



Speaker Conditioning of Voice Activity Detection via Implicit Separation

Matthew Maciejewski

Human Language Technology Center of Excellence, Johns Hopkins University, USA

matt@mmaciejewski.com

Abstract

Within the domain of multi-talker recordings, many speech technologies rely on an initial segmentation step of finding when each person was talking. One common approach to this task is Target-Speaker Voice Activity Detection (TS-VAD), in which a model is supplied with a representation corresponding to a particular speaker and then identifies the temporal regions when that person was talking. As in many cases, the increased complexity of this task over regular Voice Activity Detection (VAD) imposes constraints on the data used to train such a model. In this work, we explore conversion of a pre-trained VAD model into a TS-VAD model via an implicitly-trained separation front end—decoupling the need for speaker-discriminative training data from the basic speech/non-speech data used in training VAD models—which can lead to improvements in model robustness and speech recall in the domains present only in the training data of the VAD.

Index Terms: TS-VAD, speaker diarization, speech separation

1. Introduction

The ability to focus on a particular person’s speech in an environment where multiple people are talking has long been considered one of the most difficult problems in speech processing [1]. In recent years, much attention has been given to recordings in which multi-talker conversations are captured by one or more microphones placed within the room. For example, the localization/attribution task of speaker diarization (i.e. “who spoke when?”) is the focus of the DIHARD challenges [2, 3], and the CHiME challenges [4–6] center on the task of speaker-attributed multi-talker speech recognition (i.e. “who said what when?”).

Competitive systems for these tasks can be quite complicated, comprising multiple connected modules and sub-modules, which can be either modeling-based or data-driven, i.e. leverage deep learning [7]. The top-performing CHiME systems almost always contain an explicit diarization step [8–10], with one of the most significant gains coming from the development of TS-VAD [11], based on the Personal VAD [12] method. In this approach, initial speaker identity vectors for all people in the recording are computed using a traditional clustering-based diarization method, and then the TS-VAD model is conditioned on each of these vectors to find the speech of each individual person.

In some sense, TS-VAD exists in a space defined by the rough intersection of three closely-related tasks: voice activity detection (localization of speech), speech separation/target-speaker extraction (producing a clean, non-overlapped recording of a person’s speech from a recording with interfering speech and noise), and speaker diarization (localizing and iden-

tifying the speech of multiple people speaking within a recording). Besides TS-VAD, other methods exist within this space; there have been numerous approaches to jointly performing speech separation and diarization [13–15] as well as using separation as a front-end for diarization [16, 17].

In this work, we aim to explore the task of speaker-conditioned VAD through integrating an existing pre-trained VAD with a front end that is styled after target-speaker extraction models [18, 19]. To do so, we take an approach that avoids an explicit speech separation module, aiming to avoid some of the difficulties in coupling speech separation front-ends with downstream models, as described in previous studies on separation for diarization [16, 17] and for speech recognition [20, 21].

By directly optimizing the separator from the VAD outputs, we hope the gradients will be strongest in the regions of the input features that are most informative for the VAD task, similar to how training of speech separation systems was improved by using l^2 loss on masked spectrograms rather than the masks themselves [22]. This focused learning on high-energy spectral regions (i.e. the speech signals) and not silence and noise, where the mask is irrelevant to the task and is difficult to learn.

Additionally, we hope that this approach will allow leveraging a powerful pre-trained VAD for better segmentation than would be possible to train using the speaker-labeled data required to train TS-VAD, similar in spirit to attempts within speech recognition to speaker-condition large-scale transcription models like Whisper [23, 24] that would be difficult to reproduce in multi-talker conditions.

2. System Overview

The core design of this work is to speaker-condition an existing pre-trained VAD model with an ivector [25] speaker identity embedding to produce separate voice activity outputs for both the “target” speaker captured by the ivector and the other “non-target” speakers. This is accomplished through a front end speech separation-style target speech extraction network, which masks the spectral input features of the VAD model, and is ultimately trained using supervision generated using the VAD itself. Figure 1 diagrams this system and training scheme.

The primary motivations of this design are as follows:

- A powerful pre-trained VAD may be used, such that overall speech detection performance is not limited by the available speaker-labeled data.
- Similarly, the speaker-labeled conditioning data need not be labeled for activity or any annotation beyond identity (such as transcription).
- A separation-focused approach may provide better detection of overlapped speech, in contrast to the single-class classification objective used in initial TS-VAD models.

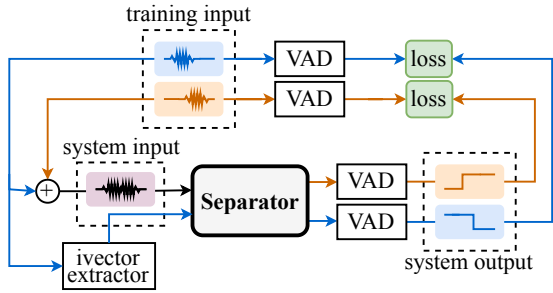


Figure 1: Training flow and basic system overview. Only the separator is updated in training. The processing paths of the target speaker and non-target speakers are colored in blue and orange, respectively, for clarity.

- Optimizing the separator using the downstream VAD output rather than conventional separation supervision focuses the separator’s learning on spectral regions of speech that are useful for VAD.

We call particular attention to the first point within the context of non-target-speaker (NTS) VAD. There is some evidence that people who contribute only small amounts of speech in a recording will have their identity lost while performing speaker diarization with clustering components, despite their speech being detected during VAD [26], suggesting that models only struggle with determining their identity. Such cases ultimately cannot be detected by TS-VAD due to an inability to extract an embedding vector for them. Our hope is that these speakers could effectively be captured by the NTS branch of our model, contrasting models that jointly model speaker identity and activity.

2.1. Model Architecture

The model architecture is diagrammed in Figure 2, featuring a spectral magnitude mask-based system inspired by the speech separation and target speaker extraction literature. It features two convolutional neural networks (CNNs) local feature extractors, after which the ivector speaker embedding is concatenated, followed by two bi-directional Long Short-Term Memory (BLSTM) layers modeling longer-term dependencies, with a final linear projection layer to the appropriate mask size and sigmoid activation to restrict the mask to the range $[0, 1]$.

A noteworthy diversion from the separation literature is that the system operates on the feature space of the pretrained VAD model, rather than the typical Short-Time Fourier Transform or learned filter-bank representations. In particular, as speech technologies frequently operate on log filter-bank energy (LFBE) features, an element-wise product of the magnitude mask would not be appropriate, and instead the log of the mask must be added to the features instead.

2.2. Data Pipeline

Our model was trained using synthetic on-the-fly mixtures generated using speech from speaker recognition corpora, augmented with noise and reverberation, which we refer to as “Walkman” datasets. In line with the motivation of our work, the dataset was designed to parallel speech separation datasets rather than diarization datasets, with the supervised training targets being generated from the pre-mixed audio signals rather than any human annotation of the speech boundaries. The use

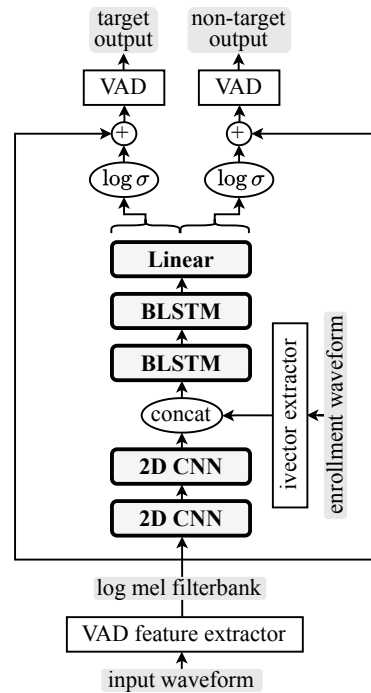


Figure 2: System architecture, with bold components representing the learned separator. The ivectors are concatenated to the post-CNN embedding at each frame. The log sigmoid sum is the log domain equivalent to a standard spectral magnitude mask.

of on-the-fly generation is to maximize the diversity of samples seen in training, including the use of sampling-without-replacement queues on the source corpora of speech, noises, and room impulse responses (RIRs), to ensure not a single piece of audio is seen twice in training until necessary. However, fixed random seeds were used in validation and testing to ensure consistent results across training runs and models.

Despite being closer in design to a separation dataset, we nevertheless aimed to mimic realistic VAD scenarios featuring speech onsets and offsets. In each sample, we potentially select an offset to be present by using the end of a source recording, placed to finish in the middle of the sample, and similarly potentially select an onset using the start of another recording, placed in the middle of the sample. This procedure is described precisely in Algorithm 1. We additionally used a version of this procedure with fully-overlapping speech for the entire sample duration more akin to typical speech separation training. The resulting two versions of the dataset are referred to as Walkman-diar and Walkman-sep respectively.

3. Experimental Configuration

The key questions we aimed to answer experimentally were:

1. Does the proposed approach improve robustness to the conditions only seen in the VAD model training?
2. Does using the VAD to generate supervision produce a better model than conventional supervision?
3. Can the model be trained effectively using the fully-overlapped separation-style (non-conversational) data?
4. Are both TS/NTS outputs necessary in training, as discovered by Ding et al. [12], since it is no longer a 3-class problem?

Algorithm 1 Walkman Pipeline for 5-Second Samples

```
1: use_reverb  $\leftarrow$  Bernoulli(0.15)
2: speech_wavs  $\leftarrow$   $\emptyset$ 
3: if Bernoulli(0.8) then  $\triangleright$  Trial will contain an offset
4:   offset  $\leftarrow$  last Uniform(1, 4) sec. of next(speech-queue)
5:   pad silence to end of offset
6:   if use_reverb then
7:     offset  $\leftarrow$  offset  $\otimes$  next(rir-queue)
8:   add offset to speech_wavs
9: if Bernoulli(0.8) then  $\triangleright$  Trial will contain an onset
10:  onset  $\leftarrow$  first Uniform(1, 4) sec. of next(speech-queue)
11:  pad silence to beginning of onset
12:  if use_reverb then
13:    onset  $\leftarrow$  onset  $\otimes$  next(rir-queue)
14:  add onset to speech_wavs
15: if Bernoulli(0.6) then  $\triangleright$  Trial will contain noise
16:  noise_wav  $\leftarrow$  random 5 sec. of next(noise-queue)
17: if Bernoulli(0.3) or speech_wavs =  $\emptyset$  then
18:  spk_emb  $\leftarrow$  random({all other ivectors})
19:  TS_wav  $\leftarrow$  silence
20: else
21:  TS_wav  $\leftarrow$  random(speech_wavs)
22:  spk_emb  $\leftarrow$  ivector(TS_wav)
23: NTS_wav  $\leftarrow$  sum(speech_wavs) - TS_wav
24: mixture  $\leftarrow$  sum(speech_wavs) + noise_wav
25: return spk_emb, TS_wav, NTS_wav, mixture
```

3.1. Models

The VAD model¹ used is from the SpeechBrain toolkit [27]. The model was trained on the LibriParty dataset, with additional augmentation from MUSAN [28] and CommonLanguage [29]. The LibriParty dataset is an artificial cocktail party/meeting scenario dataset generated from LibriSpeech [30] speech recordings, with noise from the QUT-NOISE-TIMIT [31] dataset and reverberation from Ko et al. [32], using conversational simulation as described by Fujita et al. [33].

The ivector extractor² used is the Kaldi [34] wideband VoxCeleb model, trained solely on unaugmented data from VoxCeleb 1 [35] and 2 [36]. The choice of ivectors over a more powerful deep learning model is not unusual in the literature, and is motivated by a more direct relation between the embedding and the acoustics that may be easier for the model to learn.

For the separation model, the 2D CNN layers have kernels of size (3, 3) and channel dimensions of 8 and 16 respectively, with LeakyReLU [37] activation. The BLSTM layers have 500 units in each direction. Given the 40-dimensional LFBE features used by the SpeechBrain VAD, the final linear layer contains an output dimension 80 (or 40 in experiments for question 4) and uses LogSigmoid activation with summation masking. The model ultimately contains 12M trained parameters, with the VAD being 100k parameters.

3.2. Data

The Walkman dataset was configured to generate 5 s samples using VoxCeleb 1 [35] and 2 [36] for speech, MUSAN [28] for noise (with speech and vocal music removed) and RIRs from Ko et al. [32]. As there is no inherent notion of an epoch with procedurally-generated data, we chose to define the training, validation, and test sets as containing 20k, 5k, and 3k samples respectively. Since the VoxCeleb corpora contain only train and test splits, we split the test set in half (maintaining disjoint

¹<https://huggingface.co/speechbrain/vad-crdnn-libriparty>

²<https://kaldi-asr.org/models/m7>

speaker sets) to form new validation and test sets, which retained enough data for fully-unique procedural samples. Models were trained using both Walkman-diar and Walkman-sep configurations to explore question 3, though the test set was always Walkman-diar to reflect deployment conditions.

We additionally used the dev and eval splits of the LibriParty dataset created for the SpeechBrain VAD, to evaluate experimental question 1. The dev split was monitored in training, but purely for observational purposes, with only the Walkman validation set being used for model selection.

3.3. Training Configuration

The models were trained for 300 epochs using the Adam [38] optimizer, using a learning rate schedule of $0.001 \times 0.99^{epoch}$, with Binary Cross Entropy as the loss function. The model from the epoch with the lowest validation loss was selected as the final model for evaluation.

Dropout of 0.1 was applied to the separator BLSTM layers, and all dropout layers in the SpeechBrain VAD were enabled during training. Additionally, Gaussian noise with standard deviation 0.5 was added to the ivector at each frame.

For question 2, the model was either trained using logits generated by the VAD on the ground truth separated input (VAD-Supervised, or “VS”) or conventional ground-truth annotation supervision (GT).

3.4. Evaluation Configuration

For evaluation, we used the inference mode of the SpeechBrain VAD, which involves chunking longer audios and using an algorithm to convert frame-wise likelihoods into segmentation. There is also a double-check feature, for which we used the plain VAD rather than speaker-conditioned model.

For evaluation we computed F_1 score, precision, recall, false positive rate (FPR), and false negative rate (FNR), all of which are standard binary classification metrics.

4. Results and Discussion

The results of our core experiments are shown in Table 1. One obvious conclusion is that using the fully-overlapped separation-style dataset appears to be completely ineffective, with performance being on par with simply running the VAD without any separation masking front-end. The models showed no signs of failure in the Walkman-sep training and validation datasets, which indicates that the model is simply not capable of operating on conversational-style overlap when trained in this fashion.

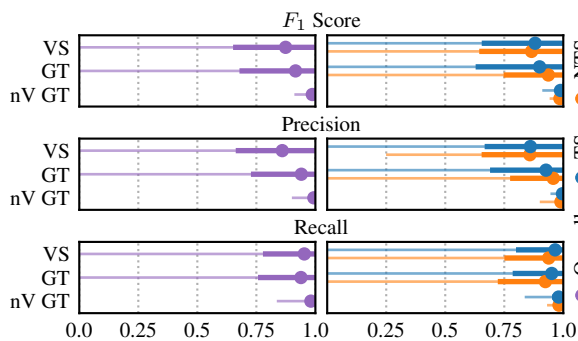
Similarly, the single-branch models performed quite poorly, confirming the findings of Ding et al. [12]. Interestingly, the TS model appears to be working to some extent, just with very poor recall, but the NTS seems to fail completely.

Overall, the best-performing model is, disappointingly, the model that did not use the pretrained VAD, and was simply trained entirely from scratch. However, it is worth noting that this model outperforms the SpeechBrain VAD when operated as a VAD model by fusing the TS and NTS outputs, which indicates that the other models are likely having their performance gated by the comparatively weak VAD backend. This is not terribly surprising, as the conditioned model has two orders of magnitude more parameters than the SpeechBrain VAD. We also see that using the ground truth annotations in training outperforms using the VAD itself for supervision, which we hypothesize is also due to the weak VAD performance, since with

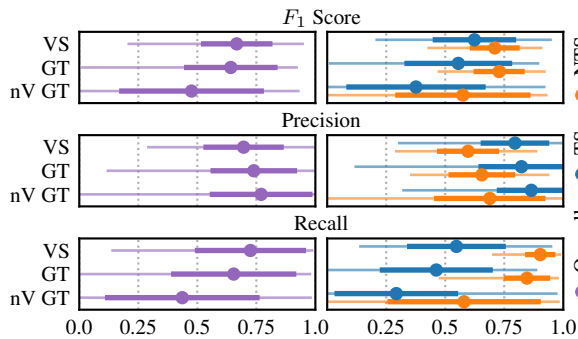
Table 1: Performance breakdown. All metrics in %. VS and GT refer to VAD supervision and ground-truth label supervision. Data style refers to the Walkman configuration. The VAD column refers to fusing the TS and NTS outputs to evaluate overall VAD performance.

	Branches	Pretrain VAD	Data Style	Train Target	Conditioned															
					Overall					TS					NTS					VAD
					F_1	Prec.	Rec.	FPR	FNR	F_1	Prec.	Rec.	FPR	FNR	F_1	Prec.	Rec.	FPR	FNR	F_1
Walkman-diar	dual	✓	diar.	VS	90.2	84.8	96.3	9.5	3.7	90.9	85.3	97.3	9.3	2.7	89.4	84.3	95.2	9.8	4.8	96.0
	dual	✓	diar.	GT	96.2	96.8	95.5	1.7	4.5	96.5	96.4	96.5	2.0	3.5	95.9	97.2	94.6	1.5	5.4	97.9
	dual	✓	diar.	GT	98.8	99.4	98.2	0.3	1.8	98.8	99.7	98.0	0.2	2.0	98.8	99.2	98.5	0.5	1.5	99.4
	dual	✓	sep.	VS	64.9	49.8	93.0	51.8	7.1	62.6	48.3	88.9	52.5	11.1	67.1	51.3	97.0	51.1	3.0	90.7
	dual	✓	sep.	GT	68.4	54.3	92.5	43.1	7.5	69.4	57.8	86.8	35.0	13.3	67.6	51.5	98.2	51.1	1.8	91.9
	mono	✓	diar.	VS	-	-	-	-	-	66.7	88.6	53.4	3.8	46.6	-	-	-	-	-	-
	mono	✓	diar.	VS	-	-	-	-	-	-	-	-	-	-	13.1	18.7	10.1	24.1	89.9	-
SpeechBrain VAD					67.6	51.2	99.5	52.4	0.5	67.6	51.2	99.5	52.4	0.5	67.6	51.2	99.5	52.4	0.5	98.2

ground truth labels the separator has the opportunity to learn to “correct” the signal in cases where the VAD decision would have otherwise been subpar.



(a) Walkman-diar Test Set



(b) LibriParty Test Set

Figure 3: Breakdown of model performance per dataset. The dot indicates mean, thick bars are \pm standard deviation, thin bars are min/max. VS indicates VAD-supervised training, GT indicates annotation ground truth supervision, and nV indicates no pretrained VAD was used and the model was fully trained.

However, we start to see the benefit of the proposed method in evaluation of the LibriParty test set, the results of which are presented in Figure 3. In this condition, we see a reversal of ranking of the functional systems, with the model not using the pretrained VAD performing appreciably worse than the rest. In particular, we see that the most of the relative drop in performance can be found in recall, especially in the NTS output, where the proposed method retains relatively strong performance. It seems that although the model still struggles to track the out-of-domain speakers when selected as the target

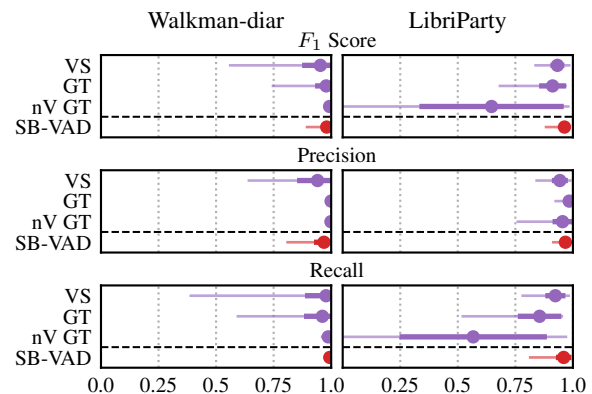


Figure 4: Breakdown of performance with model operating as a VAD, fusing the TS and NTS outputs. The dot indicates mean, thick bars are \pm standard deviation, thin bars are min/max. VS indicates VAD-supervised training, GT indicates annotation ground truth supervision, and nV indicates no pretrained VAD was used and the model was fully trained. SB-VAD is the stock SpeechBrain VAD model.

speaker, the proposed method still manages to fall back on the pretrained VAD’s performance for the non-target speech, while training the model from scratch does not grant this ability.

This is further supported when looking at the performance of the models when operated as a VAD, i.e. fusing the two output branches, as shown in Figure 4. Here we see that there is a large negative impact to the overall speech detection capabilities of the model not using the SpeechBrain VAD on the LibriParty dataset that it was developed for.

5. Conclusion

We have demonstrated the potential to effectively modify a VAD model to be a speaker-conditioned VAD model using a separation front-end module. By leveraging the VAD model as both a system component and method of supervision, the resulting model has improved robustness and recall of speech, particularly in the detection of non-target speakers.

Future lines of work to explore include leveraging a more powerful VAD model and exploring the use of neural speaker embeddings for more competitive performance, as well as developing support for multi-vector input, improving the usefulness of the non-target output for finding undetected speakers as postprocessing of conventional diarization systems.

6. References

- [1] E. C. 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, 1953.
- [2] N. Ryant, K. Church, C. Cieri *et al.*, "The second DIHARD diarization challenge: Dataset, task, and baselines," in *Proc. ISCA Interspeech*, 2019.
- [3] N. Ryant, P. Singh, V. Krishnamohan *et al.*, "The third DIHARD diarization challenge," 2020, arXiv:2012.01477.
- [4] S. Watanabe, M. Mandel, J. Barker *et al.*, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2020.
- [5] S. Cornell, M. S. Wiesner, S. Watanabe *et al.*, "The CHiME-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios," in *Proc. 7th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2023.
- [6] S. Cornell, T. J. Park, H. Huang *et al.*, "The CHiME-8 DASR challenge for generalizable and array agnostic distant automatic speech recognition and diarization," in *Proc. 8th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2024.
- [7] X. Chang, S. Watanabe, M. Delcroix *et al.*, "Module-based end-to-end distant speech processing: A case study of far-field automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 41, no. 6, 2024.
- [8] J. Du, Y.-H. Tu, L. Sun *et al.*, "The USTC-NELSLIP systems for CHiME-6 challenge," in *Proc. 6th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2020.
- [9] R. Wan, M. He, J. Du *et al.*, "The USTC-NERCSLIP systems for CHiME-7 challenge," in *Proc. 7th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2023.
- [10] A. Mitrofanov, T. Prisyach, T. Timofeeva *et al.*, "STCON system for the CHiME-8 challenge," in *Proc. 8th International Workshop on Speech Processing in Everyday Environments (CHiME)*, 2024.
- [11] I. Medennikov, M. Korenevsky, T. Prisyach *et al.*, "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. ISCA Interspeech*, 2020.
- [12] S. Ding, Q. Wang, S.-Y. Chang *et al.*, "Personal VAD: Speaker-conditioned voice activity detection," in *The Speaker and Language Recognition Workshop (Odyssey)*, 2020.
- [13] T. von Neumann, K. Kinoshita, M. Delcroix *et al.*, "All-neural online source separation, counting, and diarization for meeting analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.
- [14] C. Boeddeker, A. S. Subramanian, G. Wichern *et al.*, "TS-SEP: Joint diarization and separation conditioned on estimated speaker embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 32, 2024.
- [15] J. Kalda, C. Pagés, R. Marxer *et al.*, "PixIT: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings," in *The Speaker and Language Recognition Workshop (Odyssey)*, 2024.
- [16] X. Fang, Z.-H. Ling, L. Sun *et al.*, "A deep analysis of speech separation guided diarization under realistic conditions," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2021.
- [17] G. Morrone, S. Cornell, D. Raj *et al.*, "Low-latency speech separation guided diarization for telephone conversations," in *IEEE Spoken Language Technology Workshop*, 2023.
- [18] K. Žmolíková, M. Delcroix, K. Kinoshita *et al.*, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, 2019.
- [19] Q. Wang, H. Muckenhirn, K. Wilson *et al.*, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. ISCA Interspeech*, 2019.
- [20] J. Wu, Z. Chen, S. Chen *et al.*, "Investigation of practical aspects of single channel speech separation for ASR," in *Proc. ISCA Interspeech*, 2021.
- [21] L. Lu, N. Kanda, J. Li, and Y. Gong, "Streaming end-to-end multi-talker speech recognition," *IEEE Signal Processing Letters*, vol. 28, 2021.
- [22] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, 2017.
- [23] L. Meng, J. Kang, Y. Wang *et al.*, "Empowering Whisper as a joint multi-talker and target-talker speech recognition system," in *Proc. ISCA Interspeech*, 2024.
- [24] A. Polok, D. Klement, M. Wiesner *et al.*, "Target speaker ASR with Whisper," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.
- [25] N. Dehak, P. J. Kenny, R. Dehak *et al.*, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, 2011.
- [26] M. Maciejewski, D. Klement, R. Huang *et al.*, "Evaluating the Santa Barbara corpus: Challenges of the breadth of conversational spoken language," in *Proc. ISCA Interspeech*, 2024.
- [27] M. Ravanelli, T. Parcollet, P. Plantinga *et al.*, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.
- [28] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, arXiv:1510.08484v1.
- [29] G. Sinisetty, P. Ruban, O. Dymov, and M. Ravanelli, "CommonLanguage," Zenodo, 2021, 10.5281/zenodo.5036977.
- [30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.
- [31] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. ISCA Interspeech*, 2010.
- [32] T. Ko, V. Peddinti, D. Povey *et al.*, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.
- [33] Y. Fujita, N. Kanda, S. Horiguchi *et al.*, "End-to-end neural speaker diarization with self-attention," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.
- [34] D. Povey, A. Ghoshal, G. Boulianne *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.
- [35] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. ISCA Interspeech*, 2017.
- [36] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. ISCA Interspeech*, 2018.
- [37] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the International Conference on Machine Learning*, vol. 30, no. 1, 2013.
- [38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of the International Conference on Learning Representations*, 2015.