Whisper-TAD: A general model for Transcription, Alignment and Diarization of speech

Camille Lavigne
Université de Nancy
lavignecamille37@gmail.com

Alex Stasica
Utrecht University
a.stasica@uu.nl

Abstract

Currently, there is a lack of a straightforward implementation of diarization-augmented speech transcription (DAST), ie. implementation of transcription, diarization and alignment to the audio within one model. These tasks typically require distinct models, necessitating to stack them together for complete processing. In this study, we advocate for leveraging the advanced capabilities of the Whisper models, which already excels in automatic transcription and partial alignment. Our approach involves fine-tuning the model's parameters on both transcription and diarization tasks in a SOT-FIFO (Serialized Output Training-First In First Out) manner. This comprehensive framework facilitates the creation of orthographic transcriptions, identification of speakers, and precise alignment, thus enhancing the efficiency of audio processing workflows. While our work represents an initial step towards a unified transcription and diarization framework, the development of such a model demands substantial high-quality data augmentation and computational resources beyond our current scope. Consequently, our focus is narrowed to the English language. Despite these limitations, our method demonstrates promising performance in both transcription and diarization tasks. Comparative analysis between pre-trained models and fine-tuned TAD (Transcription, Alignment, Diarization) versions suggests that incorporating diarization into a Whisper model doesn't compromise transcription accuracy. Our findings hint that deploying our TAD framework on the largest Whisper model could potentially yield state-of-the-art performance across all mentioned tasks.

Keywords: Diarization, automatic speech recognition, Whisper

1 Introduction

Speaker diarization (SD) endeavors to ascertain "*who spoke when*" (Tranter and Reynolds, 2006). Various methodologies have been employed to annotate audio data for the purpose of identifying speakers within it. Conventionally, this task was compartmentalized into distinct sub-modules (Park et al., 2022), ranging from voice activity detection (VAD) to clustering speech segments and assigning speaker labels. However, the optimization of each module in isolation restricted overall optimization. With the advent of deep learning techniques, neural networks have been leveraged to improve the performance of these sub-modules by extracting speaker embedding (Variani et al., 2014; Heigold et al., 2016), thereby rendering models easier to train, more resilient to speaker variability, and robust under varying acoustic conditions (Zhang et al., 2019). A recent breakthrough is the adoption of fully end-to-end Neural Diarization (EEND; Fujita et al. (2019a,b)), wherein all sub-modules are replaced by a single neural network. This promising approach enables the joint optimization of model components, potentially enabling the handling of multi-speaker audio and overlapping speech. Initially implemented using bi-directional long short-term memory architectures (Fujita et al., 2019a), these models swiftly transitioned to self-attention-based networks (Fujita et al., 2019b). Nevertheless, challenges persist, including the model's limited capacity to handle a large number of speakers, the difficulty in achieving online processing, and the tendency for models to overfit the training data distribution (Park et al., 2022).

Recent advancements have demonstrated that the concurrent modeling of SD and automatic speech recognition (ASR) can enhance the performance of both tasks, as exemplified in various models (Silovsky et al., 2012; Huang et al., 2007). This integration allows SD to use both acoustic and linguistic information, resulting in superior performance compared to models relying solely on acoustic information. Furthermore, it enables not only to determine ”*who spoke when*” but also discerning ”*what*” was spoken. As discussed in Park et al. (2022), various approaches have been explored, including the introduction of speaker tag roles in transcripts (Shafey et al., 2019), MAP-based joint decoding frameworks (Kanda et al., 2019), and the emergence of End-to-End Speaker Attribution ASR (E2E SA-ASR, Kanda et al. (2020a)), which facilitates speaker counting, multi-talker ASR, and speaker identity determination without limitations on the number of speakers.

Our aim in this research is to unify the diarization and transcription task in one model. We achieved this by fine-tuning existing Whisper models, (Radford et al., 2023), which already transcribe speech with state of the art performance and align the transcription to the audio. Our fine-tuning enables the recognition of distinct speakers within the speech audio. By focusing on fine-tuning rather than extensive pre-training, we achieve transferable results even with limited data, making our model applicable to languages with minimal available resources. Thus, we introduce Whisper-TAD (Transcribe, Align, Diarize), an initial version of a versatile model that streamlines the DAST pipeline.

Our article is structured as follows; in section 2 we present our methodology, then in section 3 our experimental setup, in section 4 our experimental results and we finally discuss possible further works in section 5.

2 Methodology

2.1 Foundation model

As a foundation model we use Whisper (Radford et al., 2023). Whisper models already reach state of the art performance in orthographic transcription task. As highlighted by the authors, these models were designed in a multi-task format, also solving: translation, VAD, partial alignment, and language identification tasks. Although diarization was cited as a desirable task to solve in an ASR pipeline, the authors didn't address this in their original publication. In order to add this ability to the Whisper models, we add special tokens to the tokenizer as well as new randomly initialized embeddings for these new tokens. The new tokens are up to five speaker tokens as well as a noSpk token for VAD. We then fine-tune the models on both ASR and diarization tasks jointly.

2.2 Fine-tuning task

For the fine tuning, we used the SOT FIFO framework. SOT (Serialized Output Training) as been first introduced in Kanda et al. (2020b). It allows to train an attention-based neural network on both transcription and diarization using only one output. It is usable on data that contains multiple speakers and overlapping speech. When there is multiple speakers to classify, there are different ways to output the result of the deep neural network in one output. We choose FIFO (First In First Out) as it is the most used variant of SOT. In the FIFO approach, a distinct speaker ID is incrementally assigned to each newly detected speaker in the audio. For instance, the initial speaker detected is labeled as ”spk1,” the subsequent one as ”spk2,” and so forth. Consequently, there is no correspondence between the speaker IDs assigned to two different segments of audio, even if they contain the same speaker or speakers. For this reason our framework is a local E2E DAST model. We have not yet implemented a clustering of the speakers to recognize when speakers in different chunks of a same audio have the same identity.

One of the limitation of the SOT method that we use is that we cannot classify more than five speakers in one chunk of thirty seconds. However, the cases in which more than five speakers talk in one 30s chunk are pretty rare. It would therefore require a large amount of data augmentation to achieve decent accuracy on more than 5 speakers.

Figure 1 illustrates The SOT FIFO framework where all necessary tokens for each 30-second segment of audio (chunk) are generated by the autoregressive decoder. For every utterance of an audio chunk, the model initiates by outputting a speaker token (depicted in blue) alongside a timestamp token (depicted in green) to mark the beginning of the utterance. Following this, tokens outputted by the Whisper byte-pair encoding tokenizer (depicted in orange) are employed to transcribe the utterance. Once transcription is complete, a final timestamp token is appended to signify the end of the speaker's utterance. Furthermore, if the same speaker contributes multiple times within a single chunk, they are assigned a consistent speaker ID (ranging from 1 to 5).

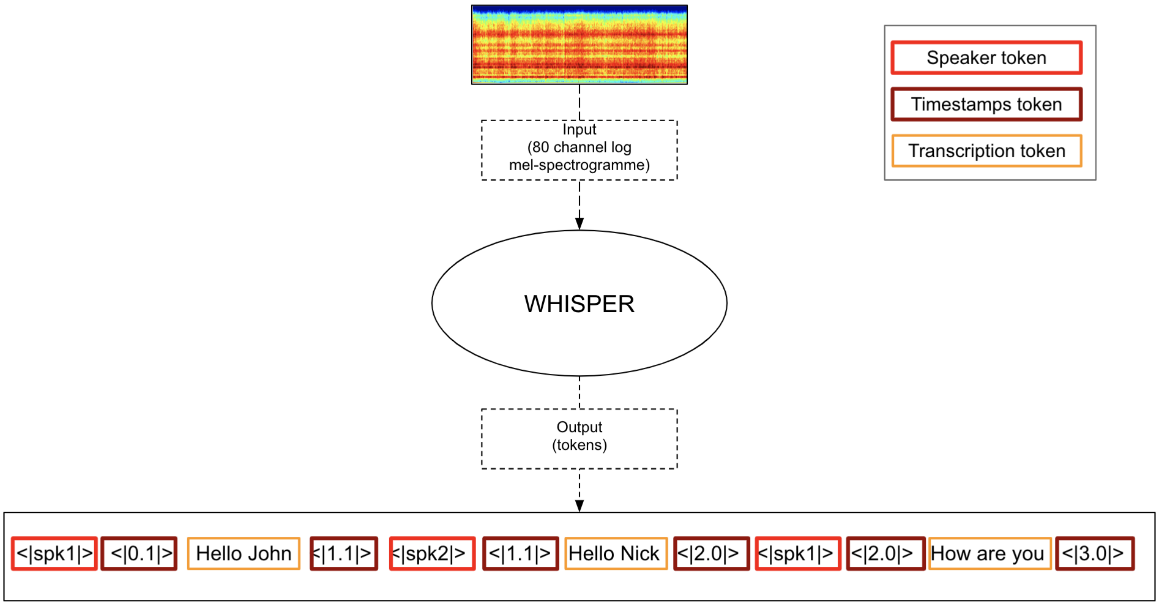

Figure 1: WHISPER-TAD framework

3 Experimental setup

3.1 Data

This study makes use of the AMI (Kraaij et al., 2005) and ISCI (Janin et al., 2003) benchmark datasets for our experiments.

The ISCI corpus regroups 75 meetings, with 4 different types of meetings with up to 10 participants, and the AMI corpus comprises 100 hours of audio from 171 meetings coming from multiple sites in which 3 to 5 participants are present. Both datasets provide the meetings transcriptions, word level alignment, and speaker labels. These datasets are suitable for the fine-tuning task as we plan on evaluating the performance of SD along with an ASR module. As we didn't found clear guidelines to split the ICSI corpus we used the full ISCI for training and validation of our model, not for testing. For the AMI corpus guidelines diverges (Landini et al., 2022). We decided to split it in train, validation and test sets as described by the official suggestions on the website of the corpus[1] as it seems to be a reliable, efficient split of the data.

Note that a few hours of audio from the AMI and ICSI corpus contains audio speech without transcription which increase the probability of hallucination at inference if the model was trained on these data. For ISCI corpus, these parts without transcription are parts where the speakers are ask to pronounce random numbers all together. These parts where removed from the training - validation - test datasets.

	AMI	ICSI	Total
Training	78	58	136
Validation	10	12	22
Test	9	0	9

Table 1: Share of the AMI and ICSI corpus in the training - validation - test sets. Shares are given in hours of audio.

3.2 Hyper-parameters

Due to limited computational power for this experiment, we only fine-tuned the base, small and medium Whisper models, but could not fine-tune the larger versions. We used a 0.05 dropout with a learning rate of $1e^{-5}$ and a batch size of 100. The optimizer used is Adam. We had access to one (24 GB ram) RTX 6000 GPU.

3.3 Metrics

Our models are trained on three distinct tasks, each requiring specific metrics for evaluation.

For the speaker diarization task, we chose the Diarization Error Rate (DER), which quantifies the accuracy of speaker diarization systems by measur-

[1] https://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml

ing the alignment between the predicted speaker segments and the ground truth. DER accounts for missed speakers, false alarms, and speaker misalignment. Specifically, we utilized its Python implementation from `pyannote` (Bredin et al., 2020). We do not use collar, as recommanded by `pyannote` guidelines.

For the transcription task, we employed the Concatenated minimum-Permutation Word Error Rate (cpWER) (Watanabe et al., 2020). Unlike traditional Word Error Rate (WER), cpWER gathers all the speech productions from a same speaker and calculate the WER per speaker. This is particularly useful in scenarios where the speech stream is segmented in a ”Diarization-style” manner, i.e., segmented by speaker.

Additionally, we employed traditional WER to compare the performance of the pre-trained models with those fine-tuned on both transcription and diarization. This comparison allows us to assess whether multitasking improves or hampers the performance of the models on their primary task.

For the Voice Activity Detection (VAD), we utilized the Equal Error Rate (EER) metric. EER represents the point where the false acceptance rate (ie. falsely classifying non-speech as speech) equals the false rejection rate (ie. falsely classifying speech as non speech), providing a balanced measure of VAD performance across different operating conditions.

4 Experimental Results

The results of our fine-tuning task on the different Whisper models are illustrated in Table 2. As to be expected, the larger the model, the better the metrics. Another notable observation is that the performance difference between the Base and the Small models is more significant than the one between the Small and the Medium. One explanation for this phenomenon is the lack of data, and the fact that we didn't do any data augmentation to mitigate this.

	DER	EER	cpWER
Base	0.498	0.655	0.548
Small	0.202	0.120	0.345
Medium	0.189	0.151	0.313

Table 2: Results of the fine-tuning task on the Base, Small and Medium Whisper models

As explained in 3.3, we also calculated the standard WER in order to demonstrate that even while adding the diarisation task, the performance of the models on the initial task they have been trained on does not decrease, but even increases as depicted by Table 3, showing that the fine-tuning on another task is also useful for the initial task, and a joint pipeline can only increases the performance for both tasks.

Base		Small		Medium	
PT	TAD	PT	TAD	PT	TAD
0.621	0.346	0.466	0.279	0.403	0.269

Table 3: WER comparison for three different sizes of the Whisper model. The models labeled as PT denote the pre-trained models, while those labeled as TAD indicate models fine-tuned for diarization

5 Further work

This study serves as a proof of concept, with further investigations required to fully evaluate the methodology's feasibility.

Firstly, a crucial step is transitioning from local E2E processing to global E2E processing. This entails enabling the model to consistently assign the same speaker ID to speakers across different audio chunks, rather than assigning new speaker IDs for each chunk as done in prior research by Cornell et al. (2024) using Wav2Vec (Schneider et al., 2019). Various approaches can be explored, such as incorporating a classification head by clustering all speakers across the entire audio, thus necessitating an additional output head for the model.

Secondly, larger versions of Whisper need to be fine-tuned to ascertain the maximum performance achievable using this methodology.

Thirdly, in order to train larger models for the diarization task, data augmentation is indispensable. Leveraging datasets like LibriSpeech (Panayotov et al., 2015) for data augmentation can enhance the training process. Additionally, data augmentation can facilitate the fine-tuning of these models for diarization tasks in languages with limited accessible resources.

6 Conclusion

This study introduces Whisper-TAD, a preliminary investigation into a versatile model designed to integrate transcription, sentence-level alignment, and diarization tasks within a unified pipeline. Employing a SOT FIFO method, special tokens are incorporated for speaker identification, enabling recognition of up to 5 speakers per 30-seconds audio chunk. Our experiments conducted on the ISCI and AMI corpora yield promising outcomes, suggesting potential applicability across languages with limited resources. Notably, our approach achieves competitive performance, even in the absence of data augmentation and without the exploration of larger models. These findings underscore the robustness and effectiveness of Whisper-TAD, offering valuable insights for future research directions in multi-task audio processing.

References

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Samuele Cornell, Jee-weon Jung, Shinji Watanabe, and Stefano Squartini. 2024. One model to rule them all? towards end-to-end joint speaker diarization and speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11856–11860. IEEE.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. 2019a. End-to-end neural speaker diarization with permutation-free objectives. *arXiv preprint arXiv:1909.05952*.

Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019b. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303. IEEE.

Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. 2016. End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119. IEEE.

Jing Huang, Etienne Marcheret, Karthik Visweswariah, and Gerasimos Potamianos. 2007. The ibm rt07 evaluation systems for speaker diarization on lecture meetings. In *International Evaluation Workshop on Rich Transcription*, pages 497–508. Springer.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, Tianyan Zhou, and Takuya Yoshioka. 2020a. Joint speaker counting, speech recognition, and speaker identification for overlapped speech of any number of speakers. *arXiv preprint arXiv:2006.10930*.

Naoyuki Kanda, Yashesh Gaur, Xiaofei Wang, Zhong Meng, and Takuya Yoshioka. 2020b. Serialized output training for end-to-end overlapped speech recognition. *arXiv preprint arXiv:2003.12687*.

Naoyuki Kanda, Shota Horiguchi, Yusuke Fujita, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. 2019. Simultaneous speech recognition and speaker diarization for monaural dialogue recordings with target-speaker acoustic models. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 31–38. IEEE.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*.

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Laurent El Shafey, Hagen Soltau, and Izhak Shafran. 2019. Joint speech recognition and speaker diarization via sequence transduction. *arXiv preprint arXiv:1907.05337*.

Jan Silovsky, Jindrich Zdansky, Jan Nouza, Petr Cerva, and Jan Prazak. 2012. Incorporation of the asr output in speaker segmentation and clustering within the task of speaker diarization of broadcast streams. In *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, pages 118–123. IEEE.

Sue E Tranter and Douglas A Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565.

Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4052–4056. IEEE.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, et al. 2020. Chime-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings. *arXiv preprint arXiv:2004.09249*.

Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. 2019. Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6301–6305. IEEE.