% for the CHiME-6 development and evaluation sets respectively, which is much better than BUT system result for DIHARD II [18] and ASR. The same data diarization is also one of the CHiME-6 Challenge [17] Track 2 tasks. This data recorded in real-life conditions contains a large amount of overlapping speech. Conventional diarization systems are not well-suited for processing highly overlapping speech, so it is not very surprising that even the best DIHARD II system developed by BUT [18] provided DER only slightly below 60%.

During the participation in the CHiME-6 Challenge, we were solving the same diarization problem [19]. Achieving high diarization accuracy was crucial for high ASR performance, which was the main challenge goal. So we started by reviewing approaches that are effective in diarizing highly overlapping speech. One of the most promising methods of such kind is the end-to-end neural diarization (EEND) [20], which performs diarization in a single stage and outputs frame-level activity probabilities for each speaker independently. Another direction that we found to be promising consists of using pre-computed features of a speaker of interest to draw the system’s attention to only their speech. This direction is represented by such approaches as Target-Speaker ASR [21], Speaker Beam [22, 23] and Voice Filter [24] aimed at the target-speaker speech extraction, etc. Moreover, the TS-ASR approach may be used for simultaneous speech recognition and diarization [25]. One more representative of this direction is the Personal VAD [26] approach allowing to detect speech boundaries for only that speaker whose acoustic “profile” is fed into the system.
The rest of the paper is organized as follows: in Section 2 we describe in more details the CHiME-6 Challenge conditions and data, then in Section 3 we talk about how a single-channel version of TS-VAD evolved and how the above-mentioned prototypes influenced this process. The multi-channel TS-VAD system is described in Section 4, and the fusion of different systems as well as post-processing of diarization results are discussed in Section 5. Section 6 provides some conclusions and suggests possible future directions.

2. CHiME-6 Challenge

The CHiME-6 Challenge continues a series of challenges on speech recognition in complex real-life acoustic environments. It is based on the same data as the previous CHiME-5, which consists of multi-channel recordings from six 4-microphone Microsoft Kinect arrays, located in three different rooms. Each recording contains an informal conversation of four persons in a dinner party scenario in three locations. Besides, the two-channel recordings from in-ear microphone pairs worn by each person are also provided for the training. Data includes 20 sessions (16 for training, 2 for development and 2 for evaluation). The goal is to create ASR system for multi-channel Kinect recordings in two tracks. In Track 1 participants are allowed to use manual boundaries of each speaker’s utterances created manually from worn microphones audio listening, while in Track 2 this information cannot be used. Since the knowledge of boundaries can help to improve ASR substantially, one of the Track 2 tasks is to obtain a high-quality diarization of Kinect recordings. This task is very similar to Track 4 of the DIHARD II Challenge. Same as in DIHARD II, the metrics for this task are DER and Jaccard Error Rate (JER), which are evaluated without using a collar and without excluding overlapping segments. Nonetheless, there are some differences, as well. Firstly, CHiME-6 organizers provided software for time-aligning different channels within a session, which was not available in DIHARD II. All our results are obtained on pre-aligned data. Also, unlike DIHARD II allowing participants to use any training data, the CHiME-6 rules limit training data for Track 2 by only CHiME-6 data and VoxCeleb data.

Initially, reference rttm-files provided by the CHiME-6 organizers were created based on utterance boundaries set manually (the same information was used in Tracks 3–4 of DIHARD II). However, such segmentation contains intra-utterance pauses, which are treated as speech, and unlabelled speakers’ introductions. Therefore, the organizers provided another reference segmentation with the exclusion of silence segments. This was carried out using triphone GMM-HMM forced alignment of reference transcriptions over manual segments. Besides, the UEM-file was provided, which makes it possible to exclude speaker introductions from scoring. The significant difference between reference rttm files in CHiME-6 and DIHARD II (about 30-40% in terms of DER) is the main reason why the direct comparison of the results is not possible.

CHiME-6 data is rather noisy and reverberated as well as contains a significant amount of overlapping speech, so it is difficult to perform an accurate diarization with clustering-based systems. Statistics of overlapping speech in the development and evaluation datasets is shown in Table 1. If one changes the reference rttms to leave only a single speaker on each overlapping segment, this will result in DER of 25.61%/21.76% (due to the miss errors only) for the development and evaluation sets, respectively. These values show the lowest achievable DERs on the CHiME-6 data for clustering-based diarization systems without special overlaps processing.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>24.05%</td>
<td>54.25%</td>
<td>17.74%</td>
<td>3.49%</td>
</tr>
<tr>
<td>eval</td>
<td>33.47%</td>
<td>51.52%</td>
<td>12.03%</td>
<td>2.47%</td>
</tr>
</tbody>
</table>

3. Single-channel TS-VAD

3.1. Single-Speaker model

Recently, a series of papers on multi-speaker speech processing was published, where models focus on a specific speaker ignoring the speech of others. These approaches include TS-ASR [21] for target speech recognition, Speaker Beam [22, 23] and Voice Filter [24] for target speech extraction, and Personal VAD [26] for target speech detection. Most of them use the acoustic footprint of a target speaker (usually i-vector), obtained during prior enrollment, to focus on speech of interest. The Personal VAD model seemed to be most appropriate for our goals since it selects each speaker’s speech independently. However, this model was trained and tested on concatenated instead of overlapped speech segments, so it was unclear if it could handle overlapping speech as well. Besides, under the CHiME-6 conditions, it was not trivial how to compute i-vectors for each speaker since the use of manual segmentation was prohibited in test time.

Thus, we started with a purely research question: given an i-vector estimated on manual non-overlapping speech segments of a target speaker, is it possible to detect their speech in overlapping conditions? The answer was generally positive. Our first single-speaker TS-VAD model was very similar to the Personal VAD with the same three targets obtained from the forced alignment, namely silence, target speech, and non-target speech. The model architecture was a simple 3-layer BLSTM with projections [27].

All the experiments were performed in the Kaldi ASR Toolkit [28]. We used the acoustic model training dataset and i-vector extractor from the Kaldi chime6 recipe. Moreover, we used a “negative” version of each utterance with an i-vector corresponding to a random speaker from the same session and device (without such “negative” examples, the model tends to detect any speech as a target). Training targets were taken from the tri3 GMM forced alignment: silence and noise phones were treated as silence class, and the rest phones as target speech or non-target speech for the original and “negative” utterances, respectively.

TS-VAD model outputs a sequence of probabilities of target speaker presence for each time frame. To convert it into the segmentation, the simple post-processing described in Section 5 was applied. After multiple experiments with i-vectors computed on non-overlapping regions of manual segmentation, we managed to obtain DER=66.81% on the development data (for all experiments on manual segmentation, we used CH1 of the reference device, i.e., the device closest to the currently active speaker).

Single-speaker TS-VAD processes each speaker independently, which is intuitively sub-optimal. So, we introduced an additional “mutual” threshold to the post-processing procedure. Considering a set of speech probabilities on the current frame, we found that there is likely no speech of those speakers whose probabilities are strongly dominated by the maximum probability on this frame. Thus, the probability was set to zero if

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>dev</td>
<td>24.05%</td>
<td>54.25%</td>
<td>17.74%</td>
<td>3.49%</td>
</tr>
<tr>
<td>eval</td>
<td>33.47%</td>
<td>51.52%</td>
<td>12.03%</td>
<td>2.47%</td>
</tr>
</tbody>
</table>
it differed from the maximum probability on the current frame by more than the threshold. This trick provided a dramatic di-
arization improvement reducing DER to 46.12%, which looked promising.

3.2. Multi-Speaker model

We also experimented with the EEND model [20], hoping it would work well for CHiME-6 data. Unfortunately, this was not the case due to several reasons. First of all, the relatively short audio pieces are required to apply EEND since the self-
attention module needs to see the whole sequence. But such block-wise processing induces the between-block speaker permutation problem. We used oracle permutation to estimate the lowest possible DER for the EEND. An oracle DER of the EEND model trained on the VoxCeleb data was comparable to DER of the baseline diarization, presumably due to the severe acoustic mismatch between the VoxCeleb and CHiME-6 data. The model trained on CHiME-6 data provided much better oracle DER (about 50%). To resolve between-block permutations, we extracted baseline x-vectors from the EEND-induced seg-
ments and applied several clustering approaches to them. We found that the simple unconstrained k-means algorithm worked best, but the obtained DERs were only slightly better than those of the baseline diarization. Nonetheless, we adopted some ideas from EEND to improve the TS-VAD model.

As the CHiME-6 diarization task is closed (the number of speakers in a session is always four), we decided to use i-vectors of all speakers in the session as inputs, instead of the target speaker’s i-vector only. The initial model was designed to esti-
mate speech probability for the speaker corresponding to the first i-vector. It was beneficial to average the predicted proba-
bilities obtained on various permutations of i-vectors. However, much better results were obtained with the TS-VAD model de-
sign to predict speech probabilities for all speakers simulta-
neously, as the EEND model does. This model with four output layers was trained using a sum of binary cross-entropies as a loss function. We also found that it is essential to process each speaker by the same Speaker Detection (SD) 2-layer BLSTM, and then combine SD outputs for all speakers by one more BLSTM layer. The model architecture is shown in Figure 1. Note that parameters of the SD block are shared across speakers, and it is trained jointly with the whole TS-VAD model.

![Figure 1: Single-channel TS-VAD scheme](image)

As we performed all the experiments in the Kaldi ASR
Toolkit [28], it was easier to use 2-class softmax instead of a sig-
moid in the output layers. Training targets were 8-dimensional vectors representing four pairs of silence and speech probabilities corresponding to four speakers. Given an utterance, tar-
gets corresponding to the current speaker were taken directly from the forced alignment. Targets for the three other speakers were obtained by averaging alignments from neighboring overlapping utterances over all the devices and channels (for non-
overlapping frames, targets for speech and silence were 0 and 1, respectively).

To provide more data variability, we performed on-the-fly random permutations of speakers (i.e., both i-vectors and tar-
gets) during the training. It was also beneficial to use the mixup data augmentation [29] using our Kaldi-compatible tools pre-
presented in [30]. Moreover, we obtained a small DER improvement (about 0.5%) by adding an 800h subset of the VoxCeleb data augmented by artificial room impulse responses. The final multi-speaker model achieved an impressive DER of 37.40% using i-vectors estimated on manual segmentation.

Note that the “mutual” threshold providing a huge DER re-
duction for the single-speaker TS-VAD turned out to be totally useless for the multi-speaker model.

3.3. Estimation of i-vectors

After the experiments on manual segments, we switched to the real CHiME-6 task and tried to compute i-vectors from the seg-
mentation provided by the baseline clustering-based diarization (DER=63.42% on the development set).

Unfortunately, the early versions of TS-VAD did not pro-
vide any tangible improvement over the baseline results. So, we firstly focused on improving the baseline diarization. To this end, we applied the improved x-vectors extractor based on a Wide ResNet (WRN), which was trained on the VoxCeleb data. The details on the extractor are given in [31]. Application of the baseline Agglomerative Hierarchical Clustering (AHC) based on PLDA scoring to these x-vectors improved DER by about 10% abs. Then we used the idea from [32] and replaced AHC clustering based on PLDA scores with Spectral Cluster-
ing (SC) based on Cosine Similarities. Using similarity matrix binarization with respect to automatically chosen threshold [32] provided the consistent DER improvement on both the develop-
ment and evaluation sets (see Table 2 for the results). Besides, using WRN x-vectors substantially improved the permutation resolution on the EEND segmentation and covered a large frac-
tion of the gap to the oracle DER.

The improved clustering-based diarization provided a good initial estimation of i-vectors for TS-VAD. The next idea was to re-estimate i-vectors iteratively, using segmentation from the previous TS-VAD iteration. However, we found that even better results can be achieved using probabilities obtained by the TS-
VAD as soft-weights for re-estimation of i-vectors. Note that, to ensure robust i-vectors estimation, for each speaker we con-
sidered only those frames where speech probability was more than 0.8 of total speech probability for all speakers on a given frame. The second iteration provided a significant gain, but the third one did not lead to any improvement. Note that such i-
vectors estimation provides as good results as i-vectors com-
puted on manual segmentation. Later we found that the same iterative procedure with the best TS-VAD model starting from the baseline diarization also improves DER by about 15% abs-
olute. However, more iterations are required for convergence.
4. Multi-channel processing

The single-channel version of TS-VAD (TS-VAD-1C) processes each channel separately, but we suggested that the multi-channel processing may be beneficial. Indeed, we found that the multi-channel Weighted Prediction Error (WPE) dereverberation [33, 34] improves the results of TS-VAD by about 1% DER absolute. Moreover, averaging of per-channel TS-VAD probabilities provides up to 2% absolute DER reduction.

To process separate Kinect channels jointly, we investigated the multi-channel TS-VAD model (TS-VAD-MC), which takes a combination of TS-VAD-1C model SD blocks outputs from a set of 10 Kinect recordings as an input. The architecture of this model is presented in Figure 2. The channels of input Kinect recordings are chosen randomly for training, while CH1 and CH4 are taken at test-time. This way of combining information from different channels is more effective than a simple averaging of probabilities, as in the TS-VAD-1C model. All the SD vectors for each speaker are passed through a 1-d convolutional layer and then combined by means of a simple attention mechanism. Combined outputs of attention for all speakers are passed through a single BLSTM layer and converted into a set of per-frame probabilities of each speaker’s presence/absence.

Finally, to improve overall diarization performance, we fused 3 single-channel and 3 multi-channel TS-VAD models by computing a weighted average of their probability streams. Fusion weights were tuned to minimize DER on the development set. The transitions from the silence state to the two-speaker states and vice versa were prohibited and other transition probabilities were shared between any pairs of states with the same number of speakers (7 tunable parameters in total). The most likely state sequence found by the Viterbi search determined the final segments.

The influence of several post-processing techniques on DER improvement is shown in Table 3.

5. Post-processing

To convert the TS-VAD output probabilities into a sequence of segments, the simple post-processing was applied. It includes 51-tap median filtering, binarization with the threshold of 0.4, combining speech segments separated by pauses shorter than 0.3s, and deleting speech segments shorter than 0.2s.

Alternatively, Viterbi decoding was applied for the post-processing. We introduced a simple Hidden Markov Model (HMM) with 11 states representing silence, speech from each of four speakers without overlaps, and overlapping speech from 6 possible pairs of speakers (overlaps of three and four speakers were neglected due to short duration). The emission probabilities in each state were induced from TS-VAD output probabilities, while the transition probabilities were tuned to minimize DER on the development set. The transitions from the silence state to the two-speaker states and vice versa were prohibited and other transition probabilities were shared between any pairs of states with the same number of speakers (7 tunable parameters in total). The most likely state sequence found by the Viterbi search determined the final segments.

We presented a novel approach for the diarization of multi-speaker conversations, which provided state-of-the-art results in a complex multi-channel dinner party scenario. The proposed Target-Speaker VAD selects speech of every conversation participant taking their i-vector as an input along with MFCC features. Since the sufficiently good segmentation is required for the reliable i-vectors estimation, we improved the baseline clustering-based diarization significantly.

It is worth noting that we also tried to replace i-vectors with more discriminative speaker embeddings like x-vectors as the TS-VAD inputs, but results got much worse. We believe the reason is that the relatively simple TS-VAD architecture is unable to match well the current MFCC features to the speaker embeddings obtained from a much more complicated model. Another possible reason is severe overfitting due to a sparse embeddings space and a small number of speakers in the training data.

Although our final solution is task-dependent (multi-channel input, a fixed number of speakers), we believe the proposed approach is flexible enough to be easily modified for other similar tasks. In the future, we plan to extend our solution to the scenario of informal meetings with an unknown large number of participants.

6. Conclusions

We are grateful to STC Voice Biometrics Team for the awesome speaker embeddings extractor and valuable discussions.

Acknowledgements

This work was partially financially supported by the Government of the Russian Federation (Grant 08-08).

Table 2: Diarization results (* stands for DIHARD II reference)

<table>
<thead>
<tr>
<th></th>
<th>DER</th>
<th>JER</th>
<th>EVAL</th>
<th>JER</th>
</tr>
</thead>
<tbody>
<tr>
<td>x-vectors + AHC</td>
<td>63.42</td>
<td>70.83</td>
<td>68.20</td>
<td>72.54</td>
</tr>
<tr>
<td>EEND + WRN x-vectors</td>
<td>52.20</td>
<td>57.42</td>
<td>56.01</td>
<td>61.49</td>
</tr>
<tr>
<td>WRN x-vectors + AHC</td>
<td>53.45</td>
<td>56.76</td>
<td>53.79</td>
<td>62.02</td>
</tr>
<tr>
<td>WRN x-channel Weighted + SC</td>
<td>47.29</td>
<td>49.03</td>
<td>60.10</td>
<td>57.99</td>
</tr>
<tr>
<td>+ TS-VAD-1C (i1) + TS-VAD-1C (i2) + TS-VAD-MC</td>
<td>39.19</td>
<td>40.87</td>
<td>45.01</td>
<td>47.03</td>
</tr>
<tr>
<td>Fusion*</td>
<td>32.84</td>
<td>36.31</td>
<td>36.02</td>
<td>40.10</td>
</tr>
<tr>
<td>Fusion</td>
<td>41.76</td>
<td>44.04</td>
<td>40.71</td>
<td>45.32</td>
</tr>
</tbody>
</table>

Table 3: DER for different post-processing. T, F, S and V stand for binarization with the threshold, median filtering, short speech/pause processing, and Viterbi decoding, respectively.

<table>
<thead>
<tr>
<th></th>
<th>Best single TS-VAD</th>
<th>Fusion T</th>
<th>Fusion V+S</th>
<th>Fusion T+F+S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-processing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEV</td>
<td>34.59</td>
<td>34.73</td>
<td>33.56</td>
<td>32.84</td>
</tr>
<tr>
<td>EVAL</td>
<td>37.57</td>
<td>37.52</td>
<td>36.63</td>
<td>36.02</td>
</tr>
</tbody>
</table>
8. References


