



# NeMo Open Source Speaker Diarization System

*Tae Jin Park, Nithin Rao Koluguri, Fei Jia, Jagadeesh Balam and Boris Ginsburg*

NVIDIA

{taejinp, nkoluguri, fjia, jbalam, bginsburg}@nvidia.com

## Abstract

We introduce an open-source speaker diarization system which is part of the NeMo conversational AI toolkit. During the Show and Tell session, we will present an interactive system which demonstrates both online and offline speaker diarization. The audience would be able to test the speaker diarization system by recording their voice. We believe that our demo session would be an excellent opportunity to learn and experience how a speaker diarization system can be implemented for real-life applications using an open source toolkit.

**Index Terms:** Speaker diarization, NeMo Toolkit, Automatic Speech Recognition

## 1. Introduction

In recent years, a rapid technological progress in automatic speech recognition (ASR) and natural language processing (NLP) has been made owing to the improvements in machine learning models, increased computational resources and available public datasets. When it comes to conversational AI, every component in the pipeline from ASR modules to natural language understanding (NLU) should perform seamlessly to understand or react to the free-form speech input. In a speech understanding system, the basic components include voice activity detection (VAD), speaker diarization, core ASR module, inverse text-normalization and punctuation module. Especially, when it comes to applications for understanding multi-speaker conversations, the importance of speaker diarization cannot be overemphasized since the understanding of human interaction can never be achieved without firmly identifying “who is speaking when”. Therefore, having a high accuracy on speaker diarization results is essential to build a reliable speech understanding pipeline.

NVIDIA NeMo [1] is an open-source modular toolkit for building state-of-the-art models for ASR, text-to-speech (TTS) and NLP. The modules in NeMo typically represent the building blocks of the conversational AI model architectures, including data layers, encoders, decoders, language models, loss functions, or methods of combining activations in neural networks. As a part of the ASR collection, speaker diarization models that can achieve state-of-the-art performance were recently added to NeMo. The speaker diarization toolkit comes with extendable collections of pre-built modules for speaker diarization with ASR.

In this proposal, we introduce our plan for demonstrating speaker diarization models and systems in the NeMo toolkit. In addition, we state the motivation of this demo and describe the core technical aspects of the NeMo speaker diarization pipeline.

## 2. Motivation

While there are numerous conversational AI open-source projects that include ASR and NLU models, the speaker diarization systems in these open-source toolkits are not providing

full-fledged integration with ASR. Since the actual speaker diarization results end-users see are always accompanied by transcription, having raw timestamps as speaker diarization output while missing ASR transcription still leaves a huge gap between the open-source systems and productized systems. On the other hand, the NeMo toolkit provides a speaker diarization system integrated with ASR where speaker labels, timestamps and ASR decoding results are all provided simultaneously. Thus, compared to other open-source toolkits, the NeMo speaker diarization system is much closer to the actual form of the deployed speaker diarization systems in real-life applications.

By demonstrating the speaker diarization system in NeMo, we would like to contribute to both industry and academia. For industry, the NeMo speaker diarization system can help build a prototype system or a production targeted speaker diarization system. For academia, the speaker diarization models and codebase will help speaker diarization and ASR research by providing a solid baseline system.

## 3. The System Architecture

Fig. 1 shows the data flow of the NeMo speaker diarization system that we demonstrate. Thanks to the solid foundation of the NeMo toolkit, the latest and greatest techniques in speech signal processing and deep learning are applied to the NeMo speaker diarization pipeline. The speaker diarization system consists of a VAD model based on MarbleNet [2] that generates timestamps of the input audio stream where speech-active regions filter out silence and background noise. Following the VAD module, speaker representations are extracted using TitaNet [3] model, which generates speaker embeddings on speech segments obtained from the VAD timestamps. These speaker embeddings are processed by a clustering algorithm to generate speaker labels with time-stamps. Finally, the time-stamps and the speaker labels created by speaker diarization module are matched with the word timestamps output from the ASR decoder to generate a transcript with speaker labels. Each of these models can be fine-tuned for improving performance or adapting to a certain domain.

## 4. Demo Description

### 4.1. Setup

At the table reserved for our Interspeech Show and Tell session, a laptop computer and a microphone are prepared to demonstrate our speaker diarization system. In terms of language, we only support English ASR and speaker diarization for this demo.

### 4.2. Offline Speaker Diarization

The offline mode represents the speaker diarization system that receives the whole input audio recording then processes it all at once. The models comprise the speaker diarization pipeline, in-

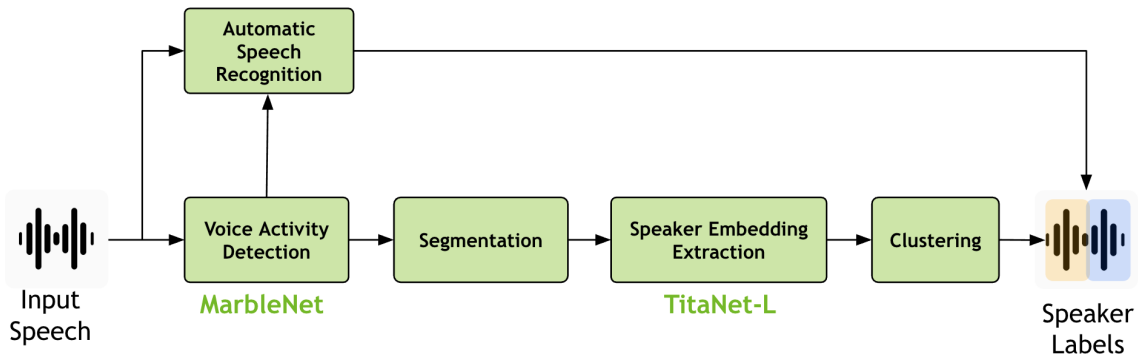


Figure 1: Data-flow of the NeMo speaker diarization system.

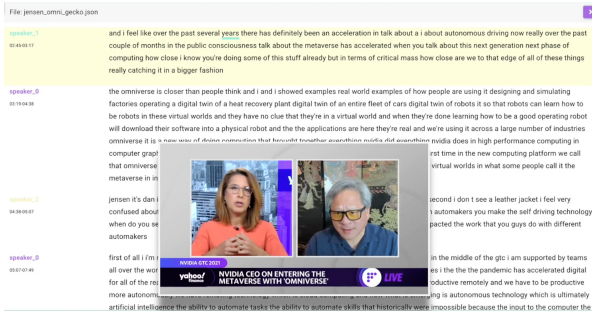


Figure 2: Screen capture of the offline speaker diarization demo video.

cluding MarbleNet and TitaNet, are customizable and all the adjustable parameters are provided in the YAML files. Along with the demo, we demonstrate that NeMo speaker diarization supports various ASR models in NeMo toolkit such as QuartzNet, CitriNet and ConformerCTC [1] which are pretrained and publicly released.

During the demo session, the audience is guided to record their voice and then feed the audio file to the NeMo speaker diarization system. After the recording is done, the user interface submits the audio recording to the diarization system and ASR transcription with speaker label is displayed on the screen. The audience can visually check the ASR result along with the speaker labels and the timestamps. If the demo system is not actively engaged by the audience, we play the demonstration videos of speaker diarization for real-life web video clips showing the actual application of the speaker diarization system on the real-world free-form speech.

In addition to the speaker diarization demo, we show how our speaker diarization system framework is compatible with annotation tools such as Gecko tool [4]. Gecko tool provides a user-friendly annotation interface for ASR and speaker diarization. The transcription and speaker diarization results from NeMo speaker diarization can be plotted, analyzed and annotated with the Gecko tool.

### 4.3. Online Speaker Diarization

Unlike offline speaker diarization, online speaker diarization processes the audio stream based on a few seconds of buffer so that the speaker diarization result is produced in real-time with a marginal delay. Online speaker diarization is often referred to as streaming speaker diarization depending on the context.

In the online speaker diarization demo, the audience are guided to record their voice to NeMo's speaker diarization sys-

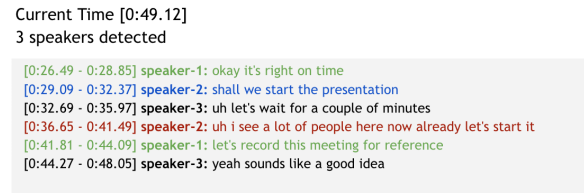


Figure 3: Screen capture of the streaming speaker diarization demo.

tem. As in Fig. 3, timestamps, speaker labels, decoded ASR results are displayed on the screen. To make a clear distinction between the speakers, each speaker's transcription is displayed in different colors and the estimated number of speakers is displayed in real-time. The audience can interactively test and examine the online speaker diarization system by testing the features of the speaker diarization system, such as speaker counting, speaker turn detection and time stamp generation. During the idle time where no audience is engaging the system, we play the recorded video of NeMo online speaker diarization.

## 5. Conclusions

In this proposal, we introduced our plan to show open-source offline and online speaker diarization systems. The audience will not only experience the live-action demo of the NeMo speaker diarization system, but also gain hands-on knowledge regarding the development, model training and evaluation of a speaker diarization system. Finally, we would like to emphasize the point that the entire speaker diarization and ASR systems are solely based on the publicly accessible NeMo open-source toolkit. We believe that our Interspeech Show and Tell session can contribute to both industry and academia in multiple aspects.

## 6. References

- [1] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Krizan, S. Beliaev, V. Lavrukhin, J. Cook *et al.*, "Nvidia nemo," <https://github.com/NVIDIA/NeMo>, 2022.
- [2] F. Jia, S. Majumdar, and B. Ginsburg, "Marblenet: Deep 1d time-channel separable convolutional neural network for voice activity detection," in *Proc. ICASSP*. IEEE, 2021, pp. 6818–6822.
- [3] N. R. Koluguri, T. Park, and B. Ginsburg, "TitaNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," *arXiv preprint arXiv:2110.04410*, 2021.
- [4] G. Levy, R. Sitman, I. Amir, E. Golshtein, R. Mochary, E. Reshef, R. Reichart, and O. Alouche, "Gecko-a tool for effective annotation of human conversations." in *INTERSPEECH*, 2019, pp. 3677–3678.