RoSSDaV: ROBOTIC SYSTEM (FOR) SPEAKER DIARIZATION AND VERIFICATION

BY

MASON DZIADULEWICZ TIPTON

BS, University of Wisconsin - Green Bay, 2022

THESIS

Submitted in partial fulfillment of the requirements for
the degree of Master of Science in Computer Science
in the Graduate School of
Binghamton University
State University of New York
2026

Accepted in partial fulfillment of the requirements for
the degree of Master of Science in Computer Science
in the Graduate School of
Binghamton University
State University of New York
2026

May (X), 2026

Shiqi Zhang, Committee Chair, Faculty Advisor
Monika Roznere, Thesis Defense Committee Member
Sujoy Sikdar, Thesis Defense Committee Member
School of Computing, Binghamton University

Abstract

Diarization, or, the problem of "who spoke when", is a critical topic for a wide variety of domains. These include, but are not limited to, telehealth, virtual meetings, media analysis, and live captioning. It also is important for robotic agents to be accurately able to recall who spoke when for further downstream processing, especially in environments where there are multiple speakers engaging in conversation. However, current solutions are limited in their application by robotic agents– solely audio-based systems fail, and audio-visual systems assume a static, egocentric camera. The main objective of this thesis is the design of a three-module system, RoSSDaV, that can leverage the mobile capabilities of a robot to carry out both audio diarization and subsequent speaker identity verification. This thesis outlines the theoretical construction of such a system. High-detail design and experiments concerning the first module in this pipeline, audio-based diarization (DiariChan), follows, as well as discussion of unique challenges to be addressed within this component.

Acknowledgments

This thesis would not have been possible without the support of my mentors and friends, both in and outside the lab. Before starting grad school, I would have never thought that this is something I could do. I thank them all for their tireless support, late-night answers to silly questions, and patience for my excited rambles. I would specifically like to highlight Professor Shiqi Zhang, who allowed me to embark on this journey and gave me the space to grow as a scientist. I would like to also thank Jake Juettner, who helped with the formatting of this thesis and letting me tag along on his adventures. Accordingly, I would also like to thank Zainab Altaweel for feedback on figures throughout this thesis. Special thanks to Ray Shi for helping me to avoid homelessness after the loss of my home.

I would most of all like to thank my crow; you given me wings in tandem.

I declare that generative artificial intelligence (AI) tools were used in this thesis as follows:

• AI Tool(s) Used: Claude • Sections/Pages: Methodology • Purpose: To construct scripts related to Python generating figures. • Verification Process: The AI-created script was manually reviewed and corrected in several places by the author.

Detailed documentation of use is available in Appendix [Enter in Appendix letter or number here] I declare that no generative artificial intelligence (AI) tools or services were used in the writing or analysis within this thesis. Claude Code was used to write scripts pertaining to routine testing, with manual correction by the author. ResearchRabbit and other AI literature search tools were used sparingly and experimentally to retrieve research, but the reading of these were done by the author. (TODO: EDIT TO CONFORM TO GUIDELINES)

Table of Contents

List of Tables viii

List of Figures x

List of Abbreviations xi

1 Introduction 1

2 Background 7

2.1 Metrics within Audio Diarization 7

2.1.1 Diarization Error Rate 8

2.1.2 Jaccard Error Rate 8

2.2 Confidence Calibration and Expected Calibration Error 9

2.2.1 Confidence Calibration 9

2.2.2 Expected Calibration Error 11

3 RoSSDaV: Methodology 12

3.1 Overview 12

4 Confidence Calibration in Audio Diarization 14

4.1 Overview 14

4.2 Literature Review 15

4.2.1 On Audio Diarization Techniques 15

4.2.2 On Confidence Calibration 17

5 Temperature Scaling: Experiments and Results 18

5.1 Methodology 18

5.1.1 Research Objectives 18
5.1.2 On Models 20
5.1.3 Procedures 20
5.2 Results and Discussion 22
5.3 Conclusion 25

6 DiariChan **26**
6.1 Uncertainty Tap 27
6.2 Speaking Tracing Buffer 29

Conclusions **31**

Bibliography **32**

List of Tables

5.1 Speaker diarization results on AMI, AISHELL4, and ALI test sets. DiariZen is evaluated offline; Sortformer is evaluated in streaming mode under four latency scenarios. DER and JER use a 0 s collar. ECE measures confidence calibration of $P(\text{any speech})$ at $T = 1$ with no scaling and after optimal temperature scaling (T^*). Lower is better for all metrics (↓). 23

List of Figures

1.1 A diagram showing a rather extreme example of upstream confusion leading to broken immersion in the downstream. 3

1.2 High Level Overview of RoSSDaV System 5

3.1 High Level Overview of RoSSDaV System 13

5.1 Overview of DiariZen audio diarization pipeline. Figure adapted from Raghav 2026. [46] 19

5.2 Offline results. 24

5.3 Effects of Temperature Scaling on Sortformer (all scenarios) and Offline DiariZen versus DER/JER and ECE. 25

6.1 AudioStream Step of DiariChan 27

List of Abbreviations

ASR Automatic Speech Recognition

DER Diarization Error Rate

ECE Expected Calibration Error

JER Jaccard Error Rate

OLA Overlap Add

RoSSDaV Robotic System (for) Speaker Diarization and Verification

SD Speaker Diarization

SotA State of the Art

STB Speaking Tracing Buffer

WDER Word-Level Diarization Error Rate

1 Introduction

The process of separating a stream of audio and ascribing what one is hearing to individual speakers is a process integral to social interaction. Without being able to perform this, complex acoustic scenes would be difficult to traverse, as listeners would no longer be able to selectively attend to a single conversational partner. The voice of said partner would be lost to competing speech streams.

Fortunately for humans, preventing this is a task that takes little to no conscious effort[1]. The brain automatically performs the process of selective auditory attention, honing in on an auditory stimulus and performing subsequent filtering while enhancing task-relevant auditory signals [7] [52]. This means that a person in a crowded bar can focus on a conversation with the bartender without necessarily having to know all the details of the conversation that the patrons next to them are having. Within psychoacoustics and cognitive neuroscience, this phenomenon is studied as the "cocktail party effect" [3].

In Human-Robot Interaction, a robotic agent being able to do the same has obvious benefits in populated settings. Being able to give personalized care in hospital waiting rooms or in sports stadiums allows an agent to carry out a conversation with a user in need while still taking inventory of who else may be waiting to talk to them. However, the internal mechanisms by which the human brain can identify a speaker do not easily map to the digital realm. Within the artificial intelligence field, this is referred to as "speaker diarization" [2]. It is a task that has generated significant research interest in the past decade [41] [51], with a wide variety of audio [21] [15] and audio-visual [12] processing-based methods being explored for online [54] and offline [41] [21] diarization.

[1]This statement is in a general sense, and is a claim that does not hold for some neurodiverse individuals. [36]

However, most audio-visual-based solutions assume a static, egocentric camera in which all speakers of note are in-frame of the camera [34]. This does not leverage the nature of a mobile robot, which can actively articulate itself in order to move its camera, which in turn can contribute to visual verification of a speaker [55]. These systems also do not often provide failsafes for this process, which can break immersion if an AI erroneously labels the user dialog with a faulty speaker identity [24]. There is little research addressing this problem specifically as a robotics capability, to the point where the application of audio diarization to robotics has no extant survey paper. The omission of this challenge from the major audio diarization survey papers [41] [51] and lack of available literature [12] makes the subtopic of integration of speaker diarization with robotics a nascent one.

In addition to the problem of camera mobility status, there are two other factors blocking audio-visual diarization solutions from being proper fits for robotic agents in the context of Human-Robotic Interaction.

One would be the hardware capabilities of the platform such a solution is running on. While certain models of robot have strong on-board compute, such as the NVIDIA Jetson Orin Nano the Unitree G1 possesses, there are older models of semi-/humanoid robot that don't have those same strengths. Pepper is semi-humanoid robot that is a good example of this. While the Unitree G1 is able to boast top-of-the-line AI performance, Pepper's computational power is sparse by comparison, only having a Intel Atom E3845 quad-core CPU and 4GB of RAM, despite being a model that is designed for social interaction. We find it to not be a compelling argument that hardware should necessarily scale to accommodate for robotics software that takes an ambitious amount of computing power. Instead, this thesis aligns itself with the philosophy that the best solution is often the simplest, both in design and in the amount of processing power it demands. A solely audio-visual solution takes a large amount of processing power. This is problematic for a social robot, as diarization is only one step in the path to speech recognition and responding in turn via a dialog module. Unnecessary power being utilized may lead to the diarization module being a bottleneck, which can break immersion by causing a delay in a robot's response.

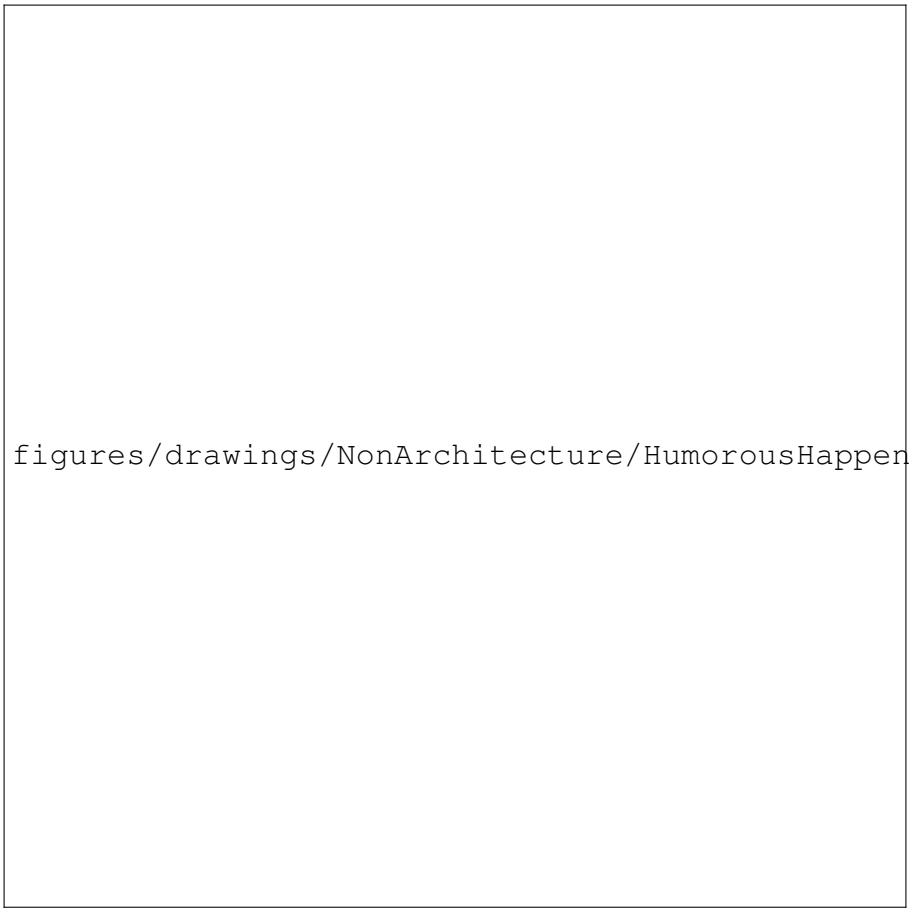


Figure 1.1: A diagram showing a rather extreme example of upstream confusion leading to broken immersion in the downstream.

The second factor to consider is reliability. Being confidently wrong about our guesses as to who spoke when in our upstream can have severe consequences for our downstream processes. Figure 1.1 illustrates such an example of this through a scenario involving a robotic agent and two users. In steps A and B, the robot engages in dialog with the users, whose responses differ in sentiment. The user that has treated the agent with respect in their dialog returns in step C. While the cited "voice deepening surgery" may seem outlandish, it reflects a wide variety of scenarios where there is some vocal difference in a returning speaker. This could be from illness, the natural effects of puberty on the adolescent voice, or a user simply talking in a different style than they employed during their initial meeting with the agent. Whatever the cause, on the acoustic level, the agent now cannot match the acoustic data with an existing identity. It is, however, similar enough to the speaker who treated it with disrespect. Based on the level of confidence the agent has in that prediction, it may decide that the best move is to assume that the blond speaker is the previous black-haired one. Step D shows the result of this– the user is (rightfully) rather confused and the immersion is broken. Thus, having accurate confidence values and deciding what to do with said values are both important.

It is clear from these two concerns that having an audio-visual solution that can confidently fail may not be the best fit for robots having interactions with humans.

Current diarization modules do not take these factors into account. In the case of audio-only solutions, the engineering approaches bifurcate into two paths. One path is the modern cloud large audio-language model (LALM) path that collapses the pipeline (Voice Activity Detection $\rightarrow$ Diarization $\rightarrow$ Speaker Identification $\rightarrow$ Transcription) from start to finish and effectively becomes a black box given the architecture of these models. There are also different failure modes associated with the speaker identification aspect of this. Either the LALM will refuse this task [31] or try but be unable to [28].

The alternative is having a local acoustic pipeline. This is beneficial on the transparency front, as they can preserve identity and local acoustic features. However, the mainstream solutions of this type (ex. diart [11], DiariZen [18]) label but do nothing with our confidence we have in a

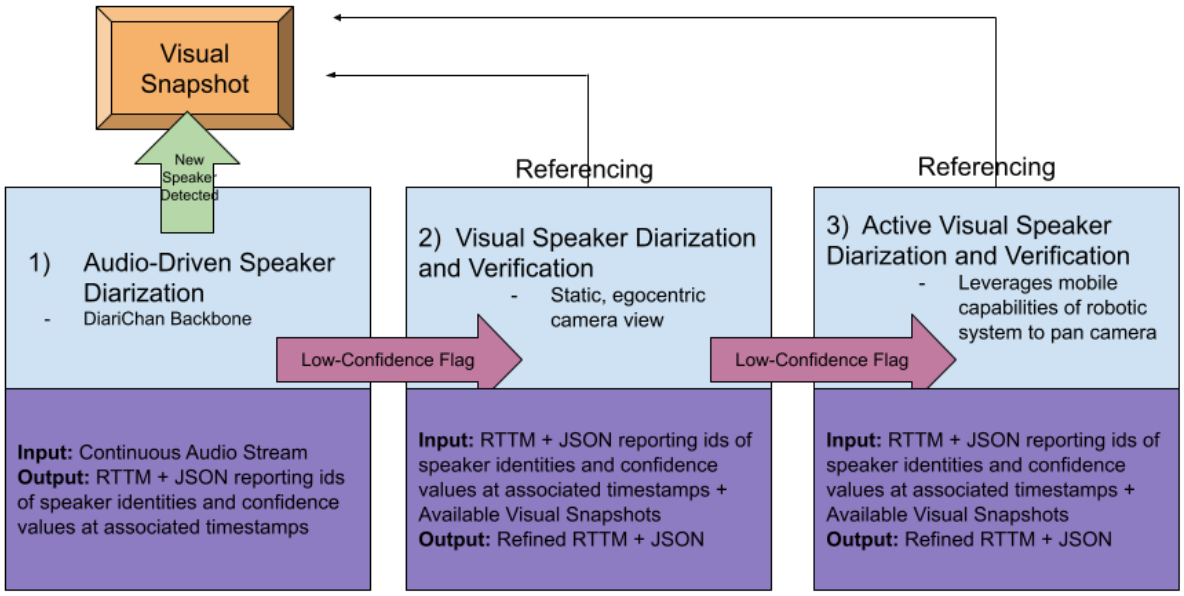


Figure 1.2: High Level Overview of RoSSDaV System

speaker label. The deployment infrastructure for this type of solution is also nearly absent in the standard Robotic Operating System (ROS).

With the gaps of existing solutions clearly exposed, we can start to outline a new approach. It is sensible of us to want a solution that benefits from the transparency a local acoustic pipeline has while still being transparent about the confidence we have in our model's predictions. This can be the difference between being confidently (and disastrously) wrong and flagging the dialog module to say, "Pardon me, who said (x) before?". Being on a robot also gives us both a constraint and a benefit. While we'd like to only leverage more computationally needy solutions when we truly require them, we can also take advantage of the fact that we are on a robot. If we opt for an audio-visual approach, we are not limited by a static camera. We are able to receive more information by becoming mobile and changing the orientation of our camera.

The proposed system in this thesis, RoSSDaV, aims to tackle the problems of speaker diarization and identification through a three-module pipeline. As can be seen in Figure 1.2, this begins with online audio diarization, with a snapshot taken of the environment when a new speaker is detected by the audio diarization module. Depending on confidence of the audio diarization system, we optionally progress to the audio-visual stage. Transition between this and the last

stage is carried out in a similar manner, hinging on the confidence the module has in the refined prediction.

We also contribute DiariChan, a streaming audio diarization module that builds on the local acoustic pipeline tradition while being explicit about its confidence values as it predicts who is speaking and when. Leveraging findings from experiments dealing in applying calibration to the confidence values of online and offline audio diarization models, DiariChan introduces *confidence gates* — thresholds that selectively invoke downstream modules only when the audio stage alone is insufficiently confident.

The remainder of this thesis explores experiments within the first component of RoSSDaV and an architectural overview of DiariChan, along with the latter's experiments as well. The other two components of RoSSDaV are discussed in a theoretical capacity, along with next concrete steps towards actualization of the full system.

2 Background

This thesis focuses on developing modules for audio diarization and audio-visual diarization for robotic agents, where confidence values are used in order to flow between these stages.

Given how little diarization is discussed within robotics, other research on this topic is sparse. Dhaussy et al. [12] proposed an audio-visual solution in 2023 within the framework of Human-Robot Interaction. While this system, which functions as a training-free solution tracking sound emission and source, achieved 19.27% DER on the Pepper corpus (robot setting, no oracle VAD), there are inherent limitations to Dhaussy's system that RoSSDaV aims to address.

Dhaussy et al.'s solution, at its simplest, aims to solve the speaker diarization problem through tracking what face is associated with noise at a given time. This is good for on-screen speaker diarization but explicitly does not address off-screen speakers, which RoSSDaV aims to provide a solution to. Speaker identity is also never verified within Dhaussy's system, only tracked over time. RoSSDaV aims to fill in this gap as well by performing speaker verification through the three stages of its pipeline.

The remainder of the chapter will focus on contextualizing vocabulary that is used throughout the thesis.

2.1 Metrics within Audio Diarization

Two main metrics are used within this thesis for measuring error rate within audio-based speaker diarization– diarization error rate (DER) [41] and Jaccard error rate (JER) [48]. These should not be confused with the error types which comprise these metrics. Other metrics, such as

WDER [49], also appear in related scientific literature. However, experiments within this thesis utilize DER and JER as main metrics.

2.1.1 Diarization Error Rate

Diarization Error Rate, or "DER"[1] is a fraction reporting the amount of time within an audio file/stream that was incorrectly diarized. It is a composite of three different error types divided by the total time, giving us:

$$\text{DER} = \frac{T_{\text{miss}} + T_{\text{fa}} + T_{\text{conf}}}{T_{\text{total}}}$$

T_{miss} refers to the time spent in which the system erroneously records no speech activity. This also can count for multi-speaker overlap where a given system does not mark all who are speaking as doing so.

T_{fa} is the opposite of T_{miss}, meaning that the system erroneously does record speech activity when there is none – a false alarm. Some definitions of this, depending on the applied task, extend this to include background noise or speech [50].

T_{conf} is confusion. This refers to instances where speech is misattributed to another person. This is dangerous for a downstream task where an AI agent responds to speech, potentially interrupting personalization [2].

2.1.2 Jaccard Error Rate

While DER is the gold standard of error rates [13] [51], it is a metric that can suffer from an adjacent problem to class imbalance in machine learning [48]. A speaker being consistently well-diarized by the system is a mark of success– unless that speaker is the main focus of the audio and other speakers that take up far less time (ex. individual reporters at a press conference)

[1] Also sometimes referred to as 'NIST DER', reflecting the term's origins in the National Institute of Standards and Technology (NIST) [1]

are poorly labeled by the system. DER alone gives no insight into this issue.

Jaccard Error Rate, or "JER", a metric popularized by the release of the DIHARD dataset [48], is a metric used to paint a better picture by weighting each speaker equally. This is based on the Jaccard index [23], and is expressed by:

$$\text{JER} = \frac{1}{N}\sum_{i=1}^{N_{\text{ref}}} \frac{\text{FA}_i + \text{MISS}_i}{\text{TOTAL}_i}$$

where N_{ref} is the number of speakers in a given reference script of an audio. Of DER and JER, DER is more standardized and popularized within scientific literature, but the downstream implications of JER make JER important to this thesis.

2.2 Confidence Calibration and Expected Calibration Error

2.2.1 Confidence Calibration

Confidence calibration is a concept that is not specific to the domain of audio diarization. A well-calibrated model is defined as one where "the output predicts probability estimates representative of the true correctness likelihood" [16]. This is a problem that predates machine learning by at least a century, having well-documented origins in discussions of accuracy of meteorological reports [35]. 1950 would bring a more precise mathematical definition to calibration (again in the same sphere of meteorology), with Brier coining the eponymous Brier Index as a measurement of "verification of forecasts expressed in terms of probability" [6].

Machine learning, in the same quest to achieve calibration, would briefly flirt with methods like isotonic regression and Platt scaling for sigmoid post-hoc recalibration. However, the first paper to conduct an empirical, systematic study of calibration across machine learning model types would come in 2005 at Cornell, where researchers Alexandru Niculescu-Mizil and Rich Caruana came to the conclusion that models, including neural networks, tended to be well-calibrated [38].

The rapid developments in deep learning that followed this paper would upset this finding as models increased their capacity dramatically. Guo et al. [16] revisited the conclusion that Niculescu and Caruana had come to over a decade before and found that their conclusion was no longer true. The cause? Advancements in the same factors that influenced calibration.

Fortunately, Guo et al. had also identified a solution to be found in *temperature scaling*, a single-parameter variant of Platt scaling. Platt scaling, a parametric approach to calibration, approaches the problem of calibration by "using the probabilistic predictions of a classifier as features for a logistic regression model" [16]. While other extensions to Platt scaling existed at this time, Guo's solution was comparatively simpler. Rather than passing the outputs of the classification layer of a neural network through a logistic regression model, Guo would look directly at the *logit vector*– the raw, unnormalized output layer of a classification model that exists as unscaled scores before they are passed through a softmax or sigmoid function.

Temperature scaling is described mathematically as:

$$\hat{q}_i = \max_k \sigma_{\text{SM}}(\mathbf{z}_i/T)^{(k)}$$

where:

- $\hat{q}_i$ is the calibrated confidence score.
- σ_{SM} is the softmax function.
- $\mathbf{z}_i$ is the logit vector (raw unnormalized output layer of a classification model).
- T is the temperature parameter ($T > 0$).
- k is an index for a specific class, with $\max_k$ representing the total number of classes.

This temperature scaling notably does not affect model accuracy, as it does not change the maximum of the softmax function.

2.2.2 Expected Calibration Error

When describing confidence calibration, Guo et al. uses Expected Calibration Error (ECE) as a primary metric, which is based on defining miscalibration as "difference in expectation between confidence and accuracy" [16]. ECE builds towards this definition by taking predictions, dividing evenly into M bins, and finding weighted bin average of calibration. [40] This is expressed as:

$$\text{ECE} = \sum_{m=1}^{M} P(m) \cdot |o_m - e_m|$$

where:

- M, as previously stated, is the number of bins.
- o_m is the true fraction of positive instances in bin m– our accuracy.
- e_m the mean of the post-calibrated probabilities for the instances in bin m– our confidence.

$P(m)$ is the empirical probability of all instances that fall into bin m. This can be further defined by:

$$\text{P(m)} = \frac{|B_m|}{n}$$

where

- B_m is the set of indices of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m-1}{M}]$ and
- n is the number of samples in the bins.

As Naeini [40] states, "(t)he lower the value (...) of ECE (...) the better is the calibration of a model".

3 RoSSDaV: Methodology

3.1 Overview

This thesis presents a three-module system for speaker diarization and subsequent speaker verification, aiming to preserve both unnecessary computational power and also leveraging the unique capabilities of a robotic agent to actively verify speaker identity. Subsequent chapters focus on each module of the system, including experiments performed and justification of the system.

A high-level overview of the system can be seen in Fig 3.1, with corresponding input and output for each module of the system's pipeline. RoSSDaV is intended to be used in settings where an unknown number of speakers are generally maintaining their relative positions. The system's design also gives some leeway in speaker presences, such as use cases where one is speaking while crossing to go to a room's water fountain. Notably, we also aim to tackle the challenge of off-screen speakers having a lower chance of being counted as speakers when audio confidence is low by allowing the robotic agent to actively pivot its camera to perform speaker verification.

We begin with an audio-driven speaker diarization backbone in the form of our creation DiariChan, optionally progressing to visual and then mobile stages depending on whether or not our system is confident in its predictions. At the time a new speaker is detected by the audio module, a picture of the robot's surroundings is taken and stored for optional processing by the visual and mobile stages, should the given situation require those modules. The visual stage employs the robot's camera to survey the area at its immediate orientation to see if any low-confidence predictions can be refined and have their confidence boosted by fusion of results from the visual detection stage. If the robot finds that it cannot resolve this, it can potentially

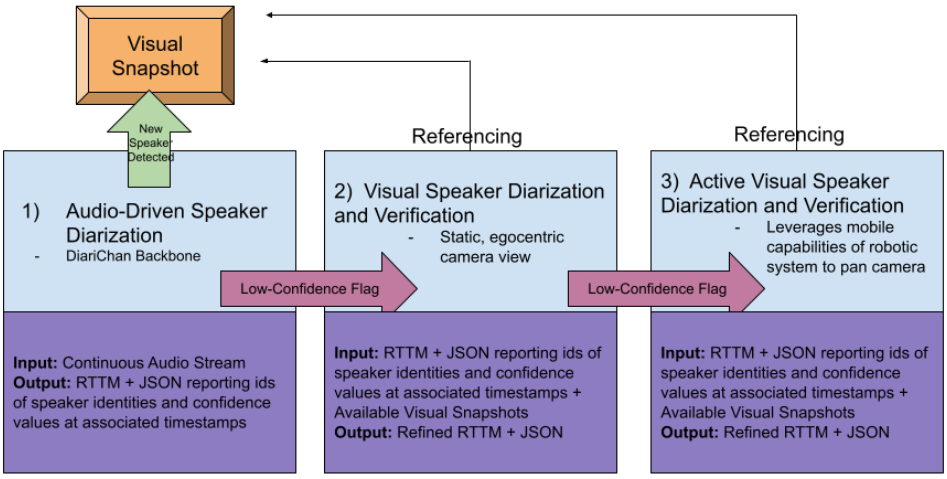


Figure 3.1: High Level Overview of RoSSDaV System

pivot to attempt to identify the speaker whose identity is in question.

This system is designed with the knowledge in mind that as a robotic agents internal processes grow more complex, there is a need for hardware to scale in turn. Swiss computer scientist Rolf Pfeifer discusses this idea in his book, “Understanding Intelligence”. Among other robotic design principles, he mentions the concept of *parsimony* (also referred to as *Occam’s razor*), wherein the best design choice is often the simplest. That is, Pfeifer states that ”A robot for obstacle avoidance that employs simple reflexes based on IR sensors is more parsimonious than one that does extensive visual analysis for the same purpose.” [43]. The principle that this quote exemplifies also applies to RoSSDaV– audio diarization alone may work quite well, and it is entirely possible for speakers to have a conversation with a robotic agent where speakers are easily identified by the audio module alone, with the visual snapshot not playing an active role in downstream speaker verification. It would be a waste of power to unnecessarily employ a visual and potentially mobile solution, especially given that a robot using RoSSDaV could be utilizing the system to hand its results off to a dialog system that also requires computational power. RoSSDaV is not a system necessarily running in isolation, and thus, the design of RoSSDaV is centered around consideration for other tasks the robotic agent may be simultaneously performing.

4 Confidence Calibration in Audio Diarization

4.1 Overview

This thesis presents a comparative analysis of SotA offline and online audio diarization systems and confidence calibration of said systems, with this chapter devoted to providing context for experiments laid out in following chapters. Accuracy is an important metric in classification problems, but the confidence we have of that reported accuracy also must be taken into consideration. For the purposes of this thesis, we turn to the findings of Guo et al. (2017) within their groundbreaking paper, "On Calibration of Modern Neural Networks" [16].

In this work, Guo et al. found that most modern neural networks at the time (including the famous *ResNet*) were not well-calibrated. This means that within a classification problem, the "probability associated with the predicted class label...(does not)...reflect its ground truth correctness likelihood". They also provided a solution for this miscalibration in the form of *temperature scaling*, a parametric solution that serves to correct miscalibration while preserving the accuracy of the model. Despite the impact of this 2017 paper, we find in this thesis that SotA audio diarization models fail to apply this to their models and that applying this technique to audio diarization models (both online and offline) results in marked improvement of confidence calibration in certain cases.

4.2 Literature Review

4.2.1 On Audio Diarization Techniques

Speaker diarization, or the problem of "who spoke when", is a task that is fundamental for downstream tasks like ASR (automatic speech recognition, where spoken word is transcribed) and speaker verification, where a speaker's identity is confirmed. Indeed, much of the early research efforts from the 1970's onward [39] for speaker diarization were motivated by the need for ASR. ASR techniques for much of this time were dominated by solutions based on HMMs (Hidden Markov Models) [45].

Markov models are used to stochastically model pseudo-random systems [53], with the simplest Markov model being the Markov chain. The Markov chain models the state of a system with a random variable that changes through time, making it useful when we need to compute a probability for a sequence of observable events [4]. A Hidden Markov Model is a Markov chain for which the state is only partially observable or noisily observable. In the case of speech recognition, the observable state for ASR is the audio signals at a given timestamp, with the hidden state being the words being spoken [45].

The explosion of deep learning popularity in the 2010s uprooted the dominance of HMMs, with a joint statement from Toronto, Microsoft, Google, and IBM in 2012 reporting that deep neural networks "have been shown to outperform GMMs[1] on a variety of speech recognition benchmarks, sometimes by a large margin." [20]

Even with the advent of deep learning, diarization systems still suffered from a few persistent problems. Diarization systems were typically based on the clustering of speaker embeddings [15], which are vectors of fixed length encoding a voice's unique qualities that map to variable-length audio data [10]. These vocal fingerprints are rich in information related to speaker identity

[1]Gaussian mixture model– a method used during the period of HMMs dominance for the task of speaker verification.

(timbre, tone, etc.), allowing diarization systems to distinguish one speaker from another. However, the clustering approach also faced difficulty when it came to speaker overlaps, as the clustering algorithms implicitly assumed separation between speakers [15].

This was a problem that had previously been side-stepped through usage of multi-channel audio data[2], but this was a method that did not fare well when it came to real audio recordings. Fujita et al.'s 2020 paper[3] offered a solution to this through development of End-to-End Neural Diarization (EEND), which avoided the need for clustering through reframing the problem of speaker diarization as a multi-label classification problem, which could then be processed by a neural network employing BLSTM[4] or self-attention. In the following years, this base model would undergo various evolutions in subsequent models, with DiariZen [18] (a hybrid method utilizing both EEND and clustering) being the latest successor to this SotA lineage of offline methods.

The EEND architecture has also been applied to streaming audio as well. Shortly after Fujita et al.'s 2020 paper, the Amazon Alexa Speech team would publish BW-EDA-EEND [17] in 2021. This transformed the EEND model into an online version, utilizing a Transformer encoder in a blockwise fashion to enable the online diarization. Only a few months later would this be outpaced by Xue et al.'s addition of the FLEX-STB [54], which allowed for a flexible number of speakers rather than the rigid limitations that were put into place by the initial EEND architecture. The mainstream streaming lineage would continue with Liang et al.'s 2023 paper on FS-EEND [30] (introducing frame-wise detection) and follow-up 2024 paper on long-form streaming [29]. For accessibility reasons, however, this thesis does not make comparison to

[2]Interestingly enough, literature at this time does not acknowledge this gap, but rather, implicitly dictates the gap as a necessity for the HMM through the fundamental equation $W^* = argmax P(W)P(O|W)$ [45] positing one acoustic stream O from one word sequence W. If one is historically-minded as to why this was accepted for decades without much questioning of the limitations of this, one could look to DARPA being the primary source of funding for most major ASR milestones from 1971 through the late 1990s [37] and note that most DARPA target domains were single-speaker or could be recorded on multiple channels.

[3]Fujita et al. has two separate papers mentioned in this thesis surrounding their SA-EEND framework. Their 2019 pre-print introduced EEND architecture, whereas their 2020 conference paper explicitly highlighted their contribution of formulating speaker diarization as a multi-label classification problem to be solved through a permutation-free objective function. For the sake of clarity and cohesion, all mentions of Fujita et al.'s EEND work will be referred to by the 2020 paper.

[4]Bidirectional Long Short-Term Memory, which is a specialized variant of a Recurrent Neural Network for sequential data.

LS-EEND but instead makes comparison to SotA of a hybrid model, DiariZen. This model, introduced by Han et al. in 2024, utilizes the Pyannote [5] framework for EEND and clustering.

4.2.2 On Confidence Calibration

During the 2010s, advancements in deep learning drastically improved neural network architecture [27]. Model capacity increased at a dramatic pace. Batch Normalization improved training time. These improvements were to the benefit of model accuracy, with empirical evidence supporting the notion that performance scales with data and model size [19]. The advancements in neural networks would also serve to make models more miscalibrated, as Guo et al. would show in their 2017 paper, "On Calibration of Modern Neural Networks".

Despite the impact of this paper, with over 10k citations according to Google Scholar results (pulled on May 4th, 2026), confidence calibration is a scarcely mentioned topic within audio diarization papers. Within the large body of audio diarization work, only three papers touch on the question of confidence calibration in any meaningful way [44] [33] [9]. None of these papers mention temperature scaling. The implications of this mirror the same concerns outlined in the original Guo et al. paper as models for diarization have grown more advanced. While temperature scaling is a simple enough fix for neural networks in the face of successive papers in confidence calibration [22] to the point where this method may not be mentioned within said literature, it is worthwhile to investigate this claim and study the effects of this fix on models.

5 Temperature Scaling: Experiments and Results

5.1 Methodology

We take a deductive, quantitative, post-positivist approach to this section of research where confidence calibration is operationalized as an objective metric. However, we acknowledge the variation in different metrics associated with confidence calibration [26] [25].[1] For the purposes of this research, ECE is treated as the metric of interest and we evaluate our hypotheses with this metric as the main focus.

5.1.1 Research Objectives

The boom in reliance on AI-driven solutions necessitates not blindly placing trust in reported metrics, as such metrics may not fully capture model reliability or uncertainty [32]. Therefore, it is critical to analyze confidence calibration techniques and their effect on model accuracy and downstream task performance. Performing in accordance with this means carrying out a careful comparison of offline versus online audio diarization techniques as well. This thesis aims to analyze their performance under controlled scenarios and assess the effect of confidence calibration on these, as well as assess modified suitability for slotting into the RoSSDaV system.

Specifically, this thesis aims to:

- **Assess the efficacy of modifying SotA offline solution DiariZen to produce results similar to that of SotA online audio solutions.** DiariZen, an offline model, performs, on average, better in terms of DER and JER than Streaming Sortformer (see Figure 5.1). We

[1] The irony of the statistical epistemological recursion at play here is not lost on the author.

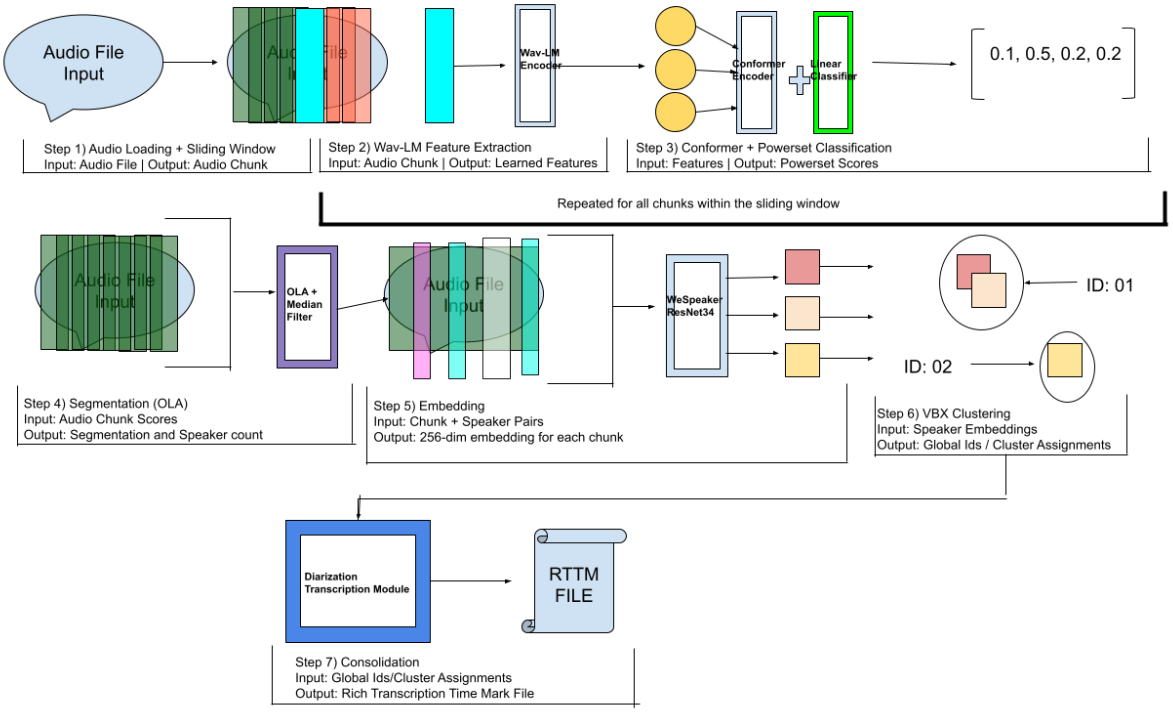


Figure 5.1: Overview of DiariZen audio diarization pipeline. Figure adapted from Raghav 2026. [46]

hope to modify DiariZen to extend the core architecture to a streaming audio diarization model in order to replicate the success of this model to the online audio domain.

- **Compare effects of temperature scaling on SoTA offline and online models.** We utilize temperature scaling to compare effects of temperature scaling on DER/JER and ECE metrics for both the offline DiariZen and the online Streaming Sortformer. A well-calibrated model is paramount for our downstream of speaker verification, and so we care about our ECE. As the two papers utilize different datasets as appropriate for their different use cases, we ran Streaming Sortformer on the same datasets utilized in the DiariZen paper.

5.1.2 On Models

This thesis makes use of two audio diarization models for analyzing the results of temperature scaling. These models are DiariZen and Streaming Sortformer respectively.

We refer to DiariZen as a model in this thesis as the DiariZen-Large-s80 is consistently utilized. However, it is more accurate, in truth, to describe DiariZen as a whole as an SD pipeline rather than an SD model, as the DiariZen system, as pictured in 5.1, is a hybrid diarization pipeline made up of multiple stages, but includes a pre-trained neural component. The pre-trained neural component, DiariZen-Large-s80, is what blocks 2 and 3 within Figure 5.1 consists of.

Streaming Sortformer, introduced by Nvidia in 2025, is specifically designed to handle the challenges of processing streaming audio, with a greater focus on latency within audio rather than DER and JER. They employ an Arrival-Order Speaker Cache (AOSC), rather than the standard speaker-tracing buffer, to store frame-level embeddings of previously observed speakers. This means that embeddings can be processed by arrival time order and dynamically updated as audio progresses, rather than the offline approach of DiariZen processing all chunks of a finished audio file. This former approach is optimized for latency through continuous learning.

5.1.3 Procedures

To attempt to answer whether or not temperature scaling has direct results on speaker diarization models, we must formally define what we mean by direct results. Our metrics of interest are ECE and DER/JER, for different reasons. ECE measures how well-calibrated a model is, whereas DER/JER is the reported diarization error rate. ECE, within the larger RoSSDaV system, matters more on the level of confidence calibration. A well-calibrated system reduces the need for unnecessary computation through deferring to visual and kinetic stages unnecessarily.

In the spirit of Guo et al.'s original study, analysis is performed on two SotA models, which are offline and online respectively. The DiariZen system was selected as our framework due to

the combination of recency, robustness, and previously achieved SotA results with the included neural model (henceforth referred to as the "DiariZen model" for clarity's sake).

To properly analyze the effects of temperature scaling on offline and online speaker diarization models, a fork of the DiariZen system was created with the goal of analyzing the datasets the DiariZen paper includes as benchmarks– AliMeeting, the AMI Meeting Corpus, and AISHELL-4. DiariZen was selected as our comparison pipeline due to the combination of recency, robustness, and convenience of the included model achieving SotA results.

Due to the differences in architecture between streaming and offline models, our temperature scaling works differently in what exactly we are scaling.

As seen in Figure 5.1, the DiariZen model utilizes powerset classification at Step 3. In the modified DiariZen model, uncertainty tap was included before Step 3 in order to obtain the raw logit vector before softmax is applied, whence temperature scaling could thus be applied before DiariZen's to_multilabel() call.

Computing ECE for the DiariZen model follows a similar pattern– the system created takes advantage of the fact that the raw pre-softmax logits are temperature-independent. We cache these, then sweep temperatures in-memory. For each temperature T and each session, we compute:

$$\text{P(SpeechAtFrameT)} = 1 - softmax(logits[t]/T)[class\_0]$$

where class 0 in the powerset space is the empty set (no speakers active = silence).

Ground truth is derived from the pyannote.database reference annotation: GT(frame t) = 1 if any speaker is active at time t, else 0.

ECE is then computed by binning P(speech) and comparing mean confidence to mean accuracy per bin:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|\text{bin}_b|}{N} \, |\text{mean\_conf}_b - \text{mean\_acc}_b|$$

Streaming Sortformer was similarly evaluated in terms of DER/JER versus ECE. Streaming Sortformer out-of-the-box includes four scenarios corresponding to balance within the latency-accuracy tradeoff– very high accuracy, high accuracy, low latency, and ultra low latency. These correspond to the lengths of chunks taken in by the Streaming Sortformer. As Streaming Sortformer operates on streaming parameters are specified in 80ms frames, latency can be given by $latency = chunkLength * 0.08s$.

However, DiariZen outputs powerset logits — a single multinomial distribution over all possible speaker-subset combinations. With 4 speakers there are 11 classes (silence + 4 singles + 6 pairs). The model jointly reasons about who is speaking in combination. Class 0 is the silence state.

Our online solution differs in the output. The original Sortformer [42], whose architecture the Streaming Sortformer similarly leverages, outputs per-speaker sigmoid probabilities. There are S independent binary scores, one per speaker channel, with no constraint that they sum to 1 and no shared reasoning between channels.

With DiariZen, we are tapping into the uncertainty point before softmax is applied, whereas with Streaming Sortformer, we scale the per-speaker binary logit through inverting the sigmoid function, dividing accordingly with $logit(p)/T$ before reapplying the sigmoid.

All experiments were run on a GeForce RTX 3070 within Binghamton University's AIR Lab.

5.2 Results and Discussion

We find that in the case of AISHELL-4 and ALI, the DiariZen model's ECE has a minimum at T=1 (no change in temperature). There is marked improvement in terms of ECE for AMI, but not the other two datasets utilized within these experiments. The offline results highlight the

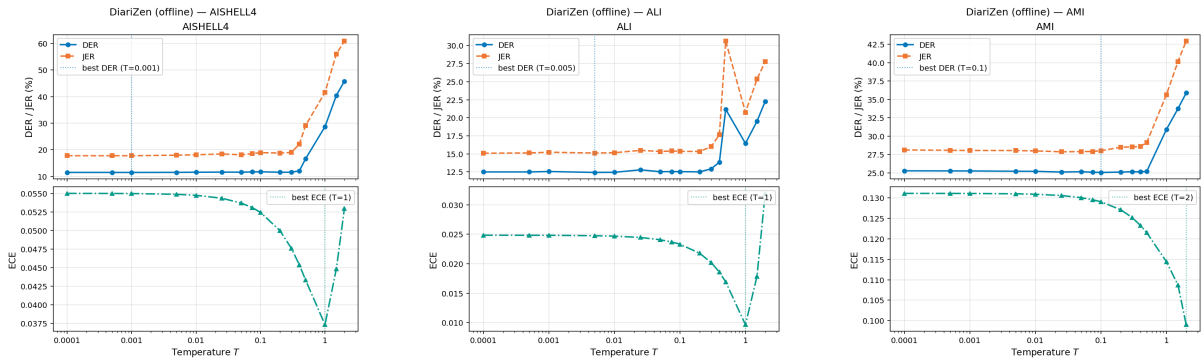
Table 5.1: Speaker diarization results on AMI, AISHELL4, and ALI test sets. DiariZen is evaluated offline; Sortformer is evaluated in streaming mode under four latency scenarios. DER and JER use a 0 s collar. ECE measures confidence calibration of $P(\text{any speech})$ at $T = 1$ with no scaling and after optimal temperature scaling (T^*). Lower is better for all metrics (↓).

System	Dataset	Latency (s)↓	DER (%)↓	JER (%)↓	$\text{ECE}_{T=1}$↓	ECE_{T^*}↓	T^*
DiariZen (offline)	AMI	—	25.06	28.01	0.1144	0.0990	2.0000
	AISHELL4	—	11.51	17.79	0.0373	0.0373	1.0000
	ALI	—	12.43	15.12	0.0097	0.0097	1.0000
Very High Accuracy	AMI	27.20	25.90	29.13	0.1433	0.1284	0.5000
	AISHELL4	27.20	25.77	48.98	0.0763	0.0528	0.4000
	ALI	27.20	13.12	14.99	0.0420	0.0184	0.5000
High Accuracy	AMI	9.92	27.16	31.35	0.1498	0.1312	0.5000
	AISHELL4	9.92	25.41	49.29	0.0821	0.0581	0.4000
	ALI	9.92	13.38	15.20	0.0460	0.0216	0.5000
Low Latency	AMI	0.48	29.62	34.71	0.1633	0.1431	0.4000
	AISHELL4	0.48	27.56	50.74	0.0960	0.0698	0.4000
	ALI	0.48	14.94	17.52	0.0530	0.0273	0.4000
Ultra Low Latency	AMI	0.24	29.85	34.12	0.1619	0.1443	0.5000
	AISHELL4	0.24	28.09	50.83	0.1099	0.0796	0.4000
	ALI	0.24	17.79	20.32	0.0585	0.0361	0.5000

inverse relationship between ECE and DER/JER metrics, particularly when it comes to the best temperature scaling value for each.

The reason for this can be found within further analysis of the datasets themselves. AISHELL-4 describes itself as "a sizable real-recorded Mandarin speech dataset collected by 8-channel circular microphone array for speech processing in conference scenarios" [14]. AliMeeting is a similar Mandarin dataset. AMI is an anomaly, but not simply because it is the only English dataset of the three.

AMI's own official documentation describes the structure of the dataset as "unusual" [8], with two-thirds of the audio data taking place through scripted scenarios, scenarios that trend towards uniform turn-taking rather than the natural meetings of AISHELL-4, AliMeeting, and the remaining one-third of AMI.



(a) Offline AISHELL-4 (b) Offline ALI (c) Offline AMI

Figure 5.2: Offline results.

A diarization model evaluated on AMI is mostly being tested on an orderly single-speaker sequence with intermittent hotspots rather than, not on the ebb and flow and overlap defining natural meetings. The interplay of ECE and DER/JER here comes from the polarizing speech types– ECE measures calibration on a global scale. When DiariZen achieves its DER-optimal performance at T=0.1 on AMI, it's exploiting that orderliness. Sharp posteriors work well because the frame classification problem is easy for most of the signal and only hard for concentrated bursts.

This also has the reverse effect. Sharpening at T=0.1 means ambiguous frames get their uncertainty effectively suppressed, leading to overconfidence within the reliability diagram. When calculating ECE, the mean confidence in the high-confidence bins exceeds the actual accuracy. Softening (T=2.0) pulls those bins back toward honesty, at the cost of introducing unnecessary uncertainty on the easy frames and degrading segmentation quality. Unlike the offline DiariZen model, Nvidia's Streaming Sortformer shows a marked improvement in ECE, but does not benefit from the temperature scaling when it comes to DER. The reason for this can be found within output representations of Streaming Sortformer.

DiariZen outputs powerset logits — a single multinomial distribution over all possible speaker-subset combinations. With 4 speakers there are 11 classes (silence + 4 singles + 6 pairs). The model jointly reasons about who is speaking in combination. Class 0 is the silence state.

However, the original Sortformer [42], whose architecture the Streaming Sortformer similarly

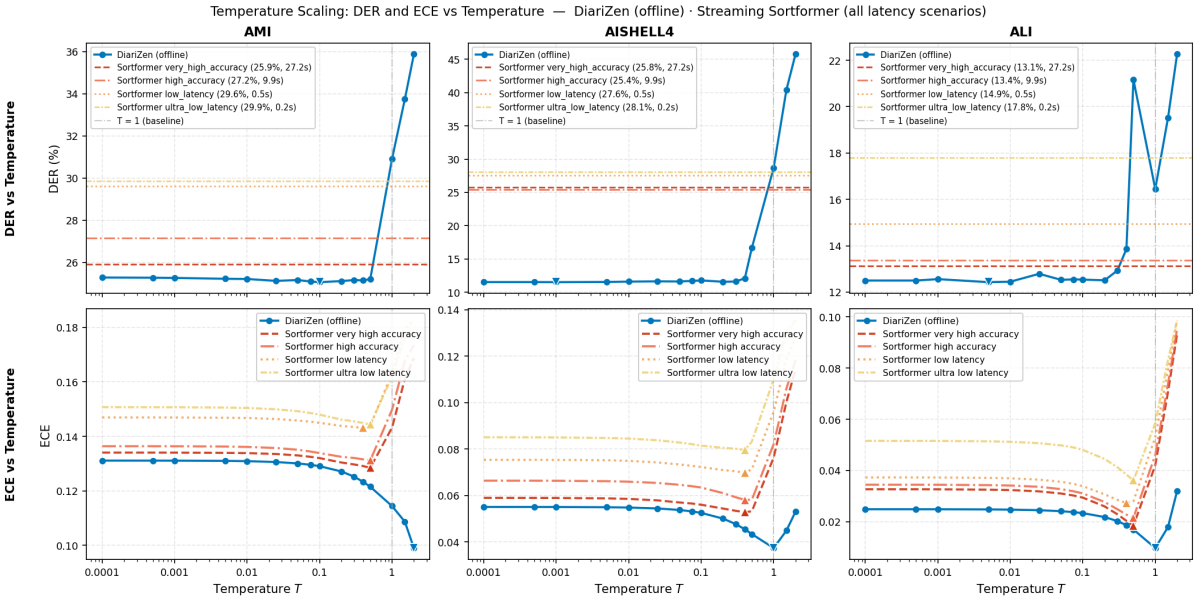


Figure 5.3: Effects of Temperature Scaling on Sortformer (all scenarios) and Offline DiariZen versus DER/JER and ECE.

leverages, outputs per-speaker sigmoid probabilities. There are S independent binary scores, one per speaker channel, with no constraint that they sum to 1 and no shared reasoning between channels.

With DiariZen, we are tapping into the uncertainty point before softmax is applied, whereas with Streaming Sortformer, we scale the per-speaker binary logit through inverting the sigmoid function, dividing accordingly with $logit(p)/T$ before reapplying the sigmoid.

This means DER does not change with temperature and the near-binary probabilities cross 0.5 in the same places regardless of T, as highlighted by the Streaming Sortfomer results. However, ECE does not care about DER/JER, but rather, if the accuracy does not line up with the confidence the model has in its predictions.

5.3 Conclusion

In conclusion, this thesis

6 DiariChan

Given Streaming Sortformer's metrics versus DiariZen's powerset classification under softmax, it is a natural desire to pursue a system that can have the best of both worlds in terms of the two systems' unique qualities.

Temperature scaling on a single, scalar logit like Sortformer provides is strictly monotone. It cannot move any output across the 0.5 decision boundary that is used to compute DER. Thus, DER is unchanged by T for all four scenarios of Sortformer, despite the influence of temperature scaling on ECE, whereas there is a tradeoff within DiariZen between ECE and DER/JER metrics.

However, there remains the fact that DiariZen is inherently offline and thus, can't be leveraged, without modification, for online streaming audio and confidence detection. Such a system would not have the luxury that the original offline system has in terms of one-shot global clustering via VBx.

In order to take a step further to creating a system that works well within the RoSSDaV pipeline, we have attempted this through the creation of DiariChan, a play on the original DiariZen model and the Japanese suffix "-chan", typically denoting something small or cute but also playing with the word "chunk", which is key to the online system.

We borrow several components from the original DiariZen system, including the offline system's same segmentation and embedding extractor. We diverge from the original system in several aspects, however. We notably replace the VBx Clustering step with an incremental, uncertainty-gated tracking loop. Processing happens on a per-chunk-basis in near real time, replacing the offline version's first step of loading in an entire WAV file. We take the continuous stream and also apply a striding window to this, where stride in this case refers to the window's advance between calls to chunk the audio (visualized in Figure 6.1).

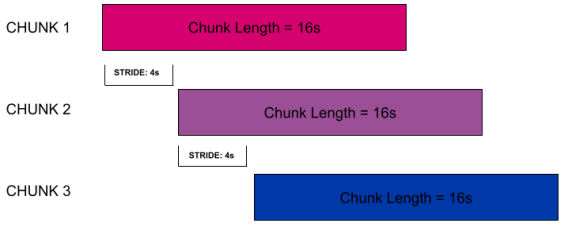


Figure 6.1: AudioStream Step of DiariChan

This chunk undergoes similar processing to that within DiariZen's original steps 2 and 3. However, similar to the temperature scaling experiments, the raw logit vector is intercepted via an uncertainty tap, outlined in a following sub-section. Within this uncertainty tap, soft and hard posteriors are derived, as well as entropy and chunk confidence. This latter value is what drives later mechanisms within our confidence pipeline.

6.1 Uncertainty Tap

Like in the case of temperature scaling, it does the uncertainty pipeline no good to simply receive probabilities after softmax like in the original Diarizen. We extract the raw logit vector following the feature extraction, conformer, and linear classification steps.

Our raw logit vector then receives a Shannon entropy calculation following our own softmax in order to give us a categorical distribution that can then be analyzed. For example, if our powerset distribution indicates we have 99% probability, our categorical distribution is peaked for that class and as such, we don't have a wide probabilistic spread. Shannon entropy is given by:

$$H = -\sum_{i}^{M} P_i \log P_i$$

where:

- H is our Shannon entropy.

- i is a given powerset class (ex. if there are 4 speakers (A, B, C, and D) and 2 speaking at any given time, a class could be A, CD, etc.)
- M is the total number of our powerset classes.

Our log base is the natural base, as we calculate via the default base of torch.log().

Because we're currently only processing this current chunk, we care quite a bit about preserving the uncertainty associated with every chunk. Thus, it is important for us to calculate both our hard posteriors (our binary multilabelling) and soft posteriors (weighted probabilities).

From there, we can assign a value to chunk confidence after filtering out silence frames. First, soft posteriors are calculated. The per-frame maximum soft posterior is given by:

$$m_t = \max_{s \in \{1,\ldots,S\}} P_n[t,s]$$

where $\tilde{p}_{t,s}$ is the soft posterior for speaker s at frame t.

From this, we can calculate our active mask, which is:

$$\mathscr{A} = t : m_t > \tau_{\text{act}}$$

where $\tau_{\text{act}} = 0.1$ within the codebase. It should be noted that this threshold is arbitrarily chosen as an indicator of speech activity with a low bar. As such, this value is open to experimentation. This active mask is simply the silence frames being suppressed.

We can then derive chunk confidence, $\hat{c}$, which is our mean m_t over our active frames. This is mathematically expressed as:

$$\hat{c} = \frac{1}{|\mathscr{A}|} \sum_{t \in \mathscr{A}} m_t$$

Following this Uncertainty Tap, we use our soft posteriors we've obtained as frame-level weight masks so we compute a posterior-weighted average of acoustic features. Frames where a speaker is more confidently active contribute more to their embedding.

On a mathematical level, $\mathrm{p}_{n,s} = P_n[:, s] \in \mathbb{R}^F$

6.2 Speaking Tracing Buffer

Following Step 3 in offline DiariZen, we continue with OLA. However, we diverge within DiariChan. Because we do not have full context at any given point like OLA inherently has within an offline model, we cannot use this component. The purpose of OLA within the original DiariZen is to average all predictions covering each output frame within a full recording to estimate number of speakers within a given frame. However, since we process on a chunk-by-chunk basis in a streaming capacity, we cannot look into future chunks like OLA can.

In lieu of OLA, we aim to proverbially kill two birds with one stone by adopting a similar approach to that of Xue et al.'s 2021 work, "Online Streaming End-to-End Neural Diarization" [54]. This work makes use of a speaker tracing buffer (STB) in order to resolve inconsistency between chunk speaker labeling. What is meant by inconsistency is that there is an inherent problem with tracing speaker identities between chunks when the number of speakers is unknown. The current segmentation model assigns speaker labels arbitrarily per chunk. A speaker deemed "Speaker 1" by the segmentation model may hold a completely separate identity in chunk n to that of "Speaker 1" in chunk n + 1. Thus, we must have a way of resolving this problem, called *speaker permutation ambiguity*.

Xue et al.'s speaking tracing buffer lives within posterior space and possesses two matrices: acoustic features and the corresponding posteriors from EEND. Their STB is trained through performing an extended search of candidates through permutation. This is then prepended to the input sequence for an EEND's transformer to operate over along with the current acoustic features at hand in a chunk. However, this has a caveat that we aim to amend. The STB has no

notion of when to trust its own buffer, meaning errors can accumulate from the model's own outputs. Early diarization errors can propagate and compound across chunks, falling into a common failure mode for autoregressive models [47].

We diverge by performing our own speaking tracing buffer post-hoc. While Xue et al. resolves permutations implicitly through the EEND's attention mechanism, we differ by resolving permutations algorithmically via the Kuhn-Munkres algorithm, which we refer to in our codebase by its alternate name, the Hungarian algorithm. This was chosen as it is well-known to be a solution for the assignment problem, operating in $O(n^3)$ time. We also maintain long-term identity via our separate Embedding Bank. We perform this algorithm on running-mean centroids from committed chunks, returning a permutation array and a new speaker mask.

Conclusions

Conclude a lot here.

Bibliography

[1] The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan. Technical report, NIST, February 2003.

[2] Xavier Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370, February 2012.

[3] Barry Arons. A Review of The Cocktail Party Effect. Technical report, MIT Media Lab, 1992.

[4] Phil Blunsom. Hidden Markov Models, August 2004.

[5] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. pyannote.audio: neural building blocks for speaker diarization, 2019. Version Number: 1.

[6] Glenn W. Brier. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*, 78(1):1–3, January 1950.

[7] D. E. Broadbent. *Perception and communication.* Pergamon Press, Elmsford, 1958.

[8] Jean Carletta. Announcing the AMI Meeting Corpus. *The ELRA Newsletter*, 11(1):3–5, January 2006.

[9] Anurag Chowdhury, Abhinav Misra, Mark C. Fuhs, and Monika Woszczyna. Investigating Confidence Estimation Measures for Speaker Diarization, 2024. Version Number: 1.

[10] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6184–6188, Barcelona, Spain, May 2020. IEEE.

[11] Juan Manuel Coria, Hervé Bredin, Sahar Ghannay, Sophie Rosset, Khaled Zaouk, Ingo Fruend, Bertrand Higy, Amit Kesari, and Yagna Thakkar. Diart: A Python Library for Real-Time SpeakerDiarization. *Journal of Open Source Software*, 9(99):5266, July 2024.

[12] Timothée Dhaussy, Bassam Jabaian, Fabrice Lefèvre, and Radu Horaud. Audio-Visual Speaker Diarization in the Framework of Multi-User Human-Robot Interaction. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes Island, Greece, June 2023. IEEE.

[13] Jonathan G. Fiscus, Jerome Ajot, Martial Michel, and John S. Garofolo. The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Steve Renals, Samy Bengio, and Jonathan G. Fiscus, editors, *Machine Learning for Multimodal Interaction*, volume 4299, pages 309–322. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. Series Title: Lecture Notes in Computer Science.

[14] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario, 2021. Version Number: 4.

[15] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe. End-to-End Neural Speaker Diarization with Self-attention, 2019. Version Number: 1.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks, 2017. Version Number: 2.

[17] Eunjung Han, Chul Lee, and Andreas Stolcke. BW-EDA-EEND: streaming END-TO-END Neural Speaker Diarization for a Variable Number of Speakers. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7193–7197, Toronto, ON, Canada, June 2021. IEEE.

[18] Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukas Burget. Leveraging Self-Supervised Learning for Speaker Diarization, 2024. Version Number: 3.

[19] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically, 2017. Version Number: 1.

[20] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[21] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu. End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors, 2020. Version Number: 3.

[22] Siguang Huang. Uncertainty Calibration).

[23] Paul Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. 1901. Medium: text/html,application/pdf,text/html.

[24] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne Van Waveren, and Joakim Gustafson. Behavioural Responses to Robot Conversational Failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 53–62, Cambridge United Kingdom, March 2020. ACM.

[25] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate Uncertainties for Deep Learning Using Calibrated Regression, 2018. Version Number: 1.

[26] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified Uncertainty Calibration, 2019. Version Number: 2.

[27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[28] Jonggeun Lee, Junseong Pyo, Gyuhyeon Seo, and Yohan Jo. SpeakerSleuth: Can Large Audio-Language Models Judge Speaker Consistency across Multi-turn Dialogues?, 2026. Version Number: 2.

[29] Di Liang and Xiaofei Li. LS-EEND: Long-Form Streaming End-to-End Neural Diarization with Online Attractor Extraction, 2024. Version Number: 2.

[30] Di Liang, Nian Shao, and Xiaofei Li. Frame-wise streaming end-to-end speaker diarization with non-autoregressive self-attention-based attractors, 2023. Version Number: 1.

[31] Yu-Xiang Lin, Chih-Kai Yang, Wei-Chih Chen, Chen-An Li, Chien-yu Huang, Xuanjun Chen, and Hung-yi Lee. A Preliminary Exploration with GPT-4o Voice Mode, 2025. Version Number: 1.

[32] Zachary C. Lipton. The Mythos of Model Interpretability, 2016. Version Number: 3.

[33] Simon W. McKnight, Aidan O. T. Hogg, Vincent W. Neo, and Patrick A. Naylor. Uncertainty Quantification in Machine Learning for Joint Speaker Diarization and Identification, 2023. Version Number: 2.

[34] Kyle Min. STHG: Spatial-Temporal Heterogeneous Graph Learning for Advanced Audio-Visual Diarization, October 2023. arXiv:2306.10608 [cs].

[35] Allan H. Murphy. The Early History of Probability Forecasts: Some Extensions and Clarifications. *Weather and Forecasting*, 13(1):5–15, March 1998.

[36] Frank E. Musiek and Gail D. Chermak, editors. *Handbook of central auditory processing disorder. Volume 1. Auditory neuroscience and diagnosis*. Plural Publishing, San Diego, California, second edition edition, 2014.

[37] National Research Council. *Funding a Revolution: Government Support for Computing Research*. The National Academies Press, Washington, DC, 1999.

[38] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pages 625–632, New York, NY, USA, 2005. Association for Computing Machinery.

[39] Douglas O'Shaughnessy. Speaker Diarization: A Review of Objectives and Methods. *Applied Sciences*, 15(4):2002, February 2025.

[40] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), February 2015.

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

[42] Taejin Park, Ivan Medennikov, Kunal Dhawan, Weiqing Wang, He Huang, Nithin Rao Koluguri, Krishna C. Puvvada, Jagadeesh Balam, and Boris Ginsburg. Sortformer: A Novel Approach for Permutation-Resolved Speaker Supervision in Speech-to-Text Systems, July 2025. arXiv:2409.06656 [eess].

[43] Rolf Pfeifer and Christian Scheier. *Understanding intelligence*. MIT Press, Cambridge, Mass, 1999.

[44] Alexis Plaquet and Hervé Bredin. On the calibration of powerset speaker diarization models. 2024. Version Number: 1.

[45] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[46] Nikhil Raghav. DiariZen Explained: A Tutorial for the Open Source State-of-the-Art Speaker Diarization Pipeline, April 2026. arXiv:2604.21507 [eess].

[47] Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning, 2010. Version Number: 3.

[48] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. The Second DIHARD Diarization Challenge: Dataset, task, and baselines, June 2019. arXiv:1906.07839 [eess].

[49] Laurent El Shafey, Hagen Soltau, and Izhak Shafran. Joint Speech Recognition and Speaker Diarization via Sequence Transduction, July 2019. arXiv:1907.05337 [cs].

[50] Mark Sinclair and Simon King. Where are the challenges in speaker diarization? In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7741–7745, Vancouver, BC, Canada, May 2013. IEEE.

[51] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1557–1565, September 2006.

[52] Anne. M. Treisman. SELECTIVE ATTENTION IN MAN. *British Medical Bulletin*, 20(1):12–16, January 1964.

[53] Joseph Xu Zhou, Xiaojie Qiu, Aymeric Fouquier d'Herouel, and Sui Huang. Discrete Gene Network Models for Understanding Multicellularity and Cell Reprogramming: From Network Structure to Attractor Landscapes Landscape. In *Computational Systems Biology*, pages 241–276. Elsevier, 2014.

[54] Yawen Xue, Shota Horiguchi, Yusuke Fujita, Yuki Takashima, Shinji Watanabe, Paola Garcia, and Kenji Nagamatsu. Online Streaming End-to-End Neural Diarization Handling Overlapping Speech and Flexible Numbers of Speakers, 2021. Version Number: 2.

[55] Jinzheng Zhao, Yong Xu, Xinyuan Qian, and Wenwu Wang. Audio Visual Speaker Localization from EgoCentric Views, September 2023. arXiv:2309.16308 [cs].