Research Master Thesis

Investigation of Scalable Audio-based Speaker Identification in the context of Communicative Robots

Szabolcs Pál

Supervisor	Piek Vossen, Luis Morgado da Costa
2nd reader	Antske Fokkens

a thesis submitted in fulfillment of the requirements for the degree of

MA Linguistics

(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Linguistics and Text-Mining Lab
Department of Language and Communication
Faculty of Humanities

Date	August 15, 2025
Student number: 2829786	
Word count: 25,565	
Student number	Student number
Word count	Word count

Abstract

This project investigates the application of scalable audio-based speaker identification within the context of a communicative robot. While existing literature primarily evaluates speaker identification models on standard benchmarks, real-world deployment imposes unique constraints that remain underexplored. To address this gap, three representational approaches were compared: (1) explicitly handcrafted speech features, (2) raw audio signals, and (3) dense embeddings extracted from transformer-based models. These strategies differ in their level of feature abstraction and potential scalability, yet their relative performance in this application context is unclear.

The study addressed two main research questions: firstly, how each representation scales in closed-set classification while maintaining performance, and secondly, their capability to perform open-set classification with conservative abstention. Two architectural paradigms, namely ensemble models and transfer learning methods, were applied within a grid-based evaluation framework, measuring precision in identifying the correct speaker across varied enrollment sets.

Results indicate that models relying on higher feature abstraction, such as handcrafted features and dense transformer embeddings, struggled to generalize under increased classification complexity, producing inseparable speaker clusters. In contrast, models trained directly on raw audio signals generated robust speaker embeddings and sustained high performance, even in heavily saturated classification scenarios.

Overall, findings suggest that minimizing input feature abstraction enhances adaptability and flexibility, enabling the model to maintain discriminative capacity as the number of enrolled speakers and variability in the data increase. These insights highlight the importance of representation choice for scalable speaker identification in dynamic, real-world applications such as communicative robotics.

Declaration of Authorship

I, author, declare that this thesis, titled *Investigation of Scalable Audio-based Speaker Identification in the context of Communicative Robots* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 15.08.2025

Signed:

Acknowledgments

I would like to express my gratitude to my supervisors, Piek Vossen and Luis Morgado da Costa, for their patience, encouragement, and invaluable help throughout this project. I am also thankful to my second reader, Antske Fokkens, for her feedback and involvement with the project. I would like to thank my family for their continuous support during the development of this research, and my friends for being a source of motivation and perspective when I needed it most. Finally, I am grateful to the CLTL research group for their guidance and expertise, which have been instrumental in shaping the direction and execution of this work.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Identifying speakers plays a fundamental role in human communication by allowing humans to refer to shared knowledge among the mutually recognised participants. Similarly, in communicative robot systems, the modelling of speaker identity allows the robot to map conversational turns to speakers, thus enabling a deeper understanding of the uttered information and the speaker's perspective. This is also the case for Leolani, which is a communicative robot (Vossen et al. (2019)) developed by the CLTL research group at VU Amsterdam. Leolani is a dialogue system which operates from multi- modal data input, and is designed to learn incrementally from conversation with its users. During conversation, this communicative robot maps utterances to their speakers, and structures the preceived information into a knowledge graph (Santamaría et al. (2022)), which is later used as the basis of output generation and argumentation. The stored perspective of the information allows Leolani to refer to previous shared knowledge with its users, and to conduct personalized conversation.

While the Leolani system includes face-based visual speaker identification, it currently lacks the ability to recognise speakers by voice. The application of audio-based speaker identification could increase the efficiency of the Leolani system in cases where the face is not recognisable or not in the view of the robot. This research project proposes a number of audio based speaker identification models to eliminate the current limitations, and compares these identification modules in their suitability to be applied within the context of a communicative robot.

1.1 Speaker Identification

Speaker identification is the task of determining who is speaking based on their unique vocal characteristics (Wang et al. (2024)). Modern audio-based speaker identification systems typically work in two stages. First, a speaker embedding model transforms audio samples into fixed-length vectors, called speaker embeddings, that capture distinctive features of a person's voice. The model learns to place embeddings from the same speaker close together, while separating those from different speakers.

In the second stage, known as enrollment, the system stores embeddings of known speakers to build a reference space. When a new audio sample is received, its embedding is compared to these stored references. The system then assigns the identity of the speaker whose embedding is most similar to the new input (Wang et al. (2024)).

1.1.1 Modeling approaches used in speaker identification

Speaker identification is a well studied task within the field of speech technology due to its numerous application contexts. The main research focus within this task lies within the speaker characterisation process, as the generation of robust and discriminative acoustic representations is essential for the successful execution of this task.

The development of effective speaker embedding models, which are used in the speaker characterization process, involves two fundamental design factors, namely the data representational strategy and computational architecture. Numerous approaches have been introduced within the literature to improve these two design parameters. Within the data representational approaches, strategies such as specific, spectral and generalised data representations have been deployed within the task to improve robustness and general performance on widely used benchmark datasets. Numerous computational approaches have also been introduced, some of which include temporal modelling of energy distribution (Snyder et al. (2018)), the convolutional computation of raw audio segments (Jung et al. (2019)), or the deployment of transformers based architecture used with an additional fine-tuned or transfer learning classification head (Hsu et al. (2021)). While these approaches have shown significantly elevated results within the task of speaker identification, their performance has mostly been demonstrated on traditional benchmark datasets. However, research into how each of these approaches benefits the modeling of acoustic speaker characteristics, and how appropriate they are in specific application contexts has not been extensively researched within the field.

1.2 Application constraints

Deploying speaker identification within a communicative robot introduces a distinct set of challenges that must be addressed. These constraints can be grouped into two main challenges; maintaining performance while deployed in scaled classification setups, and performing open classification.

The first challenge regarding the maintained performance within scaled up classification setups is characterised by two fundamental design constraints. Firstly, a communicative robot must be capable of identifying speakers independently of the content of their speech. This property is crucial in natural conversation where speaker utterances are spontaneous and highly variable in linguistic content. Secondly, the system must be robust to variations in the manner of articulation. A speaker's vocal characteristics may show high variation in mood, emotional state, fatigue, health conditions (e.g., a cold), or changes in speaking rate and pitch. These intra-speaker variations emerge naturally over time, especially in long-term human–robot interaction scenarios.

Moreover, the speaker identification system must operate as an open classification model. A communicative robot like Leolani is designed to engage with a gradually expanding set of users. It must therefore be able to recognize when a speaker is previously unseen, rather than carrying out classification in a fixed set of familiarised speakers. Failure to identify unseen speakers could result in incorrect mapping of perspectives within the robot's internal knowledge graph, leading to degraded performance in dialogue understanding and personalized interaction.

These constraints are further complicated by the heterogeneity and longitudinal nature of conversational audio data. As the robot interacts with users over time, the acoustic data it collects becomes increasingly diverse, which happens not only across speakers but within individual speaker representations as well. This evolving representation of speaker identity makes it difficult to define prototypical voice characteristics, which complicates the final classification stage of the task. Consequently, speaker identification in communicative robotics requires models that are not only accurate but also adaptive and robust to variability over time.

Previous research on communicative robots has explored how these systems can recognize and re-identify speakers during interactions, even when natural variations in a person's voice occur due to factors such as emotion, mood, or speaking style. Studies have demonstrated that it is possible to achieve robust speaker recognition in dynamic, real-world environments, showing promising results in maintaining reliable identification across varied speech inputs (Foggia et al. (2023)). Although previous research has addressed challenges such as intra-speaker variability and limited enrollment data, it has not fully investigated the long-term scalability of speaker identification in communicative robots. Key open questions remain about how performance is affected as the number of enrolled speakers increases over time, and how growing overlaps between speaker characteristics might impact reliable recognition in more complex, evolving interaction settings.

1.2.1 Representational approaches

Moreover, contemporary approaches have demonstrated a wide range of computational and data-driven decisions in their pursuit of achieving elevated results in the task of speaker identification. However, prioritising one approach to another within an application context such as a communicative robot is a challenging task, as the constraints followed by these studies are not identical to the current context. It is especially important to make a well-founded decision about the type of data that is used within speaker identification, as the upscaled ever expanding enrolment set can emerge unwanted patterns in complex classification settings. When more data is introduced to the model, patterns in the enlarged representations might appear that might remain unseen in controlled fixed size classification settings. Thus, the application of the appropriate representational data type is essential, which minimises these gradually appearing limitations.

One contrast between different data representational approaches lies in the manner indicative information is encoded within the utilised speaker embeddings. This contrast can be envisioned as a spectrum. At its center lies the raw signal, simply converted into numerical data without any assumptions about what might be relevant. From this neutral starting point, one direction takes us toward specified features, which are grounded in expert knowledge and highlight specific acoustic characteristics of the signal, making them clear and interpretable. In the opposite direction, lie the generalized representations, where transformer models learn a broad array of implicit patterns from large datasets, thus encoding complex information without explicitly defined features.

At one end of this spectrum are specified explicit representations, which explicitly encode predefined articulatory or acoustic features such as vowel formants, pitch contours, nasality, or speech rate. These representations aim to model speaker identity by isolating stable, interpretable components of speech articulation, which are founded in phonetic and physiological research. As such, these representations are well-suited for high-precision environments with controlled data conditions (Meghanani and Ramakrishnan (2018), Nasr et al. (2018)). These conditions might include biometric verification systems in banking or secure access control. In these domains, the modelling of stable vocal characteristics, and performance with high precision are prioritized over generalization.

In contrast, generalized representations are created through learned embeddings produced by deep models trained on large-scale data. These models are not explicitly trained to encode specific vocal characteristics but instead, they learn useful short and long dependency acoustic patterns through iterative unsupervised prediction objectives. One widely used objective entails the masked unit prediction task, where the model is trained to predict pseudo-phonemic units based on their surrounding context (Hsu et al. (2021)). This approach closely resembles the BERT-style architecture, as such predicted pseudo-phonemic units can be viewed as less interpretable audio based subtokens. The generated embeddings of these transformers based models often combine speaker identity with other information, such as linguistic content, background conditions, or speaker demographics (e.g., gender or accent) (Stan (2022)). Because of their large set of encoded information, generalized representations are often more robust to noisy or unpredictable environments. For example, in forensic or open-domain scenarios, where audio quality and speaking conditions vary widely, generalized embeddings offer greater flexibility and adaptability, even if they are less interpretable.

Between the two ends of the spectrum lies a third approach, where systems operate without a clear preference for either strategy. These systems typically rely on low-level signal features (e.g., MFCCs or spectrograms) and do not explicitly represent or separate speaker-specific cues from other acoustic content. Instead, the model learns the patterns that are most predictive based on the available dataset, by sourcing features directly from the signal, and not from already abstracted or pre-trained representations. While this method can yield high performance in practice, it comes at the cost of interpretability, as it is not always clear which features are responsible for a classification decision.

The choice between these encoding strategies should be guided by the demands of the application context. For instance, in communicative robots, the speaker identification system must not only function in heterogeneous and dynamic environments, but also support incremental learning, and open-set classification. These requirements introduce a tension: while generalized embeddings may offer a greater but largely varied set of encoded information for speaker identity modelling, specified representations could provide a smaller set of stable acoustic cues that appear within a diverse range of data input. Moreover, a third possibility might be that an explicit prioritisation of one data type is not necessary, as the model architecture is capable of identifying scalable speaker specific patterns from low-level audio features. Although all mentioned approaches show potential to be beneficial for the identification system in upscaled classification settings, few studies have explored the effect of these representational strategies in longitudinal, open-set contexts. This leaves an important gap in understanding which encoding strategy is better suited for scalable, dynamic systems like communicative robots.

Previous research has explored how different data representation strategies influence the performance of speaker identification models, comparing approaches based on manually extracted phonetic features, spectral inputs, and generalized embeddings learned from deep neural networks. While findings indicate that generalized embeddings often perform well under noisy and variable conditions (Brydinskyi et al. (2024)) they lack interpretability and their adaptability to long-term applications remains uncertain. Other studies suggest that phonetic features may excel in structured settings, whereas spectral approaches offer more flexibility in dynamic environments (Gendrot et al. (2019)). However, results vary significantly across datasets and evaluation setups, leaving open questions about which representational strategy is most effective for practical, scalable use in communicative robots, especially under open-set and continuously evolving conditions.

1.2.2 Open-Set and Thresholded Speaker Classification

In speaker identification research, the ability to execute open classification is a crucial requirement, particularly in applications where systems must handle the presence of previously unseen speakers. Many studies have investigated this challenge, with the most common solution being the introduction of a thresholding mechanism that separates known and unknown speakers based on the model's confidence scores (Karadaghi et al. (2014), Chakraborty and Parekh (2017), Affek and Tatara (2022)). While this approach has achieved strong results across various studies, it can be too rigid for the application context of communicative robots. In these settings, the data representations available to the model are constantly evolving, and they do not always provide clear, indicative patterns for confident classification. Forcing the system to assign every input to either a known or unknown label risks introducing errors, potentially leading to incorrect mappings of speaker identity within the robot's knowledge base. In interactive, long-term deployments, such errors can compromise both performance and reliability. This highlights the need for a more conservative modeling approach that allows the system to abstain from making a prediction when the available information is inconclusive. However, this capability has not been widely explored in speaker identification research, leaving an important gap that must be addressed to ensure safe and scalable speaker identification for communicative robots.

1.3 Gap in the Literature

Despite significant advances in speaker identification, several critical gaps remain unaddressed in the context of communicative robotic systems. First, there has been little to no research on how evolving speaker identity representations affect the performance and adaptability of robots interacting with users over time. Unlike traditional identification systems, communicative robots require models that can incrementally update and refine their internal speaker profiles as new data is observed in dynamic environments. Second, the question of which type of input representation, for instance, specified (phonetic and phisiologically based), generalized (implicit transformer-based), or spectral (spectrogram and MFCC representaions) is most appropriate for this application has not been systematically explored. Existing literature often evaluates these feature types in isolation or under closed, static conditions, without assessing their suitability for applied settings.

Furthermore, while some progress has been made in open-set speaker identification, the specific demands of robotic interaction, particularly the ability to abstain from uncertain predictions, remain unexplored. In a communicative robot systems, the ability to abstain from making inconclusive predictions about a speaker's identity is essential for avoiding incorrect mapping of perspectives within the knowledge base of the system.

Together, these gaps highlight the need for a comprehensive investigation of representational strategies, evolving identity modeling, and conservative open classification mechanisms in the design of speaker identification systems for communicative robots. The following paragraphs introduce the research questions that are investigated within this study, along with hypotheses based on previous empirical evidence.

The following fundamental research question encompasses the identified gaps in the literature.

- How should speaker identity be represented and modeled in communicative robotic systems to support accurate, adaptable, and conservative speaker identification over time, across both open-set and closed-set classification tasks?

In order to investigate the primary research question, two subquestions are posed to investigate the modeling of speaker identity within a communicative robot.

The first research question evaluates how well these models adapt, and identify patterns in speaker identities when identifying known speakers, under varying enrollment conditions. Specifically, these conditions include the varied number of enrolled speakers and the amount of enrollment data per speaker. This experiment assumes idealised settings, such as exclusive interaction with previously known speakers, which rarely occur in real-world communicative robot applications.

Thus, the second proposed research question dives into how these three modelling strategies can distinguish not only between their known speakers, but also identify input that is dissimilar from all enrolled representations. This is achieved through a two-threshold decision system that introduces a third option, abstention, when the model lacks sufficient confidence. The thresholding mechanism categorises the confidence of each prediction into three possible outcomes: high confidence leads to a known speaker prediction, low confidence to an unknown speaker label, and when the confidence falls between these two values, it triggers abstention. This allows the system to avoid making predictions which are based on inconclusive information, ensuring that high precision is maintained across both known and unknown classification labels.

1.3.1 Research Questions:

- · RQ1: How do speaker identification models based on explicit phonetic features, spectral features, and transformer-extracted embeddings compare in their ability to identify previously enrolled (known) speakers under varying ratios of number of enrolled speakers and enrollment data per speaker?

- · RQ2: How do speaker identification models based on explicit phonetic features, spectral features, and transformer-extracted embeddings compare in identifying known (enrolled) and new (not enrolled) speakers, while maintaining an option of abstention across varied ratios of number of speakers and number of enrolled instances per speaker?

1.4 Hypotheses:

The following hypotheses address how the enrollment configuration variables affect model performance. These variables represent two opposing pressures on the speaker identification task. An increase in the number of enrolled speakers introduces greater classification complexity, as the model must distinguish between more speaker profiles. This is expected to negatively affect overall performance. Conversely, a higher number of enrollment instances per speaker provides the model with more robust representations, improving generalisation and confidence in matching. This should enhance the model's ability to accurately identify speakers, even as the number of enrolled individuals grows.

The opposing nature of these two variables are the basis of the following hypotheses:

- Number of enrolled speakers is hypothesised to:
  - Correlate negatively with model performance
- Number of enrollment instances per speaker is hypothesised to:
  - Correlate positively with model performance

Hypotheses for the modelling approaches:

The expected performance patterns of the three modeling strategies are grounded in their representational foundations and the flexibility of their features under changing speaker enrollment conditions.

Phonetic feature-based models rely on structured, manually defined acoustic properties such as pitch, formants, jitter, and shimmer. While these features are grounded

in phonetic theory and offer interpretable insights, they have primarily been shown to provide indicative value in highly controlled experimental settings. As a result, these models are expected to yield lower overall performance, but their dependence on explicitly defined features may lead to more stable behaviour as enrollment conditions scale.

Spectral feature-based models, by contrast, convert raw audio into representations such as MFCCs and spectrogram bands, which are widely used in speaker recognition due to their ability to capture rich acoustic cues. These features are inherently adaptable, allowing the model to prioritize different types of information as the number of enrolled speakers and enrollment instances increases. This flexibility is expected to yield the best performance across conditions, even as the enrollment set expands.

Transformer-based models, such as those using self-supervised speech representations, encode dense and general-purpose embeddings from large-scale pretraining. These embeddings contain speaker-relevant acoustic information but are not explicitly optimized for speaker identity. As a result, transformer models are anticipated to perform better than the phonetic feature-based models, especially in smaller configurations, but they may struggle to adapt as well as the spectral models in larger, more complex setups. This is due to the inbleeding of broad pretraining information, which may limit their ability to highlight speaker-specific variation under increasing classification complexity.

1.4.1 Hypotheses for First Research Question

- Phonetic feature-based model is hypothesised to:
  - Achieve lower overall performance than other models
  - Maintain more stable performance across varying enrollment conditions
- Spectral feature-based model is hypothesised to: – Achieve the highest overall performance

  - Adapt most effectively to changes in speaker and instance count

- Transformer-based model is hypothesised to:
  - Outperform the phonetic model in smaller-scale conditions
  - Underperform relative to the spectral model as enrollment complexity increases due to reduced adaptability

1.4.2 Hypotheses for Second Research Question

In real-world communicative scenarios, speaker identification systems must operate under open classification conditions, where the identity of an incoming speaker may not have been previously encountered. To address this, the present study employs an entropy-based thresholding mechanism that allows the model to abstain from uncertain predictions, categorizing inputs as known, unknown, or null. The effectiveness of this mechanism, however, is expected to depend on the composition of the enrolled speaker space. As the number of enrolled speakers grows, the embedding space becomes increasingly populated, raising the likelihood that an unknown speaker's representation will closely align with that of an enrolled speaker. Therefore, successful open classification with abstention requires models not only to maintain high precision in selecting the top speaker but also to exhibit stable and reliable confidence patterns when ranking candidates within the gradually expanding enrollment set.

This increases the risk of false positive predictions with low entropy, thereby hindering the model's ability to identify unknown speakers with high precision in complex classification settings.

Successful open classification with abstention requires models to demonstrate not only robustness in precisely identifying the correct top-ranked speaker, but also stability in their confidence patterns across the full set of enrolled speakers. Since open classification performance is closely tied to how well a model can form robust and scalable speaker clusters, its effectiveness in this setting is inherently influenced by its performance in closed-set classification under scaled conditions. Consequently, the ability to maintain high precision in identifying unknown speakers becomes increasingly challenging as the number of enrolled speakers grows. This leads to the general hypothesis that the precision of unknown speaker classification will decrease as the speaker population increases, primarily due to a higher incidence of false positive predictions with high confidence scores. Given that open classification outcomes are heavily dependent on the scalability and separability of clusters in the embedding space, no definitive prediction can be made regarding which representational type will outperform others, as their performance in closed-set scaled classification is a determining factor in open-set success. Based on these considerations, the following hypothesis is proposed:

- · General hypothesis: The precision of unknown speaker classification will decrease as the number of enrolled speakers increases, due to the increasing number of false positive predictions with high confidence scores.

2.3 Experimental Setup

This study compares the selected speaker identification models on two key aspects: **(1)** their ability to generalize as the number of known speakers and available samples per speaker increases, and **(2)** their ability to perform open-set classification while allowing for inconclusive predictions. Both experimental phases used a grid-based setup where these variables were systematically varied to test how scaling affects model performance. Precision scores were used to assess results of the first experiment, while precision, recall and abstention ratio values were observed within the second experiment.

The first experiment focused on closed-set classification, evaluating how each model's precision changed under different enrollment sizes. Results were visualized using heatmaps to capture performance trends as the system scaled up.

The second experiment extended this setup to open-set conditions by introducing an entropy-based thresholding mechanism. This mechanism allowed models to abstain from uncertain predictions, categorizing inputs as known, unknown, or inconclusive. Thresholds were learned from the distribution of confidence scores and applied during inference to improve decision reliability. Heatmaps were used to illustrate the effect of this mechanism on precision, recall, and abstention rates.

Chapter 2

Theoretical Background

This section provides the theoretical background necessary to understand the task of speaker identification. It outlines the fundamental principles behind distinguishing speakers based on their unique vocal characteristics and discusses the types of signals traditionally used in this field. Furthermore, it reviews key computational approaches explored in the literature, highlighting how different modeling strategies have been employed to capture speaker identity. Finally, it examines the role of data representation, detailing how various representational strategies, situated on the spectrum of generalisation, influence the ability of speaker identification systems to extract and utilize speaker-specific information.

2.1 The Task of Speaker Identifiction

Speaker identification is the task of determining an individual's identity from their voice alone. In audio-based systems, the goal is to analyze a speech signal and match it to a known speaker from a pre-enrolled set of identities. This capability underpins many applications, including biometric authentication, forensic analysis, human–computer interaction, and communicative robotics. A recent survey of the field provides a comprehensive overview of the methods, challenges, and trends in this area Wang et al. (2024).

A typical speaker identification pipeline can be divided into three stages: speaker characterization, enrollment, and classification. In speaker characterization, the system extracts a speaker embedding, a condensed numerical representation that encodes the most salient vocal traits of the speaker. These embeddings are stored during enrollment to represent known speakers, and in classification, embeddings from new audio segments are compared against the enrolled set to identify the most likely match.

Two common distinctions shape system design. First, closed-set classification assumes that all speakers in evaluation are already enrolled, while open-set classification must also detect previously unseen speakers. Second, systems can be text-dependent, requiring fixed phrases, or text-independent, where identification must work for arbitrary speech. The latter is generally more challenging but aligns better with unconstrained, real-world use cases.

Among these stages, the speaker characterization phase has received the most sustained research attention, as the quality of embeddings directly impacts classification accuracy and generalizability. Loss functions play a central role in shaping the embedding space. Classification-based objectives (e.g., cross-entropy over speaker IDs) encourage embeddings of the same speaker to form tight clusters while separating those of different speakers. However, this approach is inherently speaker set dependent, the model learns to discriminate only between the speakers present during training, meaning that identifying new, unseen speakers requires retraining or fine-tuning. Metric learning objectives, such as contrastive loss and its variants Wang and Liu (2021), go further by operating on pairs of inputs—pulling embeddings from the same speaker closer together and pushing embeddings from different speakers further apart. Over time, this pairwise training process creates a structured embedding space in which each speaker occupies a distinct, well-separated region, while maintaining speaker set independency, allowing the system to generalize to speakers not encountered during training without requiring retraining.

The success of these methods also depends on the acoustic representation provided to the model. Raw audio contains the full detail of the signal but is too unstructured for most models, so it is transformed into more informative forms. Common examples include Mel-Frequency Cepstral Coefficients (MFCCs), which compactly encode vocal tract characteristics in a perceptually motivated way, and spectrograms, which capture detailed time–frequency patterns of the signal. MFCCs distill the voice into a low-dimensional summary suited for speaker discrimination, while spectrograms retain more raw information, potentially allowing finer-grained analysis of voice traits. Both aim to highlight speaker-specific information while reducing irrelevant variability, making the representation stage fundamental to the overall system.

2.1.1 Signal Representational Strategies

The effectiveness of speaker identification heavily depends on how audio signals are transformed into meaningful representations that capture speaker-specific characteristics. At its foundation lies the raw audio signal, an unprocessed waveform that contains all the fine-grained details of a speaker's voice. While this raw form preserves every nuance of the sound, it is often too complex for direct use in modeling, as it lacks structure and may not highlight the most informative aspects of speech. To make the data more meaningful and manageable, researchers transform the raw signal into specialized representations that capture speaker-relevant characteristics. Two widely used forms are Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms. MFCCs are numerical representations designed to approximate how humans perceive sound, emphasizing frequencies that the human ear is most sensitive to. They provide a compact, low-dimensional summary of the vocal tract's characteristics, making them highly effective for identifying speakers. Spectrograms, in contrast, represent how energy is distributed across different frequencies over time, providing a detailed, fine-grained view of the signal's temporal and spectral structure. While MFCCs distill the audio into a simplified, perception-driven form, spectrograms retain more raw information, allowing models to analyze subtle patterns in the voice that might be lost in more compressed representations. Both approaches serve the same purpose, transforming complex audio into data that highlights speaker-specific traits, but they differ in how much detail they preserve and how closely they mimic human auditory processing.

2.2 Architectural and Data driven approaches to Speaker Identification

2.2.1 Architectural Approaches

In contemporary work within the task of speaker identification, three fundamental computational approaches can be identified within the literature. The first is the x-vector (Snyder et al. (2018)) framework, which transforms speech into fixed-length representations by modelling and aggregating the energy distribution of sound at given points in time. The second is the ResNet-based framework (Jung et al. (2019), which relies on deep con- volutional networks that learn patterns from the audio signal in a layered, hierarchical manner. The third is the transformer-based framework, which uses self-attention mech- anisms to implicitly acquire contextual dependencies between the appearing phonemic units within utterances, capturing both short-term details and long-range dependencies important for recognizing speaker identity (Hsu et al. (2021)).

The x-vector framework

The x-vector framework is built around a time-delay neural network (TDNN), a type of feedforward architecture designed to capture temporal patterns in sequential data (Snyder et al. (2018)). It operates on short frames of raw audio, typically represented as MFCC features, which encode how energy is distributed across different frequency bands over time. Each timestamp can be viewed as vertical sample slices of sound, representing the distribution of energy and frequency at a given time. The TDNN model processes these frames while maintaining awareness of their relative temporal order, learning how speaker traits are expressed across short time intervals.

The key idea behind the x-vector approach is to aggregate frame-level representations into a fixed-length speaker embedding using a statistical pooling layer. This layer compresses temporal information by computing statistics such as mean and standard deviation across time, producing a single vector that summarizes the entire utterance. Once embeddings are generated, speaker identity is typically determined by comparing them using similarity metrics such as Probabilistic Linear Discriminant Analysis (PLDA). This framework has shown strong performance on structured datasets like VoxCeleb1 and VoxCeleb2, (Nagrani et al. (2017), Chung et al. (2018)).

The ResNet-based framework

The ResNet-based framework adopts a convolutional neural network (CNN) archi- tecture with residual connections, allowing deeper models to be trained without losing important speaker-related information (Jung et al. (2019)). These residual connections allow the network to pass information directly through layers, improving learning efficiency and helping preserve speaker-relevant patterns across depth. ResNet based models take in 3D audio representations, such as log-mel spectrograms, which represent frequency on one axis, and time on the other, while signaling the energy at the third axis. This approach of modelling speaker identity is inspired by methods used in image recognition, as the network scans these 3D inputs for recurring patterns that signal a speaker's unique vocal traits.

Residual CNNs learn to detect and abstract indicative patterns through progressively deeper layers, where low-level acoustic details are combined into higher-level speaker characteristics. These embeddings are then compressed using global average pooling and projected to a fixed-length representation through fully connected layers. At inference, speaker comparisons are typically made using cosine similarity. This strated high performance on benchmarks like VoxCeleb datasets (Nagrani et al. (2017)).

The transformer-based framework

The transformer-based framework is centered on self-attention networks, originally developed for text processing but currently widely adopted in speech based modelling tasks. These models process raw audio or low-level acoustic features by leveraging self-attention mechanisms, which allow them to model contextualised relationships between phonemic units across the entire input sequence. This makes them fundamentally different from sequential models like TDNNs or spatial models like CNNs.

The pretraining of large speech models is a critical consideration when applied in speaker identification. Models such as HuBERT (Hsu et al. (2021)) are pretrained on unlabeled speech using objectives such as masked segment prediction, where the model learns to predict hidden portions of the audio based on surrounding context. This encourages the network to develop a high-level understanding of phonetic and linguistic structure, while encoding information such as intonation and prosody. In contrast, models like Whisper (Radford et al. (2023)) are trained on large scale transcribed speech datasets using supervised automatic speech recognition (ASR) objectives, which guide the network to learn complex relations between linguistic, textual information and acoustic content. Both pretraining strategies yield embeddings that can later be fine-tuned for speaker identification.

In this context, transformer layers act as deep feature extractors, capturing voice quality, articulation patterns, and subtle prosodic cues across time. For the speaker identification task, fine-tuning typically involves contrastive learning objectives, such as binary or triplet loss, which structure the embedding space by bringing same-speaker samples closer together and pushing different-speaker samples apart. Classification is performed using cosine similarity between the extracted embedding and reference speaker embeddings. These transformer-based systems have set new benchmarks for speaker identification, demonstrating strong generalization across languages, recording conditions, and speaker variability, particularly when utilising embeddings extracted from robust pretrained encoders.

2.2.2 Data representational strategies:

Data representation plays a central role in speaker identification, shaping how information from the audio signal is captured and interpreted by computational models. Different strategies along a spectrum of generalisation have been proposed in the literature, ranging from highly specified, interpretable features to generalized, learned embeddings. Each approach offers unique advantages and limitations in terms of robustness, adaptability, and scalability. The following sections provide an overview of how these representational strategies have been explored within speaker identification, focusing first on specified physiologically based representations, then on generalized embeddings, and finally on approaches utilising non-abstracted signal representations . This discussion sets the stage for understanding how different representation types may influence the performance and practicality of speaker identification systems in communicative robots.

Specified Representations in Speaker Identification

Specified approaches to speaker identification rely on the explicit extraction of articulatory features, such as pitch, formants, voice onset time (VOT), and phoneme-aligned cepstral coefficients, to construct interpretable and physiologically grounded speaker representations. These methods are motivated by the assumption that certain vocal traits remain stable across different speaking conditions and are directly linked to the physical characteristics and articulatory habits of the speaker.

A representative example of phonetic features in isolation is presented by Meghanani and Ramakrishnan (2018), who extracted pitch-synchronous Discrete Cosine Transform (DCT) features from voiced speech segments. In this method, the pitch cycle of voiced segments serves as the temporal boundary for feature extraction, aligning the analysis with the natural periodicity of speech. The DCT is then applied within each pitch period, producing a compressed representation of the pitch contour that preserves speaker-specific prosodic and glottal characteristics. Their results showed that even without spectral features, these pitch-synchronous representations could achieve 90% accuracy on the TIMIT dataset (Garofolo et al. (1993)). This demonstrates that pitch structure alone can carry robust speaker-specific information, particularly in clean speech environments.

In contrast, a study by Nasr et al. (2018) explored the combined effect of phonetic and spectral features for speaker identification. Their system integrated normalized pitch frequency with Mel-Frequency Cepstral Coefficients (MFCCs), showing that the fusion of phonetic and spectral information significantly improved classification performance over using MFCCs alone. The authors concluded that pitch features contribute complementary information that enhances the model's ability to capture speaker identity, particularly when lexical variability or background noise is present.

In the context of communicative robots, specified approaches are especially attractive when interactions occur in relatively clean environments with limited background noise, and the goal is to ensure continuity of speaker identity across time. By encoding vocal tract-related cues, such as pitch stability and articulatory patterns, these models can maintain recognition performance even when verbal content shifts. However, in cases when phonetic features are not used in combination with other complementary representations, a key limitation remains. Their sensitivity to environmental noise and lack of robustness under highly variable conditions might limit their application to highly controlled communicative contexts.

Generalised Representational Approaches to Speaker Identification

Generalised speaker identification approaches rely on high-dimensional embeddings learned from large-scale, pre-trained transformer models. Unlike specified models that extract interpretable features like pitch or formants, these systems learn to encode speaker identity implicitly through exposure to vast amounts of raw audio data. One prominent example of such models is HuBERT (Hsu et al. (2021)), a self-supervised model trained with a masked prediction objective, originally designed for speech representation learning.

HuBERT's pretraining process employs a two-stage training approach. First, it uses unsupervised k-means clustering on MFCC features to create pseudo phoneme-like units. The resultsing cluster labels are then treated as targets for a masked prediction task. While this process is not explicitly designed to encode speaker traits, the resulting representations capture latent speaker-specific information.

This clustering mechanism can affect speaker embeddings in significant ways. Since the model organizes its internal representations around acoustically meaningful segments (often aligning with phoneme-like boundaries), it learns to capture not only linguistic information but also speaker-specific articulatory habits embedded in the phonetic realizations.

Recent studies have fine-tuned HuBERT models for speaker identification, demonstrating that the phoneme-aligned structure learned during pretraining can be leveraged to extract effective speaker embeddings. These models have been shown to perform competitively on benchmarks such as VoxCeleb and CN-Celeb (Nagrani et al. (2017)), outperforming traditional x-vector baselines, especially under open-set or low-resource conditions.

For communicative robots, this architecture offers critical advantages. Its robustness to noise, multilingual data, and spontaneous speech makes it suitable for dynamic noisy environments. However, their behaviour in incremenatlly growing enrollment sets have not been researched in the contemporary literature.

Intermediate Feature-Based Models: x-Vectors and ResNet Architectures

Situated between the specified and generalized approaches, models based on spectral features rely on unstructured acoustic inputs, such as MFCCs, log-mel spectrograms, without making explicit assumptions about which vocal traits are most informative for speaker identity. These systems are trained on raw spectral representations using deep neural architectures, allowing the model to autonomously learn speaker-discriminative patterns. The most notable approaches which use this representational data type are the x-vector (Snyder et al. (2018)) and ResNet (Jung et al. (2019)) based frameworks. Their respective modeling architectures, using temporally sensitive sequential modeling or spatially aware convolutional approaches, are well equipped to identify the emerging speaker specific patterns, due to their deep and complex architectural designs.

Both x-vector and ResNet systems can be beneficial for communicative robotics, where the audio environment is often variable and speaker input may be brief or sporadic. By learning directly from MFCC or spectrogram inputs, these models offer a balance between structured representation and adaptive learning. They support real-time inference, scale well with new data, and require no manual tuning of speaker-specific cues, making them ideal for incremental speaker enrollment in dynamic, real-world settings.

Comparative Analyses of Representational Strategies in Speaker Identification

Recent studies have examined the impact of different representational strategies on speaker identification performance however, no conclusive findings can be drawn from the contemporary findings. The results vary across datasets, tasks, and evaluation criteria, making it difficult to determine which approach is most effective in general or for specific contexts like communicative robotics. For example, Brydinskyi et al. (2024) compared several state-of-the-art embed- ding architectures, including ECAPA-TDNN, TitaNet, and WavLM (Desplanques et al. (2020), Chen et al. (2022), Koluguri et al. (2022)), across multiple benchmarks. They found that generalized embeddings, especially those from transformer-based models, consistently performed well under variable and noisy conditions, highlighting their robustness and scalability. However, these models offered limited transparency into which features influenced their decisions, and their adaptability to longitudinal application contexts remains an underresearched area.

In contrast, Gendrot et al. (2019) evaluated two systems, one based on raw spectrograms and another using 62 manually extracted phonetic features. They reported that while both systems achieved comparable overall accuracy, each excelled in different scenarios: the phonetic-based system performed better in structured environments with consistent articulation, while the spectrogram-based model showed greater flexibility and adaptability in varied experimental setups. Interestingly, their findings suggested that even models trained on raw acoustic inputs may implicitly learn pitch-related cues, pointing to potential overlaps between the different representational strategies.

Together, these studies underscore that the effectiveness of a given representation type is highly context-dependent. Without a unified evaluation across open-set, dynamic, and real-time conditions, especially those relevant to communicative robots, the field still lacks definitive guidance on the optimal strategy for speaker identity modeling in practical, scalable systems.

2.3 Open set classification:

Open-set classification is another crucial requirement for speaker identification in the context of communicative robots. Unlike closed-set systems, which assume that all speakers encountered during inference have been previously enrolled, open-set systems must handle the more realistic scenario in which unknown or novel speakers appear during interaction. This introduces the need not only for accurate identification but also detection of unseen patterns within the query inputs.

Research on open-set speaker identification has predominantly focused on thresholding mechanisms applied to distance or similarity metrics between the enrolled speaker profiles and an input query. These thresholds determine whether a speaker is recognized as known or rejected as unknown. For instance, Karadaghi et al. (2014), proposed a two-stage open-set speaker identification (OSTI-SI) system that first performs a closed-set classification and then applies a verification stage using cosine similarity of i-vectors. If the similarity score surpasses a predefined threshold, the speaker is accepted; otherwise, the speaker is rejected as unknown. Such a thresholding approach to identify previously unseen speakers have been shown to yield high performance in open classification tasks by vast amount of work in speaker identification research. (Chakraborty and Parekh (2017), Affek and Tatara (2022))

While such systems have advanced open-set speaker identification through threshold-based decision making, a notable gap remains which needs to be addressed for an open set model to be applied within the current context, the option of abstention. In communicative robots, which operate in unpredictable and socially sensitive environments, the system must not only distinguish between known and unknown voices, but should also withhold a prediction when the available information is ambiguous or insufficient. Without this option, the system may be forced to classify uncertain inputs, leading to incorrect identifications and potential corrupted representations of the learned information within the system's knowledge graph. The addition of a two-threshold system in the open-classification task could enable the system to perform the task in a conservative manner, by reducing false positive predictions and raising precision scores.

By comparing how models trained on specified (phonetically based), spectral, and generalised (embeddings extracted from transformers models) perform under this two-threshold open classification paradigm, this research contributes toward building speaker identification systems that are not only accurate, but also conservative, scalable, and well-suited for real-world deployment in interactive robotic platforms.

2.4 Interpretable Ensemble Modeling Approaches

To address two key gaps in speaker identification, how speaker identity evolves with increasing data, and how open classification can incorporate abstention, a more interpretable modeling strategy is needed. Unlike end-to-end deep models, interpretable frameworks can offer insights into the dynamics of identity formation and decision uncertainty.

Modular architectures, such as ensemble models, present a promising alternative for this purpose. By separating different feature types or modeling strategies into distinct components, ensemble systems allow researchers to observe the role of specific information streams in speaker identification.

Ensemble approaches have already been employed in speaker recognition. For instance, Zarin et al. (2024) combined RNN and DNN sub-models trained on distinct acoustic views to improve robustness under noise. Such modular strategies not only improve performance but also enhance transparency, making them well-suited for speaker identification in communicative robots where interpretability, adaptability, and cautious decision-making are essential.

To systematically investigate these questions, the present work adopts an interpretable ensemble framework in which each module is designed to model different articulatory cues relevant to speaker identity. This design allows the system to prioritise certain embedding spaces that may become more informative in upscaled scenarios, helping maintain discrimination as classification complexity increases. Moreover, the modular structure makes it possible to observe how the internal organisation of the ensemble shifts as the task scales, providing insight into which articulatory cues and embedding spaces remain stable and which degrade under increased enrollment demands.

2.5 Summary

In summary, while extensive research has explored a wide range of computational architectures, data representation strategies, and open-set classification methods for speaker identification, several gaps remain unresolved, particularly in contexts requiring scalable, adaptive, and conservative decision-making, such as communicative robots. Existing studies rarely investigate how the internal structure of a system adapts as the enrollment set scales. In addition, they do not address how such systems can maintain performance in increasingly complex classification scenarios. Furthermore, the incorporation of abstention into open-set classification remains underexplored, despite its potential to prevent erroneous identity assignments in uncertain cases. Addressing these gaps requires an approach that is both modular and interpretable, enabling targeted analysis of system behaviour while observing structural changes as task complexity grows. The following methodology section details the experimental framework developed to meet these objectives and systematically evaluate the scalability, adaptability, and cautious decision-making capacity of speaker identification systems.

Chapter 3

Methodology

This section outlines the methodological setup used to investigate the proposed research questions. It begins by describing the data used for training and evaluating the speaker identification models (2.1), followed by the preprocessing methods applied to the input audio data (2.1.1).

In the next part, the modeling approaches are detailed. After a general overview of the modeling pipeline (2.2, 2.2.5), the architectural components of the speaker embedding models are explained (2.2.3), along with the specific machine learning approaches used , and the extracted features utilized for each speaker identification model .

Finally, the section describes the experimental setups used to address the research questions. The first experimental setup focuses on the evaluation of model scalability and classification performance (2.3.1), while the second explains the methods used for entropy-based thresholding mechanisms and the evaluation metrics applied to assess the models' performance in an open classification setting (2.3.2).

3.1 Data

The LibriSpeech corpus is a large-scale, high-quality dataset of read aloud English speech, specifically designed for training and evaluating speech recognition systems (Panayotov et al. (2015)). It comprises approximately 1000 hours of audio, sourced from public domain audiobooks. These recordings are aligned with their corresponding texts from Project Gutenberg and segmented into utterances of up to 35 seconds, ensuring suitability for automatic speech processing. The dataset is partitioned into multiple subsets, such as train-clean-100, train-clean-360, and train-other-500, to support varying levels of audio quality and acoustic coverage. It features a total of over 2,300 unique speakers, with careful attention to gender balance: for example, the train-clean-360 subset alone includes 921 speakers, with roughly equal numbers of male and female participants. Each speaker contributes up to 25–30 minutes of audio per subset. The corpus's structured design, along with its studio-level audio quality and moderate per-speaker coverage, offers two key advantages for this research. First, it allows systematic investigation into the impact of enrollment size on speaker modeling performance. Second, the controlled acoustic environment minimizes the influence of noise or variable recording conditions. These features make LibriSpeech particularly well-suited for evaluating speaker identity models in a controlled and scalable manner.

The librispeech dataset is structured in the following way: It it contains separate folders for each speaker in the corpus. For each speaker, the different sources of audio recordings are listed. In this case the sources mean the books or articles they read aloud For each source, the corpus contains 'flac' audio files which are recordings clipped to the length of 36 seconds, and named consequently as they were recorded in order.

In the current study, three subsets of the LibriSpeech corpus were used to train and evaluate the speaker identification models. For training, the *train-clean-100* subset was selected due to its high recording quality. This subset contains 251 speakers, 125 male and 126 female, with a total of approximately 100 hours of speech. Each speaker contributes around 25 minutes of audio. This subset was used to train the speaker embedding models, the fusion model, and to calculate open classification thresholds.

For training the speaker embedding models, approximately 50% of the available data was used during various stages of model development. For training the fusion component of the identification system, random samples of speakers and their associated utterances were extracted. The *dev-clean* subset, comprising 40 speakers with gender balance and approximately 8 minutes of speech per speaker, was used as a validation set. This subset not only served to assess speaker embedding models' performance but also provided the basis for extracting statistical parameters used during the normalization phase.

The final evaluation was conducted on the *train-clean-360* subset, which includes 921 speakers (436 female and 432 male). It offers the same type of audio recordings as the training and evaluation set, while the speakers and their recorded instances do not overlap, which allows for the analysis of the models' performance on unseen withheld dataset. The choice of this subset was motivated by its large number of speakers and the relatively rich amount of speech per individual. In contrast, the *test-clean* subset, with only 40 speakers, was insufficient for evaluating how model performance scales with increasing speaker diversity and exposure.

3.1.1 Preprocessing

Preprocessing is a critical initial step in speaker identification, aiming to standardise the input audio while preserving speaker-specific acoustic features. However, the degree of preprocessing must be carefully managed. Excessive processing, particularly on already clean datasets like LibriSpeech, can obscure subtle acoustic cues that are essential for distinguishing between individual speakers. To mitigate this risk, a minimalist preprocessing strategy was adopted in this study to retain as much speaker-relevant information as possible.

This approach is grounded in the findings of **?**, who conducted a comparative analysis of preprocessing techniques for speaker identification. Their results demonstrated that MFCC-based systems performed best when only basic normalization was applied, while additional transformations such as time masking or frequency masking significantly degraded the performance. These results reinforce the importance of preserving raw speaker characteristics and minimizing unnecessary manipulation during preprocessing.

In line with the minimalist preprocessing strategy, only two core operations were applied before speaker embedding extraction:

1. **Voice Activity Detection (VAD)**: Voiced segments of speech were isolated by removing silent and background noise-dominated regions. This ensured that only segments containing rich speaker-specific acoustic information were used for modeling.

2. **Time Windowing**: Each voiced segment was truncated to a fixed duration of two seconds. This window length was chosen to provide a balance between capturing sufficient phonetic diversity, such as vowel quality, syllable articulation, and intonation, and maintaining consistent input size and computational efficiency.

The decision to adopt a minimal preprocessing pipeline was informed not only by the above-mentioned empirical findings but also by the inherent quality of the dataset used in this study. Thus, additional steps such as denoising, filtering, or artificial augmentation were deemed unnecessary and potentially counterproductive.

The preprocessing pipeline was implemented in Python. All audio inputs were first resampled to a standard sampling rate of 16 kHz using the Librosa (McFee et al.. Voice Activity Detection (VAD) was then applied using the webrtcvad package (Karrer (2025), configured with an aggressiveness level of 2 and operating on 30 ms frame windows. Only frames labeled as speech were retained.

Each resulting voiced segment was subsequently divided into fixed-length two-second chunks. Segments shorter than this duration were discarded to maintain consistency across inputs. The final preprocessed segments were saved as .npy files and organized into subdirectories by speaker ID, facilitating efficient data loading and speaker-specific training workflows.

3.2 Modelling approaches:

This section outlines the architectural and data-driven design choices involved in developing the speaker identification models. Two distinct modeling approaches are presented, each based on a different architectural strategy.

The first is an ensemble-based model that employs a late fusion classification head. For this model, the two data modalities, phonetic and spectral features, used to train the respective speaker embedding models are described in (2.2.1).

The second is a transformer-based model, which incorporates a transfer learning setup for speaker verification. This approach includes the use of the HuBERT model for embedding extraction, along with a Siamese neural network used during the verification stage, as explained in (2.2.5).

3.2.1 Ensemble Based Speaker Identification Model

The ensemble based speaker identification approach utilizes four stages of computation to arrive from a sample of audio recordings to predicted speaker identification labels.

The first stage of the modeling approach entails the feature extraction from the preprocessed audio recordings. This approach is different for each ensemble based speaker identification model, due to the difference between the data used in their modelling approaches.

After the feature extraction process, the resulting representations are passed into four separate speaker embedding models, which are neural networks trained to generate speaker specific vector representations. These models form the core of the ensemble approach. They generate speaker-specific embeddings that allow the likelihood of a query belonging to a known speaker to be assessed through similarity comparisons.

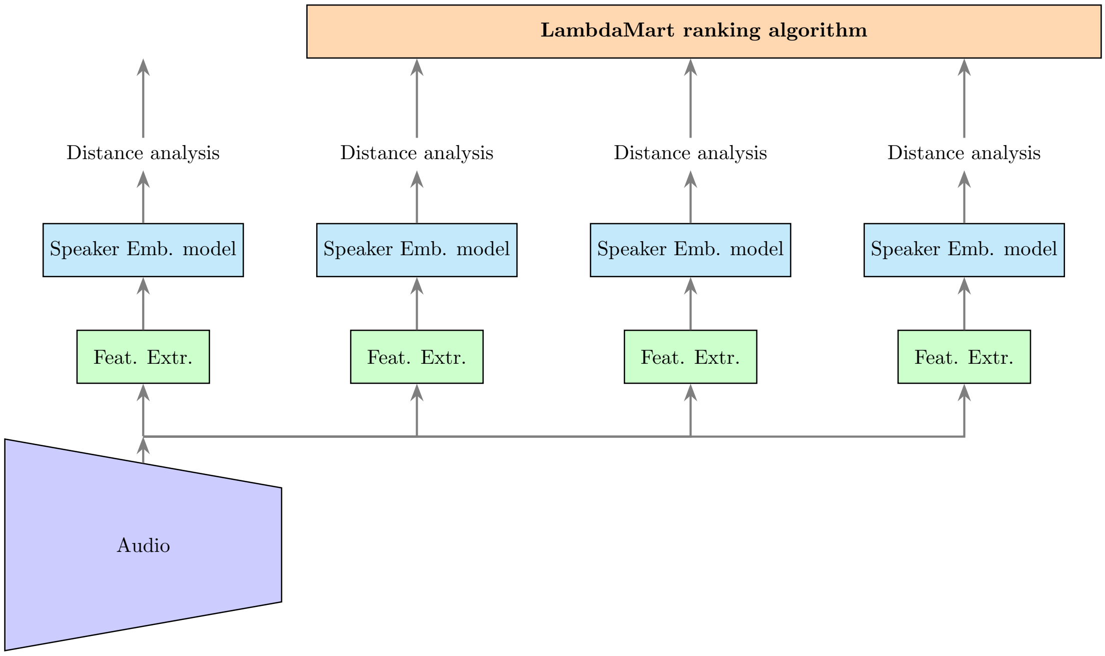
The resulting speaker embeddings generated by these models are used within an enrollment phase. This phase means that the model is familiarized with a set of speakers, and a set of instances for each speaker. This information represents the current 'knowledge' of the model.

As new instances come into the ensemble model to be classified, these instances also go through the same preceding processes, and are compared against the enrolled (known) representations. This happens by calculating distance measures from the clusters of enrolled speakers. These distance measures are calculated for each populated vector space, by the four separate speaker embedding models, and are normalized in order to make them suitable for the last stage of the classification, the fusion process.

The fusion process happens by passing these distance measures into a LambdaMart ranking algorithm (Burges (2010) is trained to generate a ranked list of candidates based on their similarity to the input of inference. This model outputs a list of confidence scores, out of which the top candidate can be regarded as the classified speaker.

This modelling approach allows for a modular design which can take into consideration different types of acoustic cues, without influencing each other within the comparison phase. Moreover, the last fusion method allows for flexible identification, as the weight of the particular information can be varied across different query instances.

In the following flow chart the ensemble-based speaker identification process is demonstrated, visualizing the above mentioned modelling processes.



Speaker Embedding models:

Speaker embedding models are neural networks trained to generate speaker specific vector representations, by modeling each speaker's distinct vocal characterestics. These models are not inherently classification models, but embedding models, and their architectural design follows appropriate machine learning approaches, and loss functions.

In this study, the speaker embedding models are implemented as shallow neural networks trained with a contrastive loss function. The architecture varies depending on the type of input features used, with Multi-Layer Perceptrons (MLPs) (far (2002)), Time Delay Neural Networks (TDNNs)(Chung et al. (2018)), and Convolutional Neural Networks (CNNs) (Villalba et al. (2019) employed across the ensemble components.

These model types have previously been shown to produce robust and discriminative embeddings for speaker identification tasks.

Within the training phase, the models get input pairs. For each of these input they generate a speaker embedding, which is compared using the contrastive loss function, in the current case cosine similarity loss function. During the training the models receive either pairs of input that originate from the same speaker, or come from different speakers. Based on this, the models weights are changed to push the generated embeddings either apart, in case of different speaker pairs, or closer to each other in case of matching speaker pairs. This contrastive loss function directly correlates with the goals of the current study, as it learns the speaker specific qualities of each speaker, while satisfying the open-set classification constraint of the current project.

Inference and Normalisation:

The generated speaker embeddings are used within the model to compare learnt representations. Once these models are employed within the ensemble approach, the ensemble goes through the enrollment process.

The inference phase starts by the generation of speaker embeddings for unseen instances. These instances are then compared against the enrolled representations.

During the comparison phase of the task, these created clusters are compared against the input of inference, otherwise named query instances, and their relation is examined through different distance metrics. Firstly, the average distance from the query instance to all of the enrolled representations of a speaker is calculated. This measure gives a simplistic overview of the most likely speakers within the enrollment setup. The second distance measure calculated is the distance from the query instance to the centroid of the speaker cluster. This allows for a more robust metric which is not influenced by the presense of outliers within the speaker populated vector space. Lastly, the minimum distance from any enrolled instance is calculated between the query instance and the enrolled representations. This measure indicated where the query might be positioned within the enrolled speaker clusters. If this measure shows a low value, it indicates that the instance is within the boundary of the speaker cluster, which makes it very likely to be belong to this speaker.

While these measures are informative, they cannot be used directly without normalization. Each of the four embedding models exhibits different sensitivity and distributional behavior, which may lead to inconsistencies in the distance scale. To address this, the distance values are normalized before being used in the final classification stage, ensuring that all embedding spaces contribute comparably to the ensemble's decision making process.

The normalization process entailed the evaluation of the speaker embedding models on a held-out developmental dataset. Within this set, the intra speaker similarities of the generated embedding were assessed. The intra speaker similarity measures that were calculated were the distance from a datapoint to the centroid of the cluster, and the average distance measures between randomized speaker embedding pairs. From these extracted distance values 10% and 90% quantile values were calculated, which indicates the values above which the instances were considered to be outliers, either due to their elevated distance, or their elevated tightness within the cluster. The average and minimum distance features were normalized using the average distance quantiles, while the centroid distance feature was normalized using the centroid distance quantiles. These instances then were used as thresholding mechanisms when normalizing the distance measures. If an distance value fell below the Q10% threshold, the instance betcame 0 as it is very likely to be part of that cluster, however, if the instance fell above the Q90% value, it is considered to be too far away of the enrolled cluster, thus elevating the possibility of introducing noisy representations within the clusters. The value of these distances became 1. The values that fell between the 10% and the 90% quantiles were scaled between these values. As such, the input data for the fusion process included these normalized distances for each speaker embedding space, resulting in 12 dimension input sequences (3 measures * 4 speaker embedding models).

Fusion process:

Once the distance measures have been extracted from the populated vector spaces, this information is used for the fusion process, which is done by employing a ranking algorithm.

The chosen ranking algorithm is LambdaMART, a ranking method that combines gradient-boosted decision trees with ranking-based loss functions(Burges (2010)) LambdaMART operates by training on pairs of candidate instances, learning to rank them by optimizing the available similarity features. Instead of predicting class labels directly, the model learns to assign scores such that more relevant candidates are ranked higher in a list. This makes it highly compatible with the speaker verification setup, where the goal is to identify and prioritize the most probable speaker identities among a number of enrolled speakers. LambdaMART is a ranking algorithm that uses a set of decision trees to learn how to sort candidates based on how likely they are to be correct. LambdaMART's ability to handle non-linear interactions, robustly integrate heterogeneous input features, and directly optimize for rank-based evaluation metrics makes it especially suitable for the late-fusion process. It allows the model to dynamically weigh contributions from multiple embedding models and distance metrics, producing an output that reflects relative confidence across candidates, rather than binary decisions.

After all speaker embedding models are trained, they undergo an enrollment process, where speaker embeddings are extracted for a specified number of speakers, with a defined number of instances per speaker. After the enrollment phase is complete, the enrollment configurations are used to infer the system, using query instances. Each query is passed through the hybrid embedding model to generate a speaker embedding. This query embedding is then compared to the stored embeddings of all enrolled speakers using the distance-based similarity measures. These features summarize the query's similarity to each speaker and form the input for the final decision-making step. In this final stage, a ranking algorithm is used to fuse the information from all speakers and generate a ranked list of likely matches, each associated with a confidence score.

The LambdaMart ranking algorithm was trained on the same amount, and exact enrollment configuration setups for both ensemble-based approaches. Each different configuration setup appeared 4 times within the training data, with different randomised speakers and instances. In the case of the training phase, the enrollment configurations generated were not cumulative, meaning a new set of random speakers and instances were chosen for each setup. In contrast, the test configuration setup used for evaluating the speaker embedding models were cumulative, thus resembling more the real life setting of scalable speaker identification. One training complete grid-based configuration set contains 49 configurations(7 number of speaker variables * 7 number of instances variables) and their combined amount of used data comprises of 25,8 hours. As each setup appeared 4 times within the training the total amount of enrolment information used for the training of the LambdaMart model equals to 103,2 hours. This number contains repeated instances and speakers, however merely between complete grid configuration setups. The enrolled instances of audio were compared to query instances. For each speaker in an enrolment configuration setup 10 instances were used. That the total amount of audio data used for the query instances for the possible 49 configurations was 0.6 hours of audio data, and as each setup appeared 4 times within the training data, the total amount of query data used in the training phase equalled 2,4 hours. The validation dataset for the LambdaMart model contained of an additional complete grid-based configuration set.

3.2.2 Modeling approaches using ensemble architecture:

The ensemble architecture was implemented in two parallel modeling approaches, each relying on a distinct set of acoustic features. The first approach utilizes explicitly extracted phonetic features, grounded in articulatory and acoustic phonetic theory. The second approach leverages spectral features**,** derived directly from the frequency content of the audio recordings. This design allows for a comparative evaluation of how theoretically informed versus spectral-based features contribute to speaker identification performance.

The phonetic model incorporates four distinct types of speaker-specific information derived from audio segments. Each of these feature sets captures a different dimension of acoustic variability, enabling the system to rely on a diverse set of cues, particularly useful when one type of feature is less informative for a given instance. The phonetic features were extracted using the Librosa (McFee et al. and Parselmouth (Jadoul et al. (2018) libraries.

3.2.3 Phonetic Model

First speaker embedding model: General speaker-specific acoustic cues

The first model is trained on a set of features generally viewed as fundamental information in articulation (Ladefoged and Disner (2012)) . These features contain information about the pitch (mean and standard deviation) and its periodical qualities such as shimmer, jitter (over the whole segment, and across 5 glottal pulses), harmonics to noise ration (HNR) and the spectral energy over the segment. Moreover, these features also utilise information about the mean of the most important fundamental frequencies such as f1, f2, and f3 . This model was trained on an input length of 11 dimensions. The feature instances that could not be extracted due to lack of information within the segment, or corrupted audio quality, were substituted with a value of 0.

Second speaker embedding model: Vowel articulation

The second speaker embedding model is trained on 9 features designed to reflect the vowel production of the speaker, including pitch (mean, standard deviation and slope) and f1 variations (mean and standard deviation), jitter (over the whole segment, and across 5 glottal pulses), voiced and unvoiced ratio, and the energy value extracted from the frequency band 0-1000Hz (Ladefoged and Disner (2012)). In order to identify vowel production specific information, the features using this model were all situated within the lower frequencies of the audio signal, thus eliminating the chance of feature overlapping with the other three frequency band specific models.

Third speaker embedding model: Approximant and Nasal consonants

The third speaker embedding model is trained on 6 features, designed to reflect higher formants of vowel production, approximant, and nasal consonant articulation. The features used for the training of this speaker embedding model included the dynamics of higher fundamental frequencies such as F2, and F3. The features included the slope of second formant values across the input segment, the variation of each of these formants, and the average frequency difference between them. Moreover, spectral features such as spectral flatness and spectral centroid were included to signal the spreading of the described formants. This information is indicative of how approximants and nasal sounds are produced, as they are highly connected to the higher fundamental frequencies of vowel production (Ladefoged and Disner (2012)). The information included within this model were sourced from the frequency band 1000-3000Hz.

Fourth speaker embedding model: Fricative articulation

The fourth model is trained on acoustic cues represented in higher frequencies, indicating voice quality in fricative consonant production (Ladefoged and Disner (2012)). This speaker embedding model was trained on 11 explicitly extracted features. These features included the spectral flatness, centroid, and bandwidth of the higher frequency spectral space. Two more advanced spectral features were also utilized such as spectral kurtosis and spectral skewness, which model the sharpness of the pronounced ficatives, and the asymmetry of the spectral information within the identified segment. Moreover, energy based features such as the mean, standard deviation, and frame-wise energy difference were also included. Lastly, the onset strength of phoneme production was included within the feature set which measures the rapidity of phonemic change within this spectral band. These features were extracted from the frequency band 3000-8000Hz.

3.2.4 Spectral model

The spectral model consisted of four different models, chosen in accordance to the phonetic model, however, in this ensemble, the features used for speaker embedding model generation, were explicit but uninterpretable representations of voice quality.

MFCC based model:

The first model utilized MFCC representations of the audio input. These representations were extracted from the 0-16000Hz frequency band. The extracted MFCC representation resulted in a 13 dimentional input sequence, which is a standard dimension within the task of speaker identification. These 13 coefficients (the dimensions of the feature representation) were calculated from 400 time frames, which were extracted every 10 ms time window over the two second segments.

This model utilized a Time Delay Neural Network architecture, which is capable of the modelling of dynamic time sensitive information such as MFCC. This combination of input data and architectural design is widely used within traditional speaker embedding models (Chung et al. (2018)). The TDNN model contained three different stages within its neural architecture. Firstly, it uses 2 time sensitive convolutional layes, which aids the model is recognizing patterns within the time-sensitive input. Then the output of the convolutional layers are statistically pooled (average pooling) in order to create a vector representation suitable for the subsequent neural layers. Then the pooled output of these layers are passed through 2 fully connected neural layers, which are aimed to help the model identify the speaker specific patterns within the contextually identified patterns. The activation function used within all of the components of the model is ReLU. The output of the fully connected layers is the final speaker embedding. This model generated an embedding with 128 dimensions. Moreover, the model was trained on a 0.001 learning rate, and utilized 6 epochs. The data fraction used for this model was kept at 10% of the available training data.

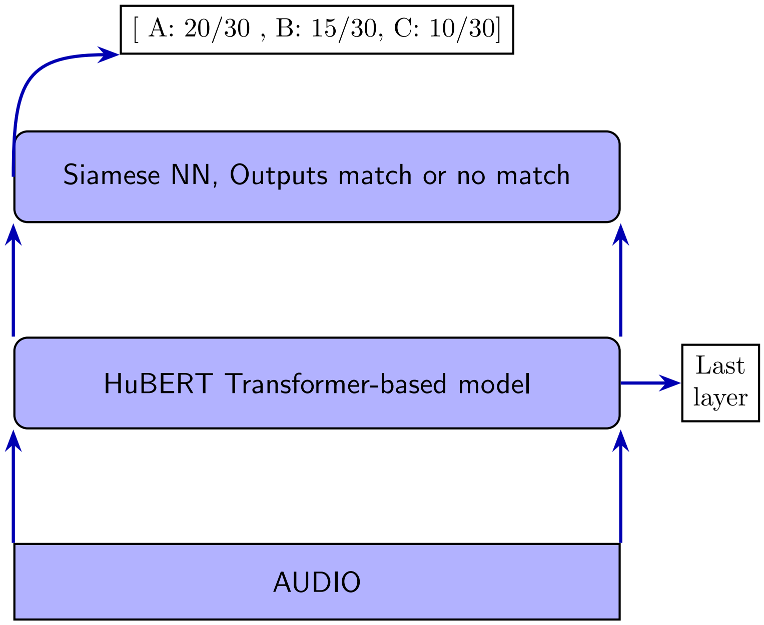
Specified Spectral Band Models:

The remaining three models in the spectral ensemble use identical Convolutional Neural Network (CNN) architectures, differing only in the frequency range of their input features. These ranges were aligned with those used in the phonetic model to support systematic comparison across both ensembles. Specifically, one model was trained on the 0–1000 Hz band, another on 1000–3000 Hz, and the third on 3000–8000 Hz.

The trained embedding models were trained using three convolutional neural layers, which took in the dimensions of frequency bins (determined by the frequency range), and the 200 timestamps which were distributed across the input sequence every 10ms in a two dimensional array. The dimensions of the subsequent convolutional layers were increasing from the initial input dimension until they reached 128 dimensions at the last layer, which was then projected to the desired output embedding . Moreover, these layers used the leaky ReLu activation function, and drop-out regularization in order to avoid vanishing gradients and overfitting. Lastly, the output of the third convolutional layer was max-pooled in order to generate the final speaker embedding. The final training configurations of the CNN models were identical to the TDNN model, thus they produced a 128 dimension output embedding, used 0.001 learning rate, and was trained over 6 epochs. The data fraction of the CNN model was also 10% of the training data, due to the high computational resources required by the architectural design.

3.2.5 Transformers based approach:

The transformer based model consists of two main computational phases. Firstly, a transformer based large speech model is used to extract embeddings for each audio segment. Then, a speaker verification head is attached to the transformer-based model, which through transfer learning utilises the embeddings to discriminate between them according to the encoded acoustic profiles. This verification step happens by the introduction of a Siamese Neural Network , which is trained for binary classification, predicting match or no match instances, in regards to the pair of instances belonging to the same speaker. Siamese Neural Networks are frequently used in speaker verification setups (Sang et al. (2022)), thus this classification head allows the current model to perform open set classification, through iterative verification over all of the enrolled speech segments. This section includes the explanation of the transformer based model, and methodology used for speaker embedding extraction, and the training of the Siamese NN verification system .


[A: 20/30 , B: 15/30, C: 10/30]
Siamese NN, Outputs match or no match
HuBERT Transformer-based model
Last layer
AUDIO

Hubert:

The transformers model used in the current project is the Hubert base model (Hsu et al. (2021)). Hubert is a transformer-based large speech model, which through unsupervised masked unit prediction, is trained to recognise both language and acoustic dependencies. Contrary to text based LLMs, during its pre-training phase, Hubert goes through an initial stage of unsupervised acoustic unit clustering. The resulting cluster labels are then used for the masked unit prediction task. The current study utilises the Hubert-base model which contains a parameter number of 95M. The specific architectural design of the Hubert model consists of convolutional waveform encoder layers, and a BERT encoder. The specific Hubert model used within this project was pre-trained on the 960 hours of audio recordings taken from the Librispeech dataset (Panayotov et al. (2015)).

Speaker embedding extraction:

The transformer based speaker embeddings were extracted from the last layer of the Hubert model. In extraction process the model received raw audio segments scaled between 0-16000Hz. The preprocessing stage for this data extraction process was identical to the ones used within the ensemble based approaches. The model received 2 second audio chunks, and after passing them through the model, the representations of the last layer were taken as the speaker embeddings.

The extracted speaker embedding representations were used as the training data for the speaker verification head. The total amount of data extracted equalled to 20 instances of 2 second segments for each speaker within the training dataset, which was 251 speakers. Out of these instances 10000 randomised embedding pairs were selected, which equals to 11,1 hours of training audio data.

Speaker Verification:

The extracted embeddings were used as the training data for the Siamese NN verification system . This system is designed as a transfer learning method, thus the weights of the transformer-based model are not altered or utilised within the verification step. This allows for an open classification setup, as required by the application constraints of the project.

The neural model chosen for this process was a Siamese NN. These models are frequenly used within the task of speaker verification (Sang et al. (2022)). The architecture of the current Siamese NN consists of a mulit-layer perceptron encoder, which maps the input into a reduced dimensional representation. As a pair of inputs are passed through the model, the resulting two embeddings from the MLP layer are compared against each other, by calculating the Euclidean distance between them. This distance then is projected into a final probability score through a sigmoid layer. The Siamese NN utilised a binary cross-entropy loss function. The model utilised the ReLu activation function, and it was trained using 0.001 learning rate, through 20 epochs.

From speaker verification to speaker identification:

The speaker verification model was used for the speaker identification task by iterative verification between the query instance and the enrolled representations. The Siamese NN produced match or no-match predictions for each pair of instances. Within each speaker cluster, the number of match instances were counted and divided by the total number of instances for that speaker. These ratios were used as the probability scores for speaker identification.

3.3 Experimental setup:

The above mentioned speaker identification models were compared in regards to two main factors; 1, their ability to generalise when scaled up, and 2, their ability to perform open-classification while keeping an option of inconclusive decisions. Both experimental phases utilised a grid-based variation setup of configuration variables. The variables, such as number of known speakers and number of instances per speaker were systematically varied across an increasing scale. Each of the variations of variables were used as a ‘knowledge-base’, or otherwise called enrolment setup for the systems, and the models performance of closed and open-set classification were assessed through the measures of precision, recall and abstention ratio scores.

3.3.1 First experimental setup:

The first experimental setup followed the core grid-based variation of configuration variables. These configurations are generated using a randomized grid generation system. This system selects a set number of speakers and samples a corresponding number of instances from each. To explore a wide range of performance conditions, the number of speakers is varied across the following values: 2, 5, 10, 20, 40, 60, and 80. The number of instances per speaker is varied across: 1, 5, 10, 20, 40, 60, and 80. These values were selected to support the systematic evaluation of how enrollment size affects the system. The only asymmetry in the configuration scaling between speakers and instances occurs at the lowest numbers. A minimum of 2 speakers is required for the system to make a distinction between known identities. However, only 1 instance per speaker is needed, since the system is not constrained by a fixed classification framework in this regard. The selected values also follow a direct growth pattern: from 5 to 80, each speaker count doubles relative to the previous one (e.g., 5 to 10, 10 to 20). This design allows for evaluation of the scalability of the model. If the system performs consistently across these increasingly large configurations, the model can be considered as highly scalable. The inclusion of 60 speakers provides an intermediate point, offering insights into system behaviour at a 50 percent increase beyond the prior configuration point. This structured scaling approach enables precise analysis of the model's capacity and robustness across enrollment sizes. Each model's precision plotted on a heatmap, designed to show the emerging impact that scaling had on performance.

3.3.2 Second experimental setup:

The second experimental setup built upon the methods used in the first experiment, however, with the addition of two entropy based thresholding mechanisms. This mechanism consisted of measuring the entropy value for the resulting list of confidence scores for any given query item. These scores were then used to calculate thresholds, through which the system can identify known, and unknown speakers at a set precision rate. The threshold calculation started with the generation of a dataset, which included query identifiers, the prediction label (whether the correct speaker had the highest confidence score) and the entropy calculated from the the top 5 speakers with the highest list confidence scores for each query. For each threshold, these datasets contains positive instances, for instance in the case of the 'known' threshold, top candidate was the correct prediction, and negative instances, where the top speaker prediction was incorrect. The threshold calculation happened through a scale from 1 to 0 with a changing rate of 0.01. At each threshold, instances were classified as positive or negative instances, according to where they were positioned relative to the threshold. Then the precision score for the positive instances was calculated. This iterative process was carried out until the found precision fell above the required value.

This threshold calibrating process was carried out for both thresholds, used to classifiy speak- ers as 'known' speaker, and 'unknown' speaker. At time of inference for the open- classification, these thresholds were used to deem the produced prediction as one of the three options, known, unknown, or null-prediction, when the entropy fell between the two thresholds. The results were plotted on a heatmap, showing the variation of precision, the recall of each label, the f1 score of the two prediction labels, and the ratio of null-predictions to all predicted instances.

Entropy calculation for different methods:

The entropy calculation differed across the two different modelling approaches, as their final predictions for each candidate speaker took different forms.

The ensemble based approaches utilised the confidence scores generated by the LambdaMart model. These scores are relative to each other within one query inference. However, the confidence scores can not be directly compared across different query instances, as they do not fall on the same scale. Due to this, the confidence scores were transformed through a normalisation process. Firstly, the speakers with the top 5 highest confidence scores were selected. These values then were min-max scaled between the highest confidence score and the value of the 6th highest confidence score. Lastly, these instances were projected onto a percentage scale, which allowed for a proportionate view on the confidence rate per candidate speaker. The entropy of these values were calculated, and then used for the thresholding mechanism.

In the case of the transformers based speaker identification model, the confidence scores used for the entropy calculation were the ratio of match instances to all known instance for the speaker. The entropy calculation of the entropy values comprised of the proportionate representation of the produced match labels. Thus, the confidence scores, were further divided by the total amount of matches the verification model produced for current query instance, and the entropy values of these normalised confidence scores were calculated.

Chapter 4

Results

This section entails the results yielded from the two conducted experiments. The first question investigated the impact of scaling variables on the performance of the three speaker identification models. In section () a brief overview of the experiment is given, alongside the proposed hypothesis for model internal and comparison behaviour. The following section delves into the specific yielded results for the modelling approaches, by examining the solitary and combined impact of the experimental variables. Lastly, the model performance is concluded and the relation of the yielded results to the previously set hypotheses is discussed.

The second part of the Results section entails the examination of the results of the second research question, which investigated the ability for open classification of the proposed speaker identification models. Similarly, to the examination of the first research question, this section is initialized by a brief overview of the experimental setup, and the previously set hypotheses. In the subsequent section the found thresholding paradigms are evaluated on the held-out test set. This research question is investigated in regards to the yielded precision metric within open and closed classification using the thresholding paradigm, and the usage of abstaining while the models make predictions.

4.1 First experiment:

The first experiment examined the effect of scaling enrollment variables on the model performance. Here two fundamental scaling factors were varied, namely, the number of speakers the model had to differentiate between, and the number of instances the model had access to in making the decision. These instances were varied across two scales, and all possible configuration setups were tested in a grid-designed experimental setup. The proposed hypotheses entailed opposing factors on the configuration variables, such as that the increasing number of speakers will result in a more complex classification setup, thus hindering the performance of the models. In contrast, the number of instances per speaker known by the model was hypothesized to aid the performance of the models, as it created more robust generalizations for the speaker profiles. Regarding the specific hypotheses posed for the modeling strategies, the spectral-based model was hypothesized to perform best, due to its adaptability to evolving speaker identities, and populated enrollment sets. The model based on phonetic features was hypothesized to show decreased performance compared to the other tested strategies, due to its low dimentionality, and sensitivity to uncontrolled data quality. Lastly, the transformers based model was hypothesized to show elevated performance in low complexity settings, with drastically dropping efficiency in high complexity enrollment sets, due to the increased amount of interfering information from the pre-training phase.

4.1.1 Spectral-based model

In the following section, firstly the spectral based ensemble model's performance is examined, then the phonetic-based ensemble model's performance is delved into, and lastly the performance of the transformers-based model is explained.

General Overview

The heatmaps in Figure illustrate the spectral model's performance across three distinct configuration sets. Although the same model was evaluated in all cases, each configuration used a different random selection of speakers and enrollment instances. This approach allowed the experiment to assess the model's behavior under varying conditions of incrementally scaled enrollment sets, simulating different levels of task complexity.

Across all configuration sets, the spectral model demonstrated consistently high performance, with precision values exceeding 0.9 in the majority of test cases. The largest observed difference in precision across the examined grids was 0.29, indicating relatively stable behavior even as the enrollment conditions varied. This stability suggests that the model successfully captured discriminative acoustic patterns, enabling reliable classification under challenging scenarios.

A substantial portion of the evaluated configurations resulted in perfect precision scores, particularly in setups involving 10 to 20 enrolled speakers. When the number of speakers exceeded this range, a slight decrease in precision was observed, generally remaining within a margin of 0.1. An exception to this trend occurred in a small subset of configurations where only a single enrollment instance was available for a large number of speakers, leading to a more noticeable drop in performance.

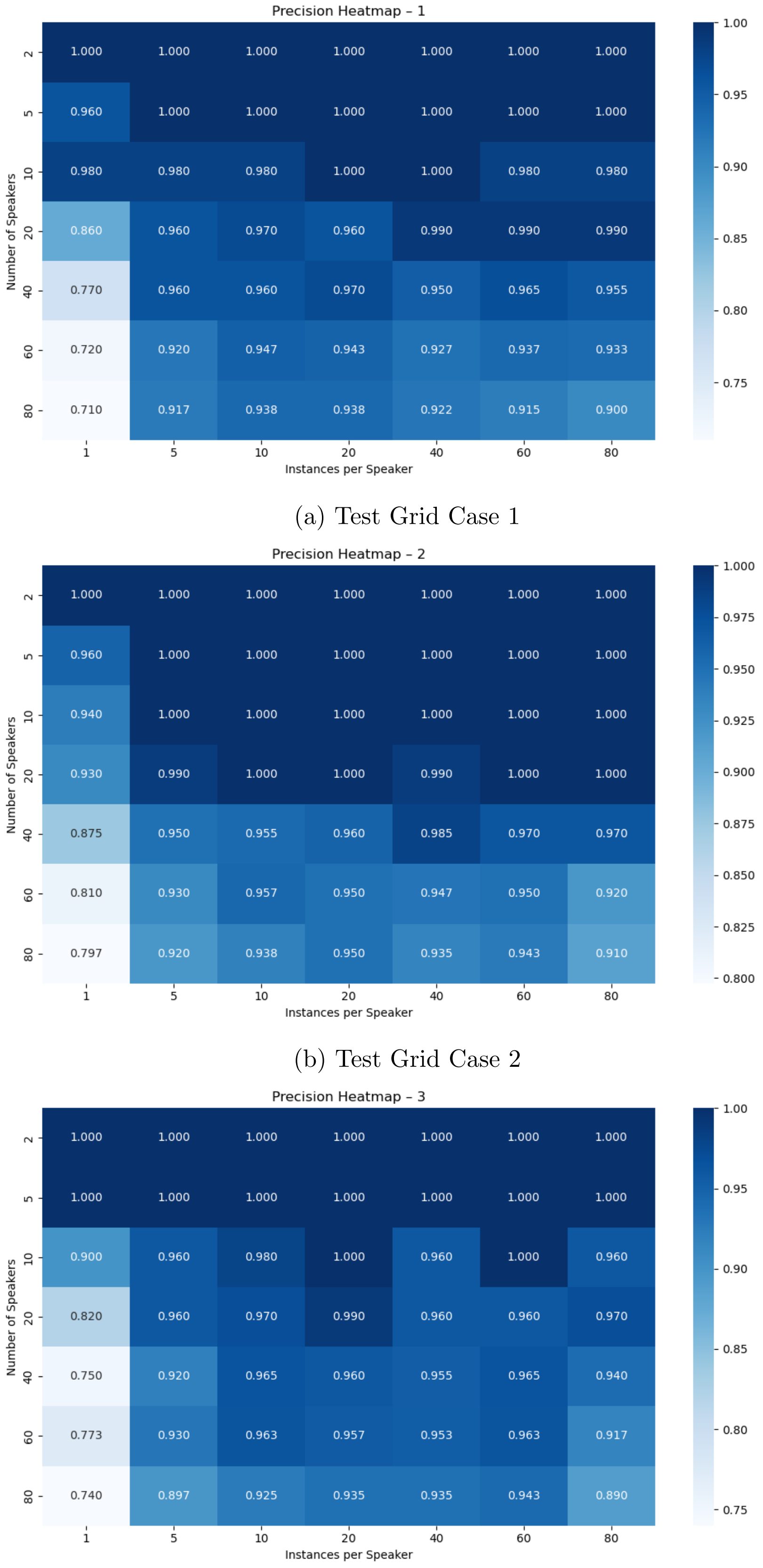
Effect of classification complexity

The number of enrolled speakers was introduced as an experimental variable to assess how increasing classification complexity affects the model's performance. As the number of speakers grows, the embedding space becomes more densely populated, reducing the separability between speaker clusters and theoretically making classification more difficult.

In the results, this effect is observable for the spectral model. Across the three evaluated configuration grids, a minor performance decline emerges beyond 10 to 20 speakers, with precision drops ranging between 0.01 and 0.06. These reductions are relatively small, suggesting that the model is capable of forming dense yet well-separated clusters in its learned representation space. Consequently, even as the classification task becomes more complex with a larger number of speakers, the spectral model maintains high performance, indicating that its embeddings are robust against increased speaker overlap.

Effect of exposure

The number of instances per speaker was varied to investigate its role in shaping the density and separability of speaker clusters in the embedding space. Intuitively, more



(a) Test Grid Case 1

(b) Test Grid Case 2

(c) Test Grid Case 3

Figure 4.1: Precision scores for the Spectral based speaker identification model within closed-set classification. The three plots show the results of the closed-set classification for each test grid case.

enrollment instances should strengthen cluster formation, improving decision boundaries and classification confidence. However, the effect of this factor in the spectral model proved to be limited and non-linear.

When only a single instance per speaker was available, clusters were minimal and underdefined, resulting in the most pronounced performance drops, particularly as the number of speakers increased. Between 10 and 60 instances, the model exhibited its highest and most stable precision values, suggesting that this range provides sufficient data to create robust and well-separated clusters.

Surprisingly, when the number of instances reached 80, a slight decline in precision was observed across all configuration grids. This suggests that overly populated speaker embeddings introduce noise or reduce inter-cluster separability, making the model less confident in its decisions. These results indicate that while increasing the number of instances generally benefits the model, there is an optimal range beyond which speaker clusters become more sparse. Thus, additional data hinders rather than helps classification performance above a certain threshold.

Combined effect of classification complexity and exposure

The interaction between the classification complexity and the exposure reveals that extreme variable conditions in the enrollment set negatively affect the spectral model's performance. When only minimal exposure is available, precision declines early, already from 5–10 speakers, as the singular representation lacks the information needed to maintain distinct cluster boundaries. At the other extreme, high saturation levels, particularly in large-scale setups with 60–80 speakers with increased exposure, also introduce performance degradation. Here, the embedding space becomes increasingly crowded, reducing separation between clusters and leading to less confident predictions.

Between these two extremes, the model maintains stable and elevated precision, indicating that moderate exposure provides the most reliable conditions for speaker identification. This suggests that the model is resilient to classification complexity overall but remains sensitive to the compounded effects of sparse data and limited cluster separation in high-complexity scenarios.

Hypotheses confirmation

The results of the spectral model largely support the original hypotheses, though with differing degrees of significance. The expected negative effect of increasing classification complexity is consistently observed. As the number of enrolled speakers grows, precision gradually declines, confirming that higher speaker counts make discrimination more challenging despite the model's overall robustness.

The hypothesized positive effect of increased exposure per speaker is only partially validated. While providing more enrollment instances improves performance compared to single-instance setups, the benefit does not scale indefinitely. At very high exposure levels, the model exhibits diminishing returns, and in some cases, precision drops. This suggests that beyond an optimal range of enrollment instances, the embedding space becomes saturated, reducing inter-cluster separation and leading to less reliable classification.

Overall, the results indicate that the spectral model is capable of maintaining high precision across most configurations. However, its performance is constrained at the extremes of sparse or overly dense enrollment conditions, showing that the two variables interact in non-linear ways to shape classification outcomes.

Representational Benefits of Spectral Information in Scalable Speaker Identification

The results highlight a key advantage of using spectral representations for scalable speaker identification: they enable the formation of dense, high-quality speaker embeddings that remain robust even under challenging conditions. With as little as a single enrollment instance, the spectral model produces reliable embeddings that effectively separate speakers, maintaining well-defined cluster boundaries across a range of classification complexities. This suggests that spectral features capture fundamental and distinctive aspects of speaker identity, allowing the model to generalize well without requiring extensive exposure to each speaker.

Moreover, the learned embedding space shows a capacity for gradual refinement as more enrollment instances are introduced. Additional exposure allows the model to consolidate speaker clusters, improving confidence and stability in most tested scenarios. However, this effect is not indefinite. At very high exposure levels and large-scale classification setups, the embedding space becomes increasingly populated, leading to reduced boundary separation and sparser clusters. This overpopulation effect introduces ambiguity in the decision space, limiting the model's ability to scale indefinitely without performance degradation.

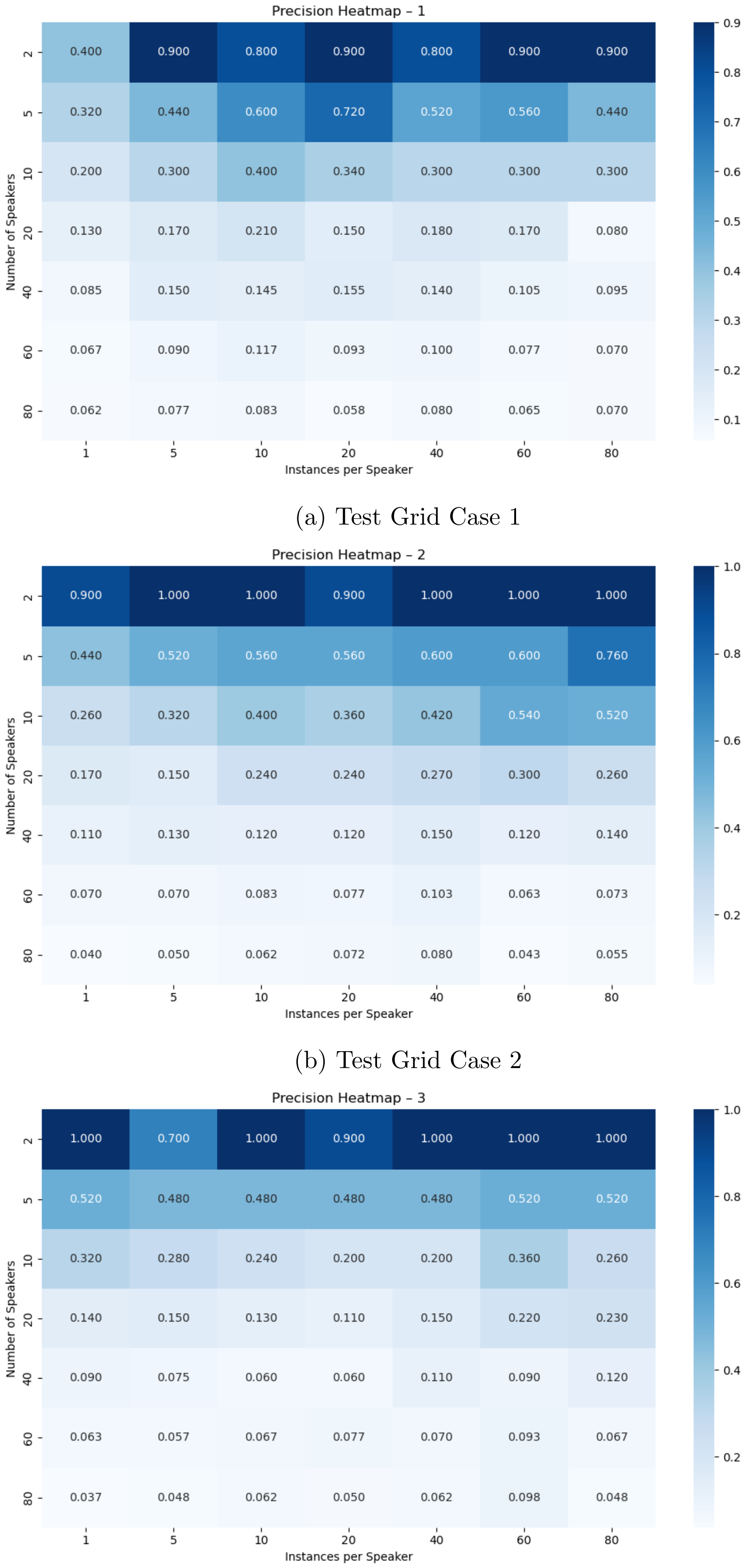
Overall, the representational benefits of spectral information lie in its ability to create compact, discriminative embeddings that remain resilient to increased classification complexity and limited enrollment data. While the approach offers strong scalability within a practical range, it ultimately faces constraints in extremely large and densely populated speaker spaces, where embedding quality and separability begin to deteriorate.

4.1.2 Phonetic Model

This section presents the results obtained from the ensemble model trained on specified phonetic features. The design of this model was motivated by its hypothesized ability to leverage stable phonetic cues, such as articulatory and prosodic tendencies, that might provide stable representations across different enrollment sets and offer discriminative power for speaker identification.

General Overview

Compared to the spectral-based model, the phonetic model yielded significantly weaker results, particularly as classification complexity increased. The model demonstrated some capacity to separate speakers in low-complexity setups (2–5 speakers), but this ability deteriorated rapidly as the number of enrolled speakers grew. Beyond 20 speakers, precision scores dropped to near-zero, indicating that phonetic representations lacked sufficient discriminatory power to maintain cluster separation in larger classification spaces. The performance of this model suggests that increasing number of speakers makes the internal representation of the model unstructured, which can not be improved by additional exposure to individual speakers.



(a) Test Grid Case 1

(b) Test Grid Case 2

(c) Test Grid Case 3

Figure 4.2: Precision scores for the phonetic based speaker identification model within closed-set classification. The three plots show the results of the closed-set classification for each test grid case.

The number of enrolled speakers had a pronounced negative effect on model performance. While low-complexity setups with very few speakers produced moderate precision scores, the introduction of additional speakers rapidly degraded precision. This trend suggests that the embedding space generated from phonetic cues becomes increasingly sparse and non-discriminative as classification complexity rises. Rather than forming well-separated speaker-specific clusters, the model appears to generate highly overlapping speaker clusters. Consequently, cluster boundaries blur at already moderate levels of complexity, leading to unstable and unreliable speaker identification.

Effect of Exposure

Varying the number of instances per speaker had no meaningful impact on the model's performance. In low-complexity configurations, exposure to additional samples neither improved nor degraded results, suggesting that the limited speaker-specific information present in the embeddings remained stable and contained few outliers. At higher complexity levels, where speaker clusters largely disappear and the model struggles to separate identities, increasing exposure similarly had no measurable effect. This indicates that once embeddings deviate from forming speaker-discriminative clusters, additional data does not refine the learned space, as it continues to encode patterns unrelated to speaker identity.

Combined Effect of Configuration Variables

The combined effect of classification complexity and exposure is not possible to identify for this model. At low complexity, performance remains relatively stable, and additional exposure does not significantly change results, indicating that the limited speaker-specific cues present are consistently encoded. As complexity increases, embeddings begin to converge into clusters that are not speaker-discriminative, making it impossible for the model to reliably separate identities. In these settings, further exposure to the same speaker labels does not improve performance, as the learned space no longer organizes data along meaningful speaker boundaries.

Hypothesis Confirmation

The first hypothesis, predicting a strong negative correlation between the model's performance and increasing classification complexity, is confirmed. As soon as the number of enrolled speakers increases beyond a minimal threshold, the model rapidly loses the ability to form clear, speaker-specific clusters, suggesting that the learned representations primarily cluster in non-speaker specific patterns.

The second hypothesis, predicting a beneficial effect of increased exposure to speaker identities, cannot be confirmed. Exposure does not improve performance even in low-complexity conditions, and at higher complexity, additional instances fail to counteract the breakdown of speaker boundaries in the embedding space. This indicates that the model lacks the capacity to strengthen or refine speaker-specific representations through repeated observations of the same speaker.

Conclusion

Specified phonetic representations fail to generate robust, speaker-specific embeddings that can scale to more complex identification tasks. While minimal classification complexity allows for some degree of speaker separation, this ability rapidly diminishes even at moderate complexity levels. As the number of speakers increases, the learned clusters begin to converge on non-speaker-specific information, blurring the boundaries between identities. This suggests that the representations encode diverse or content-driven patterns rather than stable, discriminative speaker cues. Consequently, this approach is not well suited for scalable speaker identification, particularly in settings where the diversity of the data or the limited discriminatory power of phonetic features prevents the formation of reliable speaker embeddings.

4.1.3 Transformer-Based Speaker Representations

The third evaluated model leverages a transformer-based approach to speaker identification, utilizing the HuBERT large speech model as a frozen embedding extractor. Unlike explicitly designed speaker models, HuBERT is a general-purpose self-supervised speech representation model trained on large-scale unlabeled audio. It captures rich acoustic features, including phonetic content, prosody, and speaker-specific cues, within a dense embedding space.

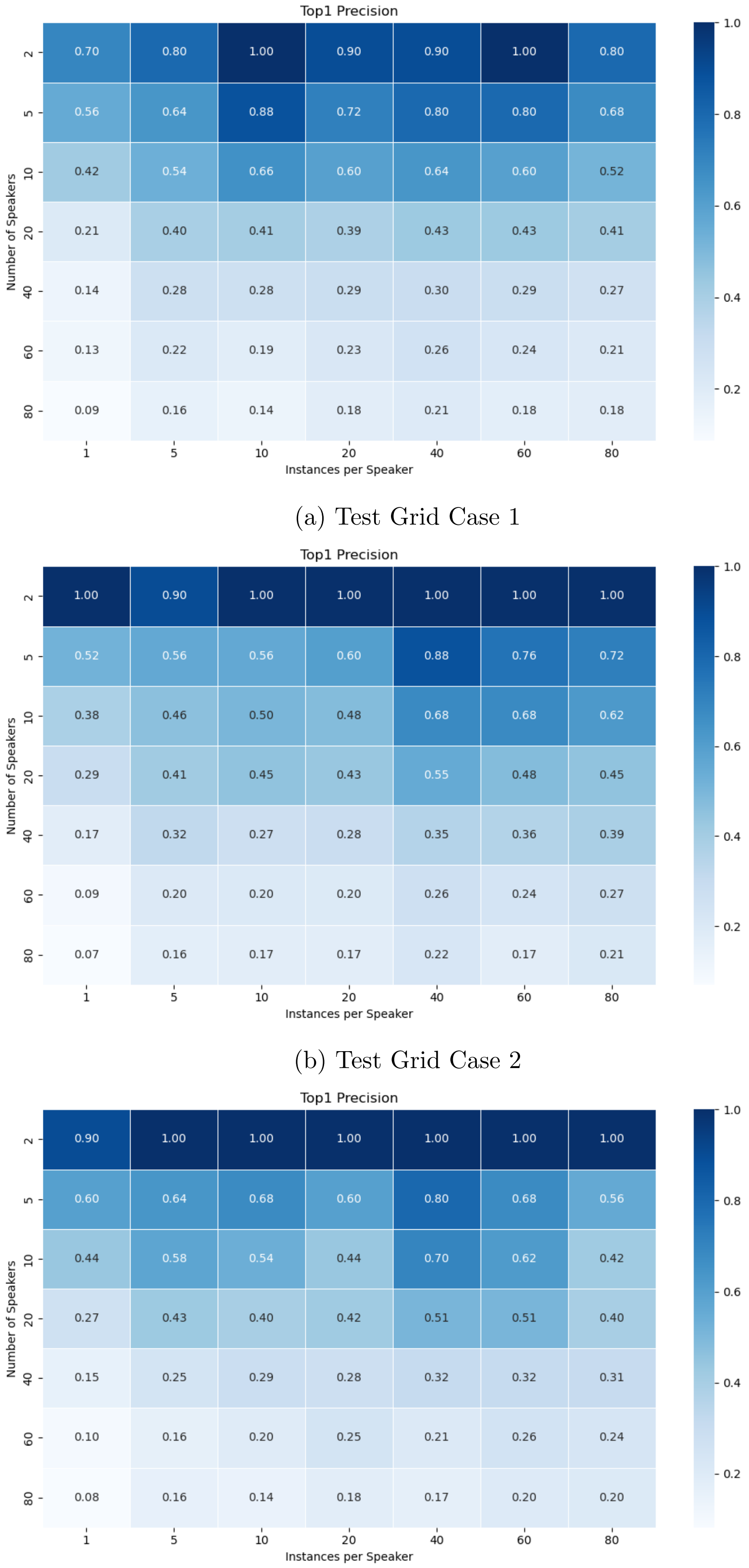
To evaluate this hypothesis, HuBERT embeddings were extracted for each speech segment and passed to a Siamese verification model that iteratively matched queries against enrolled speakers. The model's precision was assessed under varying complexity levels, defined by the number of enrolled speakers and the number of instances available per speaker. Three independent speaker sets were used to test the model, followed by an averaged result over all sets.

The transformer-based model using frozen HuBERT embeddings performs reliably in low-complexity conditions, where only a small number of speakers are enrolled. However, as the complexity of the task increases, the embeddings struggle to maintain clear speaker separation, leading to a noticeable decline in precision. These results suggest that the extracted representations mix speaker and content information, making it difficult for the model to form distinct, scalable speaker clusters as the enrollment set grows.

Effect of Classification Complexity

Across all tested configurations, the transformer-based model demonstrates a clear decline in performance as the number of enrolled speakers increases. This is shown by 4.3 While the model can form reasonably separable clusters in low-complexity conditions, adding more speakers rapidly deteriorates its ability to distinguish between identities. The results suggest that the HuBERT embeddings do not scale well in representing unique speaker characteristics when the speaker space becomes denser.

This behavior indicates that the extracted embeddings may not provide strong speaker-discriminative features and instead retain a high degree of overlap between speakers. As more speaker profiles are introduced, the boundaries between clusters blur, leading to increased confusion and lower identification precision. The model's performance plateauing at higher complexities further suggests that the representations encode weak speaker discriminative acoustic cues, limiting scalability in real-world, multi-speaker scenarios.



(a) Test Grid Case 1

(b) Test Grid Case 2

(c) Test Grid Case 3

Figure 4.3: Precision scores for the transformers based speaker identification model within closed-set classification. The three plots show the results of the closed-set classification for each test grid case.

Effect of Exposure

The results show that increasing the number of instances per speaker does not consistently improve the model's performance across different complexity levels. At lower complexity, additional exposure provides marginal yet noticeable gains in separability. For example, with five enrolled speakers, precision scores rise from around 0.60 to approximately 0.85–0.9. However the three test cases show varied ideal exposure values within this complexity level. This suggests that exposure allows the model to generate more robust speaker clusters, but its actual value is highly dependent on the individual test cases.

However, this effect does not persist when the classification task becomes more complex. As the number of enrolled speakers increases, the benefit of additional exposure quickly diminishes. Precision scores drop substantially, often below 0.2 even with 80 instances per speaker, indicating that the embeddings become more dispersed and overlapping in higher-complexity settings. This suggests that the model's representation space reaches a saturation point where more exposure does not reinforce cluster boundaries but instead contributes to a more entangled distribution of embeddings. While the 40–60 instance range appears to be an ideal exposure level in small-scale setups, this advantage vanishes as cluster separability breaks down under higher complexity conditions.

Combined Effect of Speaker Count and Exposure

The interaction between the number of enrolled speakers and the level of exposure per speaker reveals a compounding effect on cluster quality. At low complexity, clusters remain relatively compact and separable even with minimal exposure, indicating that the model encodes speaker specific information however, their discriminative power is limited. As the number of speakers increases, the embedding space becomes progressively more populated, leading to reduced separation between clusters.

Additional exposure in these high-complexity settings does not counter this effect. Instead, the increased data seems to enhance cluster sparsity, with representations dispersing and overlapping across multiple speaker boundaries. This compounding effect suggests that the model struggles to maintain distinct speaker-specific regions in the embedding space when faced with a large, diverse set of speakers. Consequently, the ability to scale to higher complexity configurations is limited by the combined degradation of both cluster separability and density.

Representational Benefits and Scalability Limitations

Transformer-based embeddings, such as those extracted from HuBERT, offer strong initial advantages for speaker identification. These representations capture rich, multidimensional acoustic patterns, allowing the model to differentiate speakers effectively under low-complexity conditions with relatively few instances. The dense and information-rich nature of these embeddings facilitates the formation of well-defined clusters, making them suitable for tasks where speaker sets are small to moderately sized.

However, as the speaker embedding space gets populated, the initially dense clusters begin to overlap, reducing the separability of clusters and leading to ambiguous decision boundaries. Additional exposure does not fully counteract this effect, as the model appears to encode overlapping content and speaker-related features simultaneously, resulting in sparse and non-speaker discriminative clusters at high complexity.

This indicates that while transformer-based representations can scale better than specified phonetic features, however, their ability to support large scale enrolment sets in speaker identification is limited.

Comparative Analysis of Model Scalability and Hypothesis Evaluation

The three evaluated models, spectral, phonetic, and transformer-based, demonstrated markedly different performance trends as classification complexity increased, revealing clear differences in their ability to scale to larger speaker identification settings. These differences closely align with the hypotheses that the representational strategy of each model plays a critical role in determining its scalability.

The spectral model, trained on broad acoustic representations, consistently showed the highest resilience to increased complexity. Its embeddings maintained well-defined speaker clusters even as the number of enrolled speakers increased, and exposure to additional instances improved the density of these clusters up to a certain point. Although performance degraded at the highest levels of complexity, the decline was more gradual compared to the other models, supporting the hypothesis that the flexibility of spectral features makes them more robust for scalable speaker identification.

In contrast, the phonetic model, relying on highly specified representations tied to articulatory and phoneme-level cues, showed the weakest scalability. While it produced clear speaker clusters in low-complexity settings, its performance dropped sharply as soon as the number of enrolled speakers increased. This suggests that the specified features encoded content or phoneme-specific information rather than reliable speaker-distinguishing characteristics, leading to sparse, poorly separated clusters under scaling. The results confirm the hypothesis that models situated at the highly specified end of the representation spectrum are more sensitive to classification complexity.

The transformer-based model using HuBERT embeddings occupied the other extreme, relying on highly generalized representations pre-trained on large-scale data. At low complexity, the model performed comparably to the spectral model, suggesting that generalized embeddings can capture some speaker-specific information. However, as the number of speakers increased, cluster separability deteriorated substantially, even at high exposure levels. The likely cause is that HuBERT embeddings encode entangled representations of speaker identity and content, leading to overlapping clusters and unstable decision boundaries under scaling.

Overall, these findings confirm both hypotheses. The spectral model outperformed the other two approaches under scalable conditions, showing the best trade-off between cluster density and separability. The phonetic and transformer-based models, performed significantly worse, than the spectral model. While the transformers based model did yield slightly elevated results compared to the phonetic model, it still lost its structured embedding space in high complexity settings. These results highlight that balanced, mid-level acoustic representations are more effective for scalable speaker identification, as they capture speaker-discriminative cues while remaining flexible enough to handle larger, more complex classification spaces.

4.2 Second Experiment

The second research question investigates how the three evaluated models perform under open-set classification conditions using a two-way thresholding mechanism that allows abstention when available information is inconclusive. This analysis does not solely focus on whether the top-ranked predicted candidate matches the true speaker but also examines how the distribution of confidence scores across the top-five predicted speakers evolves as the enrollment setup is scaled up. Particular attention is given to how these confidence patterns shift with increasing numbers of enrolled speakers and instances per speaker, revealing changes in cluster separability and overall decision certainty. Additionally, the evaluation explores how abstention operates for both known and unknown speaker labels, capturing whether models correctly abstain from making predictions when information is insufficient or avoid misclassifications under uncertainty. By comparing these dynamics across the three representational strategies, this research question aims to assess not only their raw precision but also their reliability, confidence behavior, and robustness to scaling in open-set speaker identification.

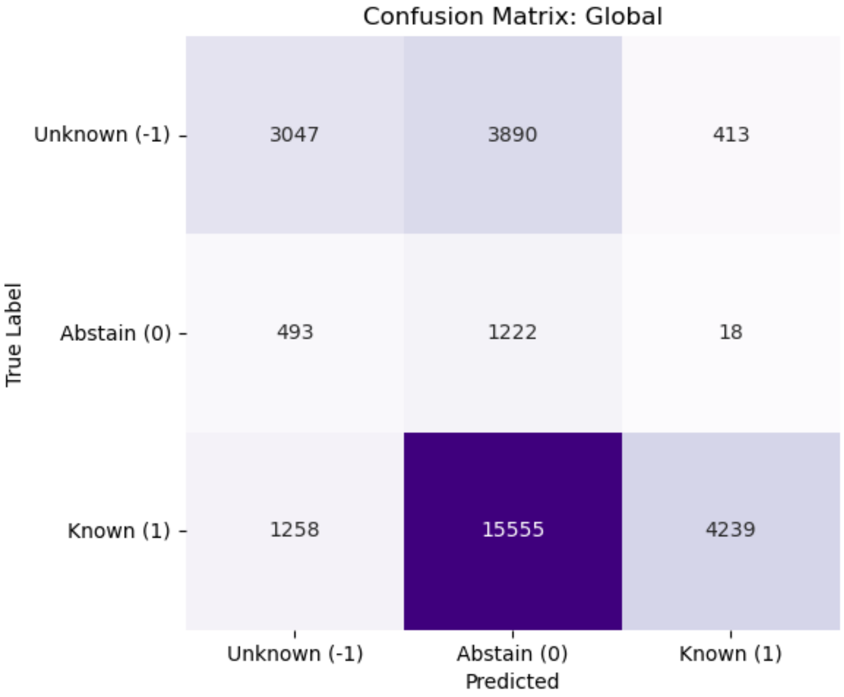


Figure 4.4: Confusion matrix of the spectral based model in open-set classification.

4.2.1 Spectral Model

The confusion matrix(4.4) provides a global overview of the spectral model's prediction behavior. While it correctly identifies 4,239 known speaker instances and 3,047 unknown speakers, the model becomes overly cautious in closed-set classification, abstaining on the majority of cases where the true label is known (15,555 abstentions). This suggests that the abstaining mechanism is heavily overused, prioritizing precision at the cost of recall. In contrast, within open-set classification, the model demonstrates overconfidence, frequently predicting unknown labels for instances that could have been abstained or correctly classified (493 abstentions and 1,258 known labels wrongly predicted as unknown). As a result, the abstaining function fulfills its intended purpose for known speakers by reducing misclassifications but fails to provide the same level of benefit for open-set scenarios. Overall, open-set classification appears inefficient for this model, with a tendency to over-rely on the unknown class and abstain on correct closed-set predictions.

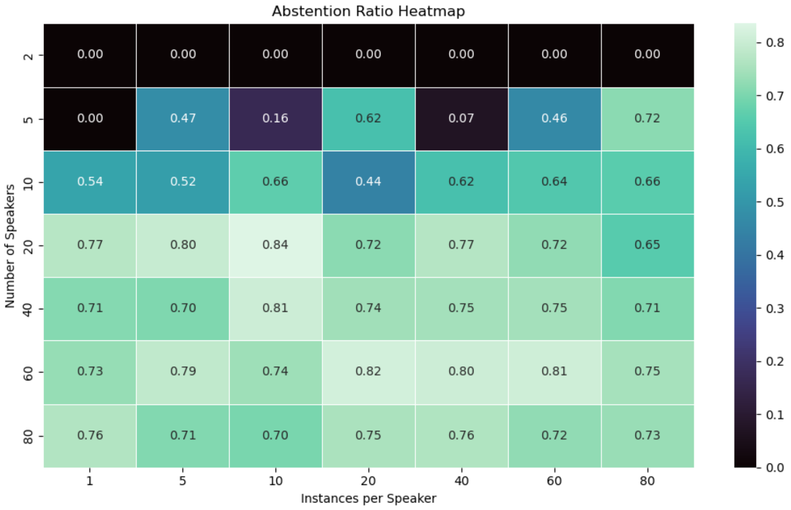


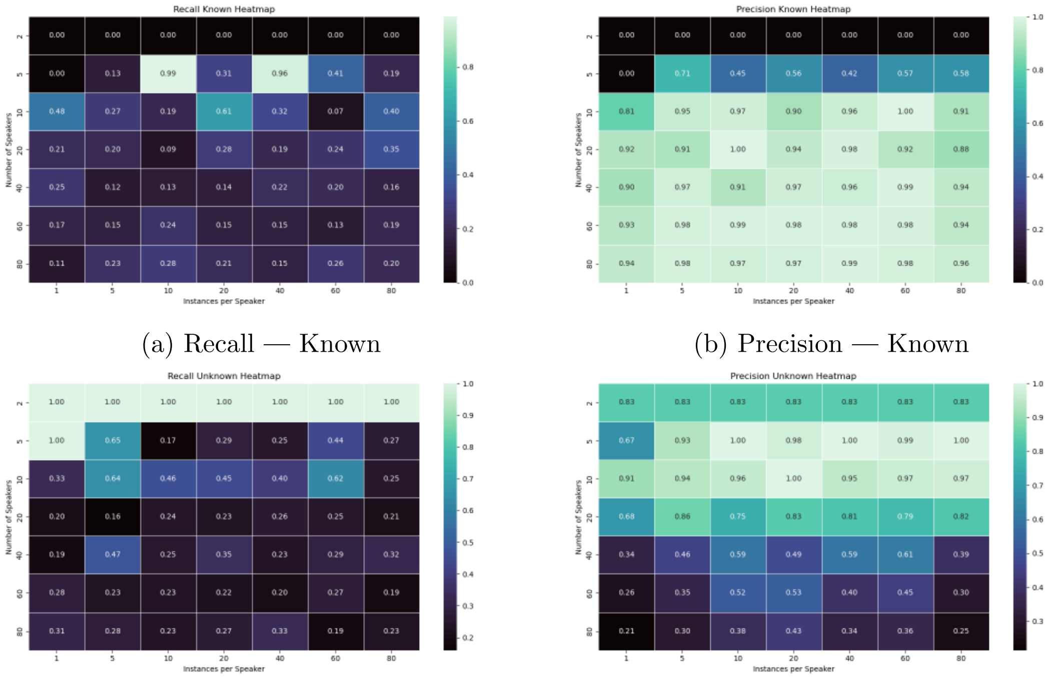
Figure 4.5: Abstention ratio heatmap for the Spectral based model.

Abstention Behaviour

As shown in the abstention ratio heatmap (4.5), the model heavily relied on abstention in closed-set classification, particularly in high-complexity enrollment settings where the number of speakers and instances per speaker increased. In many configurations, abstention rates exceeded 70%, leading to a notable drop in recall for known speakers. Interestingly, in lower-complexity setups such as 5×10 or 5×40, recall improved due to a reduced abstention rate, suggesting that abstention became disproportionately dominant as speaker clusters grew denser (see 4.6). Conversely, abstention had a minimal impact on open-set classification recall. Even in configurations with high abstention rates, unknown speakers were often misclassified as unknown with high confidence rather than being deferred, highlighting an overconfidence bias towards the unknown class and a failure of the abstention mechanism in moderating uncertain predictions for this label.

Precision-Recall Trade-off

The confusion matrix highlights two main error patterns that are also evident in the abstention and precision–recall results. Many known speaker instances are either misclassified or abstained on, especially with few instances per speaker or a larger speaker set, indicating limited discriminative power. Similarly, some unknown speakers are incorrectly classified as known, showing that the thresholding mechanism struggles in dense embedding spaces. These patterns align with the abstention heatmap, where abstention is high in low-data settings but does not consistently improve recall as data increases. Precision for known speakers reaches high values (¿0.9) with sufficient data, yet recall remains low due to frequent abstentions or errors. Precision for unknown labels drops sharply as the number of speakers grows, reflecting poor rejection of unknown inputs. Overall, the model relies heavily on abstention to maintain precision but fails to improve recall or effectively handle complex speaker distributions.



(a) Recall — Known

(b) Precision — Known

(c) Recall — Unknown

(d) Precision — Unknown

Figure 4.6: Precision and Recall values for the Known and Unknown labels from the results of Spectral based model.

Precision for known speaker identification (4.6) remains consistently high across most configurations, often exceeding 0.9. This suggests that, despite the abstention mechanism becoming overcautious, it successfully fulfilled its purpose of maintaining high reliability on accepted predictions. However, precision was noticeably lower in some low-complexity settings, likely due to the thresholding mechanism failing to establish meaningful entropy-based separation when only a few speakers were present. With limited information about how confidence scores were distributed in these sparse conditions, the model struggled to find thresholds that met the target precision set for this experiment. Recall for known speakers (4.6a) remained low throughout, particularly in configurations with more than 10 speakers, highlighting the model's strong bias toward abstention and its prioritization of precision over completeness.

For unknown speakers, precision (4.6d) was initially high in low-complexity settings but exhibited a sharp drop as the number of enrolled speakers increased. Recall for unknown speakers (4.6c) followed a similar downward trend with rising classification complexity. This degradation supports the hypothesis that the number of enrolled speakers is a dominant factor affecting open-set classification performance. As the embedding space became more densely populated, speakers with characteristics similar to the inferred new speaker likely occupied neighboring regions, reducing the separation in confidence distributions between known and unknown labels. This overlap limited the model's ability to confidently identify previously unencountered speakers, thereby lowering both precision and recall in more complex scenarios.

Entropy Dynamics and Representation Limitations

This model relies on MFCC and spectrogram-derived features, which primarily encode low-level spectral patterns such as formant structure and energy distribution. While these features are effective for clustering speakers in closed-set conditions, their limited discriminative capacity across highly diverse speakers constrains their use in open-set classification. As the number of enrolled speakers grows, the spectral representation space becomes crowded, with overlapping regions between similar-sounding voices.

Because these features capture broad acoustic similarities rather than highly distinctive speaker cues, embeddings for unseen speakers are often projected close to existing known clusters. The model, trained to favor confident assignments in such regions, produces low-entropy predictions even when the input should be identified as known. This leads to overconfidence in incorrect matches and reduces the contrast in entropy distributions between known and unknown instances. As a result, the abstention mechanism struggles to find a clear decision boundary. This limitation stems not only from the thresholding strategy but also from the representational nature of the spectral features themselves. While the generated embeddings provide strong, speaker-specific representations that excel in closed-set classification, they are less suitable for scalable open-set conditions. As the number of enrolled speakers increases, the embedding space becomes denser, reducing the margin between clusters. This results in a drop in confidence concentration for known instances, where predictions become less decisive, and a drop in confidence distribution for unknown instances, where the model fails to assign appropriately uncertain (high-entropy) scores to the enrolled speakers. Instead, the model often produces overconfident predictions even for unknown inputs, eroding the entropy-based separation that the abstention mechanism relies on. This degradation causes inappropriate rejections of known speakers or misclassifications of unknown speakers, ultimately limiting the model's ability to maintain reliable performance in high-complexity open-set classification scenarios.

4.2.2 Phonetic model

The confusion matrix (4.7) highlights fundamental weaknesses in the model's internal representations. In the closed-set task, only 653 known instances were correctly classified, while 1,673 unknown instances were misclassified as known and 4,802 incorrect known predictions were accepted with high confidence. This indicates that the embedding space does not form distinct or reliable speaker clusters, resulting in confidence scores that do not meaningfully distinguish correct from incorrect predictions.

In the open-set task, the model heavily overgeneralizes towards the unknown label, misclassifying 9,937 incorrect known instances and 1,405 true known instances as unknown. Despite this conservative tendency, false acceptances remain frequent, suggesting that the confidence distributions lack clear contrast between known and unknown labels. The abstention mechanism is therefore compensating for poorly structured embeddings, rather than leveraging well-separated confidence patterns to guide decisions.

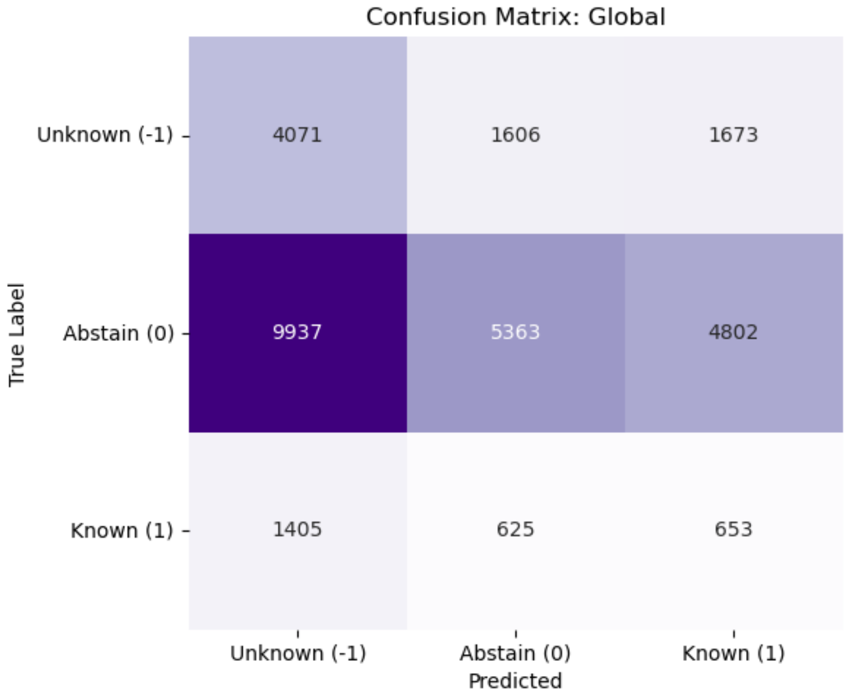


Figure 4.7: Confusion matrix of the phonetic based model in open-set classification.

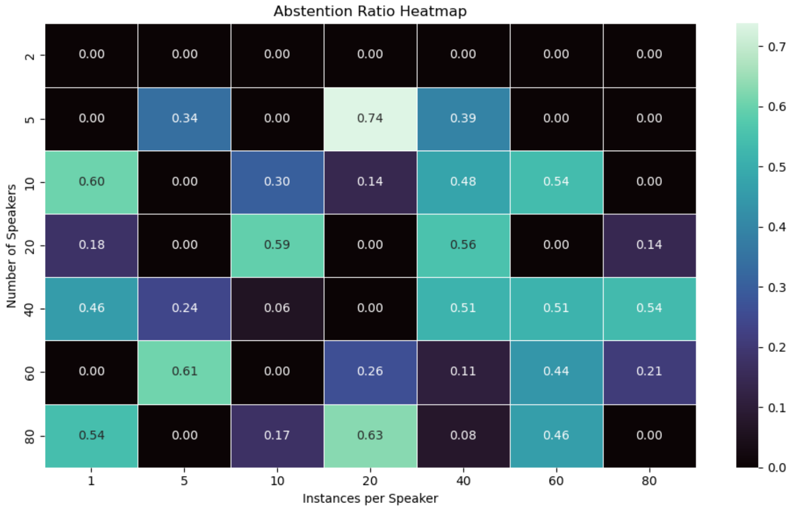
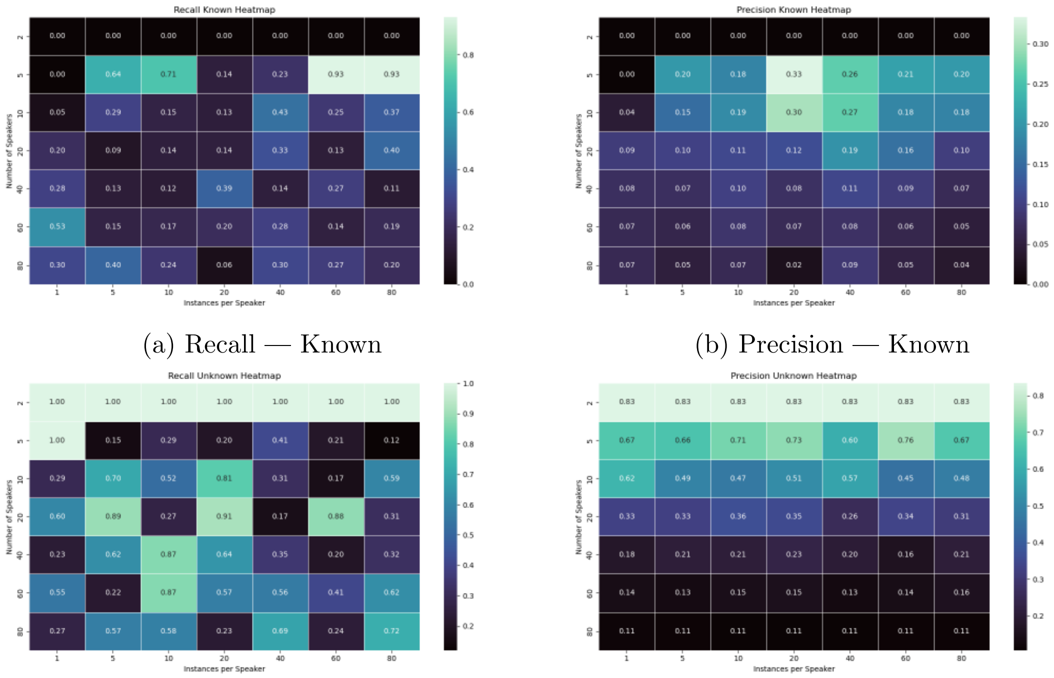


Figure 4.8: Abstention ratio heatmap for the Phonetic based model.

Abstention Behavior

The abstention ratio heatmap (5.10a) highlights the unstable decision dynamics of the phonetic-based model. Unlike the spectral model, where abstention scales predictably with task complexity, this model shows unstructured behavior, exhibiting high abstention in some low-complexity conditions (e.g., 0.74 in 5-speaker, 20-instance setups) while nearly abandoning abstention in more challenging settings, despite poor overall precision.

This inconsistency suggests that the abstention mechanism is not guided by well-formed confidence distributions. Ideally, entropy-based thresholding would separate uncertain from confident predictions, but the model's internal speaker representations fail to provide this contrast. As a result, abstention decisions become disconnected from task difficulty, signaling that the underlying embeddings do not encode reliably separable speaker information to support conservative classification.



(a) Recall — Known

(b) Precision — Known

(c) Recall — Unknown

(d) Precision — Unknown

Figure 4.9: Precision and Recall values for the Known and Unknown labels from the results of Phonetic based model.

Precision and Recall Dynamics

The precision and recall heatmaps for known and unknown speakers (4.9) illustrate the model's inability to maintain stable performance as enrollment complexity increases. Precision for known speakers fluctuates strongly across configurations and often collapses in scenarios with more than 20 enrolled speakers. Even in richer data conditions, misclassifications remain frequent, indicating that the model struggles to reliably separate speaker embeddings clusters. Recall values are consistently low, rarely exceeding 0.3 beyond small-scale setups, highlighting that the model fails to recover correct speaker identities even when abstention is not triggered.

For unknown speakers, precision initially appears promising in low-speaker configurations, where values approach 1.0, but it deteriorates sharply as more speakers are added. Recall follows a similar downward trend, suggesting that the model increasingly mislabels unknown inputs as known when the embedding space becomes more crowded. This pattern implies that the rejection mechanism lacks robustness and does not scale effectively with the number of enrolled speakers.

Overall, the phonetic model demonstrates weak discriminative power and poor boundary enforcement. Unlike the spectral model, which traded recall for consistently high precision, the phonetic model fails to achieve either, indicating insufficient representational capacity to form reliable speaker clusters or reject unfamiliar speakers effectively.

Impact of Specified Representations on Open-Set Classification

The phonetic representations used in this model appear to be poorly suited for the demands of open-set speaker identification. While designed to capture stable, speaker-specific acoustic features, these representations lack sufficient discriminative structure to separate speakers in complex enrollment scenarios. As the number of enrolled speakers increases, the embedding space becomes crowded, leading to frequent overlaps between speaker clusters. This results in high confusion rates for known speakers and an unreliable rejection mechanism for unknown inputs. Instead of supporting robust boundary formation, the representations encourage overgeneralization, causing both misclassification and ineffective abstention. Consequently, the model cannot scale effectively to more realistic open-set conditions, making these representations inadequate for the current application context.

4.2.3 Transformers based model

The confusion matrix (4.10) shows that the transformer-based model struggles to maintain consistent decision boundaries under open-set conditions with abstention. A large proportion of inputs that should have been labeled as abstain, namely incorrect known instances where the model is expected to withhold a decision, are instead classified as unknown. This indicates that the model fails to reliably separate confidence patterns between incorrect known cases and genuinely unseen speakers. Similarly, a notable share of truly known instances are misclassified as either abstain or unknown, suggesting that the model cannot anchor its representations strongly enough to differentiate between high-confidence matches and ambiguous or out-of-distribution inputs. While the model does attempt to reject uncertain cases, this behavior appears to be largely indiscriminate, collapsing both ambiguous known and unknown inputs into the same rejection category. This inability to form distinct confidence boundaries undermines its effectiveness in open-set classification, leading to a lack of reliable abstention behavior and frequent misclassifications across all label types.

Abstention Behavior

The abstention ratio heatmap (4.11) highlights inconsistent abstention behavior across different enrollment configurations. Ideally, abstention should predominantly occur in cases of uncertainty, allowing the model to withhold incorrect predictions, particularly for mislabeled known instances or genuinely unknown speakers. However, the results

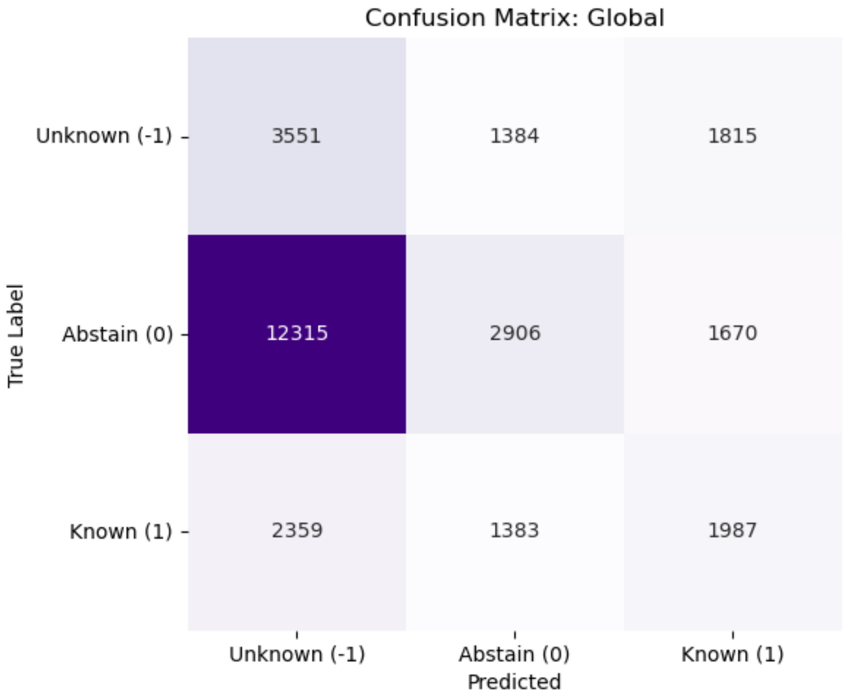


Figure 4.10: Confusion matrix of the transformers based model in open-set classification.

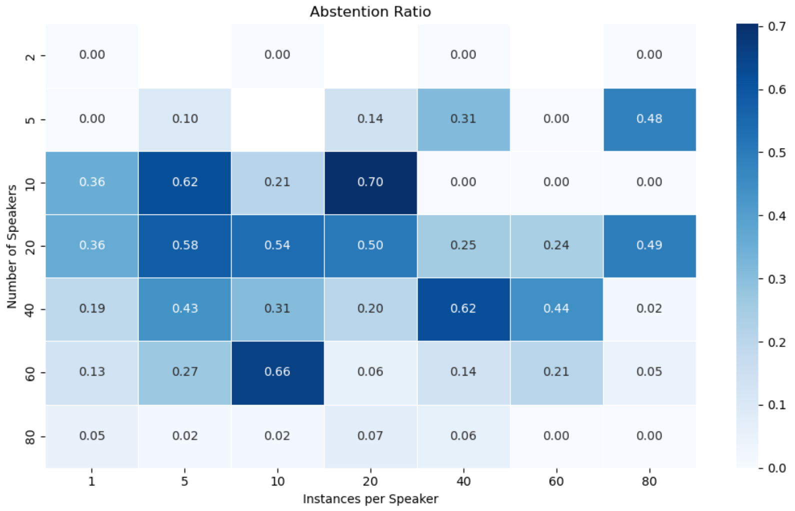
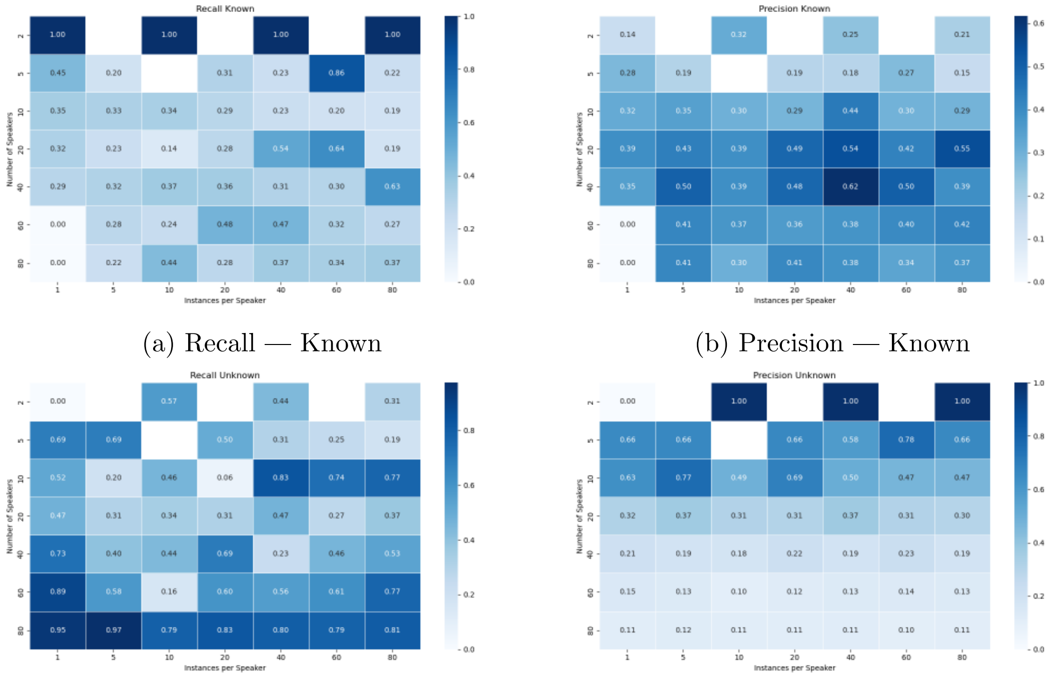


Figure 4.11: Abstention ratio heatmap for the Transformers based model.

indicate that abstention rates fluctuate sharply between conditions, without a clear relationship to the number of speakers or instances per speaker. In some settings, particularly at mid-range configurations (e.g., 10–20 instances per speaker or 20–40 speakers), abstention ratios exceed 60–70%, suggesting overly cautious behavior where the model frequently avoids making decisions altogether. Conversely, other settings show near-zero abstention, even when higher ambiguity would be expected.

This inconsistency mirrors the earlier confusion matrix results, where ambiguous known instances were often misclassified as unknown rather than being correctly abstained on. These patterns suggest that the transformer-based embeddings fail to produce well-separated confidence distributions for known, unknown, and ambiguous cases. Instead, abstention decisions appear largely configuration-dependent and erratic, undermining the reliability of the abstention mechanism in managing open-set uncertainty.



(a) Recall — Known

(b) Precision — Known

(c) Recall — Unknown

(d) Precision — Unknown

Figure 4.12: Precision and Recall values for the Known and Unknown labels from the results of Transformers based model.

Precision-Recall Trade-off

The precision and recall heatmaps reveal that the transformer-based model struggles to maintain a consistent performance in the open classification task. Precision for known speakers remains mostly below 0.6 across all configurations, even with larger amounts of enrollment data. This indicates a persistent difficulty in correctly identifying known speakers without introducing false positives, suggesting that the model's representations lack sufficient discrimination in these conditions. Precision on unknown speakers is higher in simpler setups (fewer speakers, fewer instances) but degrades rapidly as task complexity increases, implying that the model becomes less effective at rejecting nonenrolled speakers as the embedding space grows denser.

Recall values for known speakers are similarly low, rarely surpassing 0.6 and often dropping substantially in larger-scale configurations. This reflects the model's tendency to miss correct identifications altogether, abstaining or misclassifying instead. In contrast, recall for unknown speakers is somewhat better in large-scale conditions, frequently exceeding 0.6 and approaching 0.8 in high-enrollment setups. This discrepancy suggests that the model more reliably avoids misidentifying unknown speakers than it does at correctly accepting known ones.

Overall, these results indicate that while the transformer-based embeddings allow for moderate separation between known and unknown instances, they fail to create a confident and reliable decision boundary for known speaker identification. The increasing confusion with task complexity points to the model's inability to form stable speaker clusters in high-dimensional open-set conditions, leading to low precision and recall across the board.

Suitability of Generalised Representations for Open-Set Speaker Classification

The results obtained from the transformer-based model, trained on generalized representations, indicate that this representational strategy is not well-suited for open-set speaker classification. Across all tested configurations, the model struggled to maintain high precision and recall for known speakers, while its ability to correctly abstain or reject unknown speakers degraded as task complexity increased. Even in lower-complexity conditions (few speakers, few instances per speaker), precision rarely reached optimal levels, and with increased enrollment size, the system's decision boundary between known and unknown speakers became increasingly blurred.

This behavior can be attributed to the inherent qualities of the generalized data representations extracted from the transformer model. These embeddings are designed to capture a broad range of acoustic and linguistic information simultaneously, balancing speaker-specific cues with other factors such as phonetic content, prosody, and environmental noise. As a result, speaker identity is encoded in a highly compressed, mixed-information feature space, where separability between speakers is not explicitly optimized. When the classification task moves into an open-set setting, this lack of speaker-focused encoding becomes problematic: the model cannot form clear clusters for enrolled speakers, and the similarity scores between known and unknown inputs become overlapping and ambiguous.

Moreover, the generalized embeddings exhibit high intra-speaker variability and limited inter-speaker distinctiveness under challenging conditions. This is evident in the low precision scores for known speakers, where even additional enrollment data does not significantly improve discrimination. In effect, the model treats speaker identity as one of many latent factors in the embeddings, meaning that adding more samples from the same speaker does not lead to the strong, stable representations needed for reliable open-set identification.

Overall, while generalized representations offer flexibility and robustness for downstream tasks such as speech recognition or diarization, they appear fundamentally limited for open-set speaker classification. Their broad, non-specialized nature lacks the fine-grained, discriminative power required to confidently distinguish known speakers from unknown ones, particularly as the number of enrolled speakers grows and the decision space becomes more complex. This limitation suggests that open-set speaker identification systems require representations explicitly optimized for speaker distinctiveness, rather than relying on generalized acoustic embeddings.

Chapter 5

Discussion

Based on the experimental results, several patterns emerged in the performance of the three tested models as the classification task was scaled by the number of speakers and the number of instances per speaker. Models positioned at the extremes of the specified–generalized representation spectrum struggled to maintain high performance under increased complexity, likely due to their limited ability to form robust and well-separated speaker clusters as the data space became more crowded. In contrast, the model situated in the middle of this spectrum, using spectral representations achieved consistently higher performance across all configuration setups. These models showed only a minor negative impact from increasing classification complexity, maintaining strong performance even as the number of speakers grew. They also demonstrated a gradual improvement with increased exposure to speaker data; however, this improvement was not linear, eventually showing diminishing returns as the amount of available data continued to rise.

The observed results suggest that two primary negative effects influenced model performance, both stemming from the experimental variables. First, increasing classification complexity by introducing more speakers reduced the separation between speaker clusters within the embedding space. As the number of enrolled speakers grew, clusters became less distinct, leading to decreased model performance in closed-set classification.

Second, the number of enrollment instances per speaker had a more nuanced effect. In the spectral model, higher exposure initially improved cluster density, but at higher complexity levels and with large amounts of data, performance began to degrade. This pattern suggests that as exposure increases, clusters become denser up to a point, after which excessive data introduces more variability and outliers, causing clusters to become sparser and less well-defined.

Overall, the combination of decreasing cluster separability due to a growing number of speakers and increasing sparsity caused by high exposure appears to be the dominant factor driving performance degradation as the task scaled up.

While these governing factors have been identified based on the observed results, their specific effects remain hypothetical, inferred mainly from the degradation in performance as the enrollment set scaled up. To gain a deeper understanding of how the representations were influenced by these scaling variables, it is necessary to examine what actually occurred within the speaker embedding space. Specifically, this analysis aims to uncover trends and tendencies linked to the negative effects introduced by increased complexity and exposure.

5.1 Error Analysis - First Experiment

This error analysis focuses on identifying general trends in the embedding space rather than highlighting differences in individual classification instances. Since the research question examines the overall ability of the tested representational strategies in both closed- and open-set classification, focusing on isolated test cases would only reflect the specific abilities of the models in those scenarios, without revealing the broader patterns that emerge as more data is introduced into the system.

To assess these trends quantitatively, three measures are calculated directly from the embedding space to capture different aspects of cluster structure and their relationship to classification performance. The separability score measures how well the model distinguishes speakers by comparing the distance between clusters to the spread within them. For each speaker, a centroid is computed, the average distance of embeddings to their own centroid gives the intra-cluster spread, and the average distance between centroids gives the inter-cluster distance. The score is the ratio of inter- to intra-distance, with higher values indicating compact, well-separated clusters. The cluster density metric captures the compactness of each speaker's cluster by computing the average pairwise distance between all embeddings belonging to the same speaker, then averaging across speakers; lower scores indicate tighter, more internally consistent clusters, while higher scores indicate greater internal spread. The overlap score combines the effects of separability and density by measuring the proportion of embeddings that lie closer to another speaker's centroid than to their own. High overlap scores signal that clusters are either too close together, too scattered internally, or both, which are the conditions that increase the risk of misclassification.

5.1.1 Spectral Model

In this section of the error analysis, we focus on the spectral model to examine how scaling negatively affected its performance. Results indicate that the compounding impact of increased classification complexity and exposure to instances influenced this model, as it shows a slight decrease in performance within highly saturated variable conditions. While the spectral model generally performed well, forming dense, indicative, and separable speaker clusters, its performance gradually declined as the number of speakers grew and the number of instances reached a certain threshold.

Separability analysis

Looking first at cluster separability as the number of speakers increased, the MFCC-based speaker embedding model exhibited relatively low but stable separability. This stability within the separability values indicate that the MFCC based model showed overall lowered ability to generate distinct speaker clusters. In contrast, the spectrogram-based speaker embedding models maintained higher separability as the number of speakers grew, suggesting that these embeddings created more distinct speaker clusters. This property likely contributed to the overall strong performance of this model variant.

Density Analysis

When examining the effect of increasing enrollment instances per speaker, different patterns emerged. MFCC representations remained highly dense regardless of exposure,

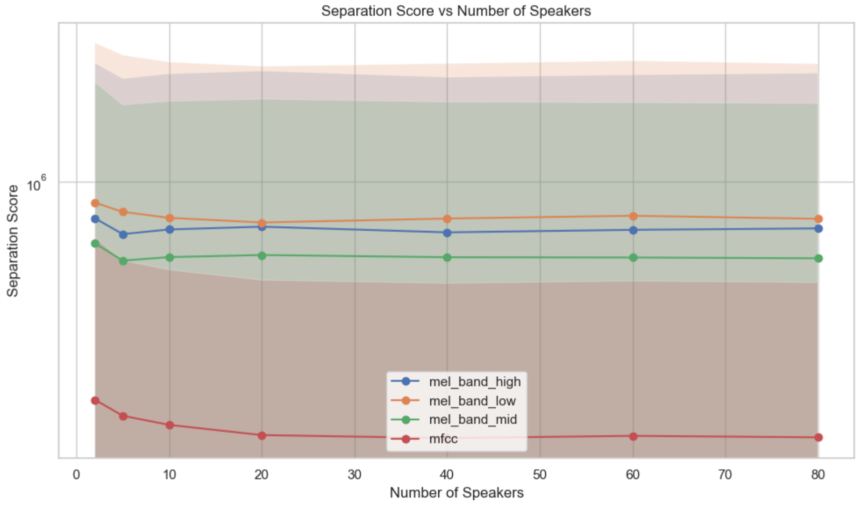


Figure 5.1: Separability score for Spectral based model.

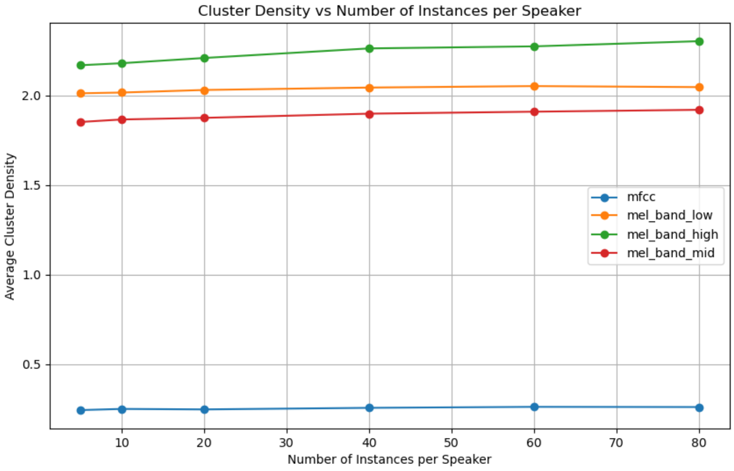


Figure 5.2: Density score for Spectral based model.

showing minimal sparsity even as more data was introduced. Spectrogram-based representations, however, displayed higher sparsity at low exposure levels, which increased further as the number of instances rose. This suggests that spectrogram models utilized a larger embedding space with more sparsely distributed clusters, while MFCC models operated in a more compact space with lower separability but consistently high density.

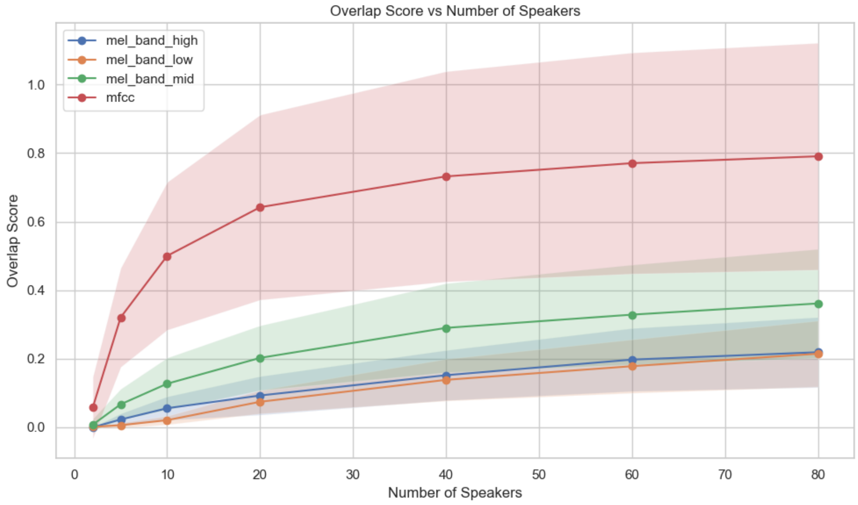


Figure 5.3: Overlap score for Spectral based model.

Overlap Measure

When examining the overlap scores between speaker clusters, a consistent relationship with the previously observed separability and density patterns becomes apparent. MFCC-based embeddings, which already demonstrated lower separability and higher density, showed the highest and most rapidly increasing overlap scores as the number of speakers grew. This suggests that while MFCC representations form tightly packed clusters, the proximity between clusters leads to substantial boundary overlap, reducing their discriminative power at scale. In contrast, the spectrogram-based models maintained much lower overlap scores, aligning with their higher separability values and sparser cluster distributions. These embeddings appear to better preserve distinct decision boundaries, even as complexity increases, though overlap still gradually rises with scaling. This interplay between overlap, separability, and density reinforces the interpretation that MFCC models operate in a compressed representational space where inter-cluster distances are limited, while spectrogram-based models leverage a broader embedding space that delays, but does not eliminate, the onset of saturation effects.

These findings indicate that the two types of spectral representations were influenced differently by the scaling factors. The combined effects of reduced separability under high complexity and increased sparsity under high exposure likely contributed to the overall performance decline of the ensemble model. Nonetheless, the tendencies observed suggest that spectral models generally provided stable and robust representations, with degradation mainly caused by saturation effects at high scaling levels. Moreover, these insights also suggest that spectral representations which undergo a limited amount of feature abstraction, utilise the embedding space better, as they show larger separability, and a slight increase in cluster sparsity as the enrolment setups are scaled up. As the current experiment only tested conditions with complexity and exposure levels up to the number of 80 instances or speakers, the saturation level did not reach the critical level for this model where cluster boundaries start to diminish and provide ambiguous enrolment scenarios. However, the initial performance decline of this effect is visible on the yielded results, thus in even higher saturation levels, the drastic decline of this model can be expected.

5.1.2 Phonetic-Based Model

The phonetic-based model showed clear signs of inability in maintaining a well-structured embedding space as scaling progressed. Patterns across the density, overlap, and separability plots indicate that the model struggled to preserve distinct and stable speaker clusters when faced with increased exposure and classification complexity.

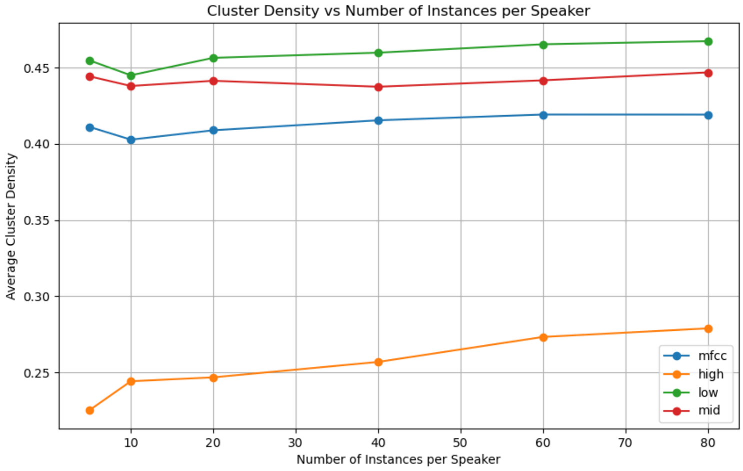


Figure 5.4: Density score for Phonetic based model.

Density Analysis

Cluster density remained relatively stable for some configurations but displayed steady increases for others as the number of instances per speaker grew. This rise in density reflects greater internal spread within clusters, suggesting that embeddings for the same speaker became less compact with additional exposure. A loss of compactness makes it harder for the model to keep speaker representations clearly defined, even if they remain somewhat consistent under lower exposure conditions.

Overlap Measure

Overlap scores increased sharply as the number of speakers rose, with all feature variants converging toward high overlap levels in high-complexity conditions. This indicates that embeddings from different speakers increasingly intruded into one another's space, reducing the margin between clusters. Such overlapping effect is a strong sign that the model's representation space was becoming saturated and less capable of separating speakers as the population grew.

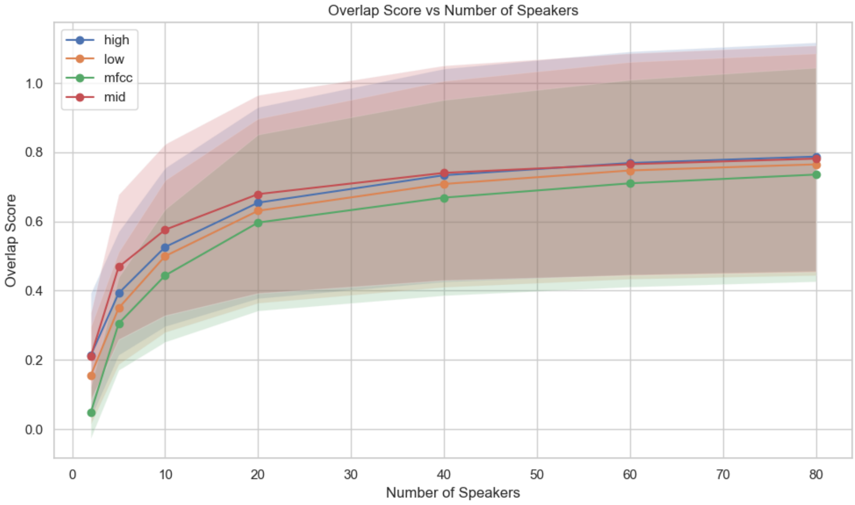


Figure 5.5: Overlap score for Phonetic based model.

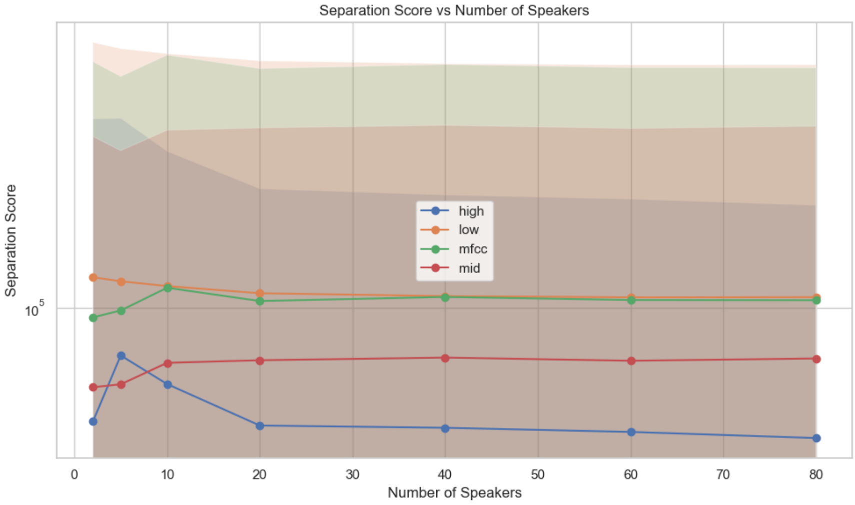


Figure 5.6: Separability score for Phonetic based model.

Separability Analysis

Separability scores followed a downward trend for all feature variants as speaker numbers increased. Initial differences between configurations diminished over scaling, with scores converging to similarly low values. This reduction shows that the relative distance between speaker clusters, compared to their internal spread, decreased substantially. The resulting loss of clear boundaries between clusters further supports the observation that the model was unable to maintain a structured, discriminative embedding space under high saturation.

These insights from the error analysis metrics clearly reveal the underlying reason for the specified model's low and sharply declining performance in closed-set classification. As scaling increased, the progressive breakdown in cluster structure, through higher internal spread, greater overlap, and reduced separability, directly eroded the model's ability to assign speakers to the correct identity, leading to the steep performance drop observed in the results.

5.1.3 Transformers based Model

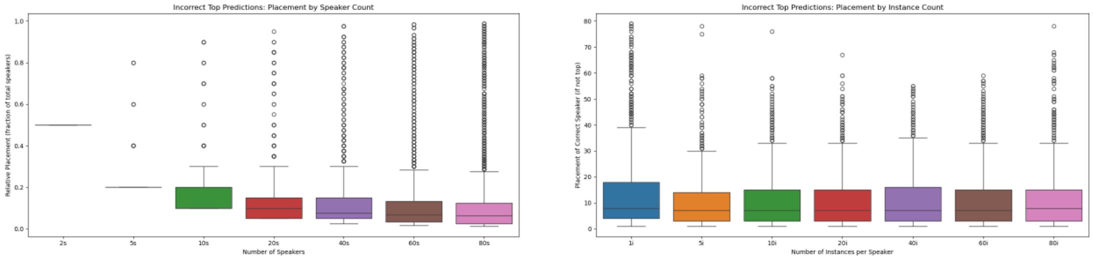
Applying the same metric-based error analysis used for the spectral and phonetic-based models to the transformer-based model presents significant challenges due to the nature of its architecture. This system employs a Siamese network trained for iterative speaker verification, meaning that its task is to identify indicative information from implicit transformer embeddings and classify speakers based on their relative match rate against other enrolled speakers. Unlike the explicit feature-based approaches, the internal representations and decision boundaries of this model are not directly interpretable, and the match decisions emerge from a complex similarity mapping process within the Siamese architecture. As a result, cluster-level metrics such as separability, density, and overlap, which are reliant on transparent embedding space geometry, cannot be meaningfully applied here.

Behavioural Analysis

Instead, the error analysis for the transformer-based model focuses on its behavioural tendencies, observed through a different metric: the placement of the correct speaker within incorrect prediction cases. This metric captures how close the model came to correctly identifying the speaker when its top prediction was wrong, offering insight into the relative ranking of the correct speaker in such scenarios. By examining these placements across different enrollment configurations, it is possible to assess whether the model's confidence, and its relative closeness to correct identification, changed as the number of instances per speaker grew. This perspective allows for a behavioural interpretation of the model's performance under scaling, even when the internal embedding structures remain uninterpretable.

The analysis of incorrect top predictions reveals that classification complexity, driven by both the number of enrolled speakers and the number of available instances per speaker, has a measurable effect on how closely the model ranks the correct speaker to the top position.

The first plot, which shows *relative placement by number of speakers*, demonstrates that the model is generally capable of placing the correct speaker within the top 20% of the ranked list across all tested complexities. This indicates that the generative embeddings contain sufficiently indicative information to consistently narrow the search



(a) By speakers (relative position)

(b) By instances (absolute position)

Figure 5.7: Placement of correct speaker in incorrect predictions, by speakers (first plot, relative position) and by instances (second plot, absolute position).

space to a small proportion of the enrolled speakers. However, as the number of speakers grows, this top-20% range still corresponds to a larger absolute placement value, meaning that in practice the correct speaker is pushed further back within the list. This pattern suggests that the model's understanding is strong enough to isolate a speaker group with similar embedding characteristics, but not consistently precise enough to identify the exact match at the top-1 position.

The second plot, which examines *absolute placement by number of instances per speaker*, provides further insight into how gradual exposure affects ranking precision. While increasing the number of available instances does result in a modest reduction in placement variance, especially in lower-instance settings, the median placement does not show a strong downward trend. This suggests that while additional examples help the model stabilise its placement behaviour, the underlying ranking mechanism still faces challenges in extracting enough discriminative information from the generative embeddings to consistently promote the correct speaker to the top-1 position. The gains from added exposure therefore appear limited to refining the pool of "close match" candidates, rather than decisively improving top-rank accuracy.

Taken together, these results show that the transformer-based model, when operating on generative embeddings in an iterative verification setting, can reliably isolate a subset of candidates that share strong similarity with the query. However, the indicative nature of the information it extracts is not sufficient to consistently elevate the correct speaker to the very top of the ranked list, particularly as classification complexity increases.

5.1.4 Results of Error Analysis - Performance in Scaled Scenarios

The patterns observed across all three data representation approaches indicate that models situated at either extreme of the representational spectrum, specified or generalized, struggle to support scalable speaker identification. Their encoded information is effective in low-complexity conditions but degrades as more speakers are introduced. In the specified model, this degradation likely stems from reliance on phonemic and articulatory patterns themselves, rather than on speaker-specific variation between these patterns. Similarly, the generalized model, while performing reasonably well in simpler settings, failed in high-complexity conditions, potentially grouping instances based on content, demographic cues, or paralinguistic factors such as mood or speech rate, rather than true speaker identity.

By contrast, the spectral model, particularly the spectrogram-based speaker embedding model, maintained better separation and density in the speaker embedding space even under increased complexity and exposure. Its advantage likely came from using less manipulated, fine-grained representations of raw audio data, which provided a richer and more flexible basis for scalable speaker identification. The superior performance of the spectrogram-based model over the MFCC-based model, which used only 13 frequency bins, supports this conclusion.

Overall, the error analysis reveals that the internal structure of the tested representations plays a critical role in how effectively a model can scale in closed-set speaker identification. Representations that maintain a well-separated, flexible embedding space enable the system to retain discriminative power as the number of enrolled speakers and the complexity of the classification task grow. Conversely, overly specified or overly generalized representations tend to produce embedding spaces where speaker clusters are less distinct or less stable, leading to performance degradation when scaling. For the context of a communicative robot, where speaker enrollment is expected to expand over time and interaction data can be highly variable, this suggests that minimally processed, fine-grained representations with sufficient dimensional capacity are better suited. Such representations allow the model to adapt dynamically to new data without collapsing decision boundaries, thereby maintaining high identification accuracy in scalable, real-world deployments.

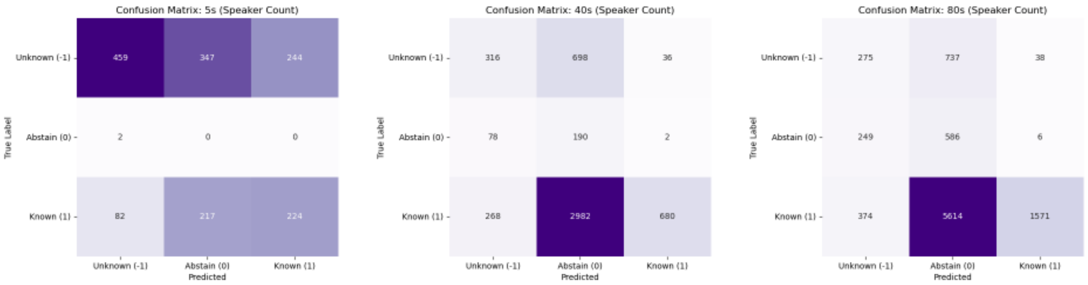
5.2 Error Analysis - Second Experiment

In the Open Classification task, similar patterns emerged: models situated at the extremes of the representational spectrum continued to perform poorly, while the spectral model, positioned in the middle, achieved comparatively better results. This outcome aligns with expectations, as the two extreme models had already demonstrated an inability to form robust and scalable speaker clusters in the closed classification task, making their lower open classification performance unsurprising.

However, the open classification experiment also revealed that the confidence rates and distributions of predictions changed noticeably with increasing complexity, even for the spectral model. Unlike in closed classification, where overall performance metrics masked these shifts, the analysis of confidence distributions among the top five speaker candidates highlighted a degradation in the model's ability to abstain from incorrect predictions. As complexity grew, entropy values between known and unknown speaker classes began to converge, reducing the distinct boundary required for reliable thresholding. This diminishing separation suggests that scaling adversely affects not only performance but also the confidence distribution essential for effective open classification.

5.2.1 Spectral Model

The confusion matrices for the 5, 40, and 80 speaker setups reveal a clear trend in the spectral model's abstention behaviour. In the smallest setup (5 speakers), abstentions on known speakers are already comparable to the number of correct classifications, showing that the model's conservative decision-making is present even at low complexity. By the time the enrollment set reaches 40 and 80 speakers, the abstention predictions on known speakers become much more frequent relative to correct classifications, indicating that the model increasingly refrains from committing to a known



(a) 5 speakers

(b) 40 speakers

(c) 80 speakers

Figure 5.8: Confusion matrices for 5, 40 and 80 speakers based on the results from the Spectral based model.

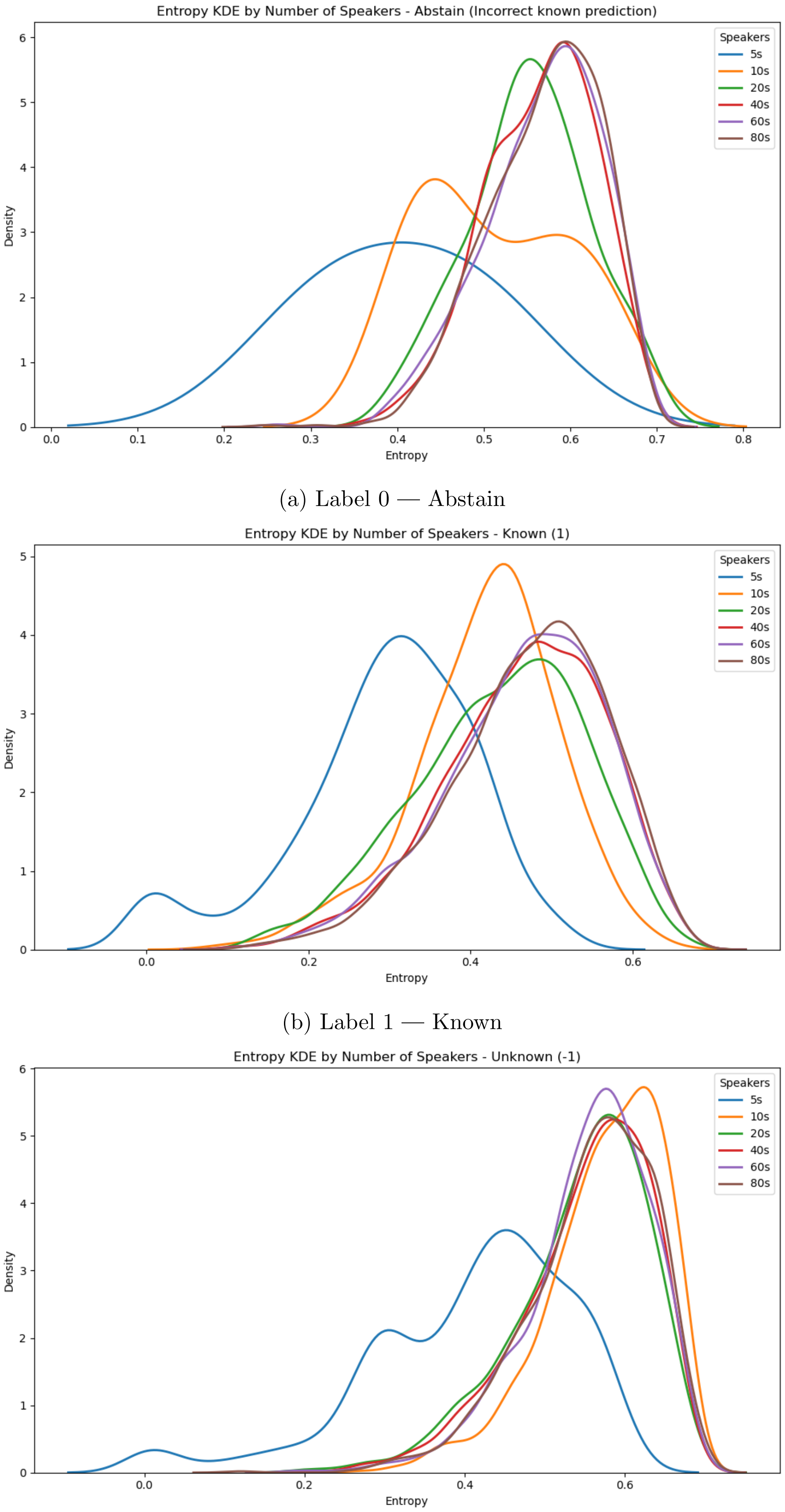
label as complexity grows. This progression suggests that, under higher classification demands, the spectral model's decision boundaries in the embedding space shift towards rejecting uncertain predictions, prioritising caution but at the cost of recall.

In open classification, the spectral model demonstrated the ability to correctly identify unknown speakers under low-complexity conditions. However, this capability diminished as complexity increased. The expected separation between entropy values for known and unknown speakers narrowed, making it harder for the model to reach the threshold required for reliable unknown classification. Analysis of the confusion matrix revealed that many false positive known predictions, incorrectly identifying a speaker as another enrolled speaker, were frequently labeled as unknown. This indicates that as complexity grows, entropy distributions for unknown speakers, correctly identified known speakers, and incorrectly identified known speakers begin to converge. Consequently, the model struggles to maintain distinct confidence patterns, leading to degraded open classification performance despite maintaining strong closed classification results.

To better understand the patterns observed in this model, specifically its inability to perform successful open classification and its overly cautious closed-set classification, it is necessary to examine how entropy distributions evolve as the number of enrolled speakers increases. The three KDE plots show entropy distributions separately for unknown instances, abstained cases that were in fact incorrect known predictions, and correctly classified known instances.

In low-complexity setups (e.g., 10 speakers), the entropy distributions for known and unknown cases are clearly separated, with unknown predictions concentrated at higher entropy values and known predictions at lower ones. Abstained instances also form a relatively distinct middle-ground distribution. However, as the number of enrolled speakers increases, these patterns progressively collapse towards a similar range. From around 20–40 speakers onward, the entropy peaks for known, unknown, and abstained cases begin to overlap heavily, reducing the model's ability to set a reliable threshold for distinguishing between these categories.

This convergence indicates that, with higher complexity, the model's generative embeddings lose the clear confidence separation seen at smaller scales. As a result, the abstention mechanism increasingly triggers on cases where the entropy does not meaningfully differ from correct known predictions, explaining why the model abstains on many instances that could otherwise have been correctly classified. The narrowing gap between the curves ultimately reflects a breakdown in the discriminative power of



(a) Label 0 — Abstain

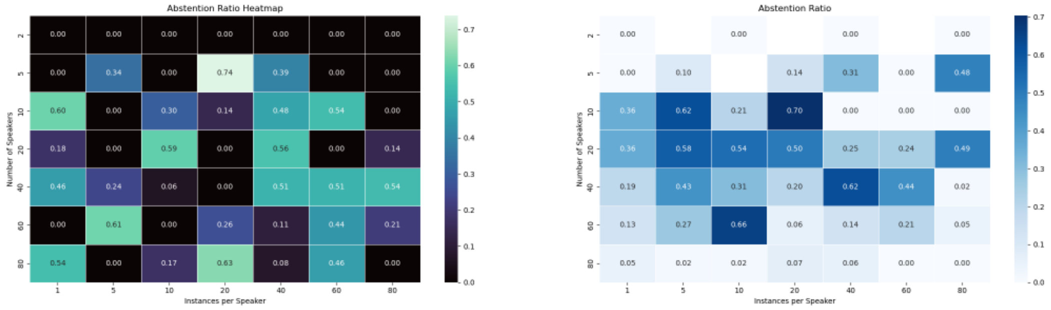
(b) Label 1 — Known

(c) Label -1 — Unknown

Figure 5.9: Entropy distribution as number of speakers are increased for the labels Known (1), Unknown (-1) and Abstain (0).

entropy as a decision signal, especially in high-complexity enrollment setups.

One possible explanation for this predictive behavior is that with more enrolled speakers, the likelihood of finding one that closely resembles an unknown speaker increases, producing higher confidence scores for incorrect matches within the top predictions. As a result, the model generates increasingly similar confidence patterns for both known and unknown instances, making it unable to maintain a clear separation threshold and thus limiting its effectiveness in open classification tasks.



(a) Phonetic based model

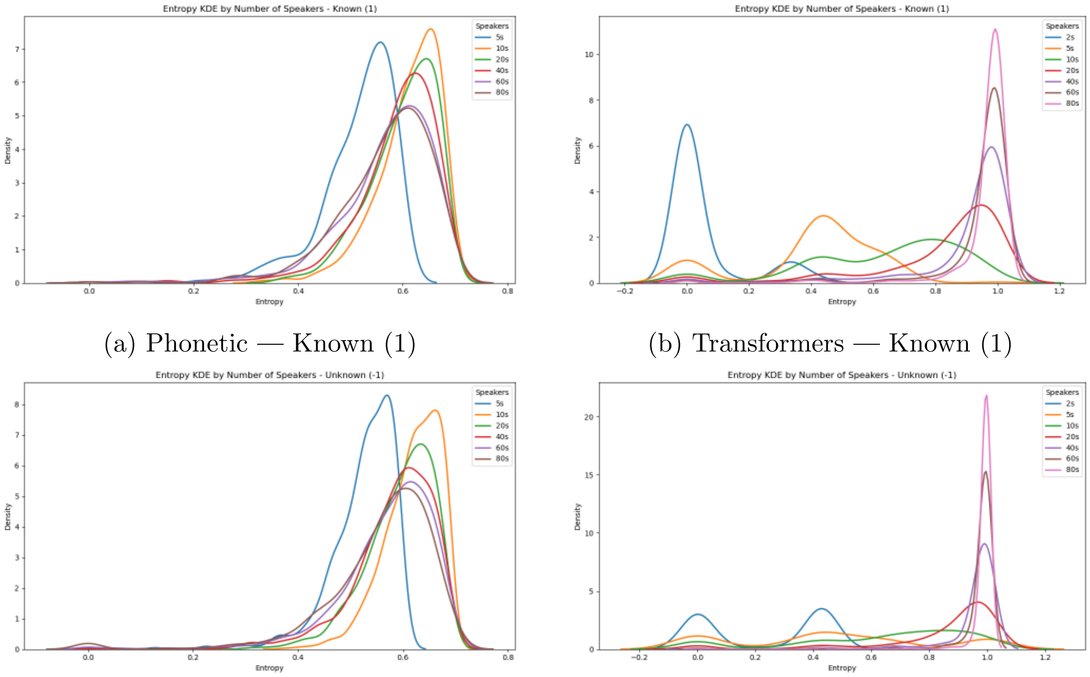
(b) Transformers based model

Figure 5.10: Abstention ratio heatmaps for the Transformers based and Phonetic based models.

5.2.2 Phonetic and Transformers based Models

For the models positioned at the extreme ends of the specified–generalized spectrum, the abstention heatmaps reveal comparable shortcomings in how rejection is applied. In the transformer-based model, abstention ratios are scattered irregularly across configurations, with no consistent relationship to the number of speakers or instances, suggesting an unstructured deployment of the mechanism. In contrast, the specified phonetic model shows sparse and inconsistent abstention, with many configurations registering almost no use of the mechanism at all. In both cases, the lack of a coherent pattern likely reflects an unorganized or weakly separable speaker embedding space. As the closed-set entropy distributions offer limited separation between correct and incorrect predictions, the abstention mechanism is unable to selectively filter errors, preventing any meaningful improvement in precision.

In open classification, both the phonetically specified and the generalized transformer-based models displayed a tendency to overgeneralize the "unknown" label, frequently assigning it to correct known instances. The entropy distributions in the first two plots (phonetic model) show substantial overlap between known and unknown cases, with peaks in similar entropy ranges, making it difficult for the model to distinguish the two categories. A similar pattern is visible in the last two plots (generalized model), where the entropy curves for known and unknown instances remain closely aligned across different numbers of speakers. This consistent similarity suggests that the threshold calibration process was inherently biased, as it was trained on distributions where misclassified known cases mirrored the entropy values of true unknowns. Without a clear separation between these groups, the thresholding mechanism could not establish a reliable decision boundary, resulting in ineffective open classification. This points to a deeper issue in both models: the absence of well-structured, discriminative speaker



(a) Phonetic — Known (1)

(b) Transformers — Known (1)

(c) Phonetic — Unknown (-1)

(d) Transformers — Unknown (-1)

Figure 5.11: Entropy distributions for the Phonetic and Transformers based models for the Known (1) and Unknown (-1) labels.

clusters, which are essential for robust and consistent classification in open-set scenarios.

Error analysis Results - Open Classification

Across the open classification experiments, the suitability of each representation type became clear. The spectral model, situated between the specified and generalized extremes, proved most capable of supporting the abstention mechanism, maintaining a degree of separation between known and unknown speakers in low-complexity settings. However, even this model suffered a progressive collapse of entropy boundaries as the number of enrolled speakers grew, leading to overly cautious behavior and reduced recall. In contrast, the phonetic and generalized models lacked coherent abstention patterns from already low complexity settings, reflecting weakly structured embedding spaces where entropy failed to discriminate between correct, incorrect, and unknown predictions. This structural limitation caused frequent misclassification of known instances as unknown, rendering threshold-based rejection ineffective. Together, these findings suggest that while intermediate representations offer some resilience, all three approaches face significant degradation in open classification performance under scaling, driven primarily by the erosion of distinct confidence distributions.

Based on these insights and results, we can reflect on the research question: *How should speaker identity be represented and modeled in communicative robotic systems to support accurate, adaptable, and conservative speaker identification over time, across both open-set and closed-set classification tasks?* The findings show that the three models behaved quite differently under scaling conditions. The models at the extremes of the specified–generalized spectrum struggled to maintain robust and well-defined speaker clusters as classification complexity increased. Their embedding spaces became unstructured when the number of speakers grew, causing cluster boundaries to disappear and limiting the benefits of gradual learning from additional exposure to speaker data.

In contrast, the spectral model produced denser and more separable clusters, allowing it to maintain high performance in closed-set classification even under high complexity. This model benefited from increased exposure up to an optimal point, beyond which performance gains diminished, indicating a non-linear ”diminishing returns” effect in relation to the number of instances per speaker.

For open classification, similar trends emerged. The models at the extreme end of the spectrum again failed to distinguish between known and unknown speakers due to their unstructured embedding spaces and overlapping confidence distributions, making abstention ineffective. The spectral model initially leveraged abstention to improve precision, but as complexity grew, it became overly conservative, abstaining on many correct predictions and suffering from low recall. Additionally, increasing complexity caused entropy values for known speakers to rise and those for unknown speakers to drop, leading to convergence in their confidence distributions. This overlap reduced the reliability of thresholding, limiting the model's ability to separate known from unknown instances effectively while preserving recall.

Based on these findings, it can be concluded that representation approaches preserving as much raw information as possible from the audio data are the most suitable for scalable speaker identification in the context of a communicative robot. Their flexibility allows the model to adapt to ever-changing patterns as the number of speakers and instances increases, enabling more robust and separable speaker embeddings.

Although the specified and generalized models contained speaker-specific information in low-complexity settings, they failed to maintain effectiveness as complexity grew. The intuitive motivation for using these hand-crafted or overly abstracted features appears valid under controlled, low-variation conditions. However, as the data and application context become more heterogeneous, with spontaneous variations introduced in large-scale scenarios, the discriminative power of these features diminishes. This leads to unstructured embeddings and poor scalability compared to models utilizing raw, fine-grained representations.

5.3 Limitations

While the results of this study suggest that spectral-based models outperform both specified and generalized data representation approaches for scalable speaker identification, several limitations must be acknowledged when interpreting these findings. These limitations relate to the experimental design, resource constraints, and dataset choices, all of which may have influenced the results and their generalizability.

Firstly, the project was conducted under significant time and computational resource constraints. This limited the development and optimization of all three models, meaning they were only implemented to the extent necessary to function and carry out the classification task. There was no opportunity to fine-tune architectures, optimize hyperparameters, or explore alternative training strategies. As a result, the observed performance differences may partly reflect implementation limitations rather than intrinsic differences between the data representation approaches.

Secondly, the models were not trained on multiple seeds or subjected to repeated experiments. This lack of replication means that the results may not be fully reliable or reproducible. Variability in model initialization, data sampling, or arbitrary elements of training could have influenced the reported outcomes. Multiple runs with different seeds would have been necessary to establish whether the observed trends are stable and not due to random chance.

Thirdly, the comparison between the three representational approaches must be interpreted cautiously, as each model was built with minimal computational capacity and without extensive optimization. In a real-world scenario or with more resources, different architectures or improved feature engineering could significantly alter the relative performance of these approaches. Consequently, this study should not be taken as conclusive evidence that spectral features are universally superior, but rather as an exploratory indication of their potential under constrained conditions.

Furthermore, the models were developed and evaluated within a specific experimental context and were not benchmarked against standard datasets commonly used in speaker identification research. The intent of the project was not to achieve state-of-the-art performance but to investigate how different data representation strategies behave as speaker identification systems are scaled up. The models should therefore be seen as experimental tools for analyzing representational differences, not as competitive speaker identification systems.

Another important limitation relates to the dataset itself. The data used consisted of audiobook segments, chosen because it provided sufficient coverage per speaker and a moderate pool of speakers suitable for simulating scaling conditions. However, this type of data lacks the spontaneity, variability, and natural conversational patterns that a real communicative robot would encounter. Audiobook recordings are typically clean, loud, and lack many of the contextual and paralinguistic variations (e.g., environmental noise, interruptions, emotional variability) that are common in real-world speech. As such, the results may not fully reflect the challenges a speaker identification model would face in a more naturalistic setting, particularly in open classification tasks.

Finally, because this project focused on exploring representational effects rather than producing deployable models, its findings should not be interpreted as recommendations about specific architectures or performance benchmarks. Instead, they should be understood as preliminary evidence that retaining fine-grained, minimally processed spectral information may offer advantages for scalability in speaker identification, while more specified or generalized representations may lose their discriminative power as complexity increases.

5.4 Future Research

While this study aimed to address key gaps in the literature regarding how speaker identification models can be integrated into communicative robots, where scalability and reliable open classification are essential, several avenues for future research remain. These could further advance the field and improve the applicability of speaker identification models in real-world deployments beyond this study's scope.

One promising direction is the exploration of interpretability. The models tested in this study relied on different types of encoded information, specified versus generalized features, yet the nature of what these models actually learned and leveraged from their embeddings remains unclear. This research primarily focused on the structural consistency of the generated embeddings and the formation of speaker clusters, rather than on understanding which specific information contributed to robust performance. Future studies could investigate which features or representations are most influential in low-complexity settings, whether these features can be amplified, and whether less relevant ones can be replaced or suppressed to improve scalability in high-complexity classification scenarios.

Furthermore, while the spectral model demonstrated strong potential for scalability in speaker identification, its inability to perform reliably in open classification and high-complexity setups highlights the need for additional development before real-world deployment is possible. One avenue to address this limitation could be the incorporation of multimodality. By integrating visually derived speaker identification embeddings into the ensemble framework, it may be possible to provide complementary information that aids the ranking model in separating known from unknown speakers more effectively. The modular nature of the ensemble approach facilitates such integration, allowing visual embeddings to be added as an additional component with minimal architectural changes. This could improve robustness in saturated audio-based embedding spaces and enhance open classification performance. Alternatively, a dedicated visual speaker identification model could operate in parallel to the audio-based ensemble, supporting speaker classification tasks specifically when audio-based uncertainty is high. Exploring these multimodal strategies may significantly improve the reliability and applicability of speaker identification in communicative robots like Leolani.

Chapter 6

Conclusion

This study investigated scalable audio-based speaker identification for communicative robots, focusing on Leolani, a dialogue system that currently relies on visual identification. Incorporating voice-based identification would allow the system to operate effectively in scenarios where visual cues, such as the face, are not available. The research addressed two key application-specific constraints: maintaining accuracy as the classification task scales to more speakers and data, and performing open classification with the option to abstain when uncertain.

Three representational strategies were compared, positioned along a spectrum from specified to generalised approaches. At one end, explicit phonetic features encode predefined articulatory or acoustic properties; at the other, dense transformer-based embeddings capture broad, implicit patterns from large-scale pretraining. Between these extremes, spectral features (e.g., MFCCs, spectrograms) provide low-level signal representations that make fewer assumptions about relevant cues. These approaches differ in abstraction level, adaptability, and scalability potential.

The models were implemented using two distinct architectural approaches. The phonetic and spectral models formed part of an ensemble architecture, combining multiple specialised embedding networks to enable modular design and interpretability at the feature level. The transformer-based model employed a transfer learning approach, integrating a transfer learning Siamese verification network into the HuBERT model to adapt pre-trained embeddings for the identification task.

This contrast in both representational and architectural design allowed the study to assess how scaling variables, such as increasing the number of enrolled speakers or enrollment instances, affect each model's internal representations and decision behaviour. Two experiments were conducted: the first tested closed-set scalability, measuring precision across varied enrollment configurations; the second introduced an entropy-based thresholding mechanism to enable abstention, evaluating precision, recall, and abstention rates in open-set conditions.

The experiments revealed substantial differences in how the three models handled scaling in both closed-set and open-set classification scenarios, with outcomes closely tied to their position along the spectrum from specified to generalised representations.

In the closed-set experiments, the spectral model consistently outperformed both the specified (phonetic) and generalised (transformer-based) models. It maintained high precision across all grid configurations, showing strong resilience to increases in the number of enrolled speakers and the resulting rise in classification complexity. In contrast, the specified and generalised models degraded rapidly as the speaker set expanded, losing structured, separable cluster formations in the embedding space early on. This prevented the gradual performance gains that might otherwise come from increased exposure to speaker data.

The open-set classification experiments, which introduced an entropy-based abstention mechanism, produced results consistent with these trends. The spectral model again achieved high precision and successfully rejected many incorrect predictions but overgeneralised abstention, leading to disproportionately low recall by abstaining on many correctly identifiable instances. In high-complexity settings, it also failed to identify unknown speakers effectively, as entropy distributions for known, unknown, and incorrect-known predictions began to overlap, making the identification of out-of-pattern confidence patterns inefficient. The specified and generalised models, already struggling to maintain robust clusters in closed-set tasks, showed inconsistent open-set behaviour, as they overused the "unknown" label due to early and extensive entropy distribution overlap.

Error analysis further revealed that the spectral model's overall high density and separability masked variation between its components: the MFCC-based submodel struggled with separation and produced overlapping clusters, while the spectrogram-based submodel leveraged a larger embedding space that, although more prone to sparsity, remained structured within the tested saturation levels. This difference was attributed to the level of abstraction used for the two representational inputs. The spectrogram representations, employing a more fine-grained strategy, produced a broader embedding space that supported favourable performance even under scaled conditions.

The phonetic model, in contrast, relied on low-dimensional, hand-crafted embeddings, which constrained the populated vector space to remain compact, preventing sustained performance as enrolment conditions scaled.

The transformer-based model occupied a middle ground in this spectrum of abstraction. Its dense, generalised embeddings captured broad acoustic and contextual patterns, enabling partial grouping of similar speakers. Ranking analysis showed that correct speakers frequently appeared within the top 20% of candidates, suggesting that the model could narrow the search space effectively. However, the absence of sufficiently fine-grained, speaker-specific cues limited its discriminative capacity in high-complexity conditions, ultimately preventing consistent identification at scale.

Overall, the hypothesis that classification complexity would negatively affect both closed-set and open-set performance was confirmed, as was the predicted correlation between increased complexity and reduced open-set accuracy. However, the expected benefits of increased exposure to speaker data were inconsistent, yielding little impact in models without well-separated clusters, and only narrow performance gains in models that did form such clusters.

In summary, spectral raw-signal models appear well-suited for closed-set classification in communicative robot applications but remain inadequate for open classification at scale, as increased complexity induces overlapping entropy patterns that undermine label discrimination. Consequently, no representation tested here fully satisfies all constraints of the intended application.

While this study provided an extensive comparison of three distinct speaker identification models, its findings should be interpreted within the context of several limitations. The comparison was conducted within a single experimental setup, and due to constraints in time and computational resources, certain conditions necessary for full reproducibility could not be met. Models were not trained across multiple random seeds, and system performance was evaluated only under the scaled saturation conditions that were also present in the training set. Additionally, all models were allocated similar computational resources, which may not have been equally sufficient across representational types. Consequently, the results cannot be taken as definitive evidence that a given approach is unsuitable for scalable identification, only that, under the current configuration and resource allocation, their applicability is limited.

The dataset used, while enabling a systematic analysis of scaling variables, consisted of controlled, high-quality speech that does not capture the variability of spontaneous conversational audio. Further research is needed to assess model behaviour in real-world communicative robot interactions, where acoustic conditions and speaker variability are less constrained.

Looking ahead, a central gap lies in the development of scalable open classification systems for communicative robots. One promising direction is the integration of multimodal speaker identification, with the modular structure of the ensemble model providing a natural pathway for incorporating visual modalities. However, the comparative benefits of early versus late fusion in this context remain unexplored. Another avenue involves interpretability analyses of the specified and generalised models, which, despite lower performance in the current setup, may encode complementary non-speaker-discriminative information. Understanding how these embeddings evolve and converge under scaling could inform implicit learning strategies for attributional cues, potentially advancing open and scalable speaker identification capabilities in communicative robots.

6.0.1 Final Conclusion

This study examined how different representational types scale within the task of speaker identification in the context of communicative robots. The findings demonstrate that models employing the least amount of feature abstraction, thus retaining fine-grained acoustic detail, performed significantly better under scaling. Such representations occupied larger embedding spaces, enabling adaptability and flexibility to incoming speaker data even as classification complexity increased.

For scalable identification in the current application context, the results suggest prioritising fine-grained representations capable of populating large-dimensional embedding spaces. These spaces provide the capacity to mitigate the “overpopulation” effect as more speakers are enrolled, helping preserve cluster separability and adaptability. In contrast, approaches that rely on additional assumptions or higher levels of feature abstraction risk constraining the embedding space, reducing flexibility, and producing inconsistent performance in large-scale conditions. While such assumptions may perform well in low-complexity settings, their utility becomes variable as systems scale, complicating their deployment in dynamic environments.

Accordingly, the study concludes that fine-grained speech representations are the most suitable choice for scalable speaker identification in communicative robots, offering the structural capacity and adaptability required for deployment in real-world, dynamic, and expanding speaker environments.

Appendix

Appendix A

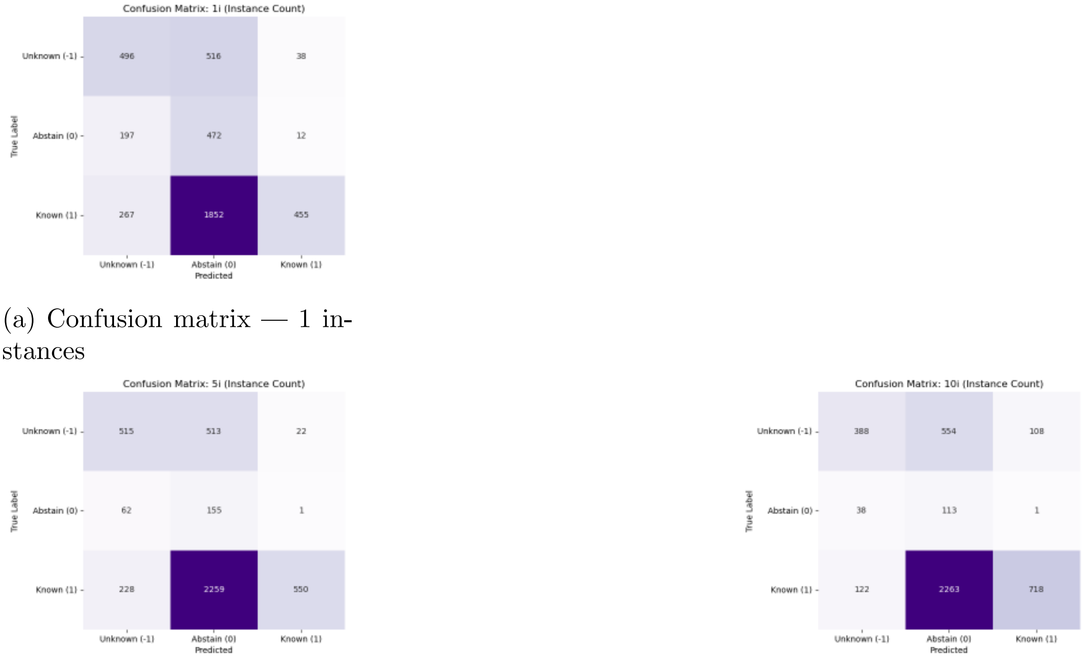
Feature	**Definition**	**Speaker Relevance**	**Citation**
Pitch (F0)	Fundamental frequency of vocal fold vibration	Encodes habitual pitch, linked to vocal fold length and tension	Ladefoged (2011)
Formants (F1–F3)	Resonant frequencies of the vocal tract (F1: height, F2: position, F3: shape)	Reflect vocal tract shape and articulatory patterns	Ladefoged (2005)
Formant Bandwidths	Width of resonance peaks in formants	Linked to articulatory precision	Sambur (1975)
Jitter	Cycle-to-cycle variation in frequency	Measures articulatory control and voicing stability	Teixeira et al. (2013)
Shimmer	Cycle-to-cycle variation in amplitude	Indicates amplitude control, change in volume	Teixeira et al. (2013)
Harmonics-to-Noise Ratio (HNR)	Ratio of harmonic to aperiodic energy	Captures breathiness and vocal clarity	Teixeira et al. (2013)
Spectral Bandwidth	Spread of energy across frequency spec-	Related to phonation type and vocal	Schwartz et al. (2018)

Appendix B

.1 Spectral confusion matrices

.2 Phonetic confusion matrices

.3 Transformers confusion matrices

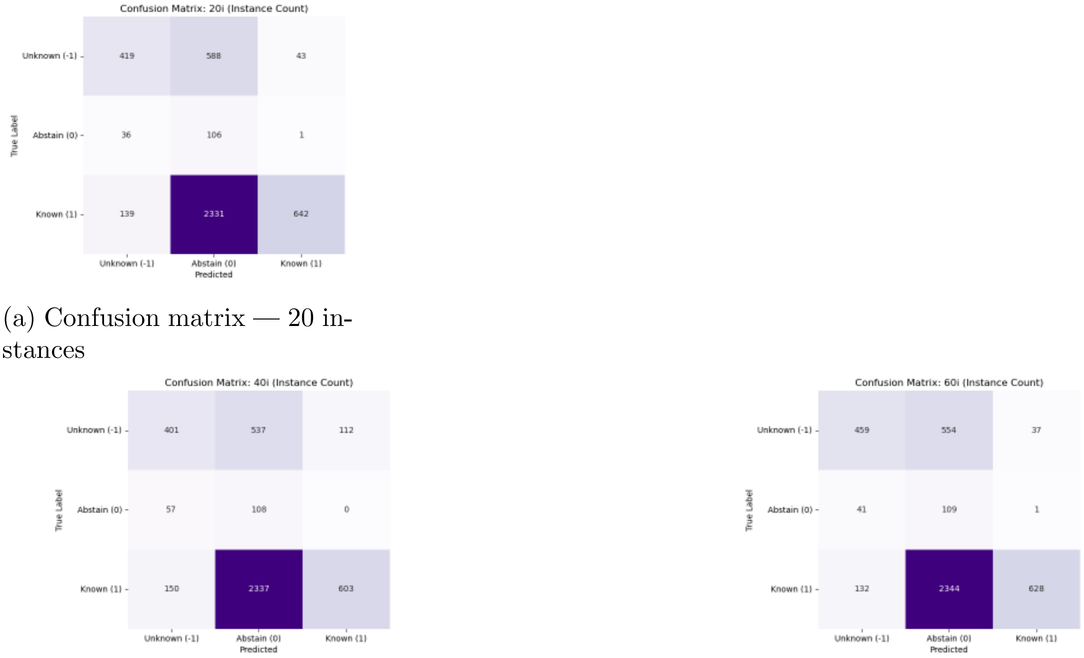


(a) Confusion matrix — 1 instances

(b) Confusion matrix — 5 instances

(c) Confusion matrix — 10 instances

Figure 1: Spectral confusion matrixes

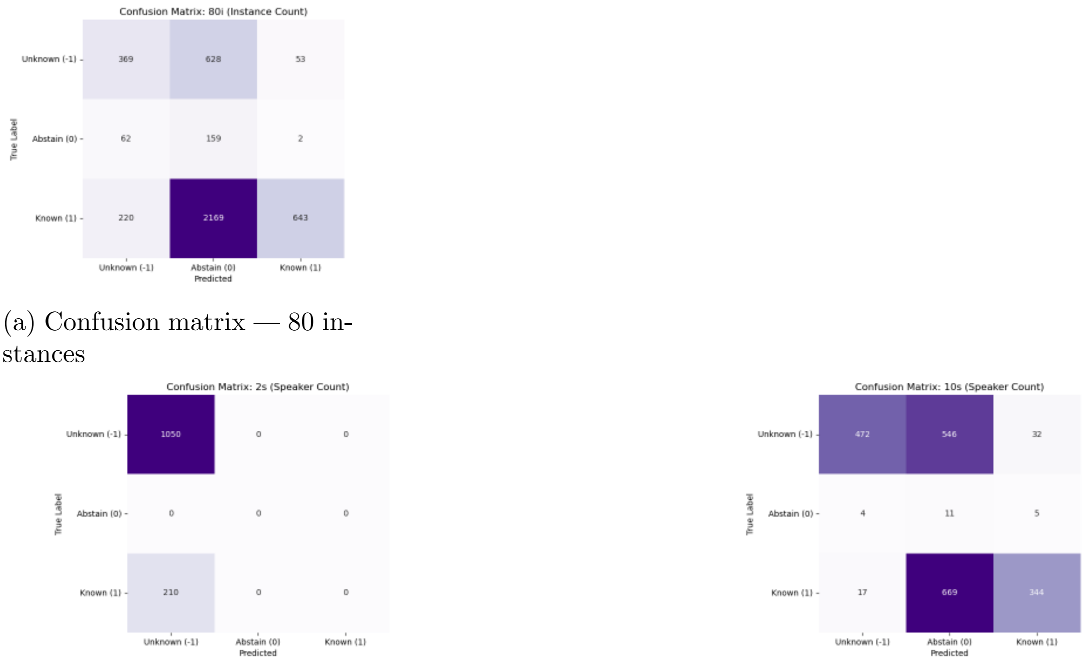


(a) Confusion matrix — 20 instances

(b) Confusion matrix — 40 instances

(c) Confusion matrix — 60 instances

Figure 2: Spectral confusion matrixes

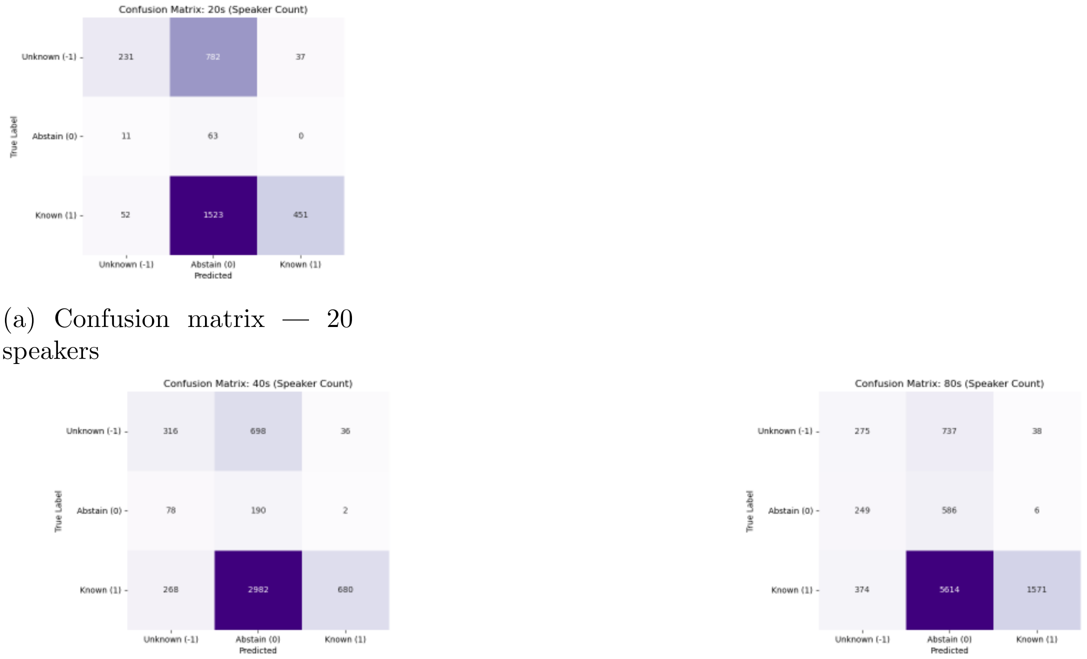


(a) Confusion matrix — 80 instances

(b) Confusion matrix — 2 speakers

(c) Confusion matrix — 10 speakers

Figure 3: Spectral confusion matrixes

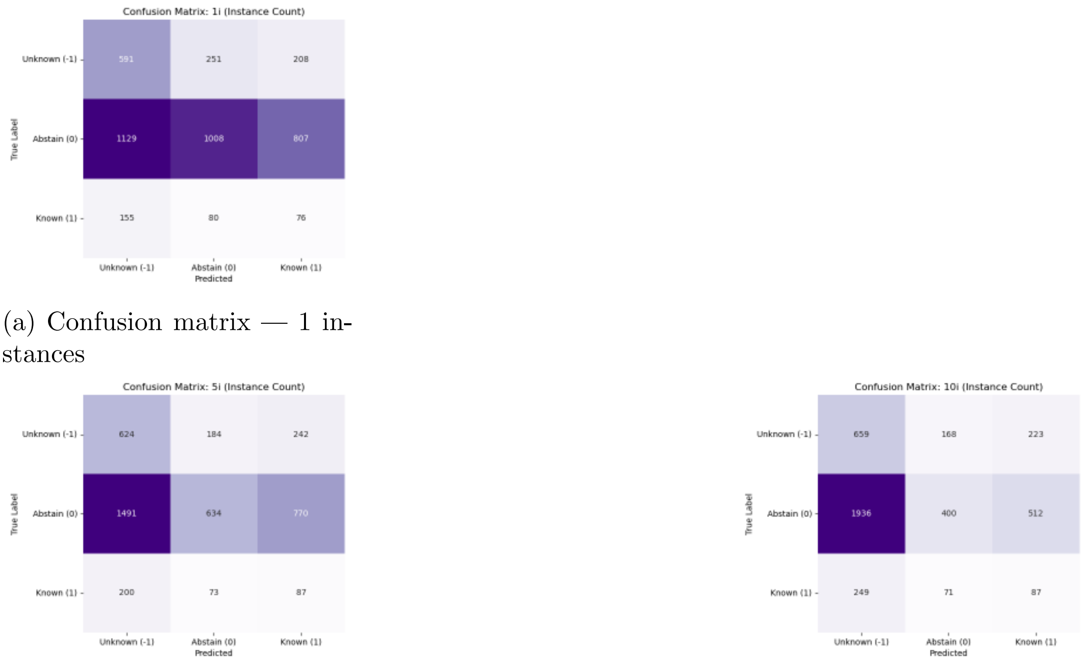


(a) Confusion matrix — 20 speakers

(b) Confusion matrix — 40 speakers

(c) Confusion matrix — 80 speakers

Figure 4: Spectral confusion matrixes

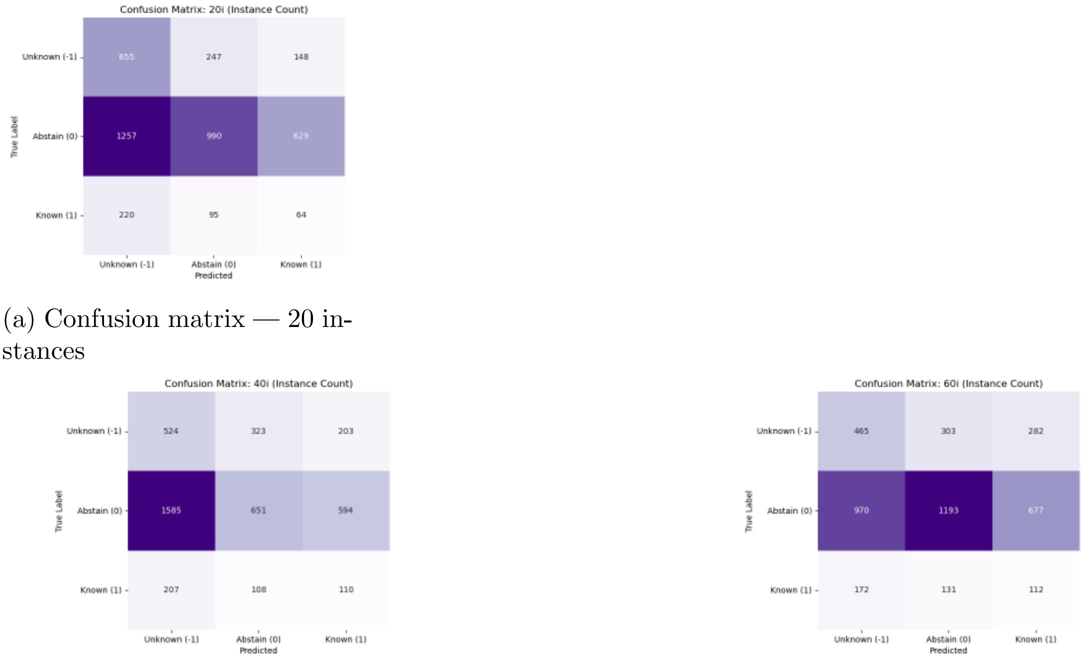


(a) Confusion matrix — 1 instances

(b) Confusion matrix — 5 instances

(c) Confusion matrix — 10 instances

Figure 6: Phonetic confusion matrixes

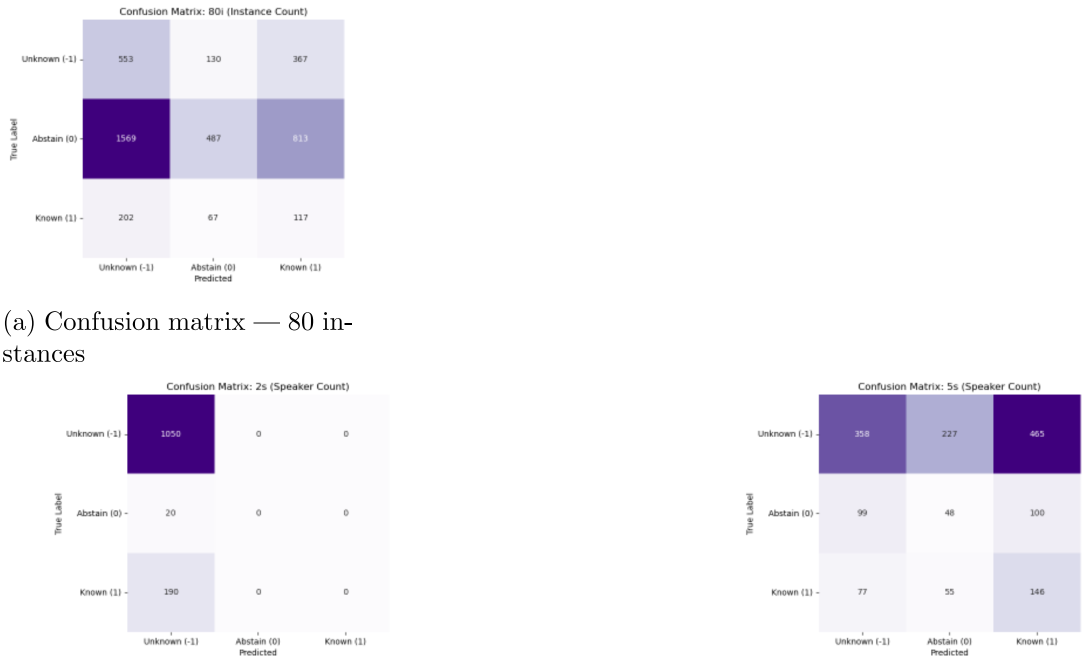


(a) Confusion matrix — 20 instances

(b) Confusion matrix — 40 instances

(c) Confusion matrix — 60 instances

Figure 7: Phonetic confusion matrixes

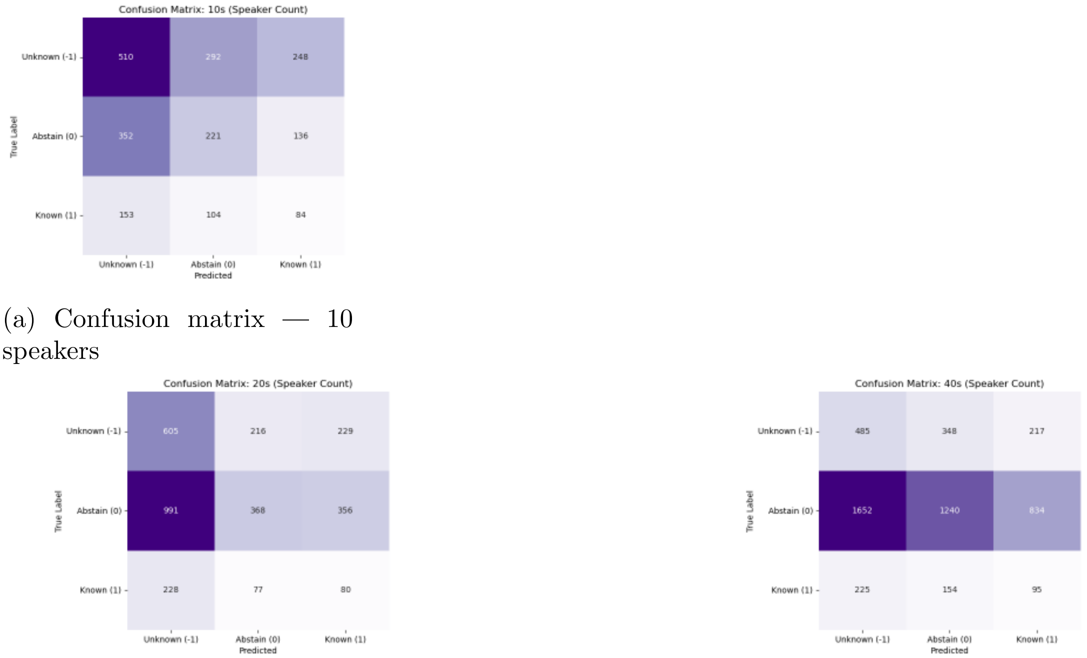


(a) Confusion matrix — 80 instances

(b) Confusion matrix — 2 speakers

(c) Confusion matrix — 5 speakers

Figure 8: Phonetic confusion matrixes

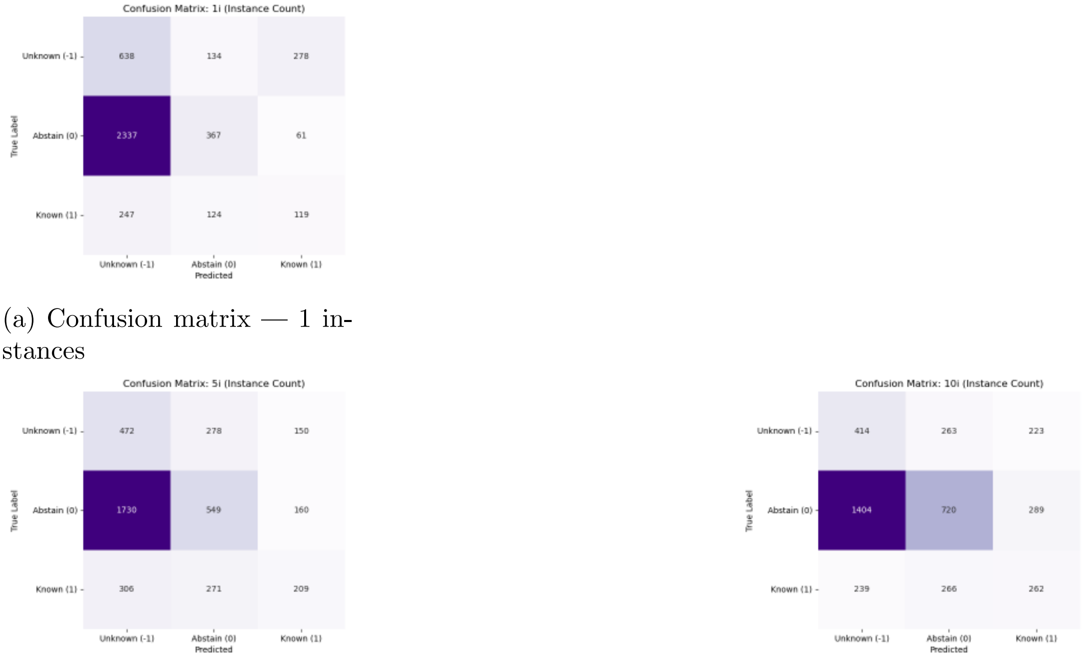


(a) Confusion matrix — 10 speakers

(b) Confusion matrix — 20 speakers

(c) Confusion matrix — 40 speakers

Figure 9: Phonetic confusion matrixes

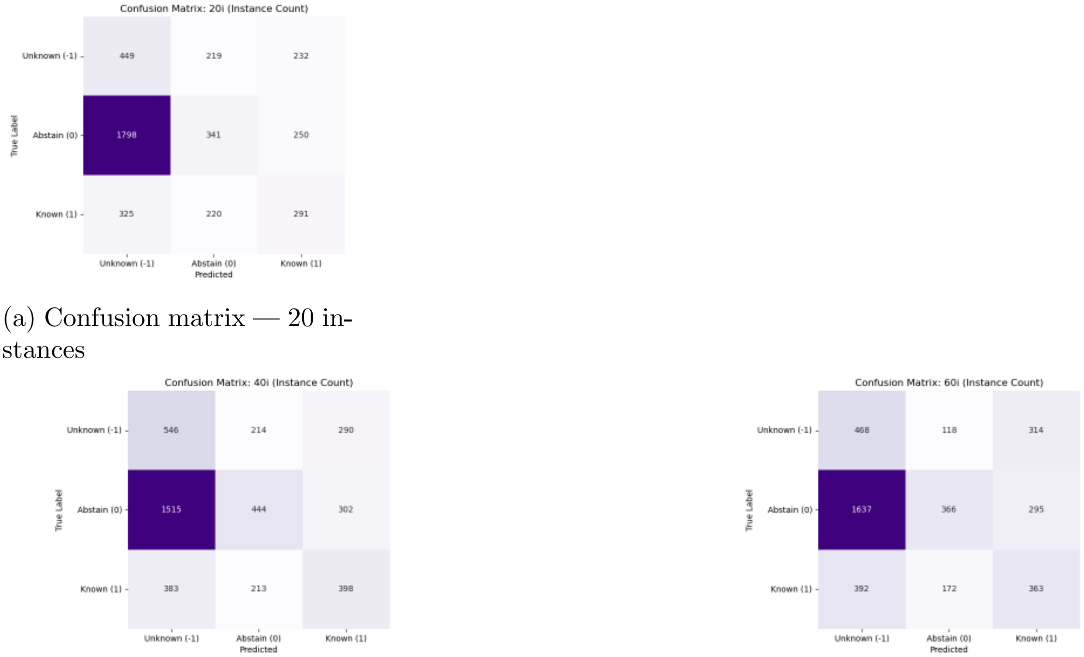


(a) Confusion matrix — 1 instances

(b) Confusion matrix — 5 instances

(c) Confusion matrix — 10 instances

Figure 11: Transformers confusion matrixes

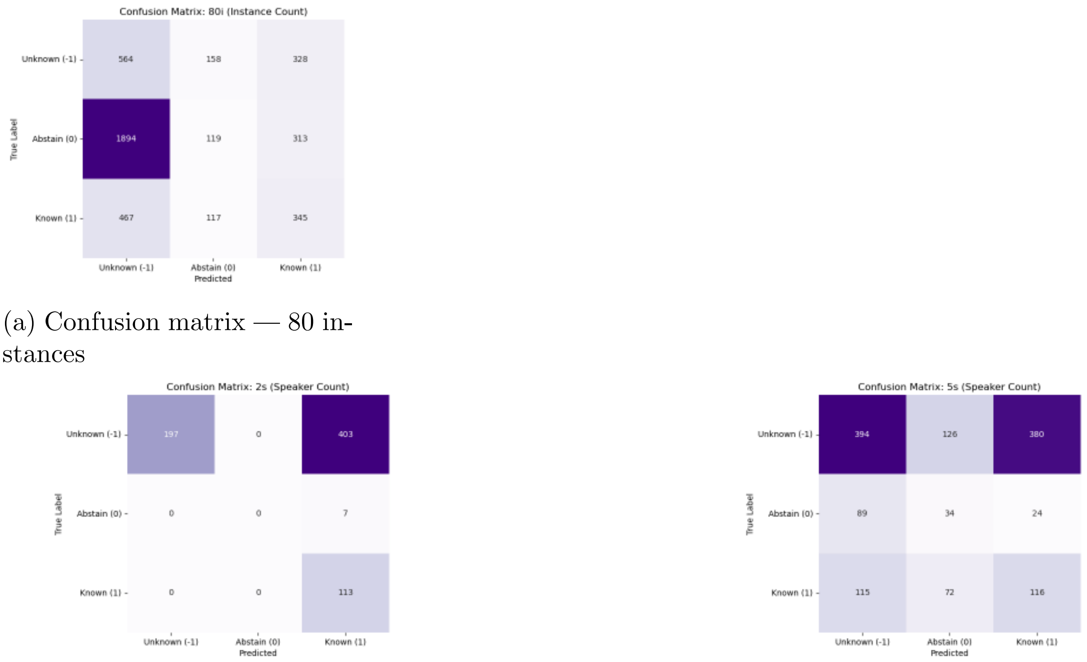


(a) Confusion matrix — 20 instances

(b) Confusion matrix — 40 instances

(c) Confusion matrix — 60 instances

Figure 12: Transformers confusion matrixes



Confusion Matrix: 80i (Instance Count)

		Predicted	
True Label	Unknown (-1)	Abstain (0)	Known (1)
Unknown (-1)	564	158	328
Abstain (0)	1894	119	313
Known (1)	467	117	345

(a) Confusion matrix — 80 instances

Confusion Matrix: 2s (Speaker Count)

		Predicted	
True Label	Unknown (-1)	Abstain (0)	Known (1)
Unknown (-1)	197	0	403
Abstain (0)	0	0	7
Known (1)	0	0	113

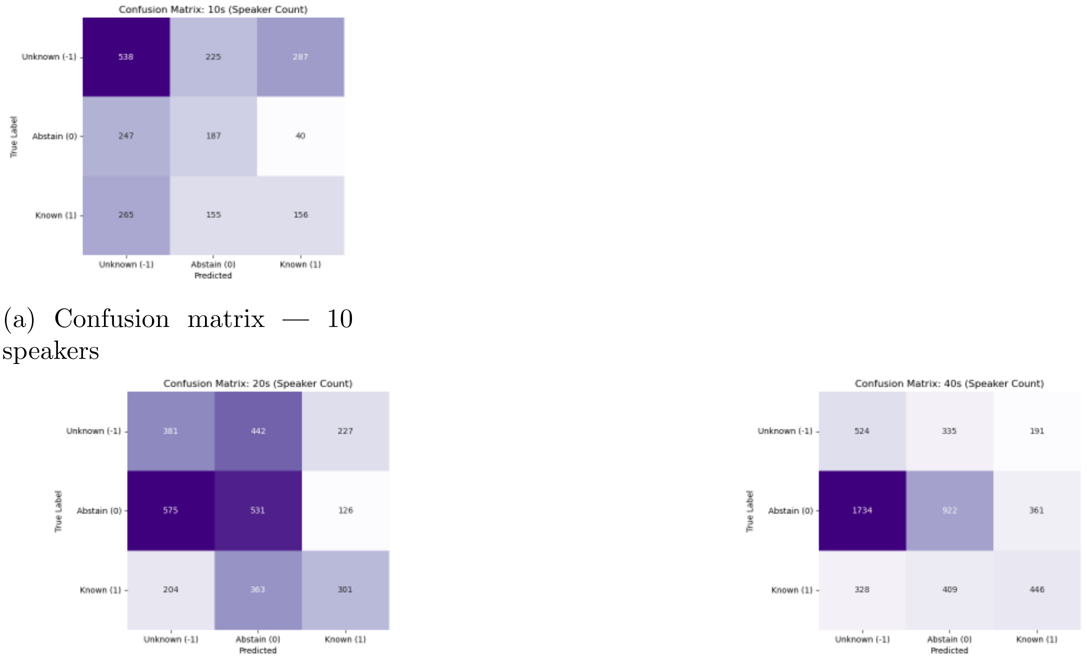
(b) Confusion matrix — 2 speakers

Confusion Matrix: 5s (Speaker Count)

		Predicted	
True Label	Unknown (-1)	Abstain (0)	Known (1)
Unknown (-1)	394	126	380
Abstain (0)	89	34	24
Known (1)	115	72	116

(c) Confusion matrix — 5 speakers

Figure 13: Transformers confusion matrixes



Confusion Matrix: 10s (Speaker Count)

		Predicted	
True Label	Unknown (-1)	Abstain (0)	Known (1)
Unknown (-1)	538	225	287
Abstain (0)	247	187	40
Known (1)	265	155	156

(a) Confusion matrix — 10 speakers

Confusion Matrix: 20s (Speaker Count)

		Predicted	
True Label	Unknown (-1)	Abstain (0)	Known (1)
Unknown (-1)	381	442	227
Abstain (0)	575	531	126
Known (1)	204	363	301

(b) Confusion matrix — 20 speakers

Confusion Matrix: 40s (Speaker Count)

		Predicted	
True Label	Unknown (-1)	Abstain (0)	Known (1)
Unknown (-1)	524	335	191
Abstain (0)	1734	922	361
Known (1)	328	409	446

(c) Confusion matrix — 40 speakers

Figure 14: Transformers confusion matrixes

References

Speaker recognition using neural networks and conventional classifiers. *IEEE Transactions on speech and audio processing*, 2(1):194–205, 2002.

M. Affek and M. S. Tatara. Open-set speaker identification using closed-set pretrained embeddings. In *International Conference on Diagnostics of Processes and Systems*, pages 167–177. Springer, 2022.

V. Brydinskyi, Y. Khoma, D. Sabodashko, M. Podpora, V. Khoma, A. Konovalov, and M. Kostiak. Comparison of modern deep learning models for speaker verification. *Applied Sciences*, 14(4):NA–NA, 2024.

C. J. Burges. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11 (23-581):81, 2010.

S. Chakraborty and R. Parekh. An improved approach to open set text-independent speaker identification (osti-si). In *2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pages 51–56. IEEE, 2017.

S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, Oct. 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3188113. URL `http://dx.doi.org/10.1109/JSTSP.2022.3188113`.

J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018, pages 1086–1090, 2018.

B. Desplanques, J. Thienpondt, and K. Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, interspeech$_2$020.$ISCA, Oct.$2020. doi : . URL `http://dx.doi.org/10.21437/Interspeech.2020-2650`.

P. Foggia, A. Greco, A. Roberto, A. Saggese, and M. Vento. Few-shot re-identification of the speaker by social robots. *Autonomous Robots*, 47(2):181–192, 2023.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus. Timit acoustic-phonetic continuous speech corpus. *(No Title)*, 1993.

C. Gendrot, E. Ferragne, and T. Pellegrini. Deep learning and voice comparison: phonetically-motivated vs. automatically-learned features. In *ICPhS*, 2019.

W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Y. Jadoul, B. Thompson, and B. de Boer. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15, 2018. https://doi.org/10.1016/j.wocn.2018.07.001.

J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.-J. Yu. Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. In *Proc. Interspeech 2019*, pages 1268–1272, 2019.

R. Karadaghi, H. Hertlein, and A. Ariyaeeinia. Effectiveness in open-set speaker identification. In *2014 International Carnahan Conference on Security Technology (ICCST)*, pages 1–6, 2014. 10.1109/CCST.2014.6986991.

R. Karrer. Google WebRTC Voice Activity Detection (VAD) module. GitHub, 2025. URL https://github.com/rafaelkarrer/mex-webrtcvad/releases/tag/v0.1.

N. R. Koluguri, T. Park, and B. Ginsburg. Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8102–8106. IEEE, 2022.

P. Ladefoged and S. F. Disner. *Vowels and consonants*. John Wiley & Sons, 2012.

B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke, S. Seyfarth, A. Malek, C. Raffel, V. Lostanlen, B. van Niekirk, D. Lee, F. Cwitkowitz, F. Zalkow, O. Nieto, D. Ellis, J. Mason, K. Lee, B. Steers, and ... and others.

A. Meghanani and A. G. Ramakrishnan. Pitch-synchronous dct features: A pilot study on speaker identification, 2018. URL https://arxiv.org/abs/1812.02447.

A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

M. A. Nasr, M. Abd-Elnaby, A. S. El-Fishawy, S. El-Rabaie, and F. E. Abd El-Samie. Speaker identification based on normalized pitch frequency and mel frequency cepstral coefficients. *International Journal of Speech Technology*, 21(4):941–951, 2018.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 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, 2015.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

M. Sang, H. Li, F. Liu, A. O. Arnold, and L. Wan. Self-supervised speaker verification with simple siamese network and self-supervised regularization. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6127–6131. IEEE, 2022.

S. B. Santamaría, P. Vossen, and T. Baier. Evaluating agent interactions through episodic knowledge graphs. *arXiv preprint arXiv:2209.11746*, 2022.

D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

A. Stan. Residual information in deep speaker embedding architectures. *Mathematics*, 10(21):3927, 2022.

J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, et al. State-of-the-art speaker recognition for telephone and video speech: The jhu-mit submission for nist sre18. *Interspeech 2019*, 2019.

P. Vossen, S. Baez, L. Bajcetic, S. Basic, and B. Kraaijeveld. Leolani: A robot that communicates and learns about the shared world. *English, in ISWC*, pages 181–184, 2019.

F. Wang and H. Liu. Understanding the behaviour of contrastive loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504. IEEE, 2021.

S. Wang, Z. Chen, K. A. Lee, Y. Qian, and H. Li. Overview of speaker modeling and its applications: From the lens of deep speaker representation learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

S. Zarin, E. Mustafa, S. K. Zaman, A. Namoun, and M. Alanazi. An ensemble approach for speaker identification from audio files in noisy environments. *Applied Sciences*, 14: 10426, 11 2024. 10.3390/app142210426.