university of groningen

campus fryslân

Speaker Identification in Mandarin Conference Speech via Transfer Learning with wav2vec 2.0

Sixing Mi

university of groningen

campus fryslân

University of Groningen - Campus Fryslân

Speaker Identification in Mandarin Conference Speech via Transfer Learning with wav2vec 2.0

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Vass Verkhodanova (Voice Technology, University of Groningen)
with the second reader being
drs. Frank Hopwood (Voice Technology, University of Groningen)

Sixing Mi (S5827094)

July 31, 2025

Acknowledgements

I would like to express my sincere gratitude towards my supervisor, for her patient and carful guidance through this year.

Deep gratitude to my parents, who always support me for all my choices and decisions and encourage me to do everything.

Many thanks to my friends, who are around me or thousands of miles away, for their physical and emotional support which get me out of difficulties for thousands of times during this year.

Thanks to the life in the University of Groningen and Voice Technology.

Abstract

In today's multilingual and digital world, speaker recognition is becoming increasingly important in real-world applications such as virtual conferencing, transcription services, and customer support. Despite significant progress in Mandarin automatic speech recognition, speaker recognition in real-world Mandarin conference speech is still imperfect due to challenges such as pitch interference, overlapping segments, and environmental noise. To further improve Mandarin speaker recognition performance, this study focuses on exploring the transfer ability of wav2vec2.0 to speaker recognition tasks in multi-person conference settings. To evaluate this, I used the AISHELL-4 corpus, which contains Mandarin conference speech with realistic acoustic variations. Specifically, my study answers the following questions: How effectively can a pre-trained Mandarin ASR wav2vec2 model be adapted for speaker recognition in real-world Mandarin conference speech? What are the effects of task and domain transfer mechanisms on its performance? This study freezes the wav2vec 2.0 encoder, adds a lightweight linear classifier on top of it, and designs two control groups: a global classification baseline model and a session-level transfer learning model. The results show that although the baseline model achieved a Top-1 accuracy of 50.3% on the entire speaker label space, the session-level model performed significantly better than the baseline model, with an average accuracy improvement of more than 20% and a maximum accuracy improvement of 36% compared to the baseline model, highlighting the superiority of the session-level model. These findings suggest that even with a small amount of fine-tuning, pre-trained ASR models can capture speaker recognition features and generalize well to noisy domains. This study provides evidence that this transfer learning strategy is effective for speaker perception systems in real-world Mandarin environments, and future directions include adaptive fine-tuning, cross-lingual generalization, and integration with speaker classification for broader applications.

Contents

1 Introduction

In today's increasingly multilingual and digitally mediated world, spoken-communication systems are expected to do more than transcribe words: they must also recognize who is speaking, under what acoustic conditions, and within which conversational role or social context. Recent surveys on meeting-analysis technology report that accurate, real-time speaker attribution has become a prerequisite for virtual conferencing, customer analytics, and automatic minute-taking, because without reliable identity tags transcripts lose accountability and downstream dialogue modelling fails to track speaker intent (Anguera et al., 2012). As voice-driven interaction spreads across virtual classrooms, customer-support hotlines, and media-accessibility services (Jha et al., 2024), the demand for systems that can simultaneously interpret speech content and distinguish between multiple speakers continues to rise. In such settings, speaker identification is no longer a secondary convenience but a core capability required to deliver coherent, auditable and personalised user experiences.

This need is especially pronounced in real-world, high-stakes communicative settings such as business meetings, academic panel discussions, and hybrid workplace collaborations. In these contexts, clear attribution of spoken content to individual speakers is crucial for maintaining accurate records, enabling speaker-specific responses, managing access control, and ensuring transparency and traceability of dialogue. Errors in speaker identification can lead to misunderstandings, misattribution of ideas, or loss of critical contextual information (Anguera et al., 2012). Furthermore, these environments often involve dynamic and unpredictable speech patterns, with participants interrupting, talking over each other, or speaking in informal, disfluent ways that diverge from scripted speech.

Mandarin Chinese, as the official language of China, is widely spoken across educational, governmental, and commercial domains. While previous research in speaker recognition has often focused on English or multilingual datasets (Vaessen & Van Leeuwen, 2022), there is a growing need to support high-accuracy speaker recognition in Mandarin-speaking scenarios, particularly in structured and spontaneous meetings. Although Mandarin is not under-resourced in the traditional sense, its unique tonal nature introduces a distinct set of modeling challenges (Tao, Tan, Yeung, Chen, & Lee, 2024). Tones are lexically contrastive, meaning that pitch variations signal different meanings; thus, the acoustic features used for speaker differentiation are partially entangled with linguistic content. This interaction increases the complexity of learning robust speaker embeddings in tonal environments (Lei, Scheffer, Ferrer, & McLaren, 2014).

Conference-style meetings, irrespective of language, are characterized by spontaneous speech, rapid floor exchanges, interruptions, filled pauses, and a wide range of individual speaking styles(Anguera et al., 2012). Mandarin-only meetings add an extra layer of complexity because tonal contours fluctuate continuously while speakers interact, increasing acoustic variability even within a single turn (Tao et al., 2024). Empirical analyses of the AISHELL-4 corpus show that more than one-fifth of Mandarin meeting segments contain two or more simultaneous speakers, while average utterance length is below three seconds and background noise levels vary with room layout and microphone placement (Fu et al., 2021). These factors—short, overlapping segments, tonal modulation and heterogeneous capture conditions—make Mandarin conference recordings a demanding test bed for speaker identification. A robust SID system therefore needs to separate speaker-specific cues from lexical-tone patterns and remain resilient to the degradations caused by overlap, reverberation and device mismatch.

Speaker identification (SID) is the task of assigning a speaker label to a given speech segment from a set of known speakers. It is fundamental to downstream applications such as speaker-attributed transcription, speaker-specific dialogue summarization, and long-form audio retrieval (Tirumala & Shahamiri, 2016). Historically, SID models progressed from Gaussian Mixture Models (GMMs) and i-vectors (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2010) to x-vectors (Snyder, Garcia-Romero, Sell, Povey, & Khudanpur, 2018) trained with deep neural networks. While effective in clean and controlled environments, these supervised methods often require extensive annotated training data and manual feature engineering, limiting their scalability.

Self-supervised learning (SSL) offers a compelling alternative. Self-supervised learning models like wav2vec 2.0 learn contextualized speech representations from massive amounts of unlabeled audio using pretext tasks such as contrastive prediction or masked reconstruction (Baevski, Zhou, Mohamed, & Auli, 2020). These models have demonstrated strong performance across diverse speech tasks, including ASR, emotion recognition, and speaker verification. Wav2vec 2.0, specifically, consists of a convolutional encoder followed by a Transformer network that captures temporal dependencies. The representations extracted from wav2vec 2.0 have been shown to preserve speaker-discriminative information, even when trained without explicit speaker supervision (Yang et al., 2021).

This study builds on the wav2vec2-large-xlsr-53-chinese-zh-cn model, a Chinese pre-trained variant of wav2vec 2.0 model. I evaluate its utility for speaker identification in Mandarin conference speech using the AISHELL-4 corpus—a large-scale Mandarin meeting dataset with spontaneous speech, speaker overlaps, and realistic background conditions. Importantly, I adopt a lightweight setup that freezes the wav2vec 2.0 encoder and trains only a simple classifier on top of pooled hidden states (Vaessen & Van Leeuwen, 2022). This enables testing whether speaker-related features learned during pretraining are sufficient for distinguishing speakers in Mandarin meeting data without any encoder fine-tuning.

By doing so, this thesis contributes to several strands of research: (1) it explores the transferability of self-supervised models from speech recognition to speaker identification; (2) it investigates speaker identification under tonal language conditions using Mandarin-only data; and (3) it offers insights into cost-effective, low-resource adaptation strategies for real-world meeting applications. The remainder of this thesis is structured as follows: Section 1.1 introduces the research questions and hypotheses. Section 2 reviews related work in speaker recognition and transfer learning. Section 3 details the dataset, model architecture, and experimental design. Section 4 presents and analyzes the results, followed by a discussion in Section 5. Section 6 concludes the study and outlines future research directions.

1.1 Research Questions and Hypotheses

In light of the preceding discussion, the central research question of this study can be formulated as follows:

> **How effectively can a Mandarin ASR-pretrained wav2vec 2.0 model be adapted to perform speaker identification in real-world Mandarin conference speech, and what are the impacts of task and domain transfer mechanisms on its performance?**

This main question can be broken down into the following sub-questions:

- How accurately can a frozen ASR-pretrained wav2vec 2.0 model classify speaker identities based solely on its learned representations?

- How does the performance of a wav2vec 2.0-based speaker identification system vary under domain mismatch between training (clean ASR data) and evaluation (noisy, spontaneous conference data)?

Based on prior findings that ASR-pretrained wav2vec 2.0 models retain substantial speaker-relevant information in their learned representations (Vaessen & Van Leeuwen, 2022), the study hypothesizes that:

A wav2vec 2.0 encoder pretrained on Mandarin ASR and kept frozen during fine-tuning can still produce embeddings that are sufficiently speaker-discriminative to enable competitive classification performance. When combined with a lightweight classifier head, the system is expected to achieve robust accuracy on noisy, multi-speaker, and overlapping Mandarin conference recordings—despite the domain mismatch between training and evaluation conditions.

2 Literature Review

This section is dedicated to providing a comprehensive review of the existing research pertaining to speaker identification, with a specific focus on Mandarin Chinese in real-world meeting scenarios. The emphasis of this thesis is on the transfer learning approach from automatic speech recognition to SID using a pre-trained wav2vec 2.0 model. This approach leverages the ability of wav2vec 2.0 to learn robust and transferable representations from large amounts of unlabeled speech data, enabling effective downstream performance on tasks like speaker identification with limited annotated resources. By conducting a thorough and critical analysis of the literature in this field, this review aims to offer valuable insights into the methods and effectiveness of applying transfer learning strategies for Mandarin speaker identification in noisy and overlapping conference speech environments.

To those ends, the section is structured as follows. To begin, I delineate the keywords used during the literature search and describe the inclusion or exclusion criteria used in selecting the literature. After that, I offer a succinct overview of the key findings and contributions of the selected papers.

I have grouped the keywords according to the topic they are related to. The topics are highlighted in bold, after which the keywords for that topic are mentioned. Thus, the topics and their corresponding keywords are:

- **Speaker identification:** speaker recognition, speaker classification, speaker identification, deep speaker embeddings;

- **Self-supervised learning:** wav2vec 2.0, self-supervised speech model, self-supervised learning in speech, speech representation learning;

- **Mandarin speech:** Mandarin speaker recognition, tonal language speaker identification, Mandarin meeting speech, AISHELL-4 dataset.

- **Transfer learning:** speech recognition to speaker identification transfer, encoder freezing, representation adaptation;

In order to ensure that the selected articles are highly consistent with the research topic, this review adopts a narrative screening strategy rather than a strict systematic review. First, the literature from 2010 to 2024 was searched in Google Scholar using keywords such as "speaker identification, self-supervised learning, wav2vec 2.0, Mandarin meeting speech, transfer learning", and a total of about 387,880 records were obtained. After a preliminary reading of the titles and abstracts, papers on tasks such as speech synthesis, speech enhancement, and emotion recognition that were not related to speaker identification were deleted, and some studies focusing on speaker modeling or self-supervised representation learning were retained. Subsequently, early works based on GMM-UBM and i-vector that lacked self-supervision or transfer learning elements were further excluded, and a list of 35 core articles was finally formed.

The above screening process is not intended to form an "exclusion count" in the statistical sense of the systematic review, but to ensure that the included studies can directly support the discussion of self-supervised transfer and Mandarin meeting scenarios in this article. Through this hierarchical screening, this review is able to focus on the latest progress of wav2vec 2.0 in speaker recognition, laying a literature foundation for subsequent method design and experimental comparison.

The final inclusion rules therefore required that studies (i) focus on speaker identification or speech recognition for under-resourced languages, specifically Mandarin, (ii) employ transfer-learning techniques within the speaker identification or speech recognition pipeline (studies that used transfer learning solely for TTS, enhancement, or unrelated downstream tasks were excluded), and (iii) make explicit use of wav2vec 2.0 or a comparable self-supervised speech model. Publications not meeting all three conditions, or appearing in languages other than English, were excluded.

Following this, I provide an overview of the key findings from the selected papers, organized by the aforementioned topics. Each subsection (2.1–2.4) delves into specific aspects such as the effectiveness of transfer learning techniques, the challenges of applying wav2vec 2.0 to Mandarin speaker identification, and the use of realistic datasets like AISHELL-4 dataset in evaluating performance. This structured approach offers a comprehensive understanding of the current state and future directions of speaker identification research for Mandarin and other under-resourced languages.

The literature review is organized into different subsections based on the general topics they cover. Subsection 2.1 discusses the literature regarding traditional speaker identification methods. Subsection 2.2 addresses the rise of deep learning approaches such as x-vectors. Subsection 2.3 presents an overview of self-supervised techniques, especially wav2vec 2.0, and subsection 2.4 focuses on Mandarin speaker identification in the AISHELL-4 dataset context.

2.1 Speaker Recognition: Progress and Challenges

Speaker recognition has gone through a long way from the statistical era of Gaussian-mixture universal background models (Reynolds, Quatieri, & Dunn, 2000) and i-vector factor analysis (Dehak et al., 2010) to today's deep neural architectures that offer real-time inference and sub-percent error rates on clean benchmarks. A decisive turning point came with the x-vector framework, where a time-delay neural network encodes frame-level features and a statistics-pooling layer aggregates them into a fixed-length utterance embedding (Snyder et al., 2018). Subsequent residual variants such as Res2Net-SE and ECAPA-TDNN introduced multi-scale convolutions and channel attention, driving equal-error rates on VoxCeleb1 down to the one-percent range (Desplanques, Thienpondt, & Demuynck, 2020) (Zhao et al., 2024). Most recently, multi-branch designs like 3D-Speaker further compact the network while improving discrimination in open-set trials (Y. Chen et al., 2025).

An equally significant shift is the move toward self-supervised pre-training. Encoders such as wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021) and WavLM (S. Chen et al., 2022) are first exposed to thousands of hours of unlabelled audio via a masked-prediction objective, learning representations that jointly encode short-term phonetic detail and longer-term speaker cues. When these encoders are frozen and paired with a shallow classifier, they already match or surpass fully supervised baselines on VoxCeleb and SITW; fine-tuning only the top Transformer blocks narrows the gap to state of the art, as summarised in the SUPERB benchmark (Yang et al., 2021). This evidence confirms that large-scale acoustic pre-training captures speaker-stable information even though it was optimised for word recognition.

Despite these advances, several long-standing obstacles still impede robust deployment. The first one is the utterance duration. Conversational segments often last fewer than three seconds, a regime where embedding variance grows rapidly and error rates can double (Kinnunen & Li, 2010). The second one is the channel mismatch. The differences in microphone frequency response, codec, or far-field capture shift embedding distributions and inflate within-speaker variability (Garcia-Romero, McCree, Shum, Brummer, & Vaquero, 2014). Then, background noise and room reverberation further smear spectral detail, while multi-talker overlap violates the single-speaker assumption that most front-ends rely on (Anguera et al., 2012). Furthermore, cross-language transfer adds another layer of complexity that phonotactic and prosodic differences can distort the statistics learned on English corpora, particularly for tonal languages where fundamental-frequency contours carry lexical content (Tao et al., 2024).

Current research mitigates these issues through domain transfer learning. Fine-tuning from clean to noisy data, synthetic augmentation with additive noise and speed perturbation, and contrastive learning on overlapped speech have each shown promise. Parameter-efficient adapters such as LoRA allow a small subset of weights to specialise to new microphones or languages while the bulk of the encoder remains frozen (Joseph & Baby, 2024). However, most published evaluations still focus on English broadcast speech (Vaessen & Van Leeuwen, 2022); how well these strategies generalise to Mandarin meetings, where tonal interference, high overlap rate and heterogeneous devices coexist, remains an open question that motivates the present study.

2.2 Transfer Learning for Speaker Recognition

Transfer learning has become a central approach in modern speech processing tasks, including speaker recognition. It refers to the use of knowledge learned in one domain or task to improve performance on a different but related task (Weiss, Khoshgoftaar, & Wang, 2016). This is particularly valuable in scenarios with limited labeled data, such as speaker identification in Mandarin conference speech. In these low-resource or high-variability environments, direct training from scratch is often infeasible, making transfer learning a practical and efficient alternative (Sherly, Pillai, & Manohar, 2024).

The past decade has seen transfer learning become an important strategy for speaker recognition systems, largely because high-quality speaker labels are costly to obtain and performance drops sharply when the recording environment differs from the training domain (Garcia-Romero et al., 2014). The transfer techniques now have three main types. The first one is task transfer. The encoders originally trained for automatic speech recognition or masked-prediction objectives are reused for speaker identification with only a small classification or similarity head appended (Cai & Li, 2024). For example, The frozen wav2vec 2.0 and HuBERT encoders could achieve sub-3% equal-error rates on VoxCeleb when they are paired with linear back-ends. This demonstrats that large self-supervised models capture speaker-stable cues despite being optimised for phonetic content (Y. Wang, Boumadane, & Heba, 2021). Selectively fine-tuning only the top Transformer layers or injecting lightweight adaptation modules such as LoRA further closes the gap to fully trained baselines while adding a small number of extra parameters (Joseph & Baby, 2024).

The second one is domain adaptation, which mitigates channel, language and acoustic mismatches between training and deployment (Farahani, Voghoei, Rasheed, & Arabnia, 2021). Early work relied on unsupervised PLDA mean-shift and whitening transforms (Garcia-Romero et al., 2014), but now deep models adopt more sophisticated methods such as multi-style training with additive noise and speed perturbation (Ko, Peddinti, Povey, & Khudanpur, 2015), adversarial feature normalisation that minimises microphone identity in the embedding space (Bhattacharya, Monteiro, Alam, & Kenny, 2019), and teacher-student distillation that transfers knowledge from near-field to far-field systems (Zhang, Wang, Lee, Xie, & Li, 2021). These approaches reduce equal-error rates effectively when moving from one domain to another, underscoring the importance of domain fine-tuning.

A third and increasingly popular one is parameter-efficient adaptation (He, Li, Zhang, Yang, & Wang, 2023), motivated by the desire to keep lowinference costs while still allowing every domain to do specialisation, which seeks to preserve the frozen encoder's speed while allowing minimal task-specific weight updates. One practical approach is the adapter bottleneck: a pair of small linear layers inserted between frozen Transformer blocks. The studies of T. Wang, Chen, Chen, Yu, and Zhu (2023) shows that inserting adapters into a HuBERT Transformer encoder and pre-training them on a mix of raw, noisy, overlapped, and noisy overlapped speech reduces the word error rate by 40% relative to the multi-label pre-trained model without adapters on ASR, while achieving comparable performance on speech separation and enhancement tasks with only a marginal increase in parameters (3.1% to 17.2%)The study confirms that parameter-efficient fine-tuning can bridge a notable portion of the domain gap while keeping computational and memory footprints suitable for real-time deployment.

Taken together, these advances show that transfer learning—whether across tasks, domains or parameter subsets—offers a pragmatic path to high-accuracy, low-latency speaker recognition in scenarios where labelled data, compute budget or recording conditions deviate from laboratory standards. They also provide the methodological scaffolding for the present work, which evaluates how a Mandarin ASR-pretrained wav2vec 2.0 encoder transfers to multi-speaker conference speech.

2.2.1 Task Transfer: From Speech Recognition to Speaker Identification

The most direct form of transfer learning in speaker recognition is reusing an encoder that was originally trained for automatic speech recognition. During self-supervised pre-training, models such as wav2vec 2.0, HuBERT and WavLM consume thousands of hours of unlabelled speech and learn frame-level features that encode both phonetic detail and speaker information (Yang et al., 2021). Once the encoder has converged, the specific output layers of ASR are discarded, a pooling operation aggregates frame embeddings into a single utterance vector, and a shallow classification or metric head is trained to map that vector to a speaker identity. Baevski et al. (2020) introduce wav2vec 2.0 and show that a frozen encoder combined with simple mean pooling can be fine-tuned for speaker identification with only a shallow classification head. Hsu et al. (2021) report similar findings for HuBERT, demonstrating that its frame-level features carry speaker-stable information even though the model is optimised for masked unit prediction. S. Chen et al. (2022) extend the approach with WavLM, confirming that discarding the ASR-specific projection layers, pooling hidden states into an utterance vector, and attaching a lightweight soft-max yields competitive speaker recognition performance on VoxCeleb and CN-Celeb benchmarks. Empirical evidence from SUPERB demonstrates that a frozen wav2vec 2.0 Base encoder combined with mean pooling achieves 75.18% identification accuracy on the VoxCeleb1 dataset, with performance approaching that of more complex setups (Yang et al., 2021).

The key adaptation step is time-scale conversion. ASR focus on 20- to 30-millisecond windows, whereas speaker recognition requires information that remains consistent across hundreds of milliseconds. Simple mean or statistics pooling suppresses phoneme-level variation while preserving long-term traits such as habitual pitch range, average formant spacing, and spectral-envelope shape. The importance of converting short-frame ASR features into utterance-level speaker embeddings was highlighted by (Snyder et al., 2018), who showed that statistics pooling (mean and variance) over 20–30 ms frame representations markedly improves speaker-discriminative power compared with frame-based scoring. Subsequent work by (Desplanques et al., 2020) in ECAPA-TDNN and by (Okabe, Koshinaka, & Shinoda, 2018) in attentive pooling confirmed that averaging or attentively weighting hidden states suppresses phoneme-level variability while retaining long-term cues such as habitual pitch range, average formant spacing and spectral-envelope shape—properties that are relatively stable over several hundred milliseconds.

2.2.2 Domain Transfer: Channel, Acoustic, and Language Mismatch

Once the deployment data deviates from the conditions during pre-training or fine-tuning, the performance of trained encoders will generally degrade. First, channel variations (microphone frequency response, distance between speaker and microphone, or differences in codecs) can amplify intra-speaker spread; Garcia-Romero et al. (2014) showed that a simple phone or broadband mismatch can significantly increase the equal error rate of NIST SRE. In addition, the diversity of languages and accents also has an impact: Tao et al. (2024) showed that differences in F_0 contours of Mandarin lexical tones can lead to pitch shifting problems in speech synthesis.

Domain transfer strategies address these gaps. For example, multi-level transfer learning from near-field to far-field in a teacher-student (T/S) framework is used to transfer the knowledge of a teacher model trained on near-field data to a student model trained on far-field data to address the domain mismatch between enrollment and test utterances in far-field speaker verification (Zhang et al., 2021).

Taken together, task transfer supplies a speaker-aware foundation, while domain transfer fine-tunes the embedding space to cope with microphone, noise, and language shifts. The collaboration of the two explains much of the recent progress in deploying speaker recognition beyond controlled laboratory conditions.

2.3 Mandarin Speaker Recognition in Conference Scenarios

The core goal of speaker identification is to accurately identify the individual to whom the current speech segment belongs from multiple candidate speakers. Traditionally, speaker recognition is usually trained and evaluated on controlled, clear, single-speaker speech data. However, in recent years, researchers have gradually turned their attention to more challenging real-world application scenarios, especially conference speech environments with multiple speakers, language diversity, and device inconsistency (Fu et al., 2021). In the context of Mandarin, this task faces a series of unique and complex challenges.

First, the language characteristics of Mandarin itself bring inherent difficulties to speaker identification. Unlike non-tonal languages (such as English), Mandarin is a language that relies heavily on pitch contours to convey lexical meaning. Every syllable in Mandarin carries a tone, and different tones can fundamentally change the meaning of a speech segment (Peng et al., 2018). Studies have shown that these tone changes are not only related to semantic expression, but are also often manifested as measurable frequency changes, which may come from both word meaning and individual characteristics of the speaker (such as physiological structure, vocalization habits, etc.) (Anguera et al., 2012). During the modeling process, this ”functional overlap” will cause speaker features to mix with language content features, making it difficult for the system to clearly distinguish which changes belong to the speaker itself and which are the result of language expression, thereby increasing the difficulty of model discrimination.

These challenges are further exacerbated when the Mandarin speaker recognition task is embedded in a real conference speech environment. Conference speech is highly realistic and complex, which is specifically manifested in four aspects: First, the speech content is often spontaneously generated, and compared with standard reading corpus or dubbing corpus, it lacks stable intonation, syntactic structure and speech flow control; second, there are many participants with different language backgrounds, expression styles and speaking habits; third, the conference process is full of natural interactions, such as frequent turn-taking, interruptions, responses, interjections, short pauses and topic jumping; fourth, the audio collection process is affected by environmental acoustic conditions, differences in microphone equipment, speaking distance and background noise, resulting in a significant decrease in recording quality and acoustic consistency (Fu et al., 2021).

In addition, the speech features in Mandarin conference speech are usually highly variable. For example, speakers may express their opinions using code switching (such as a mixture of Mandarin and English or dialects), unstructured grammar (such as omitting subjects and misusing conjunctions), semantic hesitation (such as ”that, that is, um”), and non-fluent expressions (such as pauses, repetitions, and interjections) (Fu et al., 2021). Although these real language behaviors are widely accepted in actual communication, they are a source of disturbance for systems based on acoustic consistency modeling, which will significantly reduce the stability and distinguishability of the speaker’s representation (Anguera et al., 2012).

From the perspective of system input, there are significant differences in equipment and environment between different meetings. For example, some meetings use high-fidelity recording equipment, while others rely on built-in microphones in laptops or voice acquisition modules on online meeting platforms. The frequency response range, gain settings, and pickup directions of different microphones may have a significant impact on the speaker’s voice characteristics (Araki, Ono, Kinoshita, & Delcroix, 2017). Coupled with the background noise in the conference room (such as page turning, whispering, keyboard sounds) and the volume fluctuations caused by the speaker’s movement, these factors together constitute a highly heterogeneous input environment with low signal quality, which puts higher requirements on the generalization and robustness of the speaker identification system.

A common problem is that the number of speakers in Mandarin conferences is usually large, often ranging from a few to dozens of people. Large-scale multi-speaker recognition tasks require higher model representation capabilities. The system not only needs to establish sufficient differentiation between multiple speakers, but also must tolerate the internal variation of each speaker in terms of speech rate, intonation, and prosodic features (Fu et al., 2021). In addition, due to the fragmented characteristics of conference speech, the duration of each speaker’s speech data may be very limited, and some speakers may even speak for less than ten seconds. This ”short speech modeling” problem (short utterance problem) will significantly limit the stability and recognition ability of the embedding vector, and is a difficulty in current speaker recognition research (Kinnunen & Li, 2010).

In summary, speaker recognition in Mandarin conference scenarios faces many challenges: the tonal interference of the language itself, the uncertainty of spontaneous speech, the uncontrollable quality of recordings, and the large number of speakers. These problems not only weaken the effectiveness of traditional supervision methods, but also expose the limitations of current modeling strategies in complex real-world environments. Because of this, Mandarin conference speech recognition provides a valuable testing platform for evaluating the robustness, transferability, and scalability of the new generation of speaker recognition models. Especially in the framework of transfer learning, this scenario has important research and application significance for testing whether the model can successfully generalize to complex, high-noise, and tone-sensitive speech environments without relying on large-scale labeled data.

2.4 Wav2vec 2.0 Model

The wav2vec 2.0 model was originally proposed by the Facebook AI research team and is mainly used for automatic speech recognition tasks. However, its performance in various paralinguistic tasks, including language recognition, emotion recognition, voice activity detection, and speaker identification, is also very promising. It adopts an innovative hierarchical structure and consists of two main parts: one is the front-end convolutional feature encoder (CNN Feature Encoder), which is used to encode the original audio signal into a local time-frequency representation; the other is the back-end Transformer context network (Context Network), which is used to learn to capture higher-level, context-dependent speech representations. Its training goal is mask prediction, that is, randomly masking a part of the frame in the input sequence and then predicting its true representation. This design encourages the model to learn long-term dependencies using contextual information to obtain robust, semantically relevant representations (Baevski et al., 2020).

With the continuous development of self-supervised learning methods, the application of models such as wav2vec 2.0 in speech representation learning has become an important breakthrough in the field of speech processing in recent years (Baevski et al., 2020). The main advantage of this type of model is that it can be pre-trained on a large scale of unlabeled audio corpus without a large amount of labeled data to obtain speech representations that are widely applicable to downstream tasks.

Unlike traditional acoustic modeling systems, wav2vec 2.0 achieves end-to-end raw waveform modeling capabilities. It abandons the reliance on hand-crafted features (such as MFCC or PLP) and instead directly learns multi-level acoustic representations from raw audio signals through convolution and Transformer networks. This capability not only improves the adaptability and generalization of the system, but also makes it possible to build a unified speech understanding model (Baevski et al., 2020). In this context, researchers began to explore the feasibility of migrating self-supervised pre-trained models such as wav2vec 2.0 to speaker recognition tasks, especially in real-world scenarios where data is scarce and speech is complex, such as Mandarin conference speech environments.

This paradigm shift from feature engineering to representation learning not only lowers the threshold of prior knowledge that the system relies on, but also greatly broadens the application boundaries of pre-trained models in the field of speech. On the SUPERB benchmark, a frozen wav2vec2-base encoder plus a single linear layer obtains near-state-of-the-art scores not only on automatic speech recognition but also on speaker identification, speaker diarisation and language identification (Yang et al., 2021). On the VoxCeleb1 dataset, the fine-tuned HuBERT-large-960h model achieved an equal error rate of 2.36% on the speaker verification task (Y. Wang et al., 2021). These results suggest that self-supervised encoders implicitly disentangle phonetic and speaker factors in their internal space, thus providing a strong starting point for low-resource or rapid-deployment speaker identification systems.

2.4.1 Adaptation of wav2vec 2.0 to Speaker Identification

Although wav2vec 2.0 was originally developed for automatic speech recognition, subsequent studies have found that it also exhibits strong transferability in speaker identification tasks. Specifically, researchers often add simple downstream structures (such as mean pooling + linear classifier) on the basis of freezing the wav2vec 2.0 encoder parameters to achieve speaker identity discrimination. This lightweight fine-tuning strategy has achieved excellent results in the speaker identification subtask of SUPERB, showing strong universal embedding capabilities (Yang et al., 2021).

Further research has explored combining the output embedding of wav2vec 2.0 with traditional backend systems (such as PLDA and Cosine Distance) to improve its performance in speaker verification tasks. Some scholars have also tried to incorporate contrastive learning objectives (such as SimCLR or triplet loss) into the fine-tuning process to enhance the discriminability of the embedding space. In addition to achieving high performance under controlled conditions, wav2vec 2.0 also shows remarkable robustness in noisy, low-quality real-world environments. Zhu et al. (2022) verified its noise immunity on multiple noisy speech test sets, further confirming its applicability in practical scenarios such as conference speech.

The representation learned by wav2vec2.0 is considered to contain rich acoustic, prosodic and individual voice features, and has shown the ability to preserve speaker characteristics in multiple tasks. This makes the model very suitable as a universal encoder for building domain-robust speaker recognition systems, especially when the target task labels are insufficient, and its zero fine-tuning ability is particularly outstanding.

3 Methodology

In this section, I present the methodology used to address the research question and validate the hypothesis in a structured manner. Subsection 3.1 describes the dataset used for training and evaluation. Subsection 3.2 and 3.3 focus on the model architecture employed in this study. Then, in subsection 3.4, I detail the evaluation metric and its justification. Finally, subsection 3.5 outlines the ethical considerations.

3.1 Dataset

The main dataset used in this study is AISHELL-4, an open Mandarin corpus for real-world conference speech processing tasks released by Fu et al. (2021). This corpus is designed for multi-speaker, multi-channel, and noisy speech technology research, especially for speaker identification, speaker diarization, and automatic speech recognition. The reason why this study chose AISHELL-4 dataset is based on its modeling capabilities and task diversity support for real conference environments, especially its ability to accurately simulate speech technology application scenarios under actual deployment conditions.

AISHELL-4 dataset contains a total of 211 hours of recording data from 118 independent meetings. The number of participants in each meeting ranged from 4 to 8 speakers, covering a variety of communication forms from formal speeches to free discussions, from planned speeches to natural interruptions. This diversity of language interaction not only increases the breadth of model training, but also greatly improves the robustness of evaluating speaker recognition models under dynamic turns. It is particularly noteworthy that the dataset is designed to deliberately retain common phenomena in natural language interaction, such as speech overlap, interruption, disfluency, and ambient noise, providing a challenging experimental platform for this study.

In terms of recording settings, AISHELL-4 dataset uses an 8-channel microphone array with a circular arrangement of microphones to simulate the spatial pickup structure in a real conference room. This multi-channel recording method has a stronger sound field capture capability than a single-channel device, and can record distance changes, echoes, and reverberation effects caused by changes in the speaker's position. These spatial characteristics are key factors in modeling the speaker-device interaction mode and provide important verification conditions for the actual adaptability of the speaker recognition system.

The speakers used in the dataset come from multiple regions of China and are all native Mandarin speakers, but their ages, genders, occupations, and expression styles are diverse. This demographic diversity helps to enhance the model's generalization ability to different sound features, thereby alleviating the gender bias or age bias problems that speaker recognition systems may encounter in real-world applications. In addition, the corpus provides a complete metadata index file, including the audio path of each speech segment, the conference it belongs to, the speaker number, the start and end time of the speech segment, the channel number, etc., allowing researchers to flexibly construct training sets, validation sets, and test sets.

In summary, AISHELL-4 dataset is a high-quality conference speech dataset with language representativeness, task adaptability, and structural complexity. It can provide sufficient training resources and evaluation dimensions for this study, ensuring the systematic verification of wav2vec2.0 model's transfer learning capabilities under real conference conditions.

3.2 Model Architecture - wav2vec 2.0

The model architecture used in this study is based on wav2vec 2.0, an advanced self-supervised speech modeling framework proposed by Facebook AI Research in 2020. The core goal of wav2vec 2.0 is to solve the problem of traditional ASR systems' dependence on a large amount of manually annotated data, build high-quality speech representations using unlabeled speech data, and migrate them to a variety of downstream tasks, such as automatic speech recognition, speaker recognition, and speech emotion recognition. This model successfully introduced the concept of end-to-end learning into the field of speech representation modeling and achieved performance that exceeds traditional supervised methods.

As shown in Figure 1, the wav2vec 2.0 architecture mainly includes the following three key components:

- Feature Encoder
- Quantization Module
- Context Network

The front end of wav2vec 2.0 is a feature encoder consisting of multiple 1D temporal convolutional layers, which is responsible for extracting preliminary low-level acoustic features from the raw waveform input $X \in R^T$. The encoder's task is to identify local patterns and short-term dependencies in the audio, such as syllable boundaries, formant changes, and other basic building blocks in speech signals. After convolution, the raw waveform is converted to a latent representation $z \in R^{T' \times d}$, where T' is the number of frames after downsampling and d is the feature dimension.

In the pre-training phase, in order to design self-supervised learning tasks, the model uses a vector quantization module. This module discretizes a portion of the frames in z into a limited number of codebook entries. Specifically, it uses the Gumbel-Softmax method to select the best matching codebook vector from multiple embedding vectors, and these quantized vectors are used as the training targets of the model. The learning task of the model is to identify the correct codebook from several negative samples through the context vector, and then optimize the contrastive loss function.

The context network is a stack of self-attention layers based on the Transformer architecture. It takes the latent sequence z as input and outputs a higher-level contextualized representation c, capturing long-range dependencies across time. The contrastive loss is applied by masking parts of the input and training the model to identify the correct latent prediction from a pool of negatives, improving the robustness of learned representations under varying acoustic conditions.

Throughout the architecture, GELU (Gaussian Error Linear Unit) is used as the activation function in the Transformer blocks. It is defined as:

$$\mathrm{GELU}(x) = x \cdot \Phi(x)$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. The GELU function provides a smooth non-linearity that balances linear and nonlinear behavior, and has been shown to improve convergence in deep Transformer models.

In the pre-training stage, the model masks the input frame and trains the context network to predict the quantized vector corresponding to the masked frame from the unmasked context. This

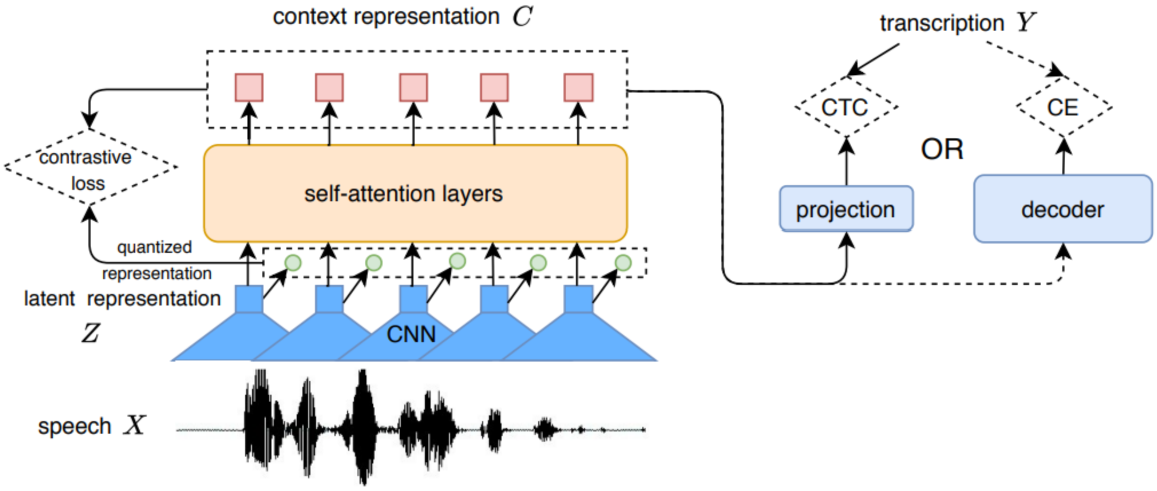

Figure 1: Overview of wav2vec 2.0 architecture

contrast objective strengthens the model's perception of speaker and context changes, thus having the potential for cross-task migration.

It is worth noting that this study did not use the quantization module or contrast loss function when migrating to the speaker recognition task, both of which only play a role in the pre-training stage. The downstream task only uses the context network composed of the encoder and Transformer as a feature extractor to extract high-level embedding representations and provide input for the speaker recognition classifier. This approach makes the migration process lighter and the recognizability of the pre-trained representation can be evaluated separately.

3.3 Large-Scale Cross-Lingual Model - XLSR-53

In order to maximize the generalization ability of multilingual pre-trained models, this study selected a variant based on wav2vec2-large-xlsr-53-chinese-zh-cn for downstream speaker recognition experiments. This model was implemented by Jonatas Grosman and hosted on the HuggingFace platform. It is a Chinese fine-tuned version of Facebook's original XLSR-53 (Cross-Lingual Speech Representations) model. XLSR-53 is a key achievement of wav2vec2's expansion into the field of multilingual speech understanding. It is pre-trained on multi-source speech corpora in 53 languages, and strives to capture the common acoustic structure of languages in the model.

XLSR-53 uses the same network structure as wav2vec2-large, including: 24 layers of Transformer encoding, 1024-dimensional hidden representation per layer, 16 attention heads, and a total model parameter count of over 300 million. During the pre-training phase, the model receives audio input from multiple language resources such as CommonVoice and Multilingual LibriSpeech, enabling it to not only learn language-specific features such as intonation, speech rate, and voice timbre, but also capture cross-language consistent speaker representation structures. This cross-language modeling capability provides theoretical support for dealing with complex accents, speech rates, and style differences in Mandarin conferences in this study.

Since jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn has been fine-tuned in Chinese ASR tasks, its acoustic representation is more adaptable to Chinese intonation patterns and prosodic changes. With the help of this model, this study attempts to verify its transfer ability in speaker recognition tasks, especially in Mandarin conference speech environments with complex background noise, frequent speech overlap, and diverse speech styles. In this study, the encoder weights are completely frozen during training to retain pre-training knowledge. The advantages of this strategy are low resource consumption, high training stability, and strong experimental reproducibility, and it is particularly suitable for fast migration scenarios in low-resource target domains. At the same time, this method can also be used to evaluate the speaker discrimination ability implicit in the pre-trained model to determine whether its representation is identifiable, stable and robust.

A lightweight linear classifier is used on top of the frozen encoder. The frame-level hidden states extracted from wav2vec 2.0 are aggregated by mean pooling to obtain a fixed-length utterance embedding. The embedding is passed through a linear layer to predict the speaker identity. This simple architecture provides an effective way to evaluate the quality of the extracted features in speaker recognition tasks.

3.4 Evaluation Method

In this study, the evaluation of the speaker identification (SID) system is based on Top-1 Accuracy, a commonly used metric in classification tasks. Top-1 Accuracy measures the proportion of predictions where the model's most confident (i.e., highest probability) output corresponds exactly to the ground-truth speaker label. Formally, given a set of N utterances, the Top-1 Accuracy is defined as:

$$\textbf{Top-1 Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i = y_i)$$

- $\hat{y}_i$ is the predicted speaker label for the i-th utterance,
- y_i is the ground-truth speaker label,
- 1 is the indicator function that returns 1 if the prediction is correct, and 0 otherwise.

This study uses Top-1 accuracy as the core metric to quantify the model's speaker recognition ability under Mandarin conference speech conditions. The reason for choosing this metric instead of more complex metrics is mainly based on three considerations. First, Top-1 accuracy is intuitive and easy to interpret: it directly calculates whether the model's most confident prediction is consistent with the true label, and it also has a clear meaning for non-technical readers. Second, the conference speech task is essentially an N-way closed-set classification problem, and each utterance must be attributed to a unique speaker in a fixed set of candidates; in this scenario, the system only returns one identity prediction after going online, so the Top-1 accuracy remains completely consistent with the actual deployment requirements. Finally, this study focuses on whether the pre-trained representation still retains speaker information after freezing the encoder. Top-1 accuracy is sufficient to answer this targeted question, while more fine-grained metrics (such as EER or mAP) have no additional advantages in evaluating simplicity, intuitiveness, and readability.

There are three advantages to using this metric: First, a single value can measure the overall recognition success rate, which is convenient for intuitive comparison with baselines or other models; second, it is also effective for multi-class classification and is not affected by changes in the number of categories; third, it does not require threshold adjustment or curve integration, and has the lowest computational cost in small batch and fast experiments. It should be noted that the Top-1 accuracy ignores the model's confidence in the suboptimal candidate and cannot evaluate uncertainty. In more open scenarios or scenarios that require the return of multiple candidates, Top-5 accuracy, confusion matrix analysis, and even Equal Error Rate can be introduced as a supplement; however, in the closed set setting of this study, the Top-1 accuracy is sufficient to reflect the key behavior of the model, so it was selected as the core metric.

3.5 Ethical Considerations

3.5.1 Data Ethics and Privacy

This study is based entirely on the public Mandarin conference speech dataset AISHELL-4, without additional data collection or subject participation, and therefore does not involve recruitment, informed consent, or ethical approval.

From a privacy perspective, AISHELL-4 dataset has been de-identified before release, retaining only anonymous speaker IDs, speech boundaries and timestamps, transcripts, and microphone channel numbers, without any information that can be used to directly identify individuals, such as names, ID numbers, contact information, or image data, so the overall risk is low. However, the removal of direct identifiers from the data itself does not mean that the model output is absolutely fair. When the demographic distribution is uneven or the sample is underrepresented, the model may still perform poorly on specific groups (such as dialect users, women, or elderly speakers), thereby amplifying structural biases in reality. Although AISHELL-4 dataset has covered different genders, occupations, and speech speeds during the construction phase, its coverage is still limited. To mitigate potential unfairness, more balanced data can be introduced in subsequent studies, fair evaluation indicators such as Equal Opportunity can be adopted, and stratified analysis of specific groups can be performed in experimental design to clarify the source of bias. It should be noted that all evaluations in this study use objective indicators, such as Top-1 accuracy, and do not rely on subjective scoring, thus avoiding personal bias in the evaluation process.

All codes are hosted in a public GitHub repository(See the project on GitHub) and can be opened on demand, and the AISHELL-4 conference dataset is also available for free, thus ensuring the reproducibility of the results and transparency of external review.

3.5.2 Potential Abuse and Responsibility Limits

In addition to academic and commercial purposes, speaker recognition technology may also be deployed in highly sensitive scenarios such as security monitoring or law enforcement. In these scenarios, if there is a lack of proper supervision and authorization, the relevant technology may be used as a tool for illegal monitoring, privacy infringement or social manipulation. Therefore, any actual deployment should strictly comply with local data protection regulations, fully inform and obtain user authorization before collecting and processing voice, and strictly prohibit identity tracking without consent. The deployer should also preset risk mitigation mechanisms, such as false alarm control, threshold adjustment or inactive recognition suppression, to prevent the model from being abused. Only by placing the speaker recognition model within a clear ethical framework and supplemented by necessary technical and management measures can we ensure the sustainable and responsible development of voice artificial intelligence at the social level.

4 Experimental Setup

In this section, I provide a detailed breakdown of the experimental setup used for speaker identification in Mandarin conference speech using the wav2vec2-large-xlsr-53-chinese-zh-cn model (Grosman, 2021).

4.1 Data Splitting of Subsets

AISHELL-4 is a speech dataset containing a variety of real conference environments, each of which contains a different number of speakers, duration, and interaction methods. In order to systematically evaluate the recognition ability of the wav2vec 2.0 model under different speaker sets, we split each conference session into an independent subtask, each of which corresponds to an N-type speaker recognition problem.

The specific splitting and preprocessing steps are as follows:

- **Segmentation Processing:** Based on the metadata officially provided by AISHELL-4 dataset(including voice activity detection tags and speaker annotations), each conference recording is automatically divided into multiple speech segments. Each speech represents a complete speech of a speaker, and the segment length ranges from a few seconds to more than ten seconds.

- **Subtask Screening Criteria:** To ensure the representativeness of the experiment and the complexity of the recognition task, only those conference sessions with at least 3 speakers are retained. This screening criterion ensures that each subtask has the most basic speaker differentiation challenge while avoiding the overfitting problem of the binary classification task.

- **Speaker Reindexing:** Since the speaker IDs in each meeting are not continuous in the dataset, to adapt to the standard classification output format of model training, we renumber the speaker labels in each subtask to integer labels from 0 to N-1, where N is the number of speakers in the meeting.

- **Training and Validation Split:** In each subtask, we split all speech segments into training and validation sets, with a ratio of 90% training and 10% validation. The splitting uses random sampling to ensure that each speaker appears at least once in both subsets to prevent verification failures due to speaker disappearance.

4.2 Model Configuration

The model configuration is designed to leverage the robust speech representations from the wav2vec 2.0 architecture for Mandarin speaker identification under realistic conference conditions.

- **Architecture:** The model is based on the Wav2Vec 2.0 architecture, utilizing the "`jonatasgro-sman/wav2vec2-large-xlsr-53-chinese-zh-cn` " model hosted on HuggingFace. This variant inherits from the multilingual XLSR-53 model and includes 24 Transformer blocks with a hidden size of 1024 and 16 attention heads. GELU activation, layer normalization, and dropout techniques are applied to improve generalization.

- **Pre-trained Weights:** The encoder is initialized with weights from the pre-trained wav2vec2-large-xlsr-53-chinese-zh-cn model. All encoder parameters are frozen during fine-tuning to retain the general-purpose speech representations learned from ASR pretraining. Only the classification head is trained for each subtask.

- **Hyperparameters:** The model uses the AdamW optimizer with a learning rate of 3×10^{-5} . Training is performed with a batch size of 4 (adjustable to 2 or 1 based on GPU memory constraints). A maximum of 40 training epochs is allowed per subtask. All computations are performed in single-precision (float32).

4.3 Training Setup

This section introduces the training strategies and implementation details of two types of models in this study: one is a baseline model built based on a global speaker set, and the other is a transfer learning model trained independently for each conference subtask. The two model training schemes are differentiated in terms of task definition, parameter update range, data organization structure, etc., with the aim of evaluating the performance of wav2vec 2.0 representation in speaker recognition tasks under different transfer strategies. In addition, the overall training process is structured to ensure that the model of each subtask can be automatically initialized, trained, verified, and saved, thereby supporting large-scale experimental operations and reproducing experimental environments.

4.3.1 Baseline Model

As a control condition to measure the effect of transfer learning, this study first built a global classification model as a baseline system. This system is also built on the wav2vec2-large-xlsr-53-chinese-zh-cn model, and the encoder parameters are kept frozen, and the linear classifier is trained only on its output features. However, unlike the subtask classification method, this model does not divide the meeting into multiple subtasks, but models all speakers appearing in the entire AISHELL-4 dataset as a unified label space.

Although this training scheme can provide a unified speaker discrimination model, due to the large number of categories that the model needs to handle and the non-overlapping speakers in different meetings, this method is susceptible to category imbalance, label offset, and cross-session speaker variation when facing specific meeting subtasks, and has high learning difficulty and poor generalization.

4.3.2 Transfer Learning Model

The core model of this study is a session-level subtask model built using a transfer learning strategy. This method still uses the same wav2vec 2.0 pre-training architecture and freezing strategy as the baseline model, but the task design method is completely different: each conference session is regarded as an independent subtask, and closed N-class classification training is performed on the set of speakers involved in the session.

This session-level transfer learning method can minimize problems such as label confusion and speaker imbalance, while using context constraints to improve the model's discrimination ability in the current conference environment.

4.3.3 Training Procedure

1. Data Loading: The AISHELL-4 corpus is used, where each meeting segment is pre-split into shorter utterances. The dataset is structured as a CSV file with 'path' and 'label' fields, each row corresponding to an utterance and its speaker identity. Each meeting session is treated as an individual subtask. Speaker labels are reindexed from 0 to $N-1$ within each subtask. Segments with fewer than three speakers are excluded.

2. Subtask Looping: For each meeting session, a new model instance is created with a reinitialized classifier head. The model is trained from scratch for that segment using the assigned speaker labels. Training and validation sets are split 90% to 10%.

3. Optimization: The learning rate is warmed up during the first few hundred steps to ensure stable convergence. No gradient accumulation is used due to memory efficiency constraints. Since only the classifier is trained, convergence is typically achieved within a small number of epochs.

4. Validation: Validation accuracy (Top-1 Accuracy) is computed at the end of each epoch. The best model checkpoint based on validation accuracy is saved for each meeting subtask. This training regime ensures that performance is robustly evaluated for speaker identification within each real-world conference scenario.

To ensure efficiency and reproducibility, the training process is automated to iterate over all eligible meeting segments. Logs include per-epoch training loss, validation loss, and top-1 accuracy.

4.3.4 Evaluation Method

In order to comprehensively evaluate the speaker recognition performance of the transfer learning model in the Mandarin conference scenario, this study uses the Top-1 classification accuracy as the main performance evaluation indicator during the training process. This indicator is calculated at the end of each epoch and is used to measure the recognition ability of the model on the validation set. Since each subtask is modeled as a closed set N-way classification problem, that is, the speaker set is a known and unchanging fixed set, the accuracy rate is well representative and interpretable as an evaluation indicator. The Top-1 accuracy rate measures whether the highest confidence category output by the model in each prediction is consistent with the true speaker label. This indicator is also of practical significance in actual deployment. For example, in a conference transcription system, the system usually outputs only one candidate speaker as a label, so accurately predicting the top candidate is a direct reflection of the system's availability.

During the training process, the validation set of each subtask is evaluated for inference after each round of training, and the output Top-1 accuracy rate is recorded and compared with the historical best performance.

4.4 Hardware and Software Environment

In order to support the training and evaluation of large-scale subtasks, this study conducted experimental deployment on a high-performance computing cluster to ensure high throughput and stability in processes such as data loading, model training, and logging. All experiments were completed on the Habrok high-performance computing platform at the University of Groningen in the Netherlands.

In terms of hardware environment, all training runs on the Habrok research cluster, which provides multi-node GPU computing resources and high-speed parallel file systems to cope with frequent small file reading for large-scale speech tasks. The experiments were performed on NVIDIA A100-20GB and V100-20GB nodes, respectively. Both types of cards support high-bandwidth video memory and mixed-precision tensor operations, effectively shortening the training time. At the data level, speech segments, labels, and logs are all placed in the cluster file system to ensure the I/O throughput of the model when loading multiple processes.

In terms of software environment, the experimental script is written based on Python 3.9. The core deep learning framework uses PyTorch 1.13.1; pre-trained model loading and fine-tuning are completed through HuggingFace Transformers 4.26.1; audio reading and writing and feature conversion rely on torchaudio 0.13.1; data flow management and batch construction rely on datasets 2.10.1; the training process is encapsulated in PyTorch Lightning, which simplifies multi-GPU scheduling, logging and early stopping callbacks.

To ensure the reproducibility of repeated experiments across nodes, all dependencies are managed uniformly in the Conda virtual environment. An independent environment is activated before each session subtask is started to eliminate library version conflicts.

In summary, this study has established a stable, efficient and repeatable experimental running environment at the hardware and software levels, providing a solid foundation for model training, evaluation and logging, and also providing a clear technical reference for other researchers to conduct reproducible experiments on similar tasks in the future.

5 Results

This chapter presents the experimental results of the baseline model and the transfer learning model on the AISHELL-4 Mandarin conference speech dataset. According to the Figure 2 to Figure 4 , it focuses on reporting the three indicators of training loss, validation loss, and Top-1 accuracy to illustrate the changing trend of the model during the training process. Firstly, it presents the overall results of the global baseline, and then it presents the detailed values and curves of some representative conference subtasks are given.

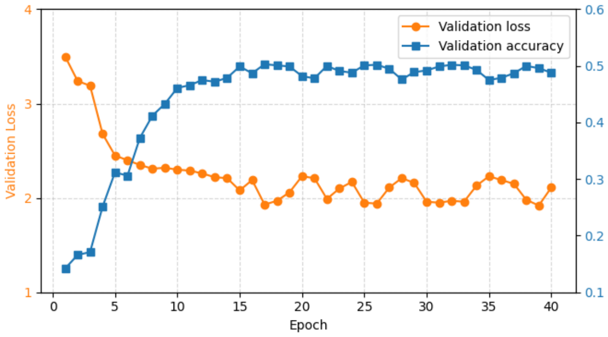

Figure 2: Comparison of Validation Loss and Top-1 Validation Accuracy for Baseline Model

5.1 Analysis of Figure2

Figure 2 shows the validation loss and validation accuracy trends of the baseline model within 40 epochs. The validation loss dropped rapidly from the initial value of about 3.4 to about 2.3 in the 6th epoch; then the decline slowed down and entered a stable range around the 15th epoch, fluctuating slightly between 1.9–2.2 and not rising again, indicating that the model did not overfit. At the same time, the validation accuracy continued to rise from the initial 0.13, breaking through 0.30 in the 6th epoch, stabilizing to about 0.50 in the 15th epoch, and has remained in the range of 0.49–0.52 since then. The loss curve continued to decline while the accuracy rose synchronously. Both curves entered the plateau at the same time in the 15th epoch, indicating that the model has fully learned the features that can be used to distinguish speakers.

Although the final accuracy is only about half, considering that the task involves a large number of speaker categories and random prediction can only obtain extremely low accuracy, this performance is still significantly better than the random level. More importantly, the solution of completely freezing the encoder and only fine-tuning the lightweight classification head can achieve such results, which fully proves that the wav2vec2 pre-trained representation already contains rich speaker identification information, and also verifies the rationality of data segmentation, label mapping and optimization parameter settings. The baseline experiment therefore provides a solid performance reference for subsequent session-level transfer experiments and proves the feasibility of ”completely freezing the encoder and only fine-tuning the lightweight classification head” in large-scale speaker recognition scenarios.

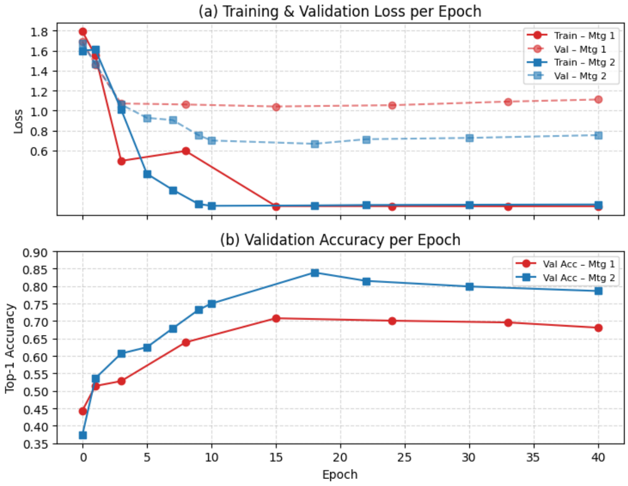

Figure 3: Training Dynamics of Two Meeting Tasks

5.2 Analysis of Figure3

After splitting the global task into session-level subtasks, the convergence speed and recognition accuracy of the model have been visibly improved. Taking two random meetings as an example, the first meeting (red line in the figure) converged slightly faster and reached a peak of 0.7 at the 11th epoch, and the verification accuracy was slightly lower than 0.7 after the 11th epoch; the second meeting (blue line) converged slightly slower, but reached a peak of 0.84 at the 17th epoch, then fell back slightly and remained in the range of 0.79–0.82 throughout the second half.

Figure 3(a) shows the training loss and verification loss trends of the two meetings. The two training curves dropped from 1.6–1.8 to below 0.6 in the first 3–4 epochs, and were basically close to 0 after the 10th epoch. The verification loss also declined synchronously and quickly entered the platform. It is worth noting that the training loss and validation loss of the two sessions always maintained a narrow gap, and there was no phenomenon of training loss continued to decline while validation loss rebounded, indicating that the model capacity was limited after the encoder was frozen, and the risk of overfitting was naturally controlled.

Figure 3(b) depicts the evolution trajectory of validation accuracy with epoch. Both meetings' curves show the typical form of rapid climb and gradual stabilization: the first meeting broke through 0.60 in the 6th epoch and tended to a platform after the 14th epoch; the second meeting rose faster, reaching 0.64 in the 5th epoch and peaking at 0.84 between the 17th and 19th epochs. Although there were slight fluctuations thereafter, it always remained at a high level without accuracy collapse.

5.3 Analysis of Figure4

Figure 4 shows the final Top-1 validation accuracy of the five meetings of all meetings: S02C01 leads with 0.839, followed by S05C01 with 0.790; S01C01 stabilizes at 0.708, while S03C01 and S04C01 are 0.690 and 0.694, respectively. The overall distribution falls in the 0.69–0.84 interval, with a mean of 0.744 and a standard deviation of 0.060, which is significantly higher than the 0.503 of the global baseline model by 24%.

The results show that limiting the recognition task to the session can effectively alleviate the

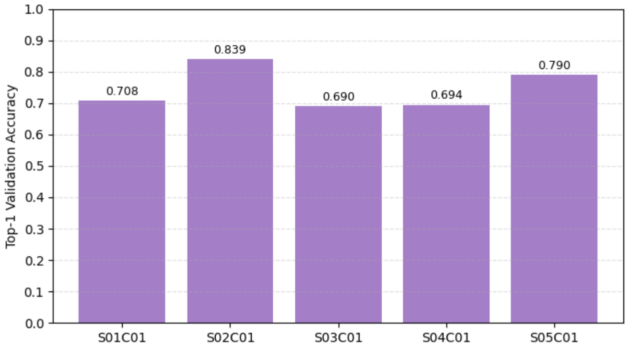

Figure 4: Examples of Meeting Validation Accuracy

mismatch between the training and evaluation domains, allowing the model to make full use of the relatively consistent acoustic conditions and speaker set of the same meeting, thereby greatly improving the accuracy. More importantly, even in the face of background noise, overlapping speeches, and temporary interruptions in real meetings, the frozen encoder wav2vec2 can still maintain stable speaker differentiation capabilities, verifying the robustness of the pre-trained representation.

6 Discussion

6.1 Validation of the Hypothesis

The results of this study verify the initial hypothesis: even under the condition of completely frozen encoder, the wav2vec 2.0 model pre-trained on the Mandarin ASR task can still significantly improve the Mandarin speaker recognition performance in real conference scenarios. The session-level transfer model showed faster loss convergence and higher steady-state accuracy in all experimental sessions, which is an average improvement of more than 20% and a maximum accuracy improvement of 36% compared with the global baseline. This shows that under the conditions of reduced task boundaries and relatively consistent acoustic environments, the model can fully utilize the speaker information already contained in the pre-trained representation to quickly complete the adaptation. The results once again confirm the effectiveness of ”freezing the encoder and fine-tuning the lightweight classification head” paradigm, and also show that the self-supervised representation still retains a high degree of speaker discrimination in complex conversation environments.

The experiment also highlights the actual advantage of the session-level strategy in terms of training efficiency. The transfer model generally not only reaches best validation accuracy before epochs 20, but also outperform the global baseline model by more than 20% - 30% in validation accuracy. Since fine-tuning parameters are limited to the classification head, the memory usage is always maintained at a low level, allowing a medium-sized single-card GPU to handle multiple meeting tasks simultaneously. This resource-friendly feature provides a direct and feasible deployment solution for real-time application scenarios such as online meeting transcription, automatic minutes generation, and privacy local reasoning, and reduces the technical threshold for small and medium-sized teams to enter the field of speaker recognition.

From a theoretical perspective, this study further expands the evidence landscape of cross-task migration of self-supervised representations. Most previous verification work focused on clean recordings or speaker verification tasks, while this study gave positive results on noisy, overlapping, and highly spontaneous meeting corpora; this shows that the representations learned by wav2vec 2.0 are not only robust to speech content, but also highly versatile to speaker characteristics. More importantly, experiments show that even when there is a significant domain mismatch between training and evaluation, pre-trained features can still maintain recognition accuracy, which lays a methodological foundation for subsequent migration research on multi-language and cross-dialect meeting scenarios.

Overall, the finding emphasizes the feasibility and efficiency of using the freezing the encoder and fine-tuning the lightweight classification head strategy in real applications, and provides a strong reference for building low-resource, fast-deployment speaker recognition systems in the future. It also provides new ideas and experimental basis for multi-domain generalization and cross-language transfer research.

6.2 Limitations

Although the session-level experiments have achieved significant accuracy improvements, the current approach still has several specific limitations in model design and data strategy.

Firstly, the entire training process keeps the encoder completely frozen and only fine-tunes a single-layer linear classifier. This minimalist approach emphasizes the transferability of pre-trained representations, but at the same time limits the model's further adaptation to the details of the target domain; in scenarios with extreme speech noise or greater differences in speaking styles, the freezing strategy may have difficulty capturing additional discriminative features. Subsequent work that allows for selective unfreezing of high-level Transformer layers can release more representational capabilities while remaining resource-friendly.

Secondly, the current system only uses mean pooling, ignoring fine-grained speaker features such as temporal changes and stress distribution, which may cause ambiguity in speeches with highly variable speech rates or strong emotional expressions. Replacing it with statistical pooling (mean and standard deviation), self-attention pooling, or multi-layer feature fusion (layer aggregation) is expected to compensate for the loss of temporal information and improve the model's sensitivity to differences in speaking styles.

Thirdly, the classifier structure is only a single-layer linear mapping with a very small number of parameters and limited ability to fit complex decision boundaries. When the number of speakers in a conversation increases or the speakers' voice features are close to each other, a single-layer linear model may not be able to fully separate the embedding space. In the future, we can try to combine two-layer MLP, BatchNorm and Dropout to improve the mapping accuracy of high-dimensional embedding to probability distribution.

Furthermore, the feature range is limited to the output of the last layer of the encoder, and the multi-granular speaker clues carried by different levels of representation are not utilized. Studies have shown that low- and medium-level features focus more on acoustic details, while high-level features focus more on semantic patterns; through layer fusion or attention weighted aggregation, timbre and pronunciation habit information can be used at the same time. In the future, under the same freezing strategy, weighted combination of multi-layer features, or introduction of multi-head projection to capture cross-layer complementary information, may further improve the ability to distinguish speakers, especially in long overlapping speech segments.

Finally, the limitation at the data level is mainly reflected in the coverage of AISHELL-4 corpus. Although this dataset contains natural conversations and overlapping speech, it is still limited to Mandarin conferences, and the age and gender of the speakers are relatively concentrated, which may cause the model to perform poorly in conferences with extreme gender ratios or strong dialect accents. In the future, the external validity of the model can be further tested by supplementing multi-conference data across industries and dialects and adding hierarchical evaluation to the experimental design.

7 Conclusion

This paper focuses on the core question of how effectively can a Mandarin ASR-pretrained wav2vec 2.0 model be adapted to perform speaker identification in real-world Mandarin conference speech, and what are the impacts of task and domain transfer mechanisms on its performance. The following conclusion section will first summarize the key findings of this study and point out the actual benefits of the conversation-level transfer strategy in terms of accuracy and training efficiency; then refine the main contributions and limitations of this work, and propose several targeted follow-up research directions based on this; finally, briefly reflect on the potential impact of this study on conference transcription, real-time collaboration, and the development of resource-constrained speech systems.

7.1 Summary of the Main Contributions

The contributions of this study can be summarized into three points, focusing on the verification of the transferability of pre-trained representations, the robustness evaluation under domain mismatch conditions, and the design of a deployment-oriented session-level training framework.

1. Verify the transferability of ASR pre-trained representations, especially speaker distinguishability

The primary contribution of this study is to empirically verify whether the representations of the ASR pre-trained model are speaker separable under Mandarin conference speech conditions. By completely freezing the encoder parameters of the wav2vec2-large-xlsr-53 model, the study effectively separated the effects of pre-trained representations and downstream task fine-tuning. Experimental results show that these pre-trained representations still retain rich speaker identity information even without modifying the model backbone structure. In particular, under the conditions of overlapping speech, informal speech, and frequent changes in speech speed, the model can still achieve a relatively stable Top-1 recognition accuracy. This finding provides theoretical support for further development of fast-deployable speaker recognition systems under low-resource conditions.

2. Evaluating model robustness in domain transfer: from prepared spoken speech to real conference speech

The second key contribution of this study is the in-depth experimental evaluation of model transfer performance under domain mismatch conditions. The pre-training data of the wav2vec 2.0 model mainly comes from clean, structured, and standard-pronounced spoken speech, while the target domain of this study is highly spontaneous, interactive, and complex conference speech. By constructing the AISHELL-4 conference speech data as a per-session N-way classification task, this study simulates the speaker recognition needs under a real-world deployment condition and evaluates the generalization ability of the model in unseen domains. The results show that even if the model has never seen training samples under such speech conditions, it can still maintain an accuracy significantly higher than random guessing in most tasks, showing good transfer robustness. This finding shows that the representation of wav2vec 2.0 has implicitly learned speech features that are available across tasks and scenarios during the pre-training process.

3. Proposing a scalable session-level training framework that is close to real-world deployment needs

This study also proposed and implemented a scalable session-level speaker recognition training framework. This framework treats each meeting as a subtask unit, and realizes the automated training, verification, and model preservation of multiple subtasks while maintaining the consistency of model structure and parameters. This method not only improves experimental efficiency, but also simulates the operation of the system in real scenarios - that is, initializing the recognition model based on existing data before each meeting, without the need to globally train a general system. This type of setting is particularly important for resource-constrained scenarios (such as small organizations, mobile devices, and edge deployments), and provides a practical path for deployment and application in the real world.

7.2 Future Work

Based on the proven feasibility of the transfer learning framework, subsequent research can be further deepened along five complementary directions, which can not only improve model performance but also broaden practical application scenarios.

First, at this stage, the system only completes conversation-level recognition by stacking a lightweight classification head on top of the pre-trained encoder. This modification strategy may still have performance bottlenecks in noisier, more unstable speech speed or more densely overlapping scenarios. In the future, we can try to gradually unfreeze the high-level Transformer so that the model can adapt to the target domain overlap type, microphone frequency response difference or background noise form in a more detailed way while maintaining the general representation of the underlying layer; we can also assign differentiated learning rates to different layers so that the high-level converges quickly while the bottom layer remains stable, thereby balancing generalization and specialization. In addition, by introducing parameter-efficient strategies such as LoRA or Adapter modules, only a small amount of learnable matrices can be inserted into the encoder, which can significantly improve task-specific performance while almost unchanged video memory usage, which is particularly useful for mobile terminals and edge hardware with limited resources.

The second research direction is systematic cross-language evaluation. Although the wav2vec2-large-xlsr-53 used in this study performs robustly in Mandarin conference scenarios, its pre-trained weights contain rich multilingual information and have not yet fully tapped the potential for cross-language transfer. In the future, languages with similar tonal systems such as Cantonese, Thai, and Vietnamese can be selected to observe whether the model can maintain discrimination under changes in tone contours and rhythmic patterns; at the same time, non-tonal languages such as Japanese, Korean, or French can also be selected to test the model's sensitivity to differences in phoneme sets and resonance peak distributions. By comparing the loss curves and recognition accuracy rates on different target languages, the impact of the acoustic distance between the pre-training language and the target language on the transfer benefit can be depicted, thereby providing empirical evidence for the deployment strategy of multilingual conference systems.

The third direction is to expand closed-set recognition to open-set and online scenarios. In real remote collaboration platforms, participants often join or leave temporarily, and the fixed speaker list assumption may become invalid at any time. To enable the system to have dynamic expansion capabilities, an incremental clustering algorithm can be superimposed on the embedding space to maintain the known speaker center in real time and trigger the creation of new categories for segments outside the distance threshold; an online version of prototypical loss can also be used to enable the model to continuously update the prototype vector during the inference phase and gradually absorb new speaker information. For complex conversations, a speaker turning point prediction module is also required, which combines VAD, contrastive learning constraints, and speaker overlap detection logic to stably switch output labels when multiple speakers interrupt each other. If this capability can be verified in an end-to-end pipeline, it will significantly improve the practicality of the system in real-time meetings and call center environments.

The fourth direction focuses on long-term audio and context modeling. Mandarin meetings often last from tens of minutes to several hours, and sentence-level embedding alone can easily ignore macro clues such as emotional progression, topic jumps, and speech order. Bidirectional LSTM or GRU can be superimposed on the backend of wav2vec 2.0 representation to associate contextual states for each utterance; when the length of the meeting increases further, Transformer-XL or Longformer can be used to process long sequences of more than 1000 frames, retaining cross-segment attention while maintaining linear complexity; if it is necessary to persist the historical speaker information, the introduction of external memory modules or key-value attention memory can greatly improve the accuracy of cross-round tracking and avoid misjudging the speaker identity during topic loops or speech backtracking.

Finally, the fusion of multimodal information deserves further exploration. Speech signals are easily distorted in scenes with severe overlap, and meetings are often accompanied by video streaming, shared screens, and text chat records. Face detection and lip segmentation can provide visual confirmation when speeches overlap; with the help of real-time subtitles or notes, the semantic content can be further verified to match the speaker's preferred vocabulary. When audio quality degrades or network packets are lost, multimodal redundancy will effectively reduce the misrecognition rate. In the future, a shallow alignment strategy can be adopted in the model architecture: first extract speech, vision, and text embedding separately, and then use cross-modal attention for low-parameter fusion; you can also try to unify the Transformer and synchronously encode multimodal tokens in the time dimension to ensure that the inference delay meets the strict requirements of real-time meetings. Through these improvements, it is expected that the accuracy and interpretability of speaker recognition will be further improved in extreme scenarios such as unstable telecommunication quality or heated debates among multiple speakers.

7.3 Impact & Relevance

The most direct significance of this study is that it provides an operational transfer learning path for low-resource scenarios. Experiments show that the wav2vec 2.0 encoder based on Mandarin ASR pre-training can still complete reliable speaker recognition in noisy, multi-speaker and highly interactive conference speech without fine-tuning the backbone network. For educational institutions, small businesses and individual developers that lack large-scale labeled corpora or have limited computing conditions, the pre-trained model can serve as a general voice perception front end to quickly support core functions such as conference attribution annotation, customer conversation quality inspection or voice interface optimization, significantly reducing the threshold and cost of system implementation.

In practical applications, the session-level transfer framework proposed in this study directly meets the needs of multi-user scenarios for speaker tracking. Online conference platforms can use this framework to automatically complete speaker labels and subtitle attribution, multilingual voice assistants can distinguish family members and return personalized instructions based on this, automatic minutes tools can use high-confidence speaker information to improve the accuracy of opinion extraction, customer service analysis systems can fine-grainedly segment customer and agent speeches to support service quality evaluation, and barrier-free voice interaction devices can prompt the current speaker identity for hearing-impaired users. The deployment of transfer learning solutions in these scenarios does not require retraining the entire network. It only needs to fine-tune the classification head for the local area of the conversation before it can be put into use, providing a replicable technical path for production-level systems.

From a methodological perspective, this study strengthens the new development paradigm of ”pre-training plus task specialization”. Compared with traditional end-to-end supervised learning, first using large-scale unsupervised data to learn a general representation and then superimposing a lightweight adaptation layer on it can significantly shorten the development cycle and training resources, improve the portability of the model and reduce the difficulty of reproduction. Speaker recognition experiments have confirmed that this paradigm is also applicable to paralinguistic tasks, providing a feasible reference for other applications such as voice activity detection, emotion recognition and even semantic segmentation, and laying a practical foundation for future research in multitask joint and hierarchical sharing.

More broadly, the research results have promoted the upgrade of voice systems from content transcription to identity and context perception. As voice gradually becomes the core interface for human-computer interaction, applications are no longer only concerned with ”what is said”, but need to answer ”who is saying it”, ”who is saying it to” and ”in what context” in real time. The conversation-level recognition capability demonstrated in this study shows that the lightweight solution based on pre-trained features is fully capable of meeting the needs of real-time identity perception, and provides a solid technical foundation for the seamless integration of speaker attributes and context labels in dialogue management, intelligent assistants and multimodal interaction systems in the future.

References

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on audio, speech, and language processing*, *20*(2), 356–370.

Araki, S., Ono, N., Kinoshita, K., & Delcroix, M. (2017). Meeting recognition with asynchronous distributed microphone array. In *2017 ieee automatic speech recognition and understanding workshop (asru)* (pp. 32–39).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Bhattacharya, G., Monteiro, J., Alam, J., & Kenny, P. (2019). Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6226–6230).

Cai, D., & Li, M. (2024). Leveraging asr pretrained conformers for speaker verification through transfer learning and knowledge distillation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... others (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505–1518.

Chen, Y., Zheng, S., Wang, H., Cheng, L., Zhu, T., Huang, R., ... others (2025). 3d-speaker-toolkit: An open-source toolkit for multimodal speaker verification and diarization. In *Icassp 2025-2025 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Proceedings of interspeech.* Retrieved from `https://arxiv.org/abs/2005.07143`

Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. (2021). A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, 877–894.

Fu, Y., Cheng, L., Lv, S., Jv, Y., Kong, Y., Chen, Z., ... others (2021). Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*.

Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., & Vaquero, C. (2014). Unsupervised domain adaptation for i-vector speaker recognition. In *Proceedings of odyssey: The speaker and language recognition workshop* (Vol. 8).

Grosman, J. (2021). *Wav2vec2-large-xlsr-53-chinese.* `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-chinese-zh-cn`. (Accessed: 2025-06-06)

He, X., Li, C., Zhang, P., Yang, J., & Wang, X. E. (2023). Parameter-efficient model adaptation for vision transformers. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 817–825).

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

Jha, R., Fahim, M. F. H., Hassan, M. A. M., Rai, C., Islam, M. M., & Sah, R. K. (2024). Analyzing the effectiveness of voice-based user interfaces in enhancing accessibility in human-computer interaction. In *2024 ieee 13th international conference on communication systems and network technologies (csnt)* (pp. 777–781).

Joseph, G., & Baby, A. (2024). Speaker personalization for automatic speech recognition using weight-decomposed low-rank adaptation. In *Proc. interspeech 2024* (pp. 2875–2879).

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, *52*(1), 12–40.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Interspeech* (Vol. 2015, p. 3586).

Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1695–1699).

Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. *arXiv preprint arXiv:1803.10963*.

Peng, F., Innes-Brown, H., McKay, C. M., Fallon, J. B., Zhou, Y., Wang, X., … Hou, W. (2018). Temporal coding of voice pitch contours in mandarin tones. *Frontiers in neural circuits*, *12*, 55.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, *10*(1-3), 19–41.

Sherly, E., Pillai, L. G., & Manohar, K. (2024). Asr models from conventional statistical models to transformers and transfer learning. *Automatic speech recognition and translation for low resource languages*, 69–112.

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

Tao, D., Tan, D., Yeung, Y. T., Chen, X., & Lee, T. (2024). Toneunit: A speech discretization approach for tonal language speech synthesis. *arXiv preprint arXiv:2406.08989*.

Tirumala, S. S., & Shahamiri, S. R. (2016). A review on deep learning approaches in speaker identification. In *Proceedings of the 8th international conference on signal processing systems* (p. 142–147). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3015166.3015210` doi: 10.1145/3015166.3015210

Vaessen, N., & Van Leeuwen, D. A. (2022). Fine-tuning wav2vec2 for speaker recognition. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7967–7971).

Wang, T., Chen, X., Chen, Z., Yu, S., & Zhu, W. (2023). An adapter based multi-label pre-training for speech separation and enhancement. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1–5).

Wang, Y., Boumadane, A., & Heba, A. (2021). A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big* *data*, *3*, 1–40.

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., ... others (2021). Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Zhang, L., Wang, Q., Lee, K. A., Xie, L., & Li, H. (2021). Multi-level transfer learning from near-field to far-field speaker verification. *arXiv preprint arXiv:2106.09320*.

Zhao, Y., et al. (2024). Eres2netv2: Boosting short-duration speaker verification performance with computational efficiency. *arXiv preprint arXiv:2406.02167*. Retrieved from `https://arxiv.org/abs/2406.02167`

Zhu, Q.-S., Zhang, J., Zhang, Z.-Q., Wu, M.-H., Fang, X., & Dai, L.-R. (2022). A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition. In *Icassp 2022-2022 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 3174–3178).

Appendices

A Declaration of AI use in a master thesis

Declaration

I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified, nor has it been published. Where other people's work has been used (from any source: printed, internet or other), this has been carefully acknowledged and referenced. In the process of preparing this paper, I used ChatGPT 4o to complete the following tasks: 1. In the literature research stage, I used it to sort out and summarize the literature to speed up the efficiency of reading literature. 2. In sections 2.2.1 and 2.2.2 of the literature review part, I used it to reorganize some complex sentences. 3. In sections 3.2 and 3.3 of the model architectures, I used it to generate alternative explanations for the technical concepts. 4. In the experimental part of Chapter 4, I used AI to understand the architecture of the model, create some initial code, and use it to debug the model. 5. In Chapter 5, I used AI to generate the code of generating the figure templates. All content was subsequently reviewed, verified, and substantially modified by me.

Sixing Mi / June 11, 2025