Imperial College London
Department of Computing

Enhancing Speech Intelligibility through Paralinguistic Features

Xiangheng He

Supervised by Prof. Björn Schuller

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing and the Diploma of Imperial College London.

2025

Declaration of Originality

I hereby declare that this thesis and the research it describes are my own work, and all research by other authors has been appropriately and explicitly referenced. This thesis has not been submitted for any other purpose.

Copyright Declaration

Abstract

Speech is fundamental to human communication, distinguishing it from other forms, such as gestures, body language, and facial expressions, which are shared with other primates. A key factor in effective communication is speech intelligibility, which enables individuals to clearly convey and comprehend thoughts, emotions, and intentions, fostering meaningful interactions and social connections. Despite the contributions of paralinguistic features in human speech, their role in enhancing speech intelligibility, particularly in the context of human-machine communication, remains inadequately studied. This gap highlights the need for a deeper investigation into incorporating paralinguistic features into speech synthesis and analysis systems to improve speech intelligibility and enhance the effectiveness of human-machine communication.

This thesis aims to enhance speech intelligibility by enabling machines to better interpret speaker intentions in human speech and produce more comprehensible machine-generated speech through the incorporation of paralinguistic features into speech processing systems. To achieve this, this thesis first introduces a proposed unsupervised, flexible, and perceptually aligned prosody modeling approach for text-to-speech synthesis (TTS). This approach enables the development of a prosody-aware TTS model that achieves natural prosody, strong generalizability, and fine-grained prosody control, significantly enhancing speech intelligibility compared to existing TTS systems. This thesis then explores the role of emotion-related paralinguistic tasks in enhancing a machine's ability to interpret speaker intentions, proposing a novel training strategy that prioritizes high-value paralinguistic tasks while mitigating negative transfer, resulting in a model that better interprets speaker intentions. This thesis further proposes an emotional voice conversion model that leverages disentangled emotional features to achieve clean and distinct emotional representations, reducing distortions in converted speech while preserving emotional clarity and enhancing intelligibility. Extensive experiments conducted with various open-source datasets demonstrate that these proposed models are superior to the current state-of-the-art models.

Acknowledgements

This thesis marks the culmination of my doctoral journey – an incredible and memorable trip filled with moments of joy and challenges, flowers and tears. I would like to express my heartfelt gratitude to the many people who have supported, guided, and encouraged me along the way.

First and foremost, I would like to express my gratitude to my supervisor, Prof. Björn Schuller, for providing me with the opportunity to study at Imperial College London, for inspiring my research in the field of speech analysis and generation, and for offering me a full-time position at the University of Augsburg. Your guidance, support, and kindness over the past four years have been invaluable to my academic and personal growth.

Moreover, I would like to thank all my colleagues and collaborators at Imperial College London and the University of Augsburg. In particular, I am deeply thankful to Dr. Anton Batliner for his patient and tireless efforts in helping me refine my understanding of linguistic and paralinguistic concepts, as well as for his invaluable advice. I also sincerely thank Andreas Triantafyllopoulos, Georgios Rizos, Jing Han, Zixing Zhang, Meishu Song, Zijiang Yang, Shuo Liu, Xin Jing, Manuel Milling, Alican Akman, and many others for the discussions, shared ideas, and mutual encouragement. Their insights and support have been instrumental in shaping and enriching my research journey.

Above all, I am profoundly grateful to my parents for their unconditional love and unwavering support, and to my boyfriend, Junjie, for his constant encouragement and companionship. Their love and strength have been the driving force that empowered me to overcome challenges and persevere through difficult times.

To My Beloved Mother and Father.

献给我亲爱的妈妈和爸爸

Publication List

During this doctoral research, the author has (co-)authored several peer-reviewed journal articles [TSİ+23, KHK+22] and conference papers [HCZS25, HCS24, HCRS21, HTK+22, KTH+22, CHM22, CHBM24, CHM24, CHMB25], some of which are discussed in this thesis. The following provides a exhaustive list of these publications.

Journal Papers

- **An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era**

  Andreas Triantafyllopoulos, Björn W Schuller, Gökçe İymen, Metin Sezgin, **Xiangheng He**, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, Ruibo Fu, Jianhua Tao

  Proceedings of the IEEE (IF: 23.2) 111 (10), 1355-1381, 2023

- **Personalised Depression Forecasting Using Mobile Sensor Data and Ecological Momentary Assessment**

  Alexander Kathan, Mathias Harrer, Ludwig Küster, Andreas Triantafyllopoulos, **Xiangheng He**, Manuel Milling, Maurice Gerczuk, Tianhao Yan, Srividya Tirunellai Rajamani, Elena Heber, Inga Grossmann, David D Ebert, Björn W Schuller

  Frontiers in digital health 4, 964582, 2022

Conference Papers

- **ProsodyFM: Unsupervised Phrasing and Intonation Control for Intelligible Speech Synthesis**

  Xiangheng He, Junjie Chen, Zixing Zhang, Björn W Schuller

  Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2025

- **Task Selection and Assignment for Multi-Modal Multi-Task Dialogue Act Classification with Non-Stationary Multi-Armed Bandits**

  Xiangheng He, Junjie Chen, Björn W Schuller

  International Conference on Acoustics, Speech, & Signal Processing (ICASSP), 2024

- **An Improved StarGAN for Emotional Voice Conversion: Enhancing Voice Quality and Data Augmentation**

  Xiangheng He, Junjie Chen, Georgios Rizos, Björn W Schuller

  Conference of the International Speech Communication Association (Interspeech), 2021

- **Depression Diagnosis and Forecast based on Mobile Phone Sensor Data**

  Xiangheng He, Andreas Triantafyllopoulos, Alexander Kathan, Manuel Milling, Tianhao Yan, Srividya Tirunellai Rajamani, Ludwig Küster, Mathias Harrer, Elena Heber, Inga Grossmann, David D Ebert, Björn W Schuller

  44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022

- **Journaling Data for Daily PHQ-2 Depression Prediction and Forecasting**

  Alexander Kathan, Andreas Triantafyllopoulos, **Xiangheng He**, Manuel Milling, Tianhao Yan, Srividya Tirunellai Rajamani, Ludwig Küster, Mathias Harrer, Elena Heber, Inga Grossmann, David D Ebert, Björn W Schuller

  44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022

- **Modeling Syntactic-semantic Dependency Correlations in Semantic Role Labeling using Mixture Models**

  Junjie Chen, **Xiangheng He**, Yusuke Miyao

  Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022

- **Unsupervised Parsing by Searching for Frequent Word Sequences among Sentences with Equivalent Predicate-Argument Structures**

  Junjie Chen, **Xiangheng He**, Danushka Bollegala, Yusuke Miyao

  Findings of the Association for Computational Linguistics (ACL), 2024

- **Language Model Based Unsupervised Dependency Parsing with Conditional Mutual Information and Grammatical Constraints**

  Junjie Chen, **Xiangheng He**, Yusuke Miyao

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 2024

- **Improving Unsupervised Constituency Parsing via Maximizing Semantic Information**

  Junjie Chen, **Xiangheng He**, Yusuke Miyao, Danushka Bollegala

  The Thirteenth International Conference on Learning Representations (ICLR), 2025

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Objectives

Human speech is the cornerstone of human communication, serving as the primary means through which individuals convey thoughts, emotions, intentions, and cultural values. It allows for immediate and dynamic interaction, facilitating real-time feedback and adaptation during conversations [Lev93]. Central to effective communication is speech intelligibility, which refers to the clarity and comprehensibility of spoken language. Speech intelligibility ensures that messages are accurately transmitted and understood between speakers and listeners.

Speech intelligibility is a multifaceted concept that involves the following layers of cognitive processing, which may occur either simultaneously or sequentially [MDBS12].

- The listener needs to perceive the auditory signals as speech, recognizing phonemes and words by matching sounds to their stored linguistic knowledge;
- The listener interprets the literal meanings of words and sentences using grammar and vocabulary;
- Beyond literal meanings, the listener must consider prior knowledge, contextual clues, and non-verbal signals to infer the speaker's underlying intention, thereby adjusting their

understanding and determining an appropriate response.

Paralinguistic features play a crucial role in facilitating the above three cognitive processes. Prosodic components like prominence and intonation aid in the recognition of individual words [VS10a], while phrasing indicates the phrasal organization within a sentence, allowing listeners to accurately discern its structure [HG14]. Paralinguistic features, including voice quality and speaker-specific traits, help listeners distinguish between individuals, reducing cognitive load in multi-speaker environments [Lav94]. Additionally, paralinguistic elements like non-verbal vocalizations—such as sighs and laughter—are important indicators of the speaker's intentions [SC03]. These paralinguistic cues regulate the flow of conversation by signaling turn-taking and topic shifts [EH05a, GH11]. They also convey the speaker's emotional state and social signals, providing additional context beyond the literal words [Cah90]. This enriched information enhances speech intelligibility by making communication more nuanced and comprehensible.

With the advancement of technology, speech has transcended human-to-human interaction and become a critical component in human-computer interaction. Effective communication between humans and machines relies on the ability of machines to both comprehend natural speech and synthesize speech responses that are clear and intelligible to humans. However, despite their importance in human speech, the significant roles that paralinguistic features play in enhancing speech intelligibility within human-computer interaction have not been well studied. For example, machine-synthesized speech often suffers from unnatural prosody and has not yet achieved a level of realism indistinguishable from human speech [LHR+24]. Likewise, machines such as customer service robots still struggle to accurately interpret the true intentions behind human speech, which hampers their ability to provide efficient and appropriate responses [WPK+18]. This highlights the need for further research into incorporating paralinguistic features into speech-processing systems to improve speech intelligibility and enhance the effectiveness of human-computer communication.

In this thesis, we focus on enhancing speech intelligibility by enabling machines to better understand human speech and making machine-generated speech more comprehensible, through the incorporation of paralinguistic features into speech processing systems. Specifically, we

address this through three key areas with three objectives:

- We develop a prosody-aware text-to-speech synthesis (TTS) system. By integrating prosodic components such as intonation and phrasing into the TTS process, our objective is to produce machine-generated speech that is more human-like and intelligible;
- We investigate emotion-aided speech dialogue act classification (DAC). Recognizing that emotions are a crucial paralinguistic phenomenon and that speech dialogue acts reflect the speaker's intentions, our objective is to enhance the machine's understanding of underlying intentions, thus capable of providing efficient and appropriate responses;
- We explore disentangled emotional features for enhancing Emotional Voice Conversion (EVC). By disentangling and accurately incorporating emotional features, our objective is to reduce distortions in machine-generated speech and improve its intelligibility.

1.2 Challenges and Problem Statement

The three areas and objectives presented above pose several challenges. Firstly, despite the dynamic and variable nature of speech prosody, current transcription systems for prosody have significant limitations. They often lack flexibility and do not align well with human perception of prosody, hindering their ability to accurately capture prosodic nuances. Secondly, although existing studies have explored incorporating paralinguistic features to enhance speech dialogue act classification via multi-task learning, the utility of emotion-related tasks remains unclear, and negative transfer phenomena have been observed, which can degrade performance. Thirdly, current EVC models are prone to transferring emotion-independent information that should be preserved, leading to severe distortions in the speech after conversion.

In light of these challenges, the contributions of this thesis are to address three primary research questions (RQ) and two supplementary RQ, which are summarized as follows.

- **RQ-1: How can we describe and model prosody to achieve more intelligible TTS?**

To address this question, instead of using existing prosody labeling systems, this thesis (in Chapter 3) focuses on exploring a more perceptually aligned, flexible, and unsupervised approach to model prosody within TTS systems.

- **RQ-1a (side): Can the prosody in TTS synthesized speech be customized?**

  To tackle this question, the modeling of prosody in Chapter 3 is not entirely within an opaque latent space; rather, it partially employs tokens that are interpretable and controllable by humans. This approach allows for the customization of prosody in the generated speech, enabling users to freely customize the prosodic features of the synthesized speech.

- **RQ-2: How can we adjust the training strategy to let paralinguistic features better facilitate the machine's understanding of the speaker's intentions?**

  To deal with this question, this thesis (in Chapter 4) models the training problem in multi-task learning as a non-stationary multi-armed bandit problem. This method enables the selection of useful tasks at different stages of training and helps to avoid negative transfer phenomena.

- **RQ-3: How can emotional features be accurately incorporated into the EVC model?**

  To address this question, this thesis (in Chapter 5) explores the use of explicitly disentangled emotional features to obtain clean emotional representations. This approach prevents information interference when transferring these emotional features, reducing distortions in the synthesized speech and improving its naturalness and intelligibility.

- **RQ-3a (side): Can the speech after conversion be used for data augmentation?**

  To tackle this question, this thesis (in Chapter 5) investigates using the converted speech for data augmentation, evaluating its impact on the performance of Speech Emotion Recognition (SER) models.

1.3 Thesis Structure

This thesis is structured into six main chapters as follows.

Chapter 1 presents an introduction to the thesis, outlining its motivation, objectives, key challenges, and research problems, followed by an overview of the thesis structure and publications included in the thesis.

Chapter 2 provides the general background of the thesis, covering the definition of paralinguistics and paralinguistic features, their roles in enhancing speech intelligibility, their existing labeling or modeling approaches, and the research gaps associated with these approaches.

Chapter 3 presents a prosody-aware TTS system designed to improve the intelligibility of synthesized speech by incorporating prosody. The chapter begins with background information on the Flow Matching generative modeling framework and the alignment estimation task essential for TTS. It then discusses the challenges of prosody modeling, including break labeling and intonation modeling issues, and introduces a novel, perceptually aligned, flexible, and unsupervised approach for prosody modeling. The proposed training objectives, model components, and information flow during training and inference are detailed. Experimental results, including model performance, model generalizability, an ablation study, and case studies, are reported.

Chapter 4 introduces a model for emotion-aided speech dialogue act classification, focusing on improving the machine's ability to understand speaker intentions through paralinguistic features. The chapter begins with an overview of the non-stationary multi-armed bandit problem and its potential to address limitations in current multi-task learning approaches. It proposes a novel strategy to dynamically prioritize useful tasks during training while mitigating negative transfer phenomena. This chapter outlines the proposed model architecture, training strategies, and experimental details. The experimental results on task selection, model performance, model stability, and misclassification analysis are reported.

Chapter 5 presents an EVC model that incorporates disentangled emotional features to reduce distortions in synthesized speech while preserving emotional clarity. This chapter begins by

providing background information on feature disentanglement techniques and the StarGAN framework, followed by an analysis of the research gaps in emotion modeling for EVC. It proposes a solution to explicitly disentangle emotional features, capturing clean emotional representations and minimizing interference during conversion. The chapter details the model architecture, training, and inference processes and explores the use of converted speech for data augmentation, evaluating its impact on the SER task. The experimental results on model performance and data augmentation are reported.

Chapter 6 summarises the achievements, limitations, and future perspectives. In addition, this chapter outlines the ethical considerations.

1.4 Included Publications

The publications included in this thesis are as follows.

- **ProsodyFM: Unsupervised Phrasing and Intonation Control for Intelligible Speech Synthesis**

  Xiangheng He, Junjie Chen, Zixing Zhang, Björn W Schuller

  Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) 2025

- **Task Selection and Assignment for Multi-Modal Multi-Task Dialogue Act Classification with Non-Stationary Multi-Armed Bandits**

  Xiangheng He, Junjie Chen, Björn W Schuller

  International Conference on Acoustics, Speech, & Signal Processing (ICASSP) 2024

- **An Improved StarGAN for Emotional Voice Conversion: Enhancing Voice Quality and Data Augmentation**

  Xiangheng He, Junjie Chen, Georgios Rizos, Björn W Schuller

  Conference of the International Speech Communication Association (Interspeech) 2021

Chapter 2

General Background

2.1 Paralinguistics

2.1.1 Definition

'Paralinguistics' means 'alongside linguistics' (from the Greek preposition $\pi\alpha\rho\alpha$) [SB13]. It refers to phenomena that are distinct from typical linguistic phenomena, such as the structure of a language, its phonetics, its syntax, its morphology, or its lexical semantics. Paralinguistics is the study of speech and language, both of which are primary means of communication. Loosely speaking, paralinguistics focuses on *how* something is said rather than *what* is said [SB13].

The definition of paralinguistics has undergone long-term development and refinement. In this thesis, the definition follows the lines of [SB13, BHS20, BNB+23, Bat24]. As illustrated in Figure 2.1, paralinguistics encompasses three vocal and verbal aspects:

1) [-vocal, +verbal], which refers to the connotations of words and phrases in (written) language, such as sentiment analysis;

2) [+vocal, +verbal], which involves elements modulated onto or entailed in speech, such as prosody;

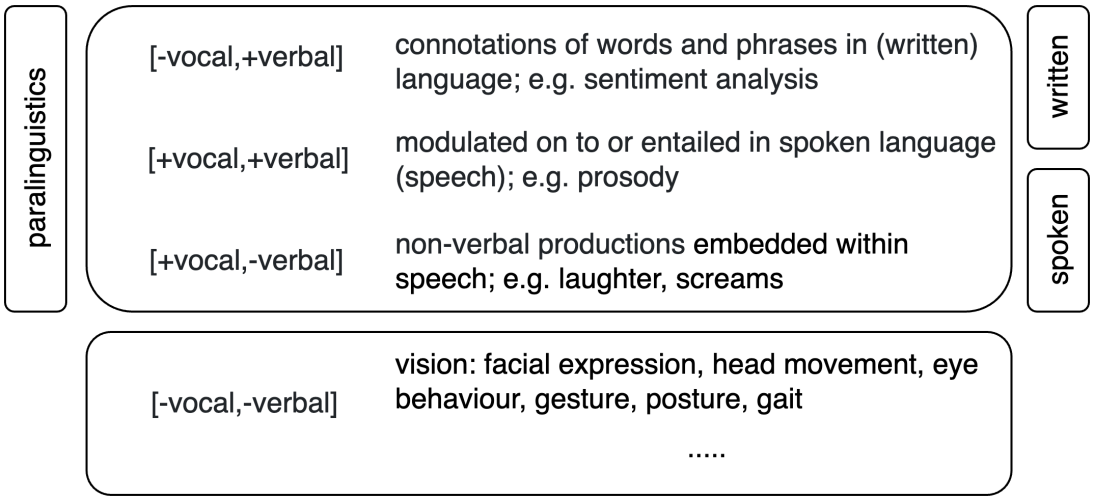


Figure 2.1: The definition of paralinguistics (adapted from [Bat24]).

3) [+vocal, -verbal], which includes non-verbal productions embedded within speech, such as laughter or screams.

In comparison to the definition adopted in this thesis, the literature presents both broader and narrower definitions of paralinguistics. The broader definition [HRS08] includes all non-verbal manifestations in conversation, i. e., it also encompasses the [-vocal, -verbal] aspect, such as facial expressions and gestures in the visual modality. The narrower definition [Cry11] limits paralinguistics to only [+vocal] aspects, disregarding the tight links between phonetics and linguistics.

2.1.2 Research Scope of This Thesis

Paralinguistics is an umbrella term that includes a broad field. Paralinguistic features in this thesis refer to those that serve paralinguistic functions. This thesis focuses on three key paralinguistic features: prosodic, emotional, and speaker intention features, along with their interactions.

In Chapter 3, we study how modeling prosody, which falls under the [+vocal, +verbal] aspect of paralinguistics, improves the intelligibility of synthesized speech.

In Chapter 4, we investigate how incorporating emotional features from both text and speech

enhances the machine's understanding of underlying speaker intentions. Text-based emotional and speaker intention features fall under the [-vocal, +verbal] aspect, while speech-based features fall under the [+vocal, +verbal] aspect of paralinguistics.

In Chapter 5, we explore disentangled emotional features to enhance Emotional Voice Conversion. These features are extracted from the spectral envelope, which falls under the [+vocal, -verbal] aspect of paralinguistics.

2.2 Prosody

2.2.1 Definition

Prosody refers to the rhythmic and melodic aspects of speech [N+97], specifically the patterns of:

1) intonation, i. e., pitch levels on individual syllables and their combination into contours;

2) phrasing, i. e., the grouping of words into meaningful units or chunks that help convey the structure and meaning of an utterance;

3) prominence (stress), i. e., the emphasis placed on particular syllables or words within an utterance;

4) rhythm, i. e., the pattern of stressed and unstressed syllables that give speech its flow and timing;

5) duration, i. e., the length of time a syllable, word, or phrase is produced;

6) loudness (intensity), i. e., variations in volume across an utterance.

Different prosodic components and their combinations can serve both linguistic and paralinguistic functions. Linguistically, prosodic components assist listeners in the recognition of lexical words, providing cues to sentence type, syntactic structure, and grammatical boundaries, such

as using intonation to distinguish between questions and statements. Paralinguistically, on the other hand, prosodic components can be used for conveying non-literal information, such as a speaker's emotions, intentions, or social cues, without altering the grammatical content.

2.2.2 Prosody Enhancing Speech Intelligibility

In the context of enhancing speech intelligibility discussed in this thesis, the boundary between prosody serving linguistic functions and paralinguistic functions is difficult to delineate clearly. For example, prosodic components like intonation can aid in recognizing the grammatical structure and word identity while also conveying the speaker's emotional state or attitude, such as sarcasm or emphasis. This overlap highlights the complex role of prosody in enhancing speech intelligibility, where linguistic and paralinguistic functions work in tandem to support both the literal interpretation and the deeper, more nuanced understanding of the message being communicated. It can be reflected in the following four aspects.

1) Prosody can help resolve ambiguities in speech [ST03]. Prosodic components like phrasing help listeners segment the continuous speech stream into meaningful units. These segmentations imply the phrasal organization in the sentence, allowing listeners to accurately discern the syntactic structure of the sentence and deduce its correct meaning [HG14];

2) Prosody can aid in the recognition of individual words [VS10a]. Correctly stressed syllables, combined with syllables that have appropriate durations, can help listeners identify words more quickly and accurately, improving the overall intelligibility of speech;

3) Prosody can regulate the flow of conversation, indicating turn-taking, topic shifts, and speaker intent through variations in prosodic elements like intonation, stress, and loudness [EH05a]. It helps listeners anticipate what comes next and provides cues about when the speaker is finished or wants to emphasize a point, contributing to a more fluent and comprehensible exchange;

4) Prosody can convey the speaker's emotional state, intentions, and social cues. This additional information provides context beyond literal words, which makes speech more intelligible.

2.2.3 ToBI Prosody Transcription System

In order to model and analyze prosody effectively in speech analysis and speech synthesis, it is essential to provide both qualitative and quantitative descriptions of speech prosody. The Tones and Break Indices (ToBI) transcription system [SBP+92] provides a standardized annotation framework for labeling the prosody in English speech. A standard ToBI transcription comprises four tiers of labels aligned with the speech signal according to time:

1) an orthographic tier, which captures the textual content;

2) a tonal tier, used to annotate pitch-related events such as accents and boundary tones;

3) a break index tier, which marks the perceived degree of juncture or the strength of association between each pair of words;

4) a miscellaneous tier, reserved for supplementary comments or remarks about the utterance.

Next, we will provide a detailed explanation of the tonal tier and the break index tier. It is important to note that the following discussion pertains exclusively to the English language.

The tonal tier models how speakers use pitch to highlight important information, structure utterances, and regulate the flow of conversation. It has three key components: pitch accents, boundary tones, and phrase accents.

Pitch accents are localized pitch movements that occur on stressed syllables, making them prominent within the speech signal. These pitch movements signal emphasis, contrast, or key information, helping listeners interpret meaning and structure in speech. In the ToBI system, pitch accents are labeled using combinations of high (H) and low (L) tones, including:

- H*: A high-pitch accent marked by a rise in pitch on the stressed syllable;
- L*: A low-pitch accent signifying a drop or sustained low pitch;
- L+H*: A complex pitch accent starting with a low tone and rising to a high tone, often indicating contrast or focus;
- H+L*: A high tone followed by a falling pitch, typically signaling strong emphasis or a completed statement.

Boundary tones mark the edges of intonational phrases and indicate the speaker's intention to complete or continue speaking. In the ToBI system, the primary boundary tones are:

- H%: A high boundary tone, which typically indicates incompletion or that the speaker intends to continue speaking;
- L%: A low boundary tone, signaling the conclusion of a thought or a complete statement.

Phrase accents occur between the final pitch accent and the boundary tone, acting as a transitional element that maintains the intonational contour across the phrase. In the ToBI system, phrase accents are typically categorized as follows:

- H-: A high phrase accent, signaling a rise in pitch, often conveying continuation or uncertainty;
- L-: A low phrase accent, indicating a pitch fall, typically associated with a more conclusive tone.

The Break Index tier labels the perceived degree of disjuncture between words, phrases, or larger prosodic units, implying the phrasing and rhythmic structure of speech. The break indices range from 0 to 4, each representing a different level of perceptual boundary.

- Break Index 0: No perceptible boundary. It commonly occurs in cases of liaison or cliticization, where words are tightly connected and pronounced without a perceptible break;
- Break Index 1: Normal word boundary. It is the most common break index and represents a typical boundary between words;
- Break Index 2: Perceived juncture with no intonation effect. It indicates a weak yet noticeable disjuncture between words, suggesting prosodic marking without signaling a full intonational boundary;
- Break Index 3: Intermediate phrase boundary. It marks a distinct prosodic break. It is typically linked with phrase accents in the tonal tier, signaling a noticeable separation without concluding the full intonational phrase;
- Break Index 4: Full intonational phrase boundary. It represents a full intonational phrase boundary typically occurring at the end of an intonational phrase or an utterance, indicating the completion of a thought or a sentence.

2.2.4 Limitations of ToBI

Although ToBI provides a standardized framework to measure speech prosody, it has been argued in the literature that ToBI has certain drawbacks [BDKG12].

1) Human perception of prosodic events does not strictly align with the categorical distinctions defined by the ToBI system. Some studies [Dil05, Red03] suggest that listeners often perceive either more or fewer prosodic categories than those represented in ToBI. For instance, evidence from various human perception studies [Cal07, WTG08, Dil10] indicates that human listeners do not perceive contours corresponding to H* and L+H* as categorically distinct, which leads to issues with inter-transcriber reliability;

2) ToBI lacks a transparent mapping between labeling distinctions and perceptual events, often requiring annotators to assign labels even in the absence of clear perceptual cues.

For example, ToBI requires that annotators assign break indices of '3' or '4' even when there is no perceptual disjuncture, solely based on the presence of a phrase accent or boundary tone in the tonal tier [Wig02];

3) ToBI fails to label certain distinctions that are critical for linguistic and cognitive studies [NR03]. For example, ToBI is constrained by a binary system of prominence (accented vs. unaccented), which does not allow for coding rhythmic patterns of speech, both of which are important to language acquisition and the perception of mature language [BCF+07].

4) The assumptions of the ToBI system do not generalize well across languages [LMS00], posing challenges in applying it to languages with different prosodic structures, making it difficult for labelers to adopt and use consistently.

The drawbacks of (1) and (2) motivate us to explore a more flexible approach to describe and model prosody in the text-to-speech synthesis system discussed in Chapter 3, aiming to better align with human perceptual paradigms.

2.3 Emotion and Speaker Intention

Emotion and speaker intention are important phenomena in paralinguistics. They play a crucial role in enhancing speech intelligibility by providing additional cues that complement the linguistic content, thereby facilitating the listener's comprehension. In order to model and analyze these phenomena effectively in speech analysis, it is essential to offer both qualitative and quantitative descriptions of them.

2.3.1 Modeling Emotional Features

Previous studies have extensively explored the features of speaker emotions. Two primary approaches dominate the field: categorical representations of discrete emotion classes and dimensional models based on affective continua.

Discrete Emotion Classes

The discrete emotion framework categorizes emotions into distinct classes, such as anger, happiness, sadness, fear, and neutral. These categories are often defined based on universally recognized facial expressions, physiological responses, and vocal cues [Ekm92]. In speech research, such discrete classes usually operate on labeled datasets [BBL+08, LR18] to enable machine learning models to identify specific emotion classes from acoustic and prosodic features [DZMS13, ZCDS14, TTN+17]. The discrete emotion framework provides clear interpretability and is consistent with the intuitive human understanding of emotions. However, it has been critiqued for its inability to capture the complexity and overlaps of human emotions, as evidenced by the theory of constructed emotion [Bar06, Bar17] and meta-analysis [PWTL02], both highlighting that emotions are constructed experiences without distinct neural or categorical boundaries.

Valence-Arousal-Dominance Framework

The Valence-Arousal-Dominance (VAD) framework offers a dimensional perspective, describing emotions along three continuous axes: valence (the pleasantness of a stimulus), arousal (the intensity of emotion provoked by a stimulus), and dominance (the degree of control exerted by a stimulus) [WKB13]. This model is able to capture the subtle variations in emotional expression. For example, while both anger and rage are unpleasant emotions (low valence), rage has a higher intensity or a higher arousal state. The VAD model's quantitative nature allows for finer distinctions between emotions and better accounts for ambiguous or mixed emotional states [GP10]. However, the interpretation of valence, arousal, and dominance values is subjective, influenced by individual and cultural differences, and requires robust methods to accurately map speech features to the three-dimensional space.

In Chapter 4, we integrate both discrete emotion recognition and VAD regression tasks as complementary information sources to provide emotional features for enhancing speech dialogue act classification. In Chapter 5, we examine clean (disentangled) emotional features to mitigate

distortion in speech synthesized by the emotional voice conversion model.

2.3.2 Modeling Speaker Intention Features

Speech intelligibility, the clarity and accuracy with which spoken language is understood, is a core aspect of human communication. It relies heavily on the ability to interpret a speaker's intentions beyond their words' literal meaning. A primary way these intentions are conveyed is through speech acts [Tom23] and their conversational counterpart, dialogue acts [SRC+00], which are fundamental to interaction. While speech acts focus on the intent behind individual utterances, dialogue acts expand this concept to the broader conversational context, considering the role of each utterance within the flow of a conversation [Wei16]. Dialogue acts cover various communicative functions, such as statements, questions, backchannels, and apologies, and are essential for understanding the speaker's intention and the dynamics of a conversation.

Previous studies proposed various dialogue act tagsets to describe dialogue acts. The DAMSL (Dialogue Act Markup in Several Layers) framework provides a dimensional annotation system that captures multiple aspects of dialogue functions, such as 'Inform' and 'Request-Information' [CA97]. The SWBD-DAMSL (Switchboard DAMSL Tagset), adapted from DAMSL for the Switchboard Corpus, offers 42 discrete classes tailored to spontaneous conversational speech [SRC+00]. The HCRC Map Task Corpus Tagset focuses on directive and interactional acts and provides discrete labels like 'Instruct', 'Explain', and 'Query-YN' [ABB+91]. The AMI (Augmented Multi-party Interaction) Corpus Tagset, designed for multi-party meeting dialogues, introduces discrete categories such as 'Evaluate' and 'Clarify' for collaborative decision-making tasks [CAB+05]. The ISO 24617-2 Annotation Standard, established to ensure consistency across domains and languages for dialogue act annotation, includes core communicative acts like 'Statement', 'Question', 'Promise', and interaction management acts [BAC+12]. These tagsets have significantly advanced the annotation and analysis of dialogue acts across various domains and applications.

Modeling dialogue acts involves identifying and classifying dialogue act tags of each utterance within a conversation. Early studies employed probabilistic methods to predict dialogue acts,

leveraging features such as lexical content, syntactic structure, and prosodic cues [SSJ+98, SRC+00, BGH+11]. With advancements in computational power, neural networks have been increasingly employed to classify dialogue acts by training on annotated corpora [SRC+00, SPSB20]. Recent advancements also include integrating other related tasks to capture complex patterns in dialogue data. For example, [FGEB21] incorporated intent detection and slot-filling tasks to enhance dialogue act classification, while [SPSB20, SGSB21] introduced emotion-related tasks for the same purpose.

However, despite being crucial paralinguistic features for enhancing speech intelligibility, the interaction between speaker intention features and emotional features remains underexplored. In Chapter 4, we investigate the role of different emotional features in improving dialogue act classification across different training stages.

Chapter 3

Prosody-aware Text-to-Speech Synthesis

Prosody contains rich information beyond the literal meaning of words, which is crucial for the intelligibility of speech. Current TTS models still fall short in many prosody aspects, especially in phrasing and intonation; they not only miss or misplace breaks when synthesizing long sentences with complex structures but also produce unnatural intonation. In this chapter, we propose ProsodyFM, a Prosody-aware TTS model with a flow matching (FM) backbone that aims to enhance the phrasing and intonation aspects of prosody. ProsodyFM introduces two key components: a Phrase Break Encoder to capture initial phrase break cues, followed by a duration predictor to realize flexible adjustment of break durations; and a Terminal Intonation Encoder which integrates a set of intonation shape tokens combined with a novel pitch processor for more robust modeling of human-perceived intonation change. ProsodyFM is trained with no explicit prosodic labels and yet can uncover a broad spectrum of break duration and intonation patterns. Experimental results demonstrate that ProsodyFM can effectively improve the phrasing and intonation aspects of prosody, thereby enhancing the overall intelligibility compared to four state-of-the-art (SOTA) models. Out-of-distribution experiments show that this prosody improvement can further bring ProsodyFM superior generalizability for unseen complex sentences and speakers. Our case study intuitively illustrates the powerful and fine-grained

controllability of ProsodyFM over phrasing and intonation.

This chapter is based on [HCZS25] *ProsodyFM: Unsupervised Phrasing and Intonation Control for Intelligible Speech Synthesis*, a conference paper accepted by AAAI 2025 on August 15, 2024. I am the first author of this paper. I identified key research gaps, proposed improvement strategies, designed the components of the ProsodyFM model, conducted the entire experimental process, and wrote the conference paper. The second author, Junjie Chen, contributed by discussing and refining the initial ideas and experimental design with me and provided insightful input during the revision of the paper. The third and final authors, Prof. Zixing Zhang and Prof. Björn Schuller, supervised the project throughout, offering valuable suggestions and guidance on manuscript development.

3.1 Introduction

As outlined in Chapter 2.2.1, prosody encompasses several key components of speech, including intonation, phrasing, prominence, rhythm, duration, and loudness, which convey information beyond the literal meaning of words. It plays a vital role in enhancing speech intelligibility, as detailed in Chapter 2.2.2. Although recent TTS systems have achieved great progress in synthesizing intelligible speech, they still lack in many prosody aspects. The Tones and Break Indices (ToBI) transcription system provides a standardised framework for measuring speech prosody and offers a potential reference for modelling prosody both qualitatively and quantitatively in TTS systems. However, as described in Chapter 2.2.4, ToBI system has certain limitations, such as the misalignment between human perception of prosodic events and the categorical distinctions defined by ToBI, and its lack of flexibility in considering human annotators' perceptual experiences. Given these limitations and the dynamic and variable nature of speech prosody, we seek a more perceptually aligned, unsupervised approach to model prosody within TTS systems. In this chapter, we specifically focus on two aspects of speech prosody in English: phrasing and intonation.

Phrasing refers to grouping words into chunks. An intonational phrase contains a chunk of

Sentence “I saw the man with the telescope.”

Sentence “I saw the man [break] with the telescope.”

Figure 3.1: Ambiguity caused by different phrasing.

words with their own intonation pattern. **Phrase break** in this chapter refers to a perceivable acoustic pause at the end of intonational phrases (namely, Break Index 4 in the ToBI system). Phrase break plays an important role in enhancing speech intelligibility [FPYT21]. It implies the phrasal organization in the sentence, allowing listeners to accurately discern the syntactic structure of the sentence and deduce its correct meaning. For example, as shown in Figure 3.1 [1], the sentence “I saw the man with the telescope.” can be interpreted differently depending on whether there is a break after the word “man”. When the sentence is spoken without the break, “with the telescope” modifies “the man”, suggesting that the man observed by the speaker had a telescope. When the break is introduced after “man”, it implies the speaker used a telescope to see the man. This demonstrates that incorrect phrasing can lead to incorrect interpretation of the sentence, thereby impairing speech intelligibility. However, due to the difficulty in obtaining break labels and the variability of break duration, current TTS systems usually miss or misplace breaks when synthesizing complex sentences.

Intonation, especially terminal intonation, is essential for synthesizing intelligible speech. We refer to the intonation pattern of the last word in an intonational phrase as the **terminal**

[1] These two images are generated by DeepAI AI Image Generator.

intonation. Terminal intonation carries many linguistic and paralinguistic information in English. A rising terminal intonation at the end of a sentence usually signals uncertainty or a request for clarification, while a falling intonation typically indicates certainty or is used to make statements and assertions [Lib75]. A rising terminal intonation in the middle of a sentence usually indicates the speaker is not finished yet, while a falling tone indicates the end of the thought [Bol98]. This information of intonation change can be represented through the change in the pitch contour [CST22]. However, instead of modeling the relative change in the pitch contour, previous TTS systems directly model the absolute pitch value. This design choice hinders their ability to accurately capture natural intonation, as pitch tracking and predicting absolute pitch values is inherently challenging.

To address these issues, we propose ProsodyFM, a novel Prosody-aware TTS model based on a Flow Matching (FM) backbone that enhances both phrasing and intonation aspects of prosody in an unsupervised manner, resulting in more intelligible synthesized speech. For the break labeling issue, we introduce a Phrase Break Encoder to capture initial break cues, followed by a duration predictor to adjust break durations, enabling flexible and accurate modeling of phrase breaks. For the intonation modeling issue, we employ a set of intonation shape tokens combined with a novel pitch processor, which effectively mitigates pitch tracking errors, enables more robust modeling of pitch shapes, and aligns more closely with human perception of intonation changes. ProsodyFM is trained without any prosodic labels and yet can uncover a wide range of break duration and intonation patterns.

3.2 Background

In this section, we begin by presenting the details of the key technique underlying ProsodyFM: Flow Matching. We then introduce text–speech alignment as a critical task in TTS and provide a detailed explanation of the Monotonic Alignment Search method proposed in Glow-TTS [KKKY20], as ProsodyFM employs the same technique. Next, we present causal evidence that phrasing and intonation directly improve listeners' memory and comprehension of spoken

language and enhance ASR performance. We then compare diffusion- vs. flow-matching–based generative architectures and non-autoregressive vs. autoregressive frameworks, and justify our architectural choices. Finally, we discuss TTS evaluation and explain the rationale behind our evaluation design.

3.2.1 Flow Matching Overview

Generative Modeling

Generative modeling is a machine learning paradigm that aims to capture the underlying probability distribution of data. This data of interest can be complex, encompassing high-dimensional and multimodal structures such as natural images, speech signals, and textual data, as well as intricate biological and physical systems. By learning the intrinsic patterns and structures present within the observed data, generative models enable the synthesis of new data points that are statistically consistent with the original data.

Formally, the generative modeling task can be described as having data samples $x_1, x_2, \ldots, x_n$ drawn from an unknown target distribution $p_1(x)$. The objective is to leverage these samples to learn a probabilistic model that approximates $p_1(x)$. Most generative models can be uniformly conceptualized as the process of learning a transformation $\phi : \mathbb{R}^d \to \mathbb{R}^d$ that maps samples from a known simple distribution $p_0(x)$ to the target distribution $p_1(x)$. In this way, by sampling from the simple distribution $x_0 \sim p_0(x)$ and applying the learned transformation ϕ, the model can generate new samples $\phi(x_0)$ that are approximately distributed according to $p_1(x)$.

Normalizing Flows

Normalizing Flows (NFs) are a class of generative models that operate by transforming a simple distribution through a series of invertible mappings, thereby generating a more complex and potentially multimodal distribution.

Let $\mathbb{R}^d$ denote the data space. Let $\phi : \mathbb{R}^d \to \mathbb{R}^d$ be a invertible continuously differentiable

function that maps elements of $\mathbb{R}^d$ onto $\mathbb{R}^d$. Its inverse function $\phi^{-1} : \mathbb{R}^d \to \mathbb{R}^d$ is also continuously differentiable. Let $p_0(x)$ be a known distribution on $\mathbb{R}^d$ and $p_1(y)$ be the induced distribution of y obtained after transforming the sample x of $p_0(x)$ by the mapping ϕ.

$$x \sim p_0 \quad y = \phi(x),$$

The density of $p_1(y)$ can be computed through the change-of-variable theorem:

$$p_1(y) = p_0(\phi^{-1}(y)) \left| \det \left[\frac{\partial \phi^{-1}}{\partial y}(y) \right] \right| \tag{3.1}$$

where the $\frac{\partial \phi^{-1}}{\partial y}$ is the Jacobian matrix of the inverse map.

In practical applications, it is often necessary to transform a simple distribution p_0 (e. g., a Gaussian distribution $\mathcal{N}(0, I)$) into a complex distribution that approximates the data distribution of real images or speech signals. A single, simple transformation ϕ is typically insufficient to achieve this objective. Consequently, neural networks are employed as flexible transformations, denoted as a flow ϕ_θ. Let the parametric density induced by the flow ϕ_θ be denoted as $p_1 \triangleq [\phi_\theta]_{\#} p_0$.

The central task then becomes optimizing the neural network's parameters θ. A straightforward objective for learning the parameters $\theta \in \Theta$ is to maximize the log-likelihood (or equivalently, minimize the negative log-likelihood) over the training dataset $\mathcal{D}$.

$$\arg\max_\theta \mathbb{E}_{y \sim \mathcal{D}} \left[\log p_1(y) \right]$$

Optimizing this objective requires computing the determinant of the Jacobian matrix of ϕ_θ^{-1}. However, computing this determinant is generally computationally expensive for a single highly complex transformation ϕ_θ. To alleviate the computational complexity, residual normalizing flows [BGC+19] represent the Jacobian as a perturbation $u_k(x)$ with a scaling factor δ of a simplified base transformation.

$$\phi_k(x) = x + \delta u_k(x) \tag{3.2}$$

Low-rank residual normalizing flows [BHTW18] utilize the low-rank perturbation $u_k(x)$, offering high computational efficiency at the cost of reduced expressiveness. In contrast, full-rank residual normalizing flows [BGC+19] employ a full-rank perturbation $u_k(x)$, providing a more balanced trade-off between expressiveness and computational efficiency. One can construct a new flow $\phi_{K:1}$ by composing multiple full-rank residual flow modules, each of which allows efficient computation of the determinant.

$$\phi_{K:1} = \phi_K \circ \ldots \circ \phi_2 \circ \phi_1.$$

The log-likelihood of the model is determined by aggregating the Jacobian determinants of each residual flow module.

$$\log p_1(y) = \log p_0(\phi^{-1}(y)) + \sum_{k=1}^{K} \log \det \left[\frac{\partial \phi_k^{-1}}{\partial x_{k+1}} (x_{k+1}) \right]$$

where $x_{k+1} = \phi_K^{-1} \circ \cdots \circ \phi_{k+1}^{-1}(y)$.

Continuous Normalizing Flows

In the residual flow formulation (Equation 3.2), the number of residual flow transformations K serves as a hyperparameter that requires tuning. Notably, by setting $\delta = \frac{1}{K}$ and taking the limit as $K \to \infty$ under certain conditions (u_k to be Lipschitz continuous and continuous by the Picard–Lindelöf Theorem), the composition of residual flows $\phi_K \circ \cdots \circ \phi_2 \circ \phi_1$ can be expressed as an ordinary differential equation (ODE).

$$\frac{d\phi_t}{dt} = u_t(\phi_t(x_0)) \tag{3.3}$$

where x_0 denote the initial condition. The flow $\phi_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ maps x_0 to x_t at time t. The value of x_t can be obtained by solving the ODE:

$$x_t \triangleq \phi_t(x_0) = x_0 + \int_0^t u_s(x_s)\,\mathrm{d}s.$$

According to the Continuity Equation (also known as the Transport Equation), the distribution induced by the flow ϕ_t satisfies the following formula.

$$\frac{\partial}{\partial t} p_t(x_t) = -\left(\nabla \cdot (u_t p_t)\right)(x_t) \tag{3.4}$$

where u_t is a vector field $u_t = [u_t^1, u_t^2, \cdots, u_t^d]^\top$ mapping $\mathbb{R}^d \to \mathbb{R}^d$, p_t is a scalar field mapping $\mathbb{R}^d \to \mathbb{R}$, and x_t is a vector $x_t = [x_t^1, x_t^2, \cdots, x_t^d]^\top$ in $\mathbb{R}^d$. Combining Equation 3.3 and Equation 3.4, the total derivative of the induced density in log-space yields:

$$\frac{d}{dt} \log p_t(x_t) = -\left(\nabla \cdot u_t\right)(x_t). \tag{3.5}$$

The complete proof process is as follows.

$$\begin{aligned}
\frac{d}{dt} \log p_t(x_t) &= \frac{1}{p_t(x_t)} \frac{d}{dt} p_t(x_t) \\
&= \frac{1}{p_t(x_t)} \left[\frac{\partial}{\partial t} p_t + \frac{\partial}{\partial x_t} p_t \frac{d}{dt} x_t \right] \\
&= \frac{1}{p_t(x_t)} \left[\frac{\partial}{\partial t} p_t + \langle \nabla_{x_t} p_t, \frac{d}{dt} x_t \rangle \right] \\
&= \frac{1}{p_t(x_t)} \left[-\left(\nabla \cdot (u_t p_t)\right)(x_t) + \langle \nabla_{x_t} p_t, \frac{d}{dt} x_t \rangle \right] \\
&= \frac{1}{p_t(x_t)} \left[-p_t(x_t)(\nabla \cdot u_t)(x_t) - \langle \nabla_{x_t} p_t, u_t(x_t) \rangle + \langle \nabla_{x_t} p_t, \frac{d}{dt} x_t \rangle \right] \\
&= -\left(\nabla \cdot u_t\right)(x_t)
\end{aligned}$$

The second equation is derived from the application of the chain rule for the total derivative. The third equation follows from the definition of the inner product and the gradient of the scalar field, where $\nabla_{x_t} p_t$ is given by $\nabla_{x_t} p_t = \left[\frac{\partial}{\partial x_t^1} p_t, \frac{\partial}{\partial x_t^2} p_t, \ldots, \frac{\partial}{\partial x_t^d} p_t \right]$. The fourth equation is

obtained by substituting the continuity equation (Equation 3.4). The fifth equation results from the application of the product rule for divergence, and the final equation is derived from Equation 3.3.

One can parameterize a vector field neural network $u_\theta : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ and solve the corresponding joint ODE to obtain the induced log density $\log p_t$.

$$\frac{d}{dt}\begin{pmatrix} x_t \\ \log p_t(x_t) \end{pmatrix} = \begin{pmatrix} u_\theta(x_t) \\ -\left(\nabla \cdot u_\theta\right)(x_t) \end{pmatrix} \tag{3.6}$$

Just as in discrete NFs, Continuous Normalizing Flows (CNFs) can be trained by maximizing the log-likelihood over the training dataset $\mathcal{D}$.

$$\arg\max_\theta \mathbb{E}_{x_1 \sim \mathcal{D}}\left[\log p_1(x_1)\right]$$

This process entails integrating the time evolution of both x_t and its log density $\log p_t(x_t)$ as described in Equation 3.6. This requires computationally expensive numerical ODE simulations during training and efficient estimators for the divergence to handle high-dimensional spaces. Typically, the CNFs training can be significantly slow due to the ODE integration at each iteration. Consequently, researchers have explored "simulation-free" methods to train CNFs, eliminating the need for integration during training.

Flow Matching and Conditional Flow Matching

Flow matching is a simulation-free way to train the CNFs models. Back to Equation 3.6, both ODEs are a function of the parametric vector field $u_\theta(x_t)$. FM seeks to directly formulate a regression objective with respect to $u_\theta(x_t)$.

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1]} \mathbb{E}_{x_t \sim p_t}\left[\left\|u_\theta(x_t) - u_t(x_t)\right\|^2\right] \tag{3.7}$$

where $u_t(x_t)$ is a ground truth vector field that induces a probability path p_t, namely, they satisfy the Continuity Equation 3.4. However, this ground truth vector field $u_t(x_t)$ is unknown. Conditional Flow Matching (CFM) constructs a known conditional probability path $p_{t|1}(x_t|x_1)$ by conditioning on data samples $x_1 \sim p_1$.

$$p_t(x_t) = \int p_1(x_1) p_{t|1}(x_t \mid x_1)\, dx_1$$

the conditional probability path $p_{t|1}(x_t|x_1)$ has certain boundary conditions:

$$p_{t=0|1}(x \mid x_1) = p_0, \quad p_{t=1|1}(x \mid x_1) = \mathcal{N}(x; x_1, \sigma_{\min}^2 \mathbf{I}) \quad \text{with } \sigma_{\min} > 0 \text{ and } \sigma_{\min} \to 0.$$

It can be shown that if the conditional vector field $u_t(x_t \mid x_1)$ induces the conditional probability path $p_{t|1}(x_t \mid x_1)$, then the marginal vector field $u_t(x_t)$ (which is of interest to us) that satisfies the following Equation 3.8 can induce the marginal probability path p_t, namely, the marginal vector field $u_t(x_t)$ and the marginal probability path p_t satisfy the Continuity Equation 3.4.

$$u_t(x_t) = \mathbb{E}_{x_1 \sim p_{1|t}} \left[u_t(x_t \mid x_1) \right] = \int u_t(x_t \mid x_1) \frac{p_{t|1}(x_t \mid x_1) p_1(x_1)}{p_t(x_t)} dx_1 \tag{3.8}$$

$$\frac{\partial p_{t|1}(x_t \mid x_1)}{\partial t} = -\nabla \cdot \big(u_t(x_t \mid x_1) p_{t|1}(x_t \mid x_1) \big) \tag{3.9}$$

Given that Equation 3.9 and Equation 3.8 hold, we aim to prove that Equation $\frac{\partial p_t(x_t)}{\partial t} = -\nabla \cdot (u_t(x_t) p_t(x_t))$ is also valid. The complete proof is shown in Equation 3.10.

This indicates that the objective of optimizing the marginal vector field can be achieved by optimizing the conditional vector field, as long as the marginal vector field satisfies Equation 3.8. Therefore, the unknown marginal vector field $u_t(x_t)$ in Equation 3.7 can be replaced by the conditional vector field $u_t(x_t \mid x_1)$ according to Equation 3.8, which forms the optimization objective of CFM.

$$
\begin{aligned}
\frac{\partial p_t(x_t)}{\partial t} &= \frac{\partial}{\partial t} \int p_{t|1}(x_t \mid x_1) p_1(x_1)\, dx_1 \\
&= \int \frac{\partial}{\partial t} \left(p_{t|1}(x_t \mid x_1) \right) p_1(x_1)\, dx_1 \\
&= -\int \nabla \cdot \left(u_t(x_t \mid x_1) p_{t|1}(x_t \mid x_1) \right) p_1(x_1)\, dx_1 \\
&= -\int \nabla \cdot \left(u_t(x_t \mid x_1) p_{t|1}(x_t \mid x_1) p_1(x_1) \right)\, dx_1 \\
&= -\nabla \cdot \int u_t(x_t \mid x_1) p_{t|1}(x_t \mid x_1) p_1(x_1)\, dx_1 \\
&= -\nabla \cdot \left(\int u_t(x_t \mid x_1) \frac{p_{t|1}(x_t \mid x_1) p_1(x_1)}{p_t(x_t)} p_t(x_t)\, dx_1 \right) \\
&= -\nabla \cdot \left(\int u_t(x_t \mid x_1) \frac{p_{t|1}(x_t \mid x_1) p_1(x_1)}{p_t(x_t)}\, dx_1 \cdot p_t(x_t) \right) \\
&= -\nabla \cdot \left(u_t(x_t) p_t(x_t) \right)
\end{aligned}
\tag{3.10}
$$

$$
\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}[0,1], x_1 \sim p_1, x_t \sim p_{t|1}(x_t|x_1)} \left[\| u_\theta(x_t) - u_t(x_t \mid x_1) \|^2 \right] \tag{3.11}
$$

It can be demonstrated that the gradient of Equation 3.7 is equivalent to the gradient of Equation 3.11 (the proof process ensures that Equation 3.8 is satisfied).

$$
\nabla_\theta \mathcal{L}_{\text{FM}}(\theta) = \nabla_\theta \mathcal{L}_{\text{CFM}}(\theta)
$$

This implies that the parametric vector field $u_\theta(x_t)$ can be efficiently trained by optimizing Equation 3.11.

The relationship between the vector fields $u_t(x_t)$, $u_t(x_t|x_1)$ and their induced marginal and conditional densities is shown in Figure 3.2.

Gaussian Conditional Probability Path

The CFM objective in Equation 3.11 is applicable to any choice of conditional vector field and conditional probability path. As a practical implementation, one can construct $p_{t|1}(x_t \mid x_1)$ and

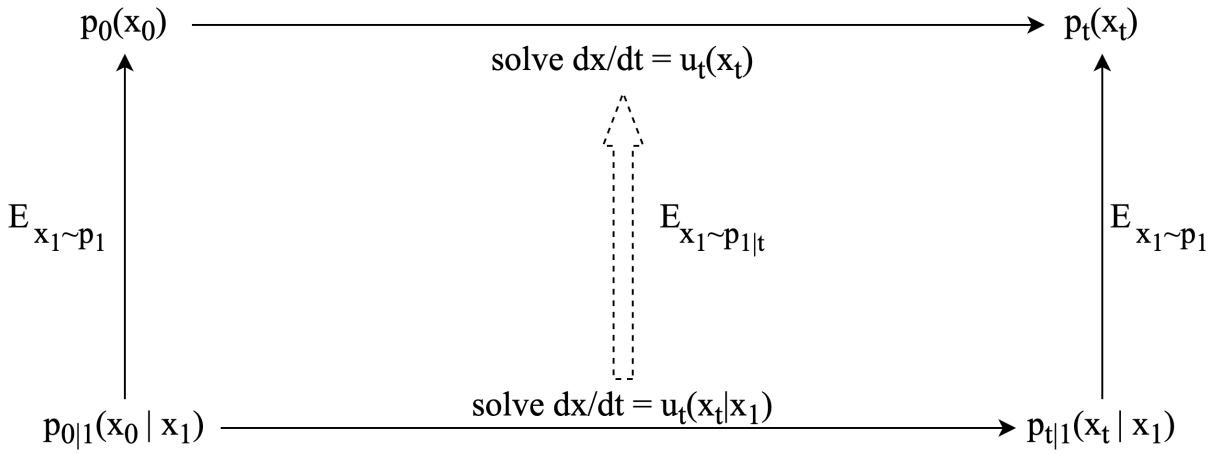


Figure 3.2: Illustration of the relationship between the vector fields $u_t(x_t)$, $u_t(x_t|x_1)$ and their induced marginal and conditional densities.

$u_t(x_t \mid x_1)$ using a general family of Gaussian conditional probability paths.

$$p_{t|1}(x_t \mid x_1) = \mathcal{N}(x_t; \mu_t(x_1), \sigma_t(x_1)^2 \mathbf{I}) \tag{3.12}$$

where $\mu_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma_t : [0,1] \times \mathbb{R} \to \mathbb{R}^+$ are the time-dependent mean and standard deviation of the Gaussian distribution. The boundary condition at time $t = 0$ and $t = 1$ can be set to:

$$p_{t=0|1}(x \mid x_1) = \mathcal{N}(x; 0, \mathbf{I}), \quad p_{t=1|1}(x \mid x_1) = \mathcal{N}(x; x_1, \sigma_{min}\mathbf{I}) \quad \text{with } \sigma_{\min} > 0 \text{ and } \sigma_{\min} \to 0.$$

There are infinitely many vector fields that can generate this Gaussian conditional probability path $p_{t|1}(x_t \mid x_1)$, such as by adding a divergence-free component to the continuity equation. While such components do not alter the underlying distribution, they introduce additional computational complexity. To minimize unnecessary computational overhead, one can opt for the simplest vector field that corresponds to a linear transformation flow $\psi_t(x_t \mid x_1)$.

$$x_t \triangleq \psi_t(x_t \mid x_1) = \mu_t(x_1) + \sigma_t(x_1)x_0, \quad \text{with } x_0 \sim p_{0|1}(x_0 \mid x_1) = \mathcal{N}(x_0; 0, \mathbf{I})$$

According to the change-of-variable theorem 3.1 (or simply the reparameterization trick of Gaussian distribution), this simple linear flow $\psi_t(x_t \mid x_1)$ pushes the initial distribution $p_{t=0|1}(x_0 \mid$

$x_1) = \mathcal{N}(x_0; 0, \mathbf{I})$ to $p_{t|1}(x_t \mid x_1) = \mathcal{N}(x_t; \mu_t(x_1), \sigma_t(x_1)^2\mathbf{I})$.

$$[\psi_t]_{\#} p_{0|1} = p_{t|1}$$

The flow's corresponding conditional vector field $u_t(x_t|x_1)$ takes these two closed forms:

$$u_t(x_t \mid x_1) = \sigma_t'(x_1)x_0 + \mu_t'(x_1) \tag{3.13}$$

$$= \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}\left(x_t - \mu_t(x_1)\right) + \mu_t'(x_1). \tag{3.14}$$

where the first equation is computed from $u_t(x_t \mid x_1) = \frac{d}{dt}\psi_t = \frac{d}{dt}[\mu_t(x_1) + \sigma_t(x_1)x_0]$; the second equation replace the x_0 in the first equation according to $x_t = \mu_t(x_1) + \sigma_t(x_1)x_0$.

Optimal-Transport Conditional Flow Matching

Optimal-Transport Conditional Flow Matching (OT-CFM) represents a specific case of the Gaussian conditional probability path. In OT-CFM, the mean and standard deviation are defined to change linearly over time.

$$\mu_t(x_1) \triangleq t x_1 \quad \text{and} \quad \sigma_t(x_1) \triangleq (1-t) + t\sigma_{\min}$$

According to Equation 3.13,3.14, the closed-form conditional vector field is:

$$\begin{aligned} u_t(x_t \mid x_1) &= x_1 - (1 - \sigma_{\min})x_0 \\ &= \frac{x_1 - (1-\sigma_{\min})x_t}{1 - (1-\sigma_{\min})t}. \end{aligned}$$

The conditional flow that corresponds to $u_t(x_t \mid x_1)$ is

$$\psi_t(x_t \mid x_1) = \left(1 - (1 - \sigma_{\min})t\right)x_0 + t x_1$$

In this case, the conditional flow is represented by the Optimal Transport (OT) displacement map between the two Gaussian distributions, p_0 and p_1. In contrast to other complex choices of vector fields, such as the diffusion vector field, the OT vector field exhibits straight-line trajectories with a constant velocity. This characteristic potentially simplifies the associated regression task.

3.2.2 Text-to-speech Synthesis based on OT-CFM

A generative TTS model simulates the distribution of real speech signals given certain conditions μ, such as the target text with speaker identity. Our goal is to pinpoint a distribution that maximizes the likelihood of the ground truth speech. Conditional Normalizing Flow [CRBD18] implicitly models this target distribution by solving an Ordinary Differential Equation (ODE) with which we can find a trajectory that transforms a random sample into speech. FM [LCB+23] is a paradigm for training Conditional Normalizing Flow by directly regressing the ODE vector field that defines this transformation. FM requires the conditional constructions of vector fields and probability paths. Optimal-Transport Conditional Flow Matching (OT-CFM) [LCB+23] provides a simple and intuitive way to construct conditional probability paths via interpolation, enhancing the efficiency of both training and sampling.

3.2.3 Alignment Estimation between Text and Speech

The estimation of alignment between text and speech is a crucial task in TTS systems. Since the target text typically contains a limited number of words, whereas the corresponding speech signal has a considerably longer frame length, poor alignment between text and speech can lead to severe errors, such as skipped phonemes, missing words, and unnatural timing in pronunciation. Previous parallel TTS models, such as FastSpeech[RRT+19] and ParaNet[PPSZ20], address this length mismatch between text and speech by extracting alignments from pre-trained autoregressive TTS models. Glow-TTS [KKKY20], however, introduced Monotonic Alignment Search (MAS) to estimate alignment without relying on external aligners. MAS operates under

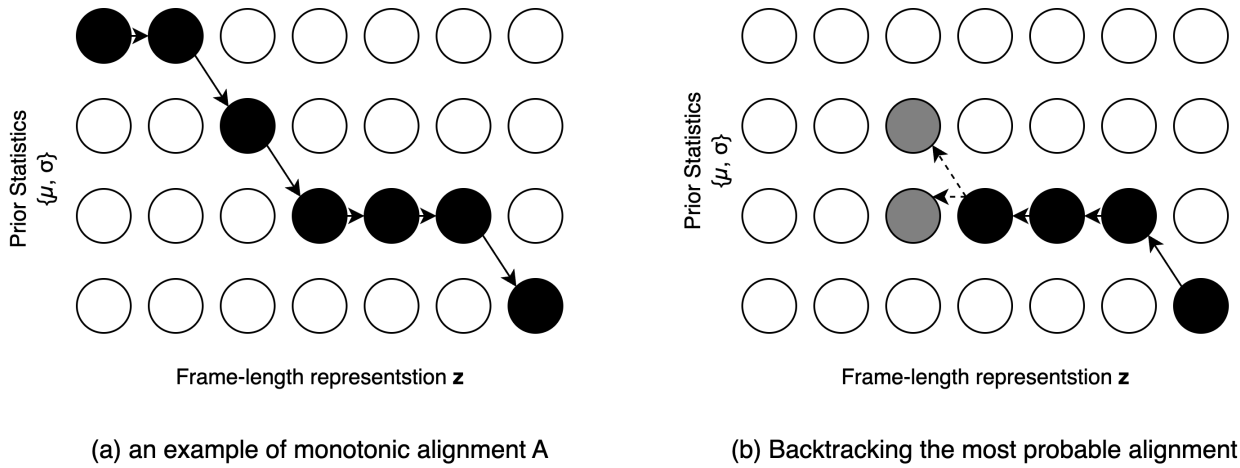


Figure 3.3: Illustrations of the MAS.

the assumption that the alignment path can only progress forward in time (from left to right) without any skip or backward steps, aligning well with the objective of the TTS task, where words or phonemes need to be generated sequentially in time without omissions or reversals.

Monotonic Alignment Search in Glow-TTS

In Grow-TTS, the target conditional distribution of mel-spectrograms, $p(x_1|c)$, is modeled by transforming a conditional prior distribution $p(z|c)$ using a flow-based decoder $f_{\text{dec}} : z \rightarrow x_1$, where x_1 represents the input mel-spectrogram, and c denotes the condition to generate the mel-spectrogram. The objective is to maximize the log-likelihood of the data samples:

$$\log p_1(x_1|c) = \log p(z|c) + \log \left| \det \frac{\partial f_{\text{dec}}^{-1}(x_1)}{\partial x_1} \right|$$

The prior distribution $p(z|c)$ is an isotropic multivariate Gaussian distribution, where the mean vector $\mu = \mu_{1:T_{text}}$ and the covariance matrix $\sigma = \sigma_{1:T_{text}}$ are obtained by the conditional encoder f_{enc}. Here, T_{text} denotes the number of words in the input text. The latent representation z should have the same length as x_1, namely, $z = z_{1:T_{mel}}$, where T_{mel} denotes the number of frames in the target mel-spectrogram. MAS searches for the most probable monotonic alignment A between the phone-level statistics of the prior distribution (μ and σ) and the frame-level latent

representation of speech z. Specifically, $A(j) = i$ means that the j-th latent representation z_j follows the distribution $\mathcal{N}(z_j; \mu_i, \sigma_i)$, where μ_i and σ_i are the statistics associated with index i.

The goal for training is to find the network parameters θ and the alignment A that maximize the log-likelihood

$$\max_{\theta,A} L(\theta, A) = \max_{\theta,A} \log p_1(x_1|c; \theta, A)$$

To avoid the computational intractability to find the global solution, this objective can be reduced to two steps: (i) finding the most probable monotonic alignment A^* for the current parameters θ, as in Equation 3.15 and (ii) updating θ to maximize the log-likelihood given A^*.

$$A^* = \arg\max_A \log p_1(x_1|c; A, \theta) = \arg\max_A \sum_{j=1}^{T_{\text{mel}}} \log \mathcal{N}(z_j; \mu_{A(j)}, \sigma_{A(j)}) \tag{3.15}$$

The MAS is introduced to solve the Equation 3.15. Let $Q_{i,j}$ be the maximum log-likelihood where the prior statics μ_i, σ_i are aligned with the latent variable z_j.

$$Q_{i,j} = \max_A \sum_{k=1}^{j} \log \mathcal{N}(z_k; \mu_{A(k)}, \sigma_{A(k)})$$

Then due to the monotonic assumption (as shown in Figure 3.3), $Q_{i,j}$ can be recursively formulated with $Q_{i-1,j-1}$ and $Q_{i,j-1}$.

$$Q_{i,j} = \max(Q_{i-1,j-1}, Q_{i,j-1}) + \log \mathcal{N}(z_j; \mu_i, \sigma_i)$$

Using dynamic programming, one can efficiently find A^* by caching Q values and backtracking from the end of the alignment, where $A^*(T_{\text{mel}}) = T_{\text{text}}$, as illustrated in Figure 3.3 (b).

3.2.4 Prosody–intelligibility Causality

In Chapter 2.2.2, we discussed how prosody facilitates enhanced speech intelligibility. In this chapter, we present causal evidence demonstrating that the phrasing and intonation aspects of prosody play a direct role in improving listeners' memory and comprehension of spoken

language, as well as contributing to improved performance in ASR systems.

Firstly, previous controlled studies provide clear causal evidence that natural phrasing and intonation reduce listeners' memory load. [RGN+04] conducted a series of controlled experiments on how phrasing (pause placement) and intonation (pitch variation) influence memory while keeping lexical and temporal properties constant. Participants listened to short narrative utterances comprising one to five intonational phrases, presented under four prosodic conditions: (i) natural speech, (ii) monotone (flattened pitch contour), (iii) pause-free, and (iv) both monotone and pause-free. Memory performance was measured through listeners' ability to recall the sentences immediately after hearing them, serving as an index of cognitive load during speech processing. For short utterances containing one or two intonational phrases, memory performance remained high across all conditions; however, as sentence length increased, memory performance declined substantially when either pitch variation or phrasing cues were removed, with the steepest decline observed in the five-phrase condition. Complementary controlled studies [OTO68, Leo73, Fra85, PTGK00] also found reliable memory costs when phrasing and intonation were selectively degraded. These findings provide clear causal evidence that natural phrasing and intonation reduce listeners' memory load by facilitating speech segmentation and information retention, particularly in longer or syntactically complex utterances.

Secondly, controlled studies show that phrasing and intonation directly shape listeners' online comprehension. When segmental content is held constant and only terminal contours are varied (e.g., H* H-H% vs. L* L-L%), listeners are rapidly biased toward question versus statement interpretations [HBGT15], indicating a close mapping between boundary tones and utterance interpretation. Follow-up eye-tracking work on boundary tone processing shows that listeners exploit these terminal contours before utterance offset [BHA+17], such that small differences in the final tune are sufficient to shift real-time comprehension without any change to the words. A review of prosodic boundaries and syntactic ambiguity [SC10] similarly summarizes controlled studies in which only phrasing is manipulated over the same string, consistently affecting syntactic attachment decisions and reducing ambiguity. Together, these findings provide clear causal evidence that natural phrasing and terminal intonation patterns directly support accurate and efficient online comprehension.

Finally, there is converging evidence that prosodic phrasing and terminal intonation also improve recognition accuracy in ASR systems. In prosody-dependent recognizers on the BU Radio News corpus, adding ToBI-style intonational phrase boundary labels to otherwise fixed acoustic and language models yields up to about 11–13% relative reductions in word error rate (WER) compared to prosody-independent baselines [CHJC+06]. Experiments in [VS08, VS10b] show that adding suprasegmental features (F0 contours, break locations, and durations) at the preprocessing or rescoring stage leads to measurable WER reductions, especially on more complex, syntactically rich utterances. Related systems for Spanish and Mandarin [CYC+12, MR03] similarly find that modeling phrase boundaries and terminal contours jointly with words yields lower WER than word-only models.

In summary, controlled studies that manipulate phrasing or terminal contours provide clear causal evidence that natural prosodic structure reduces memory load, supports efficient online comprehension, and improves ASR recognition accuracy. These findings imply that explicit phrasing and intonation controls are not just a way to improve MOS, but a principled mechanism for enhancing how both humans and recognizers process the signal: when audio is synthesized directly from text—such as long-form content—using these controls to place phrase breaks and terminal contours in line with linguistic structure should help listeners remember more, understand more easily, and yield lower WER for downstream ASR.

3.2.5 Architecture Selection

Diffusion vs. Flow Matching

Diffusion-based acoustic decoders, such as Grad-TTS [PVG+21], generate speech by simulating a reverse stochastic differential equation (SDE) that gradually denoises a noise sample over many small steps, which yields high fidelity but incurs substantial sampling cost because the steps cannot be easily reduced. In contrast, flow matching decoders learn a deterministic ordinary differential equation that maps noise to data, so the trajectory can be integrated with a much smaller number of larger steps, achieving comparable or better quality at signif-

icantly lower inference cost. Recent TTS work, including MatchaTTS [MTB+24], VoiceFlow [GDM+24], and F5-TTS [CNM+25], has shown that conditional or rectified flow matching can produce high-fidelity speech that matches or surpasses diffusion-based baselines while being significantly more efficient in terms of inference speed.

In this chapter, ProsodyFM therefore adopts a flow matching decoder following MatchaTTS. MatchaTTS trains a TTS model using optimal-transport conditional flow matching and reports that, at equal ODE steps, flow matching configurations (MAT-4, MAT-10) achieve higher MOS, lower WER, and better inference speed than the diffusion-based Grad-TTS baselines (GRAD-4, GRAD-10) in their Table 1. Building on this architecture allows ProsodyFM to retain high naturalness and robustness while keeping sampling sufficiently fast for realistic computational budgets.

Autoregressive vs. Non-audoregressive Framework

Recent speech generation systems increasingly adopt autoregressive (AR) audio language models that operate over discrete speech tokens and unify text and audio generation in a single backbone. Representative examples include Moshi [DMO+24], Kimi-Audio [DJL+25], Qwen3-Omni [XGH+25], and Step-Audio 2 [WYH+25]. AR models naturally support streaming and long-context interaction, and they integrate smoothly with large language model (LLM) capabilities such as instruction following, prompt-based control, knowledge grounding, long-context reasoning, and tool use. However, their decoding cost grows with output length, and frame–text alignments are only implicitly modeled; prior work on non-autoregressive TTS has repeatedly pointed out that AR decoders tend to be slower and more prone to alignment instabilities (e.g., skipped or repeated words) in long-form speech synthesis [RRT+19, YLH+19, EWT+24, ZKG+25].

Non-autoregressive (NAR) TTS architectures generate entire utterances (or large chunks) in parallel, conditioned on explicit frame–text alignment information. This design yields inference times that are largely independent of sequence length and provides stable, controllable frame–text alignment. Recent work has also shown that NAR components can be adapted for

streaming—for example, CosyVoice 2 [DWC+24] employs a chunk-aware causal flow matching decoder to support both streaming and non-streaming synthesis within a single framework. However, NAR TTS models do not inherently expose the broad instruction-following and long-context reasoning capabilities of LLMs; when such capabilities are required, they are typically accessed through a cascaded pipeline in which an LLM first generates text and a separate TTS model then renders it as speech, increasing overall system complexity.

In this chapter, the goal is narrower and more focused: to build a prosody-aware TTS system with explicit phrasing and intonation controls, rather than a general-purpose conversational audio assistant. Under this setting, the advantages of a NAR backbone—parallel, length-insensitive inference, explicit frame–text alignment modeling, and stable long-form generation—are more important than the tight integration with LLM-style reasoning offered by AR audio language models. Given the emphasis on efficient inference under limited computational resources and on reliable control of phrase breaks and terminal intonation, ProsodyFM adopts a NAR architecture.

3.2.6 Evaluation for TTS

Evaluating TTS systems is inherently challenging, as it involves complex, multidimensional perceptual attributes that are difficult to quantify reliably. Evaluating TTS systems has traditionally relied on a combination of subjective and objective methods. Subjective listening tests such as MOS or CMOS remain the de facto gold standard, because they directly measure human perception of naturalness, intelligibility, and overall quality. However, these listening tests are expensive to run, time-consuming, and difficult to scale. In practice, it is difficult to recruit a large and diverse listener panel, which can introduce bias and reduce the reliability of detailed system comparisons. Objective metrics, in contrast, are automatically computable, inexpensive, and scale readily to large corpora and model suites. However, widely used measures—such as ASR-based WER/CER for intelligibility or $RMSE_{f0}$ and related prosodic distances—capture only narrow signal properties and typically show at best moderate correlation with human judgments of overall naturalness [UDL+25]. As a result, contemporary TTS studies routinely

report both subjective and objective results.

Some recent work has proposed objective metrics with higher correlation to human perception, including token-based metrics such as SpeechLMScore [MPSW23] and TTScore [UDL+25] for intelligibility and prosody. However, as shown in [UDL+25], intelligibility evaluation still yields only moderate utterance-level correlations between MOS and both traditional metrics (e.g., WER; 0.34 LCC) and TTScore-int (0.43 LCC). For prosody, the situation is even more challenging: utterance-level correlations between MOS and traditional metrics (e.g., log F0 RMSE; −0.03 LCC) and TTScore-pro (0.06 LCC) remain at a poor degree. These results indicate that current objective metrics are still far from providing a reliable substitute for human listening tests, particularly for prosodic evaluation.

Within objective evaluation, an important distinction is between reference-dependent and reference-free metrics, particularly for prosody. Prosodic naturalness is inherently variable, as speakers and listeners with different linguistic, dialectal, and cultural backgrounds may prefer different yet acceptable realizations of the same text. Reference-dependent metrics, which compare synthesized speech to a specific ground-truth speech recording, mitigate this variability by anchoring evaluation to a concrete target. In contrast, reference-free metrics [MPSW23, UDL+25] assess synthesized speech only with respect to the input text and learned priors over plausible realizations, without requiring a parallel speech recording, and are therefore essential when no paired ground-truth speech is available, such as in cross-lingual speech translation, style transfer, or evaluation on open-domain text.

In this chapter, we adopt a hybrid evaluation strategy combining customized subjective listening tests with reference-dependent objective metrics. On the subjective side, we design MOS protocols that separately assess overall quality (MOS), intelligibility ($MOS_{\text{intelligibility}}$), intonation similarity ($MOS_{\text{intonation}}$), and phrasing similarity (MOS_{break}) under both parallel and non-parallel conditions: 15 utterances with different lengths are rated on a 5-point scale with 95% confidence intervals by 21 listeners, following detailed written instructions Figure 3.8 and examples for each dimension. On the objective side, we use WER as a proxy for intelligibility, $RMSE_{f0}$ for intonation, and an F1 score for phrasing, all computed against

ground-truth speech recordings; given that ground-truth speech is available in our setting, this choice of reference-dependent metrics enables fair and controlled comparison across systems. Future work can focus on developing objective metrics that exhibit stronger correlation with human perceptual judgments, particularly for prosodic aspects, and on designing reference-free formulations that remain applicable in the absence of parallel speech.

3.3 Related Works

In this section, we review prior papers in terms of the phrasing and intonation aspects of prosody in TTS and highlight the existing research gaps.

3.3.1 The Break Labeling Issue

Breaks in speech can be roughly divided into punctuation-based and respiratory breaks [HLL23]. Unlike punctuation-based breaks which are marked by punctuations, respiratory breaks have no explicit label on the text side. Most of the current TTS systems [MTB+24, LHR+24] have only considered punctuation-based breaks, resulting in many non-final phrase breaks being overlooked or misplaced [Tay09].

Some TTS systems explicitly model phrase breaks, with some employing fully supervised approaches to integrate phrase breaks into TTS. These models [ZLY+21, SAMH22, Slo23] typically rely on prosody labels derived according to the ToBI annotation system within their training datasets. However, as discussed in section 2.2.4, due to the complexity of the ToBI system, accurate annotation can only be provided by trained linguists [BDKG12], resulting in high costs for data collection and limiting the availability of large-scale annotated datasets. This limited availability and dataset size restrict the performance of supervised methods in TTS systems, particularly in this deep-learning era. Moreover, due to the misalignment between human perception and the categorical distinctions defined by ToBI, there is low label agreement among annotators, which further undermines the performance of fully supervised prosody-aware TTS

models.

In response, other models have adopted unsupervised approaches to incorporate phrase breaks into TTS. These models [HLL23, AMM+22, YKS+23] use manually designed thresholds combined with the Montreal Forced Aligner (MFA) [MSM+17] to obtain break labels in an unsupervised manner. More specifically, the MFA automatically aligns phonetic transcriptions with the corresponding speech signal, mapping each phoneme to its precise time interval. Based on this alignment, the presence and duration of phrase breaks can be derived. Subsequently, these models employ manually designed duration thresholds to categorize breaks into different classes according to their derived duration. However, previous research indicates that the frequency and duration of phrase breaks are influenced by both linguistic phrase structure and the speaker's individual speaking style [HLL23]. Given the variability in break durations and their dependence on speaker-specific factors, these handcraft threshold-based methods struggle to capture the nuances of speaker-specific variations in break durations adequately.

ProsodyFM tackles this issue by designing a fusion encoder to integrate the obtained initial break cues with relevant speaker information, and then adjusting the duration with a duration predictor, enabling flexible modeling of phrase breaks.

3.3.2 The Intonation Modeling Issue

Similar to phrase breaks, fully supervised intonation-aware TTS systems mostly rely on ToBI for labeling. These systems face the same challenges [LK19], including high annotation costs, limited data availability, low inter-annotator agreement, and labels that do not align well with human perception. Almost all the existing unsupervised intonation-aware TTS systems [RHT+21, MLYH21, HRL+22, LHR+24] directly model the absolute pitch values obtained from some pitch tracking methods. However, the performance of these systems is constrained by the following two key drawbacks.

On the one hand, pitch tracking is inherently challenging, and existing pitch tracking methods frequently yield errors like pitch doubling/halving and incorrect unvoiced/voiced flags [HdL21],

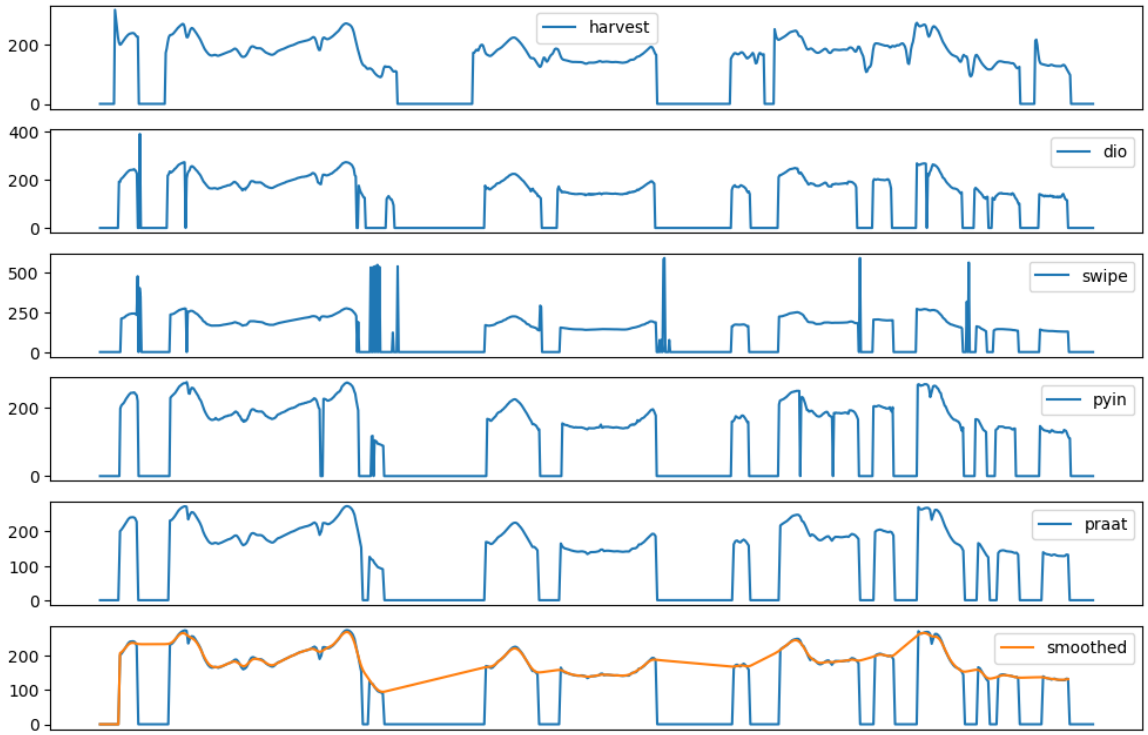


Figure 3.4: Pitch contours extracted from 5 pitch tracking methods (blue) and our pitch smoothing method (orange).

leading to unreliable results. Figure 3.4 illustrates the pitch tracking results across 5 different methods. From top to bottom are Harvest [Mor17], DIO [MKK09], SWIPE [CH08], pYIN [MD14], and Praat [Boe01]. We can clearly observe frequent prediction errors in pitch values and unvoiced/voiced flags, as well as inconsistencies across these five methods. When these pitch tracking results are used as labels in the TTS systems, they also introduce frequent errors, leading to unnatural intonation and negatively impacting the overall performance of the models.

On the other hand, absolute pitch values cannot reflect the non-linear characteristics of the perception of pitch differences, which are critical for capturing prosodic intonation change. The perception of pitch differences is inherently non-linear [HdL21]; for example, a pitch rise or fall of the same magnitude is perceived as smaller for a higher-pitched voice (such as that of a young child or woman) compared to a lower-pitched voice (such as that of a man). Similarly, when male and female speakers produce utterances that sound intonationally identical to the listener—i. e., conveying the same linguistic or paralinguistic functions—their absolute pitch range changes differ. Specifically, the pitch movements in a female voice tend to be larger, as

her pitch range is, on average, higher and wider than that of a male voice [HVG91, Gra18].

Some recent findings from human perceptual studies offer a potential basis for this issue; the authors in [CC19, CST22] have shown that **compared to the detailed pitch values, the shape of the pitch contour is more important for human perception of intonation change**.

ProsodyFM introduces the intonation shape tokens to model the pitch shape of the last word of an intonational phrase instead of directly modeling the absolute pitch values. We also employ a novel pitch processor to interpolate, smooth, and perturb the raw pitch contour to highlight its shape. The orange line in Figure 3.4 shows an example after our smoothing process. Our method alleviates pitch tracking errors, enables more robust modeling of pitch shapes, and aligns more closely with human perception of intonation change.

3.4 Methodology

ProsodyFM is designed to extract phrasing and terminal intonation patterns from reference speech and then adjust these patterns to align appropriately with the target text. ProsodyFM predicts Mel-spectrograms from raw text. The predicted Mel-spectrograms are converted to waveforms by the HifiGAN vocoder [KKB20] during the inference process.

3.4.1 Training Objective

The training objective of ProsodyFM is to find the network parameters θ and the alignment A that maximize the log-likelihood of the data samples x_1, as in Equation 3.16. Due to the computational intractability to find the global solution, we apply the iterative approach introduced in Section 3.2.3, namely, each iterative of optimization consists of two steps: (i) searching for the most probable monotonic alignment $\hat{A}$ given fixed model parameters θ, and (ii) updating the model parameters θ to maximize the log-likelihood. For step (i), following Glow-TTS

[KKKY20] (as described in Section 3.2.3), we implement the MAS algorithm to find the current optimal alignment $\hat{A}$ between text and speech, as in Equation 3.17.

$$\max_{\theta,A} L(\theta, A) = \max_{\theta,A} \log p_1(x_1|c; \theta, A) \tag{3.16}$$

$$\hat{A} = \arg\max_{A} \sum_{j=1}^{T_{\text{mel}}} \log \mathcal{N}(z_j; \mu_{A(j)}, \sigma_{A(j)}) \tag{3.17}$$

For step (ii), following our backbone MatchaTTS [MTB+24], we train the ProsodyFM with the OT-CFM procedure. So maximizing the log-likelihood of Equation 3.16 can be transformed into regressing the ODE vector field that defines a mapping from a random sample to a real mel-spectrogram, which is denoted as $\mathcal{L}_{OT-CFM}$ (Equation 3.19). We introduce a prior loss $\mathcal{L}_{prior}$ as an explicit maximum log-likelihood objective for the conditional encoders. The conditional encoders, denoted as f_{c_enc}, include the Text, Phrase Break, Terminal Intonation, Speaker, and Fusion encoders. Intuitively, providing a condition that is closely related to the generation target as input can reduce the training burden on the vector field estimator. The prior loss $\mathcal{L}_{prior}$ is given by:

$$\mathcal{L}_{prior} = -\sum_{j=1}^{T_{mel}} \log \mathcal{N}(z_j; \mu_{\hat{A}(j)}, I) = -\sum_{j=1}^{T_{mel}} (\frac{n}{2} \log(2\pi) + \frac{1}{2} \|z_j - \mu_{\hat{A}(j)}\|^2), z = f_{c\_enc}(w, x_1) \tag{3.18}$$

where z_j is the j-th frame of the f_{c_enc} output z and n represents the dimensionality of z_j. The difference between Equation 3.17 and Equation 3.18 is that Equation 3.17 aims to find the optimal alignment $\hat{A}$ given the current f_{c_enc} parameters, while Equation 3.18 seeks to optimize the f_{c_enc} parameters with $\hat{A}$ fixed.

Additionally, two auxiliary losses are introduced that do not interfere with the primary objective of maximizing log-likelihood: the duration loss $\mathcal{L}_{dur}$ (Equation 3.20) and the text-pitch alignment loss $\mathcal{L}_{tp\_align}$ (Equation 3.21). These auxiliary losses are applied exclusively to train

the Duration Predictor and BERT. These two modules are incorporated to address potential mismatches between the reference speech and the target text during inference. A stop-gradient operation is used to cut off the gradient propagation to other parts of the model.

The four losses are marked with double lines in Figure 3.5 and Figure 3.6. The total loss for training ProsodyFM is as follows.

$$\mathcal{L}_{ProsodyFM} = \mathcal{L}_{OT-CFM} + \mathcal{L}_{prior} + \mathcal{L}_{dur} + \mathcal{L}_{tp\_align}$$

We present the training algorithm of ProsodyFM in Algorithm 1.

OT-CFM Loss

The OT-CFM model introduced in Section 3.2.1 operates as an unconditional generative model. However, to meet the requirements of ProsodyFM, it needs to be extended to a conditional generative model. ProsodyFM leverages target text alongside the speaker and prosody information—derived from a reference speech—as condition c to guide the speech synthesis process. The objective of our conditional generative modeling is to produce new samples x_1 that are approximately distributed according to $p_1(x_1|c)$, where both x_1 and c are random variables. This task can be reformulated as training a neural network to model $p_1(x_1|c_1)$, where $c_1 \sim p(c)$. By providing the neural network with sufficient samples c_1 from the dataset, and using parameter sharing across conditions, the network effectively approximates $p_1(x_1|c)$.

Given a particular mel-spectrogram sample $x_1 \sim p_1$, it is associated with a specific condition $c_1 \sim p(c)$. We can replace the random variable x_0, x_t, and x_1 in the unconditional OT-CFM to $x_0|c_1$, $x_t|c_1$, and $x_1|c_1$, where x_0, x_t, x_1 are random variables and c_1 is the specific condition sample. Then we can construct the Gaussian conditional probability path as

$$p_{t|1}(x_t \mid x_0, x_1, c_1) = \mathcal{N}(x_t; (1 - (1 - \sigma_{\min})t)\, x_0 + t x_1, \sigma_{min}^2 I)$$

The objective of the conditional OT-CFM for ProsodyFM is

Algorithm 1 Training Algorithm of ProsodyFM

Input: the target text w, the reference mel-spectrogram x_1, the mel-spectrogram length T_{mel}, the text length T_{text}

Output: optimal vector field predictor u_{θ^*}

Function: TrainStep(w, x_1, x_0)
    Input the target text w and the reference mel-spectrogram x_1 into the conditional encoder f_{c_enc} and obtain the output prior statistics μ_i;
    Search for the optimal monotonic alignment $\hat{A}$ using MAS (Equation 3.17);
    Align the prior statistics μ_i with the reference mel-spectrogram x_1 based on $\hat{A}$ to get the condition c_1;
    Compute $\mathcal{L}_{prior}$, $\mathcal{L}_{tp\_align}$ and $\mathcal{L}_{dur}$ according to Equation 3.18, Equation 3.21 and Equation 3.20;
    Sample time $t \sim \text{Uniform}[0, 1]$;
    Sample $x_t \sim \mathcal{N}(x_t; (1 - (1 - \sigma_{\min})t)\, x_0 + t x_1, \sigma^2_{min} I)$;
    Compute $\mathcal{L}_{OT-CFM}$ according to Equation 3.19;
    Gradient descent on loss $\mathcal{L}_{OT-CFM} + \mathcal{L}_{prior} + \mathcal{L}_{dur} + \mathcal{L}_{tp\_align}$ and obtain new model parameters $\hat{\theta}$
end Function:
while Perform Training **do**
    Take a batch of target text w and the reference mel-spectrogram x_1;
    Sample $x_0 \sim \mathcal{N}(0, I)$;
    Call TrainStep(w, x_1, x_0);
end while

$$\mathcal{L}_{\text{OT-CFM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1], x_1\sim p_1, x_0\sim p_0, x_t\sim p_{t|1}(x_t|x_0,x_1,c_1)} \left[\| u_\theta(x_t, c_1, t) - [x_1 - (1 - \sigma_{min})x_0] \|^2 \right] \quad (3.19)$$

In this particular sample (x_1, c_1), the vector field estimator u_θ is trained to estimate the conditional vector field $u_t(x_t|c_1)$. By providing it with sufficiently diverse data samples x_1 with c_1 from the dataset and parameter sharing, the u_θ approximates $u_t(x_1|c)$ which is the real vector field that generates the target probability path $p_t(x_t|c)$. By solving the joint ODE of Equation

3.6, we can obtain the mel-spectrogram x_1 given any c as a condition.

Duration Loss

During training, the input reference speech is matched with the target text. So we can directly perform the MAS algorithm introduced in Section 3.2.3 to search for the most probable alignment A^* between phone-level (also with breaks in ProsodyFM) prior statistics and the frame-level speech representation. However, during inference, the input reference speech may not match the target text. To estimate the best monotonic alignment A^* at inference, we need to design a Duration Predictor.

As shown in Figure 3.5, we incorporate the Duration Predictor f_{dur} on top of the Fusion Encoder f_{fus}. It follows the architecture of the Duration Predictor in MatchaTTS. It is trained with the mean squared error loss (MSE) in the logarithmic domain, as described in Equation 3.20. To avoid the gradient affecting the maximum log-likelihood objective (Equation 3.16), we stop the gradient propagation between f_{fus} and f_{dur}.

$$\begin{aligned} \mathcal{L}_{dur} &= MSE(f_{dur}, [d_1, ..., d_i, ..., d_{T_{text}}]), \\ d_i &= \sum_{j=1}^{T_{\text{mel}}} 1_{A^*(j)=i}, \quad i = 1, \ldots, T_{\text{text}} \end{aligned} \tag{3.20}$$

3.4.2 Proposed Model Components

The overall structure of ProsodyFM is shown in Figure 3.5, where the parts outlined in the yellow shaded area are the components we propose for ProsodyFM to improve the MatchaTTS model. More details of four key components are shown in Figure 3.6. The Pitch Processor extracts robust pitch shape segments; the Phrase Break Encoder predicts initial phrase break cues, which are fused with speaker information and refined by the duration predictor; the Text-Pitch Aligner estimates intonation patterns from the target text, guiding the selection

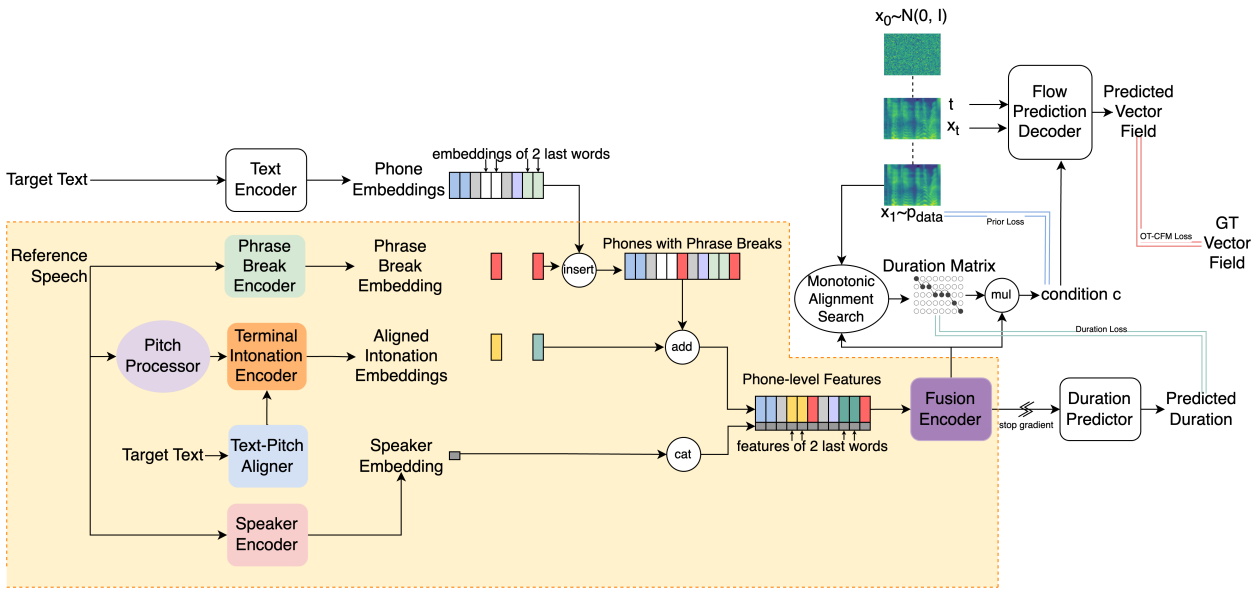


Figure 3.5: The model architecture of the proposed ProsodyFM. The components outlined in the yellow shaded area are unique to ProsodyFM and differ from those in MatchaTTS. During inference, the duration matrix is replaced by the predicted duration from the duration predictor; the predicted vector field is passed to an ODE solver to obtain the predicted mel-spectrogram. The three components in ellipses represent that these components have no training parameters. The operations "insert", "add", "cat", and "mul" in the four circles in the figure represent "insert", "add", "concatenate", and "matrix multiply" respectively.

of appropriate reference intonation patterns; and the Terminal Intonation Encoder models terminal intonation patterns that are properly aligned with the target text.

Pitch Processor

The Pitch Processor (the pink box in Figure 3.6) aims to obtain robust pitch shape segments of the last words. The raw pitch contours extracted by the pitch tracking method are discrete and unreliable. To reduce prediction errors in pitch tracking, we interpolate and smooth the discrete pitch values into a continuous pitch contour. Both interpolation and smoothing are performed using functions within Praat. Specifically, discrete pitch values are firstly smoothed to reduce prominent tracking errors (such as pitch doubling/halving), then interpolated for a continuous contour, followed by a second smoothing to accentuate pitch movement patterns.

To better highlight the pitch shape, we subtract a random value (uniformly distributed within the range $[f_{min}, f_{max}]$) from each point in the pitch contour to perturb its specific pitch value

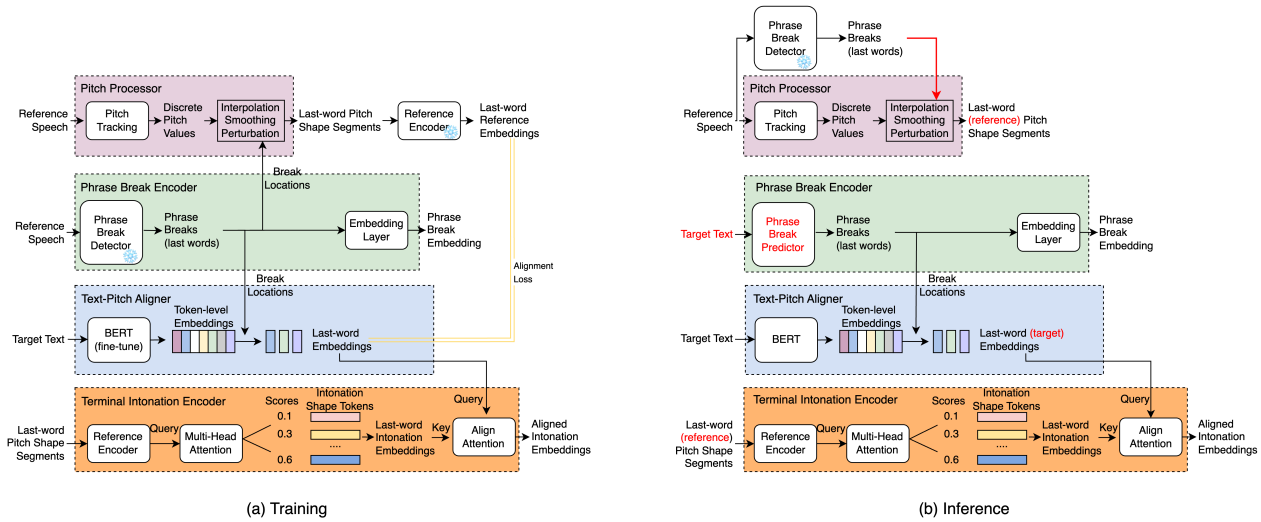


Figure 3.6: The key components of the proposed ProsodyFM in the training (a) and inference (b) phrases. The red markings highlight the differences. The snowflake mark means the module is frozen during training.

information while preserving its shape patterns.

Phrase Break Encoder

The Phrase Break Encoder (the green box in Figure 3.6) is designed to predict phrase break cues, locate the last word of each intonational phrase, and generate the phrase break embedding. The predicted locations of the last words help the Pitch Processor and the Text-Pitch Aligner select the pitch shape segment and word embedding corresponding to each last word.

During training, the phrase break encoder detects phrase breaks from the reference speech through an independent phrase break detector, which remains frozen during training. During inference, since we do not have speech aligned with the target text, the phrase break encoder predicts phrase breaks from the plain target text through our designed phrase break predictor. This phrase break predictor is fine-tuned from T5 [NHAC+22]. We report the performance of the phrase break predictor in Section 3.6.5.

Text-Pitch Aligner

The training objective of the Text-Pitch Aligner (the blue box in Figure 3.6) is to align the BERT-derived word embedding of the last word e_k with the corresponding reference intonation feature r_k in the same feature space. The reference intonation feature is extracted by the reference encoder, which is the same reference encoder as the one in the terminal intonation encoder but detached from the entire computation graph to stop gradient propagation. This alignment allows BERT, during inference, to create word embedding that can estimate the intonation patterns of the target text, even without matching speech with the target text. This estimated embedding will further guide the selection of appropriate reference intonation patterns. We use the L2 loss as the alignment loss $\mathcal{L}_{tp\_align}$

$$\mathcal{L}_{tp\_align} = -\sum_{k=1}^{T_{lword}} \|\mathbf{e}_k - \mathbf{r}_k\|_2^2, \tag{3.21}$$

where T_{lword} is the number of last words.

Terminal Intonation Encoder

The Terminal Intonation Encoder (the orange box in Figure 3.6) aims to obtain the terminal intonation patterns that are proper to the target text. The reference encoder first compresses the input pitch shape segment of each last word of the reference speech into a fixed-length reference intonation feature. Then this feature, as a query vector, is passed to the multi-head attention module. The multi-head attention module is trained to learn a similarity measure between the reference intonation feature and each token in a bank of randomly initialized embeddings. These embeddings, which we designed to describe different pitch shapes, are called **intonation shape tokens**. These tokens are jointly trained with the rest of the model with only OT-CFM loss and thus require no extra intonation labels. The multi-head attention module generates scores (weights) for these tokens, and the weighted sum of these tokens forms the last-word intonation embeddings of the reference speech.

However, during inference, the reference speech may not be aligned with the target text, re-

sulting in a different number of last words in the reference speech compared to the target text. We use scaled dot-product attention (the align attention) to align them, treating the last-word intonation embeddings (of the reference speech) as the key and the last-word embeddings (of the target text) as the query. After adjustment by the Text-Pitch Aligner, the last-word embeddings (of the target text) and the last-word reference embeddings representing pitch shape characteristics are in the same space. By using the last-word embeddings as queries, the Align Attention module can select the most suitable intonation embeddings for the last word of the target text. This enables our ProsodyFM to autonomously choose the terminal intonation pattern based on both the reference speech and the target text during the inference phase.

3.4.3 Information Flow During Training and Inference

Information Flow During Training

During training, ProsodyFM follows the standard teacher-forcing training mechanism. The reference speech is the ground truth speech and the target text is consistent with the transcript of the reference speech. The information flows through the model as follows:

1) The Text Encoder (directly inherited from MatchaTTS) analyzes the target text to extract phone embeddings;

2) The Phrase Break Detector (developed from an ASR model) detects phrase breaks from the reference speech and provides the location of each last word (just before each phrase break) to the pitch Processor and the Text-Pitch Aligner;

3) These last words (from ASR) are then aligned with the target text (real transcript) to locate their positions within the phone embeddings, where the phrase break embeddings are subsequently inserted;

4) The Pitch Processor processes the reference speech and clips the pitch shape segments of the last words based on their locations;

5) The Terminal Intonation Encoder takes these pitch shape segments as input and generates the aligned intonation embeddings. These embeddings are added to the phone with break embeddings at the position of the last word, while other positions remain unchanged;

6) The Speaker Encoder extracts the speaker embedding from the reference speech, which is then concatenated with the other embeddings;

7) The Fusion Encoder integrates the phrase break and aligned intonation embeddings with speaker and phone embeddings. It outputs the phone-level prior statistics;

8) The MAS searches for the optimal alignment to get the frame-level condition c;

9) The Flow Prediction Decoder, also inherited from MatchaTTS and illustrated in Figure 3.7, takes the condition c, a uniformly sampled time t, and x_t as inputs, producing the predicted vector field.

Information Flow During Inference

During inference, the target text does not need to be consistent with the transcript of the reference speech. Most of the information flows through the model are the same as the training process:

1) The Text Encoder analyzes the target text to extract phone embeddings;

2) The Phrase Break Predictor (instead of the Phrase Break Detector during training) predicts the phrase breaks from the target text and provides the location of each last word (just before each phrase break) to only the Text-Pitch Aligner;

3) The Phrase Break Detector detects phrase breaks from the reference speech and provides the location of each last word to the Pitch Processor;

4) The Pitch Processor processes the reference speech and clips the pitch shape segments of the last words based on their locations;

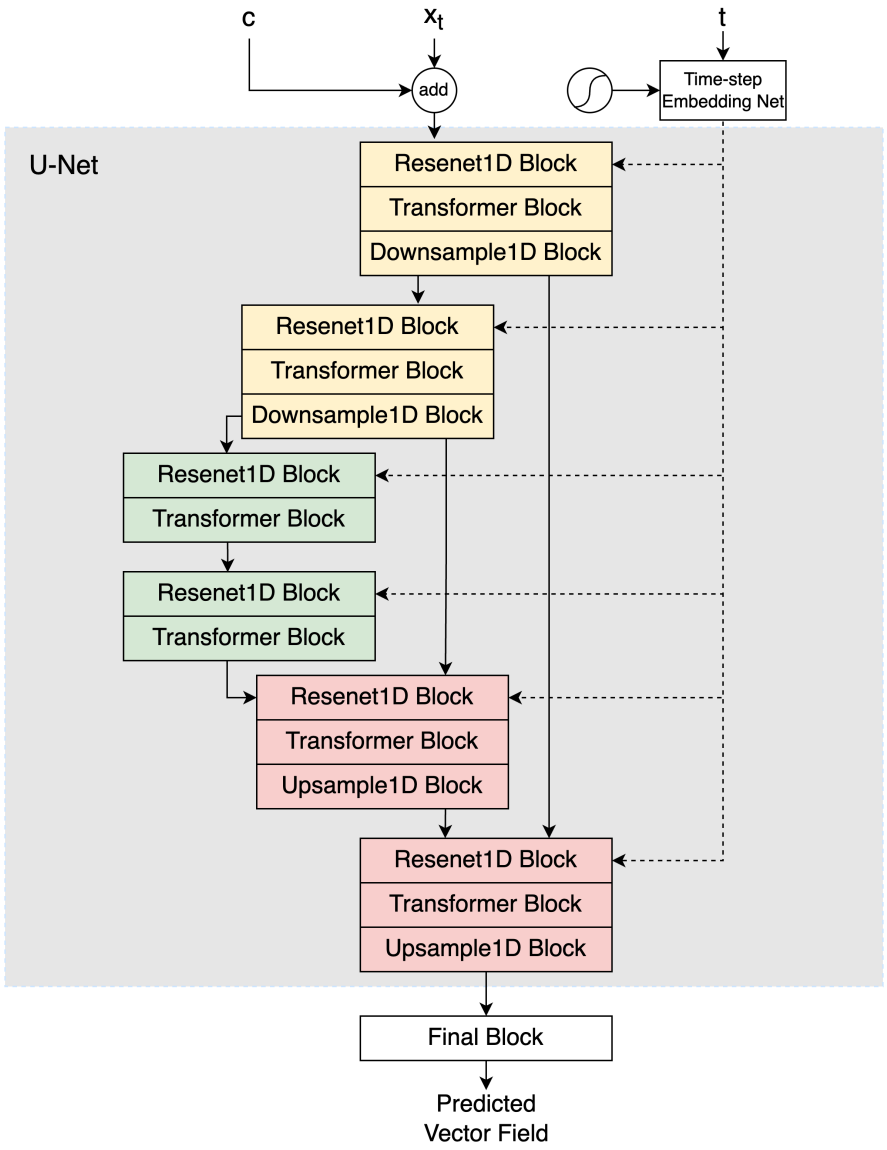


Figure 3.7: The architecture of the Flow Prediction Decoder.

5) The Terminal Intonation Encoder takes these pitch shape segments as input and generates the aligned intonation embeddings. These embeddings are added to the phone with break embeddings at the position of the last word, while other positions remain unchanged;

6) The Speaker Encoder extracts the speaker embedding from the reference speech, which is then concatenated with the other embeddings;

7) The Fusion Encoder integrates the phrase break and aligned intonation embeddings with speaker and phone embeddings. It outputs the phone-level prior statistics;

8) The Duration Predictor (rather than the MAS during training) predicts the optimal

alignment to get the frame-level condition c;

9) The Flow Prediction Decoder takes the condition c, a uniformly sampled time t, and x_t as inputs, predicting the target vector field;

10) The ODE solver produces the Mel-spectrogram based on the predicted vector field. The Mel-spectrogram is finally transformed into waveforms by the HifiGAN vocoder.

3.5 Experimental Details

3.5.1 Model Configurations

For a fair comparison, we utilize the same model architecture and hyperparameters as MatchaTTS [MTB+24] except for the following added modules. For our Terminal Intonation Encoder, we employ the attention module in [WSZ+18] with 4 attention heads and 6 64-D tokens. We replace the complex reference encoder in [WSZ+18] with a single-layer LSTM with 128-D hidden size to speed up training. For the phrase break detector, we use the released checkpoint of PSST [RGT23] without fine-tuning. The phrase break detector remains frozen during training. For the phrase break predictor, we fine-tune T5 [NHAC+22] independent from ProsodyFM using LoRA [HSW+22] with 16 ranks and consider the phrase breaks obtained from the PSST as the ground truth labels when fine-tuning. For the text-pitch aligner, we initialize BERT[2] with pre-trained weights, using its original tokenizer to process input text and obtain 768-D token-level embeddings. We then select the tokens corresponding to each last word, average their embeddings, and pass it through a fully connected layer to produce a final 192-D embedding for each last word. For the pitch processor, we use Praat [Boe01] to extract discrete pitch values and use a customized Praat script modified from [Can15] to interpolate and smooth them into a continuous pitch contour. For the speaker encoder, we extract the same external speaker embedding as in [CWS+22] for each speech sample and add two fully connected layers to transform the 512-D d-vector to the final 64-D speaker embeddings.

[2]https://huggingface.co/google-bert/bert-base-uncased

ProsodyFM and its ablated variants in Table 3.5 are trained for 350 epochs on an NVIDIA A100 GPU with 80GB VRAM with batch size 64 and learning rate 1e-4.

3.5.2 Datasets and Preprocessing

We perform the experiments in Table 3.1, Table 3.2, Table 3.3, Table 3.5, and Figure 3.10 on the LibriTTS corpus [ZDC+19]. We randomly split (speakers-independent) the audio samples in the train-clean-100, dev-clean, and test-clean sections of LibriTTS into 40421, 839, and 839 samples for our training, validation, and testing sets, respectively. The LibriTTS dataset has, in total, 71 hours of audio signals and 326 speakers. The audio format is 16-bit PCM with a sample rate of 24000 Hz.

For the experiments in Table 3.4, we train the models on the VCTK corpus [YVM19] with the same training set (43470 samples) as in [KKS21] and test the models on our LibriTTS testing set. The VCTK dataset has, in total, 44 hours of audio signals and 110 speakers. The audio format is 16-bit PCM with a sample rate of 48000 Hz.

We resample all audio to 22050 Hz and extract Mel-spectrograms with a 1024 FFT size, 256 hop size, 1024 window length, and 80 frequency bins.

3.5.3 Objective Evaluation Metrics

We conduct objective evaluations with the log-scale F0 Root Mean Squared Error ($RMSE_{f0}$), the F1 score of the break classification ($F1_{break}$), and the Word Error Rate (WER).

1) $RMSE_{f0}$ measures the pitch error. Following [BZ20], we use it to evaluate the **intonation** aspect of prosody. We leverage dynamic time warping (DTW) to extract the pitch values and measure the portion where both the ground truth speech y_i and synthesized speech y'_i are voiced.

$$\text{RMSE}_{f_0} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}\left(\log\left(\frac{y_i}{y'_i}\right)\right)^2} \tag{3.22}$$

Here, T denotes the number of voiced frames.

2) $F1_{break}$ evaluates the **phrasing** aspect of prosody. We use PSST [RGT23] to obtain phrase breaks from the ground truth speech and collect the corresponding last words as the list l, which serves as our ground truth labels. We then apply PSST to the synthesized speech to detect phrase breaks, generating a list l' of the corresponding last words for comparison. The true positives tp are the count of last words that match exactly in both l and l'. The false positives fp are the number of last words present in l' but absent from l, while the false negatives fn are the number of last words present in l but missing from l'. The $F1_{break}$ is computed as follows:

$$F1_{break} = \frac{2 \cdot tp}{2 \cdot tp + fp + fn} \tag{3.23}$$

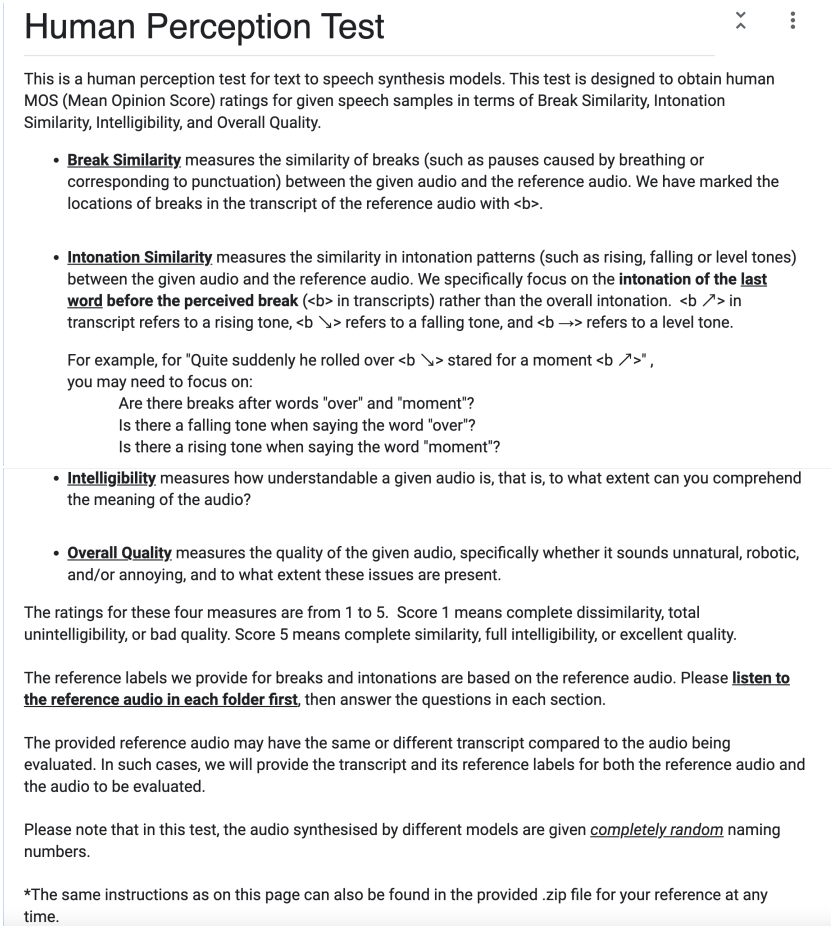
3) WER correlates well with the **intelligibility** of synthesized speech [TR21, MTB+24]. We use Whisper-small [3] [RKX+23] to obtain transcripts of both the ground truth and synthesized speech and compute the WER.

3.5.4 Subjective Evaluation Metrics

We conduct a crowd-sourced Mean Opinion Score (MOS) human listening test to assess four aspects of synthesized speech, including the phrase break similarity (MOS_{break}), the terminal intonation similarity ($MOS_{intonation}$) between the synthesized and a reference speech, the intelligibility ($MOS_{intelligibility}$) and the quality (MOS) of the synthesized speech. Each MOS is assessed using a 5-point scale with 95% confidence intervals, where score 1 indicates dissimilarity, unintelligibility, or poor quality, whereas score 5 signifies full similarity, intelligibility, or excellent quality. We randomly select 15 utterances (3 groups of 5) with different lengths in the testing set, and each sample is rated by 21 testers. Our testers are PhD students from four universities specializing in Computer Audition, with native languages including English, German, Chinese, and Turkish. They are all fluent in English.

To account for the possibility of non-expert testers, we provided detailed explanations of the

[3] https://huggingface.co/openai/whisper-small



Human Perception Test

This is a human perception test for text to speech synthesis models. This test is designed to obtain human MOS (Mean Opinion Score) ratings for given speech samples in terms of Break Similarity, Intonation Similarity, Intelligibility, and Overall Quality.

- **Break Similarity** measures the similarity of breaks (such as pauses caused by breathing or corresponding to punctuation) between the given audio and the reference audio. We have marked the locations of breaks in the transcript of the reference audio with <b>.

- **Intonation Similarity** measures the similarity in intonation patterns (such as rising, falling or level tones) between the given audio and the reference audio. We specifically focus on the **intonation of the last word before the perceived break** (<b> in transcripts) rather than the overall intonation. <b ↗> in transcript refers to a rising tone, <b ↘> refers to a falling tone, and <b →> refers to a level tone.

  For example, for "Quite suddenly he rolled over <b ↘> stared for a moment <b ↗>" ,
  you may need to focus on:
  Are there breaks after words "over" and "moment"?
  Is there a falling tone when saying the word "over"?
  Is there a rising tone when saying the word "moment"?

- **Intelligibility** measures how understandable a given audio is, that is, to what extent can you comprehend the meaning of the audio?

- **Overall Quality** measures the quality of the given audio, specifically whether it sounds unnatural, robotic, and/or annoying, and to what extent these issues are present.

The ratings for these four measures are from 1 to 5. Score 1 means complete dissimilarity, total unintelligibility, or bad quality. Score 5 means complete similarity, full intelligibility, or excellent quality.

The reference labels we provide for breaks and intonations are based on the reference audio. Please **listen to the reference audio in each folder first**, then answer the questions in each section.

The provided reference audio may have the same or different transcript compared to the audio being evaluated. In such cases, we will provide the transcript and its reference labels for both the reference audio and the audio to be evaluated.

Please note that in this test, the audio synthesised by different models are given *completely random* naming numbers.

*The same instructions as on this page can also be found in the provided .zip file for your reference at any time.

Figure 3.8: The instruction page of our Mean Opinion Score human listening test.

four MOS metrics with clear definitions and examples. Figure3.8 shows the instruction page of our crowd-sourced Mean Opinion Score human listening test. These instructions are given to the testers prior to the commencement of their rating task.

Considering text consistency between the reference audio and the target text to be synthesized, we perform both parallel and non-parallel MOS tests. Given the subjective nature of prosody evaluation, where individuals from different linguistic backgrounds can have varying perceptions of what constitutes 'appropriate' phrasing and intonation for a target text [GJD87], we adopt specific assumptions for our parallel and non-parallel subjective evaluations:

1) For the parallel subjective evaluation, we assume that the phrasing and intonation derived from the reference speech represent the 'appropriate' prosody. We provide labels for breaks and intonation based on the reference speech and ask testers to assess the **similarity** of the phrasing and intonation to these labels, rather than their appropriateness.

2) For the non-parallel subjective evaluation, a reference speech is still needed for similarity assessment. To maintain the relevance of prosody while accommodating different text content, we modify the words in the reference speech (the same 15 samples) transcript, ensuring that the sentence semantics and structure remain as close as possible to the original [4]. We then transfer the break and intonation labels from the reference speech to the new target text (provided in Figure 3.9). Testers are again asked to assess the **similarity** of the phrasing and intonation to these labels, rather than their appropriateness. This evaluation rests on the assumption that two sentences with similar semantics and structure should share the same phrasing and intonation labels.

Figure 3.9 presents the labeled transcripts provided in the human listening test for 15 random samples under parallel and non-parallel settings. The $<$ b $\nearrow$ $>$, $<$ b $\searrow$ $>$, and $<$ b $\rightarrow$ $>$ describe a phrase break with rising, falling, and level intonation on the last word, respectively. The terminal intonation and phrase break labels we provided are based on the actual pitch contour of the speech signal, supplemented by reference to the perceptual judgments of two human annotators.

3.5.5 Comparative Models

To evaluate the performance of our model, we compare ProsodyFM with four SOTA models. These four SOTA models we use are all from the official implementation of corresponding papers, along with the officially provided checkpoints.

1) StyleSpeech [MLYH21] [5]: the expressive multi-speaker TTS model built on FastSpeech2 [RHT+21];

2) GenerSpeech [HRL+22] [6]: the TTS model towards high-fidelity style transfer, also extended from FastSpeech2 [RHT+21];

[4] In practical usage, word-level intonation transfer should be applied only when appropriate; it need not be used when the target text's context or pragmatic intent differs substantially from the reference speech.

[5] https://github.com/KevinMIN95/StyleSpeech

[6] https://github.com/Rongjiehuang/GenerSpeech

File Name in LibriTTS	Labeled Transcripts in the Parallel Setting	Labeled Transcripts in the Non-parallel Setting
1363_139304_000009_000005	Quite suddenly he rolled over <b> > stared for a moment <b> > and struggled into a sitting position <b> >	All of a sudden he turned over <b> > looked around briefly <b> > and forced himself into a seated posture <b> >
6529_62554_000032_000000	Captain Harding <b> > replied Ayrton <b> > I can give you no better advice in this matter <b> >	Captain Harding <b> > responded by Ayrton <b> > I cannot offer you any better suggestions on this issue <b> >
2002_139469_000016_000005	As yet <b> > western europe was uninfected <b> > would it always be so <b> >	So far <b> > western europe remained untouched <b> > would it stay that way <b> >
8468_294887_000023_000005	I stated my errand to him <b> > with a description of the young man <b> >	I explained my intention to him <b> > with a portrayal of the youth <b> >
7302_86815_000052_000000	Well madam <b> > it will be a laudable action on your part <b> > and I will thank you for it <b> >	Well madam <b> > it would be commendable on your side <b> > and I would be grateful to you for it <b> >
4680_16041_000028_000000	Moreover <b> > on both sides <b> > the fury <b> > the rage <b> > and the determination were equal <b> >	Furthermore <b> > on both parties <b> > the anger <b> > the wrath <b> > and the resolve were the same <b> >
5163_39921_000035_000005	You need some refreshments after your long walk <b> > Missis Cropper will bring you in something <b> >	You require some snacks when finish your lengthy stroll <b> > Mrs. Cropper will help you prepare something <b> >
4970_29095_000005_000000	Well <b> > mother <b> > said the young student <b> > looking up <b> > with a shade of impatience <b> >	Well <b> > mom <b> > remarked the young pupil <b> > glancing up <b> > with a hint of impatience <b> >
3235_28433_000003_000001	An English family consisting of the mother <b> > one son <b> > and a daughter <b> > were to accompany me <b> > and we had spent weeks in making our preparations <b> >	A British family composed of one mother <b> > one son <b> > and one daughter <b> > were set to join me <b> > and we had devoted weeks to organizing our arrangements <b> >
4160_14187_000047_000003	You remember I said that four hypotheses were possible <b> > that the robbery was committed either by Rubin <b> > by Walter <b> > by John Hornby <b> > or by some other person <b> >	You recall that I mentioned four potential hypotheses <b> > the robbery was carried out either by Marry <b> > by Tom <b> > by Leo Pawson <b> > or by another individual <b> >
6836_61803_000012_000000	The judge is the same <b> > the jury the same <b> > and the spectators as before <b> > though with very different feelings in regard to the criminality of the accused <b> >	The judge remains unchanged <b> > the jury unchanged <b> > and the audience remains the same <b> > though their sentiments about the accused's guilt differ greatly <b> >
7505_258958_000019_000000	Further <b> > in the economic world <b> > there is much wealth <b> > that never can gratify any want directly <b> > many forms of wealth <b> > never can be consumption goods <b> >	Moreover <b> > in the realm of economics <b> > there exists a significant amount of wealth <b> > that can never directly satisfy any need <b> > various types of wealth <b> > never can be used as consumer products <b> >
6818_68772_000032_000002	It is a disgrace to this district that Mister Forbes allows his girlish campaign <b> > to be run by a lot of missus who should be at home darning stockings <b> > or if they were not able to do that <b> > practicing their music lessons <b> >	It is a shame for this area that Mister Forbes permits his girly campaign <b> > to be managed by a group of married women who ought to stay at home mending socks <b> > or if they cannot do that <b> > working on their musical instruments <b> >
1841_150351_000011_000000	The fact <b> > that the california indians could support themselves without any great exertion <b> > undoubtedly had the effect of making them indolent <b> > while in the desert regions of the Great Basin <b> > the struggle for something to eat was so severe <b> > that it kept the natives in a degraded condition <b> >	The reality <b> > that the Native Americans in California could sustain themselves with minimal effort <b> > indubitably contributed to their laziness <b> > whereas in the arid areas of the Great Basin <b> > the battle for food was so intense <b> > that it maintained the natives in a diminished state <b> >
2961_961_000019_000007	Other periods <b> > of wonderful length and complexity <b> > are not observed by men in general <b> > there is moreover <b> > a cycle <b> > or perfect year <b> > at the completion of which <b> > they all meet and coincide <b> > to this end <b> > the stars came into being <b> > that the created heaven might imitate the eternal nature <b> >	Other eras <b> > of remarkable duration and intricacy <b> > are not typically noticed by most people <b> > there is furthermore <b> > a cycle <b> > or wonderful year <b> > at the end of which <b> > they all converge and align <b> > for this purpose <b> > the stars were born <b> > so that the created heavens can reflect the eternal nature <b> >

Figure 3.9: The labeled transcripts provided in the human listening test for all 15 testing samples under parallel and non-parallel settings.

3) StyleTTS2 [LHR+24] [7]: the expressive TTS model with human-level speech quality, improved from StyleTTS [LHM22];

4) MatchaTTS [MTB+24] [8]: the fast and high-quality TTS model based on conditional flow matching.

To verify the effectiveness of our proposed modules, we compare ProsodyFM against three ablated variants:

5) w/o_intonation: remove the Terminal Intonation Encoder (along with the Pitch Processor and the Text-Pitch Aligner) from ProsodyFM;

6) w/o_break: remove the Phrase Break Encoder from ProsodyFM;

7) w/o_into_break: remove both the Terminal Intonation Encoder and the Phrase Break Encoder from ProsodyFM.

To provide a reference upper bound, we also include:

8) GT(vocoder): we extract the Mel-spectrogram from the ground truth audio and then reconstruct it using HiFiGAN.

3.6 Results and Conclusions

3.6.1 Model Performance

We conduct both objective and subjective evaluations to assess ProsodyFM and four SOTA models in terms of phrasing (break), intonation, and overall intelligibility. Table 3.1, Table 3.2, and Table 3.3 show the results.

Models	$RMSE_{f0} \downarrow$	$WER \downarrow$	$F1_{break} \uparrow$
GT(vocoder)	0.2264	2.05%	77.02
StyleSpeech	0.4477	4.30%	60.49
GenerSpeech	0.4556	7.29%	60.30
StyleTTS2	0.3120	3.25%	58.44
MatchaTTS	0.3376	3.56%	60.08
ProsodyFM	**0.3068**	**3.22%**	**62.76**

Table 3.1: Objective results on the LibriTTS testing set.

Models	MOS	$MOS_{intelligibility}$	$MOS_{intonation}$	MOS_{break}
GT (vocoder)	4.77±0.05	4.77±0.05	4.90±0.04	4.91±0.04
StyleSpeech	3.43±0.10	3.72±0.11	3.58±0.10	3.52±0.12
GenerSpeech	2.67±0.10	3.03±0.12	3.14±0.12	2.85±0.13
StyleTTS2	4.09±0.10	4.23±0.10	3.66±0.10	3.82±0.12
MatchaTTS	3.61±0.10	4.03±0.10	3.48±0.10	3.88±0.10
ProsodyFM	**4.22±0.08**	**4.47±0.07**	**4.36±0.08**	**4.42±0.08**

Table 3.2: MOS results with 95% confidence intervals on the LibriTTS testing set under the parallel setting (the transcript of the reference audio is the same as the target text).

Objective Results

As shown in Table 3.1, we observe that ProsodyFM outperforms the other four SOTA models across all three objective evaluation metrics. Additionally, the results for phrasing ($F1_{break}$) and intonation ($RMSE_{f0}$) show a positive correlation with overall intelligibility (WER). These results indicate that ProsodyFM exhibits superior performance in phrasing and intonation, which further contributes to its enhanced intelligibility.

Subjective Results

As shown in Table 3.2 and Table 3.3, compared to the other four SOTA models, ProsodyFM obtains significantly better scores in terms of $MOS_{intonation}$ and MOS_{break} under both parallel and non-parallel settings; it also achieves significantly better $MOS_{intelligibility}$ under the parallel setting. In the non-parallel setting, as shown in Table 3.3, ProsodyFM matches the

[7]https://github.com/yl4579/StyleTTS2
[8]https://github.com/shivammehta25/Matcha-TTS

Models	MOS	$MOS_{intelligibility}$	$MOS_{intonation}$	MOS_{break}
GT (vocoder)	-	-	-	-
StyleSpeech	3.39±0.10	3.72±0.11	3.53±0.10	3.74±0.11
GenerSpeech	-	-	-	-
StyleTTS2	**4.26±0.10**	4.36±0.09	3.78±0.11	4.08±0.11
MatchaTTS	3.90±0.09	4.10±0.10	3.73±0.10	4.08±0.11
ProsodyFM	4.07±0.09	**4.40±0.08**	**4.24±0.09**	**4.34±0.09**

Table 3.3: MOS results with 95% confidence intervals on the LibriTTS testing set under the non-parallel setting (the transcript of the reference audio is different from the target text).

Models	$RMSE_{f0} \downarrow$	$WER \downarrow$	$F1_{break} \uparrow$
GT(vocoder)	0.2264	2.05%	77.02%
MatchaTTS	0.4727	6.78%	55.28%
ProsodyFM	**0.4080**	**4.97%**	**59.95%**

Table 3.4: Objective evaluation results on the out-of-distribution (unseen long and complex sentences, unseen speakers) testing data. All the models are trained on the VCTK (short sentences) training set and tested on the LibriTTS (long and complex sentences) testing set.

$MOS_{intelligibility}$ of StyleTTS2 and surpasses the remaining three models significantly. In the non-parallel setting, ProsodyFM shows lower speech quality (MOS) than StyleTTS2, likely due to the smaller dataset used in ProsodyFM (71 hours) compared to StyleTTS2 (245 hours). Consistent with the objective evaluation results, we can still observe a positive correlation between $MOS_{intonation}$, MOS_{break} and $MOS_{intelligibility}$, further substantiating that ProsodyFM can effectively improve the phrasing and intonation aspects of prosody, thereby enhancing the speech intelligibility.

3.6.2 Model Generalizability

To evaluate the impact of enhanced phrasing and intonation on the generalizability of models, we conduct out-of-distribution experiments on unseen complex sentences. Both MatchaTTS and ProsodyFM are trained on the same VCTK training set (short sentences) and tested on the LibriTTS testing set (long sentences). The speakers in the testing set are unseen during training. Table 3.4 presents the results.

Models	$RMSE_{f0} \downarrow$	$WER \downarrow$	$F1_{break} \uparrow$
GT(vocoder)	0.2244	2.01%	76.04%
ProsodyFM	**0.3047**	**2.86%**	**62.51%**
w/o_intonation	0.3373	3.13%	61.25%
w/o_break	0.3355	3.10%	61.06%
w/o_into_break	0.3391	3.33%	60.25%

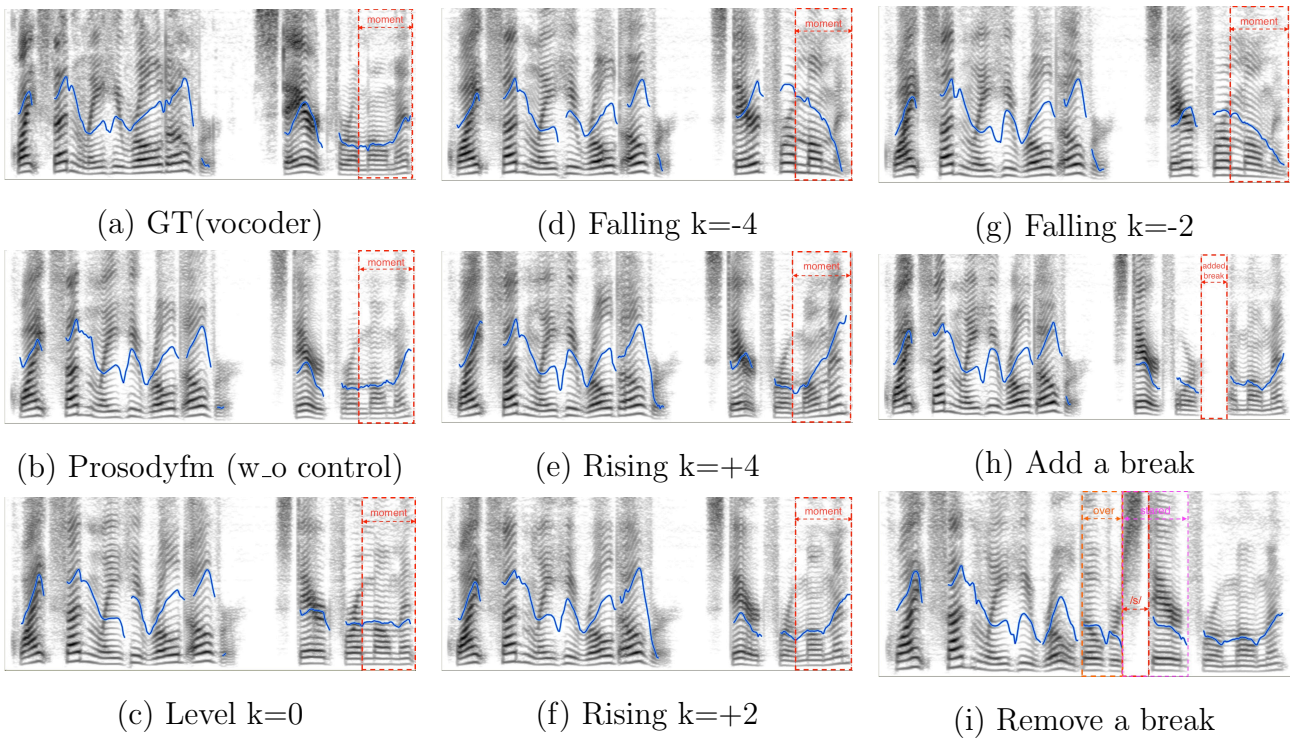
Table 3.5: Ablation study on the LibriTTS validation set.

We observe that although both MatchaTTS and ProsodyFM experience performance declines on the out-of-distribution testing set, the decrease in ProsodyFM's performance is considerably smaller than that of MatchaTTS. Notably, for the $F1_{break}$ metric, ProsodyFM in the out-of-distribution setting achieves matching performance with the four SOTA models in the in-distribution setting (as shown in Table 3.1). This indicates that enhanced phrasing and intonation can bring strong generalizability to the ProsodyFM model.

3.6.3 Ablation Study

To verify the necessity and effectiveness of our proposed Phrase Break Encoder and Terminal Intonation Encoder, we compare ProsodyFM against its three ablated variants. Table 3.5 shows the results on our LibriTTS validation set.

We observe an improved $F1_{break}$ score in both w/o_intonation and w/o_break compared to w/o_into_break, which suggests that both the break encoder and intonation encoder inject partial phrase break information. Those break information may be complementary, leading to a further boost in $F1_{break}$ when we combine the two encoders in ProsodyFM. The $RMSE_{f0}$ for three ablated variants exhibit no substantial differences, which likely stems from the $RMSE_{f0}$ metric being calculated only for voiced segments. When both intonation and break information are present, ProsodyFM achieves considerably better $RMSE_{f0}$. These observations suggest that both the Phrase Break Encoder and Terminal Intonation Encoder are essential for synthesizing highly intelligible speech.



(a) GT(vocoder) (d) Falling k=-4 (g) Falling k=-2

(b) Prosodyfm (w_o control) (e) Rising k=+4 (h) Add a break

(c) Level k=0 (f) Rising k=+2 (i) Remove a break

Figure 3.10: Part of spectrograms and pitch contours (blue line) of a reference speech and speech synthesized by ProsodyFM with controlled intonation and phrasing.

3.6.4 Case Study: Prosody Controllability

We present a case study to visually demonstrate the ability of ProsodyFM to control prosody, specifically in terms of intonation and phrasing, which are the focus of this chapter. Figure 3.10 displays the spectrograms of audio samples synthesized with controlled intonation and phrasing (can be found on our demo page). The reference speech is from our LibriTTS testing set (1363_139304_000009_000005.wav). Due to the space limit, we here show part of the spectrograms in Figure 3.10. The corresponding transcript with the original break and intonation labels is "Quite suddenly he rolled over (falling tone) stared for a moment (rising tone)".

Intonation Control

For the controllability of the terminal intonation, we manually modify the reference pitch shape segment of the last word in an intonational phrase and synthesize the corresponding speech using ProsodyFM (we modify the input "Last-word(reference) Pitch Shape Segments" of the Terminal Intonation Encoder in Figure 3.6 (b)). We apply linear adjustments to the pitch

values to create rising, falling, and level tones, controlling the magnitude of these adjustments through the slope: a slope of $k = +4$ represents a rapid rising tone, $k = +2$ represents a gradual rising tone, $k = -4$ indicates a rapid falling tone, $k = -2$ indicates a gradual falling tone, and $k = 0$ corresponds to a level tone. In Figure 3.10 (c-g), we modified the reference pitch shape segment of the last word "moment".

We observe that when a level tone is provided as a reference, the pitch contour corresponding to the word "moment" in the synthesized speech remains essentially flat. Conversely, when rising or falling tones are used as references, the pitch contour for "moment" exhibits the corresponding upward or downward movement. Additionally, the pitch contour slopes in Figures 3.10 (d) and (e) are noticeably steeper than those in Figures 3.10 (g) and (f). This indicates that the pitch contour in the synthesized speech accurately reflects the same degree of reference rapid or gradual shape. These results demonstrate that our proposed Terminal Intonation Encoder effectively captures the reference intonation pattern, allowing ProsodyFM to achieve precise and fine-grained control over intonation.

Phrasing Control

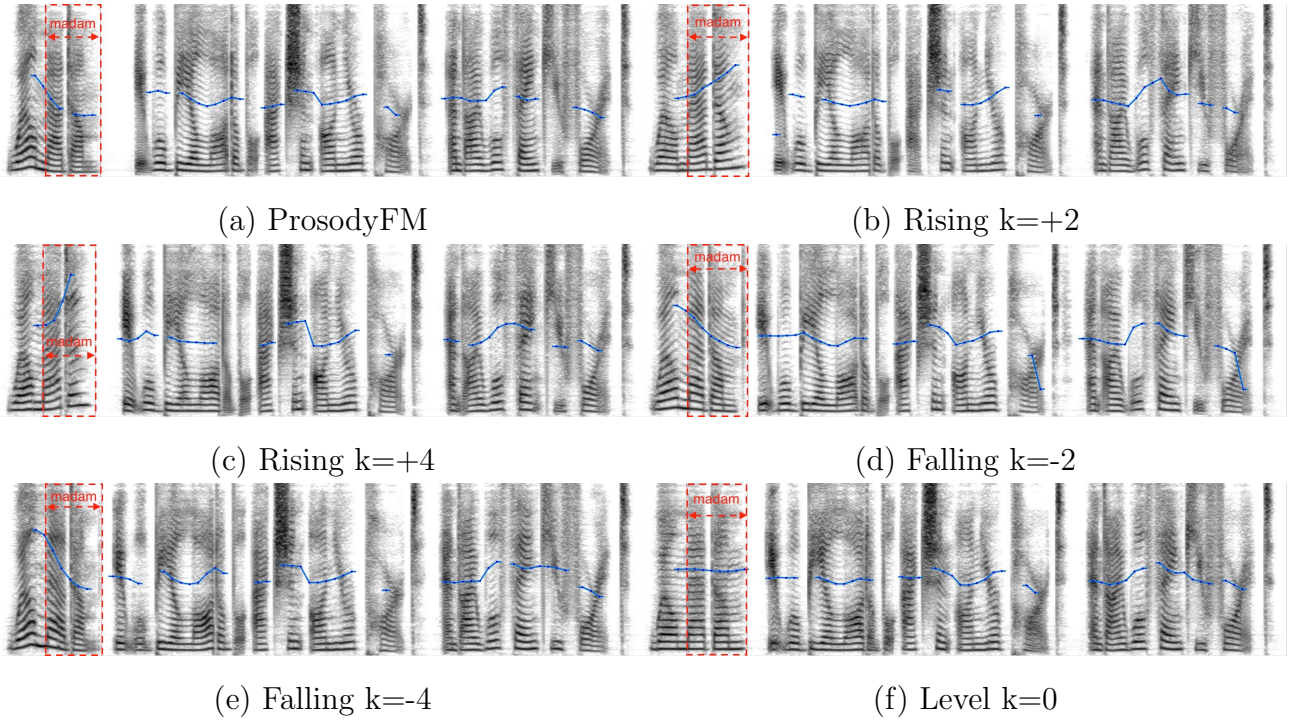
For the controllability of the phrase break, we manually add or remove a phrase break and synthesize the corresponding speech using ProsodyFM (we modify the "Phrase Breaks(last words)" in the Phrase Break Encoder in Figure 3 (b)). In Figure 3.10 (h), we added a phrase break after the word "for," and in Figure 3.10 (i), we removed the phrase break between "over" and "stared".

We observe that when a break is added after "for", as shown in Figure 3.10 (h), the spectrogram of the synthesized speech displays a noticeable blank space between "for" and "a moment." As shown in Figure 3.10 (i), when the break between "over" and "stared" is removed, a previously existing blank space disappears. This demonstrates that ProsodyFM exhibits excellent control over phrasing.

	Precision	Recall	F1
LibriTTS (w final-break)	93.96%	86.36%	90.00%
LibriTTS (w_o final-break)	91.04%	80.51%	85.45%
VCTK (w final-break)	99.15%	92.86%	95.90%
VCTK (w_o final-break)	94.74%	66.67%	78.26%

Table 3.6: The performance of the phrase break predictor on the LibriTTS and VCTK validation sets. "w final-break" and "w_o final-break" refer to the inclusion or exclusion of the sentence-final break, respectively, when calculating precision, recall, and F1 scores.

More Cases



(a) ProsodyFM (b) Rising k=+2

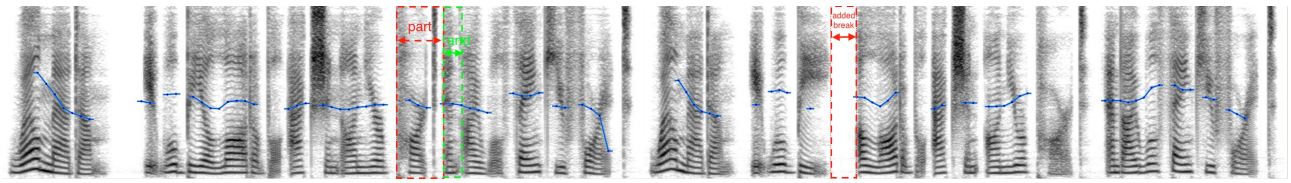
(c) Rising k=+4 (d) Falling k=-2

(e) Falling k=-4 (f) Level k=0

Figure 3.11: The terminal intonation control results for the 7302_86815_000052_000000.wav.

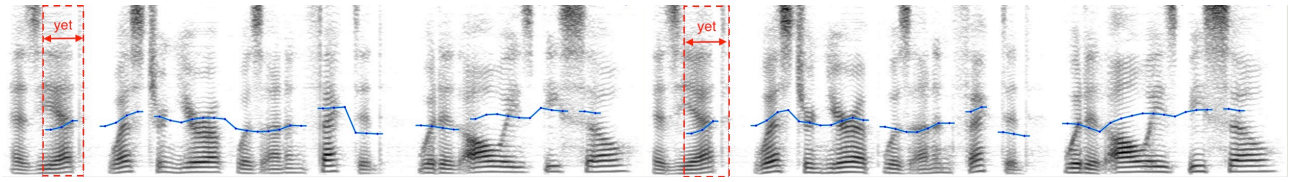
Figure 3.11 and Figure 3.12 show the intonation and phrasing control results of a speech sample 7302_86815_000052_000000.wav from the LibriTTS testing set. The original labeled transcript is "Well madam (falling tone) it will be a laudable action on your part (falling tone) and I will thank you for it (falling tone)". For the intonation control, in Figure 3.11 (b-f), we linearly adjust the reference pitch shape segment of the last word "madam" with different slope. For the phrasing control, in Figure 3.12 (g) we remove the break between "part" and "and". In Figure 3.12 (h) we add a break after "be".

Figure 3.13 and Figure 3.14 show the intonation and phrasing control results of a speech sample

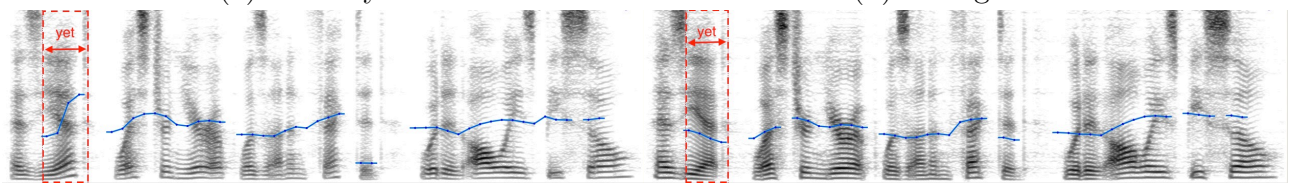


(g) Remove a break (h) Add a break

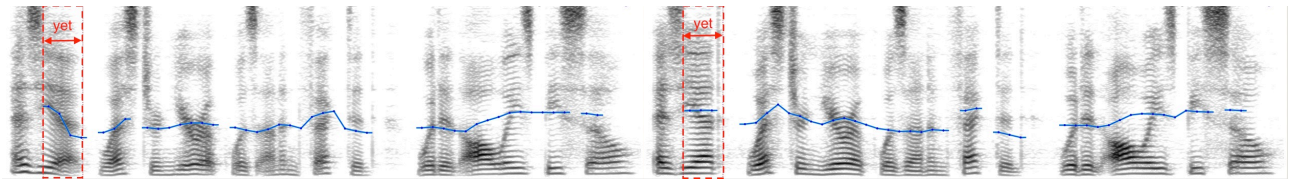
Figure 3.12: The phrase break control results for the 7302_86815_000052_000000.wav.



(a) ProsodyFM (b) Rising k=+2

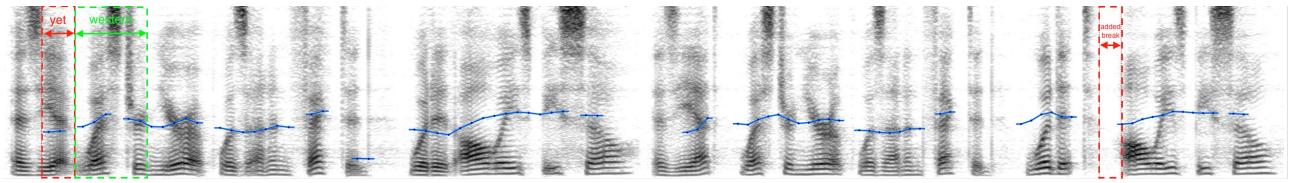


(c) Rising k=+4 (d) Falling k=-2



(e) Falling k=-4 (f) Level k=0

Figure 3.13: The terminal intonation control results for the 2002_139469_000016_000005.wav.



(g) Remove a break (h) Add a break

Figure 3.14: The phrase break control results for the 2002_139469_000016_000005.wav.

2002_139469_000016_000005.wav from the LibriTTS testing set. The original labeled transcript is "As yet (rising tone) western Europe was uninfected (falling tone) would it always be so (rising tone)". For the intonation control, in Figure 3.13 (b-f), we linearly adjust the reference pitch shape segment of the last word "yet" with different slope. For the phrasing control, in Figure 3.14 (g) we remove the break between "yet" and "western". For Figure 3.14 (h), we add a break after "it".

3.6.5 Performance of the Phrase Break Predictor

The phrase break predictor is trained independently from ProsodyFM on the LibriTTS training set. We fine-tune T5 and consider the phrase breaks obtained from the PSST as the ground truth labels when fine-tuning. We present its performance on the LibriTTS and VCTK validation sets in Table 3.6. Given that each sentence inherently concludes with a sentence-final break, we also report results excluding sentence-final breaks. We can observe a noticeable performance decline on the VCTK dataset when these final breaks are excluded. This decline can be attributed to the relatively short sentences in VCTK, which contain fewer intra-sentence breaks.

3.6.6 Conclusions

We proposed ProsodyFM, a novel prosody-aware TTS model designed to enhance phrasing and intonation without requiring any prosodic labels, resulting in more intelligible synthesized speech. We addressed the intonation modeling issue by training a set of intonation shape tokens combined with a novel pitch processor for more robust modeling of human-perceived intonation changes. We tackled the break labeling issue by designing a Phrase Break Encoder for initial phrase break cues and then adjusting the variable break durations with a duration predictor. Our performance experiments demonstrated that ProsodyFM effectively improved the phrasing and intonation aspects of prosody, thereby enhancing overall intelligibility compared to four SOTA models. Our out-of-distribution experiments showed that this enhanced prosody further brought ProsodyFM strong generalizability on unseen complex sentences. Our ablation study verified the effectiveness of our proposed modules. Our case study visually demonstrated ProsodyFM's powerful, precise, and fine-grained control over phrasing and intonation.

3.7 Contributions

The main contributions of this chapter are as follows:

- We propose ProsodyFM, a prosody-aware TTS model with strong generalizability and fine-grained prosody control, capable of synthesizing speech with natural phrasing and intonation, leading to greater intelligibility than existing systems.
- We provide novel and effective solutions for the break labeling issue and the intonation modeling issue.
- We release our demo, code, and model checkpoints to facilitate further research.

Chapter 4

Emotion-Aided Speech Dialogue Act Classification

Speech intelligibility involves more than just recognizing spoken words; it encompasses understanding the speaker's intended meaning and the communicative function of their utterances. Speech dialogue acts represent these functions—such as questioning, asserting, commanding, or expressing apology—and are essential for interpreting speech correctly. Accurate Dialogue Act Classification (DAC) enhances speech intelligibility by reducing misunderstandings and enabling systems to respond more naturally, especially in applications like virtual assistants and automated customer service. It also benefits related tasks such as Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU), bringing machine interpretation closer to human comprehension.

In this chapter, we investigate the impact of integrating paralinguistic features—specifically emotional information—into Speech Dialogue Act Classification models. We propose a novel multi-task training strategy that selectively incorporates beneficial paralinguistic features into DAC models while filtering out irrelevant features during different training stages. Our experimental results demonstrate that this selective incorporation of paralinguistic features into DAC significantly improves model performance. By enhancing the accuracy of DAC through paralinguistic information, we contribute to improved speech intelligibility.

This chapter is based on [HCS24] *Task Selection and Assignment for Multi-Modal Multi-Task Dialogue Act Classification with Non-Stationary Multi-Armed Bandits*, a conference paper accepted by ICASSP 2024 on April 14, 2024. I am the first author of this paper, responsible for identifying key research motivations, proposing improved multi-task training strategies, designing the model architecture, conducting the entire experimental process, and writing the conference paper. The second author, Junjie Chen, collaborated with me on the development and refinement of initial ideas and experimental design, offering valuable input during the paper's revision. The final author, Prof. Björn Schuller, provided ongoing supervision, offering guidance and feedback throughout the project's development and manuscript preparation.

4.1 Introduction

Given the supplementary role of paralinguistic features in conveying non-verbal information and their potential value for enhancing speech intelligibility, in this chapter, we explore the impact of incorporating paralinguistic features into Speech Dialogue Act Classification task. To introduce paralinguistic features, a natural choice is to adopt the Multi-task learning (MTL) framework, which aims to enhance the performance of a primary task by leveraging the knowledge of related auxiliary tasks [Car98]. MTL has achieved success over many research areas [ZWS19] [LJD19]. However, previous work [LLG19] [LYH16] found that using auxiliary tasks without careful selection can have a negative impact on performance, a phenomenon referred to as negative transfer. Therefore, implementing effective MTL faces two challenges: task selection, defined as the need to identify truly useful tasks, and task assignment, defined as the need to decide how to allocate those tasks.

DAC is the task of determining the communicative intention a speaker is holding in an utterance—whether they are questions, statements, commands, or expressions of apology. This task serves as the basis for many tasks related to speech intelligibility such as ASR [CMS+23] and NLU [YCHT+17]. For example, knowing the dialogue act can inform ASR models about the expected vocabulary and syntactic structures associated with specific communicative functions,

thereby improving overall recognition accuracy [CMS+23]. Incorporating dialogue act knowledge can also enhance human-computer interaction. In applications like customized dialogue system[HIM+14] [HSN06] and conversational agents [WP01], systems with accurate DAC can select more appropriate responses that match the conversational intent. This leads to interactions that feel more intuitive to users, reducing misunderstandings and improving efficiency.

Early work usually focused on single-task single-modal DAC [KGN16], that is, it uses chat transcripts with only the text mode to detect dialogue acts. Recently, thanks to the knowledge of the potential value of paralinguistic features to speech intelligibility and the development of multi-modal open-source datasets [SPSB20] [SGSB21], multi-task multi-modal DAC has gained attention. However, existing methods [SPSB20] [SGSB21] were based on the traditional MTL, which applied a random task selection strategy. Our results show that this strategy can produce negative transfer phenomena.

This chapter focuses on the multi-task multi-modal DAC task, where the primary focus is on DAC (8-class) supported by five auxiliary tasks that provide paralinguistic features. These auxiliary tasks include four emotion-related components—emotion classification (5-class), arousal regression (range [1,5]), valence regression (range [1,5]), and dominance regression (range [1,5])—as well as one speaker-related task, speaker classification (10-class). We propose a novel MTL strategy for selecting and assigning tasks with different kinds of paralinguistic features based on non-stationary multi-armed bandits (MAB) with discounted Thompson Sampling (TS) using Gaussian priors. Our experimental results show that in different training stages, different kinds of paralinguistic features (tasks) have different utility. Our proposed method can effectively identify the task utility, actively avoid useless or harmful tasks, and realize the task assignment during training. Our proposed method is significantly superior in terms of UAR and F1 to the single-task and multi-task baselines with p-values <0.05. Further analysis of experiments indicates that for the dataset with the data imbalance problem, our proposed method has significantly higher stability and can obtain consistent and decent performance for minority classes. Our proposed method is superior to the current state-of-the-art model.

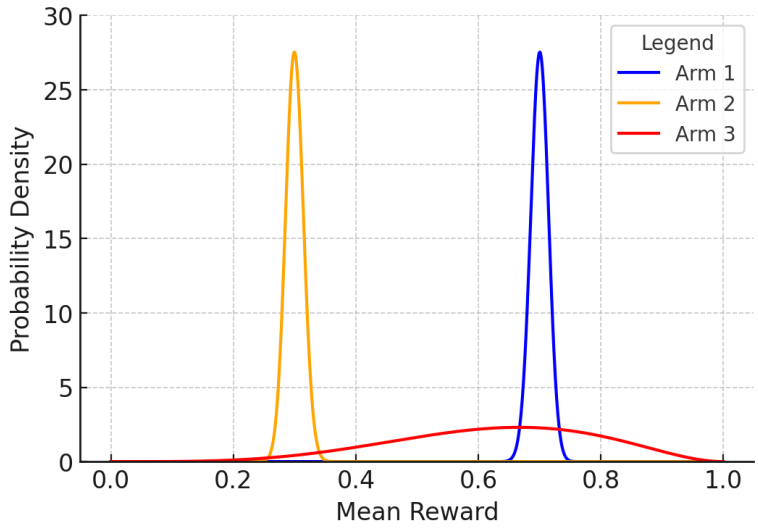


Figure 4.1: Thompson Sampling can balance exploration and exploitation. The greedy algorithm would always choose arm 1, while the Thompson Sampling algorithm has a chance to explore arm 3.

4.2 Background

4.2.1 The Multi-armed Bandit Problem

The multi-armed bandit (MAB) framework offers a potential solution to the above-mentioned two challenges in MTL through its sequential decision-making properties. In the MAB problem, we are given a slot machine with K arms, represented by $I := \{1, 2, \ldots, K\}$. Over a finite time horizon T, at each discrete time step $t := \{1, 2, \ldots, T\}$, a decision is made to select and play one of the K arms. When an arm i is played, it generates a random reward—a real-valued outcome determined by an unknown probability distribution unique to that arm, with support in $[0, 1]$. The reward distribution for each arm might differ in mean μ_i, which represents the expected reward for each arm. The reward distribution can be either stationary, meaning it remains unchanged over time (Stochastic Bandits), or non-stationary, meaning it changes over time (Non-Stationary Bandits). The rewards from repeated plays of a given arm are assumed to be independently and identically distributed (i.i.d.), and they are independent of the outcomes from other arms. The reward for each play is observed immediately after the arm is played.

The goal of an algorithm solving the MAB problem is to decide which arm to play at each time

step t, utilizing the reward information from the previous $t-1$ plays. A common objective is to maximize the cumulative reward over the T periods $Reward = \sum_{t=1}^{T} r_t$, where r_t denotes the reward obtained at time t.

4.2.2 Thompson Sampling

Algorithm 2 General Thompson Sampling

Input: a set of arms $i = 1, 2, ..., K$, the prior reward distributions for each arm $p_1, p_2, ..., p_K$, the number of round (time) T

Output: a sequence of chosen arms $\{a_1, a_2, ..., a_T\}$ and their corresponding rewards $\{r_1, r_2, ..., r_T\}$

1: **for** $t = 1, ..., T$ **do**
2: **for** $i = 1, ..., K$ **do**
3: sample $\theta_t(i)$ from the reward distribution p_i
4: **end for**
5: Choose arm $a_t = \arg\max_i \theta_t(i)$.
6: Play arm a_t and observe the reward r_t.
7: Update the posterior reward distribution p_{a_t} for the selected arm based on the observed reward r_t.
8: **end for**

In the MAB problem, the "explore-exploit" trade-off is a core challenge. Exploration involves selecting different arms to gather information about their reward distributions, which is essential when these distributions are initially unknown. Exploitation, on the other hand, focuses on selecting the arm with the highest estimated reward to maximize cumulative gain. Balancing exploration and exploitation is critical; in a limited time T, excessive exploration can lead to suboptimal cumulative rewards, while premature exploitation risks missing potentially higher-reward arms. An effective MAB algorithm optimally navigates this trade-off to achieve maximum long-term reward.

Thompson Sampling (TS) provides a principled approach to the explore-exploit dilemma by using Bayesian inference to maintain and update a posterior distribution over each arm's ex-

pected reward. As described in Algorithm 2, at each time step, instead of directly selecting the arm with the highest mean reward (as in the greedy algorithm), the TS algorithm **samples** from these posterior reward distributions and selects the arm with the highest sampled reward. Figure 4.1 demonstrates how TS effectively balances the trade-off between exploration and exploitation. In this case of three arms, the greedy algorithm would always choose arm 1, as it has the highest mean reward. While arm 2 is clearly inferior to arm 1 due to its lower mean reward and similar variance, arm 3, despite its high variance, has a mean reward very close to that of arm 1. This high variance suggests that arm 3 may be under-explored. With TS, there is a chance that a sampled reward from arm 3 could exceed that of arm 1, allowing arm 3 to be explored further.

In the following two subsections, we present two variants of TS for stochastic (stationary) bandits: one using Beta priors with a Bernoulli likelihood, and the other using Gaussian priors with a Gaussian likelihood.

Thompson Sampling with Beta priors

Algorithm 3 Thompson Sampling with Beta priors

Input: a set of arms $i = 1, 2, ..., K$, the number of round (time) T, the prior parameters for each arm $\alpha_i = 0, \beta_i = 0$.

Output: a sequence of chosen arms $\{a_1, a_2, ..., a_T\}$ and their corresponding rewards $\{r_1, r_2, ..., r_T\}$.

for $t = 1, ..., T$ **do**
  for $i = 1, ..., K$ **do**
    sample $\theta_t(i)$ from the reward distribution Beta$(\alpha_i + 1, \beta_i + 1)$
  end for
  Choose arm $a_t = \arg\max_i \theta_t(i)$.
  Play arm a_t and observe the reward r_t.
  If $r_t = 1$, then update $\alpha_{a_t} \leftarrow \alpha_{a_t} + 1$; else if $r_t = 0$, update $\beta_{a_t} \leftarrow \beta_{a_t} + 1$.
end for

TS with Beta priors is well-suited for problems involving binary rewards, such as success (reward = 1) or failure (reward = 0) outcomes. This algorithm begins by assigning each arm i a prior distribution of Beta$(1, 1)$, which is a uniform distribution over the interval (0,1). This prior reflects an initial assumption of equal likelihood for all possible success rates μ_i, where μ_i represents the probability of obtaining a reward = 1. The Beta distribution is employed due to its conjugacy with the Bernoulli distribution, which facilitates straightforward posterior updates. Specifically, if the prior distribution is Beta(α, β), observing a Bernoulli trial results in a posterior distribution that remains within the Beta family. The posterior is updated to Beta$(\alpha + 1, \beta)$ following a Bernoulli trial with a success result, or Beta$(\alpha, \beta + 1)$ following a Bernoulli trial with a failure result. Algorithm 3 shows the whole process.

Thompson Sampling with Gaussian priors

Algorithm 4 Thompson Sampling with Beta priors

Input: a set of arms $i = 1, 2, ..., K$, the number of round (time) T, the prior parameters for each arm $\hat{\mu}_i(0) = 0, k_i(0) = 0$.

Output: a sequence of chosen arms $\{a_1, a_2, ..., a_T\}$ and their corresponding rewards $\{r_1, r_2, ..., r_T\}$.

1: **for** $t = 1, ..., T$ **do**
2: **for** $i = 1, ..., K$ **do**
3: sample $\theta_t(i)$ from the distribution $\mathcal{N}(\mu_i(t); \hat{\mu}_i(t), \frac{1}{k_i(t)+1})$
4: **end for**
5: Choose arm $a_t = \arg\max_i \theta_t(i)$.
6: Play arm a_t and observe the reward r_t.
7: Update $\hat{\mu}_i(t) \leftarrow \frac{\hat{\mu}_i(t)k_i(t)+r_i(t)}{k_i(t)+2}$ and $k_i(t) \leftarrow k_i(t) + 1$.
8: **end for**

Thompson Sampling with Gaussian priors is particularly suited for problems with continuous or normally distributed rewards. In this case, both the prior and posterior are modeled as Gaussian distributions, which also exhibit conjugacy under normal likelihoods. This algorithm assumes that at time t, the likelihood $p(r_i(t)|\mu_i)$ of reward $r_i(t)$ given parameter μ_i is Gaussian

$\mathcal{N}(r_i(t); \mu_i, 1)$ and the prior $p(\mu_i)$ of each arm i for μ_i follows Gaussian $\mathcal{N}(\mu_i(t); \hat{\mu}_i(t), \frac{1}{k_i(t)+1})$, where $\hat{\mu}_i(t)$ denotes the running estimate of the mean reward for arm i at time t and $k_i(t)$ represents the number of times arm i has been played up to time $t-1$. According to the Bayesian inference, the posterior $P(\mu_i|r_i(t))$ is

$$P(\mu_i|r_i(t)) = \frac{P(r_i(t)|\mu_i)P(\mu_i)}{P(r_i(t))} \propto P(r_i(t)|\mu_i)P(\mu_i).$$

Since both the prior $p(\mu_i)$ and likelihood $P(r_i(t)|\mu_i)$ are Gaussian, the posterior $P(\mu_i|r_i(t))$ is also Gaussian (conjugacy property). Specifically:

$$\mu_i \mid r_i(t) \sim \mathcal{N}\left(\hat{\mu}_i^{\text{new}}, \frac{1}{k_i^{\text{new}}+1}\right)$$

where the updated parameters are derived as follows:

$$\hat{\mu}_i^{\text{new}} = \frac{\hat{\mu}_i(t)k_i(t) + r_i(t)}{k_i(t)+2}$$

$$k_i^{\text{new}} = k_i(t) + 1$$

The initial assumption of prior parameters at time $t = 0$ are $\hat{\mu}_i(0) = 0$ and $k_i(0) = 0$. Algorithm 4 describes the whole process.

4.2.3 Non-stationary Multi-armed Bandit Problem

The two TS algorithms mentioned above (Algorithm 3 and Algorithm 4) are designed for stochastic (stationary) bandits. The Non-stationary Multi-armed Bandit (NS-MAB) problem is a generalization of this stochastic MAB problem where the reward distributions associated with each arm change over time. This dynamic environment introduces a more challenging exploration and exploitation balance problem, as the agent must adapt to temporal variations in the reward distribution.

To address this challenge, strategies for the NS-MAB problem generally fall into two main cat-

egories—the first category leverages change-point detection algorithms [CLLW19, BKMS22] to explicitly identify shifts in the reward distributions, enabling the algorithm to reset or adapt accordingly. The second category focuses on passively mitigating the influence of past observations, often employing techniques like sliding windows [TPRG20], weighted updates [RVC19], or discount factors [RK17]. Theoretical analysis often focuses on deriving bounds for the regret, which measures how much performance is lost because the algorithm has to adapt to changing conditions, compared to an ideal system that already knows the reward patterns over time.

The NS-MAB algorithm (Algorithm 5) [QWZ23] implemented in this chapter is based on Discounted TS with Gaussian priors. It passively adapts to changes in the reward distribution by applying a discount factor to the past observations in TS. This approach is effective in both abruptly changing settings, where the reward distribution remains constant for a period and shifts at unknown times, and smoothly changing settings, where the reward distribution evolves gradually according to unknown dynamics. The core idea of this algorithm is to adjust the sampling variance: reducing it for the selected arm to prioritize exploitation while increasing it for unselected arms to encourage exploration.

4.3 Related Works

Previous work on multi-task DAC [SPSB20] [SGSB21] demonstrated that incorporating the emotion recognition (ER) task improved performance compared to single-task baselines. However, we observed instances of negative transfer. To further investigate the influence of paralinguistic information on DAC performance, we introduce four emotion-related tasks, including the ER task and three additional emotion regression tasks, along with a speaker-related task to evaluate our algorithm's ability to filter out irrelevant information. Unlike [SPSB20] [SGSB21], which employed a uniformly random task selection strategy during training, we propose a novel MAB-based task selection and assignment approach.

[GPB19] first proposed a paradigm of task selection for MTL via MAB with TS. However, it continuously added strong bias to the primary task, thereby failing to explore the different

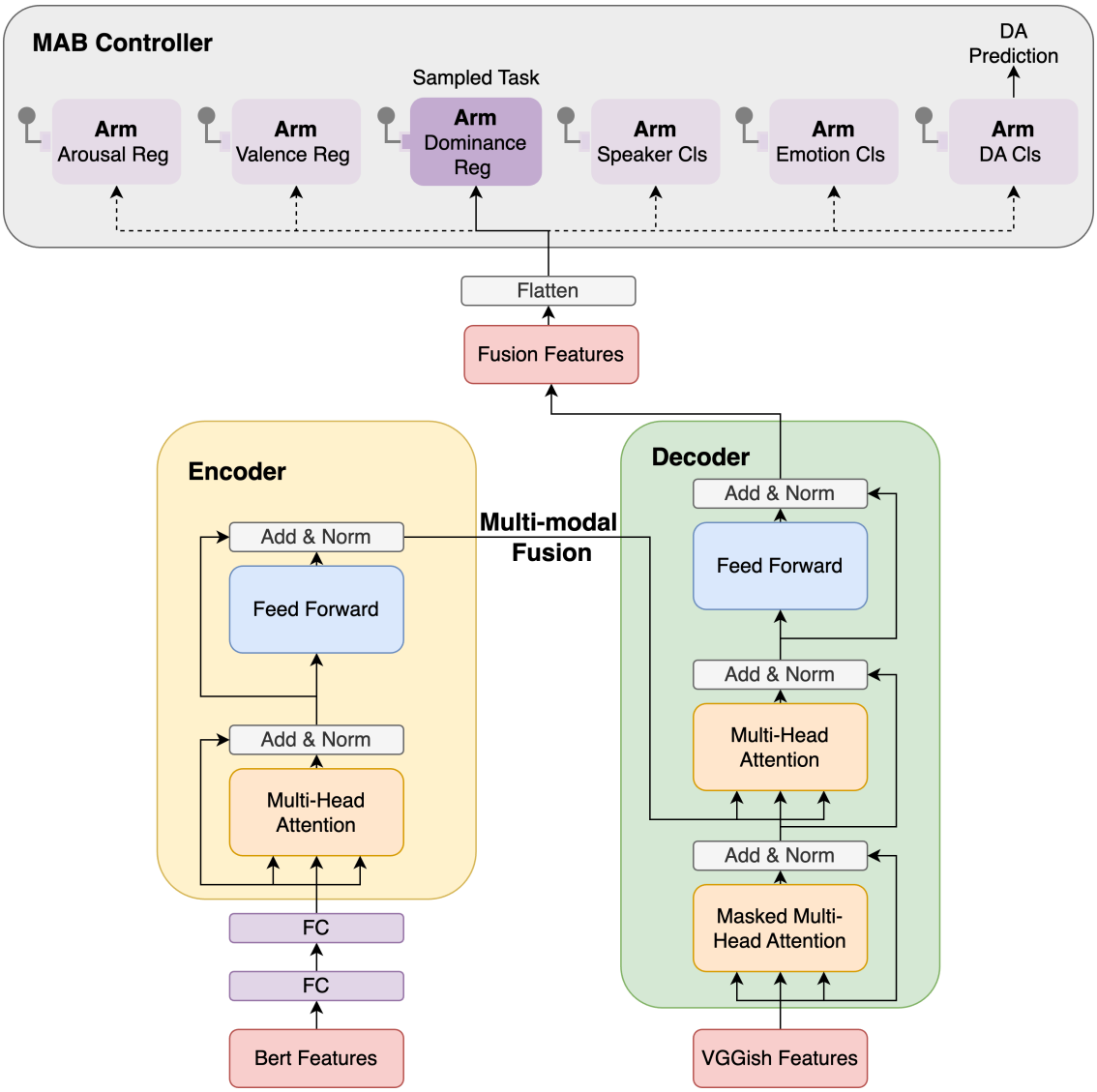


Figure 4.2: Proposed model architecture.

utility of auxiliary tasks in different training stages. It only exploited MAB for task selection and then needed to combine it with a Gaussian Process to realise task assignment. Our method can identify the utility of tasks in different training stages and can achieve the end-to-end task assignment.

4.4 Methodology

Inspired by [GPB19], we model the MTL as a MAB problem with TS. K arms correspond to K tasks. Selecting an arm to play corresponds to selecting a task to train the model. We

Algorithm 5 Non-stationary Multi-armed Bandit with discounted Thompson Sampling

Inputs:
γ, $\hat{\mu}_1(i)$, $\tilde{\mu}_1(i)$, $N_1(i)$, s, $\tau_1(i)$, τ_{max_bound}
for $t = 1, ..., T$ **do**
  for $i = 1, ..., K$ **do**
    sample $\theta_t(i)$ from $\mathcal{N}(\hat{\mu}_t(i), \tau_t(i)^2)$
  end for
  Select task $i_t = \arg\max_i \theta_t(i)$ and get reward $r_t(i_t)$
  for $i = 1, ..., K$ **do**
    $\tilde{\mu}_{t+1}(i) = \gamma\tilde{\mu}_t(i) + \mathbb{1}_{i=i_t}(i)r_t(i_t)$
    $N_{t+1}(i) = \gamma N_t(i) + \mathbb{1}_{i=i_t}(i)$
    $\hat{\mu}_{t+1}(i) = \frac{\tilde{\mu}_{t+1}(i)}{N_{t+1}(i)}$
    $\tau_{max} = \min\{st + \tau_1(i), \tau_{max\_bound}\}$
    $\tau_{t+1}(i) = \min\{\frac{1}{\sqrt{N_{t+1}(i)}}, \tau_{max}\}$
  end for
end for

define the utility of the selected task as the performance improvement after selecting this task to update the model relative to the performance of the current best model. That is, reward $r_t = V_t(i_t) - V_t^{best}$, where $V_t(i_t)$ refers to the validation performance of the model obtained after training with the selected task i_t at the round t and V_t^{best} refers to the best performance obtained on the validation set before round t. So the larger the reward, the greater the performance improvement (utility) brought by the selected task i_t.

The intrinsic utility of each task may change during training, e. g., the primary task might be more useful at the beginning, but not necessarily at the end. Therefore, the agent faces a non-stationary system where it must consistently be motivated to explore, ensuring it can adapt to shifts as the system evolves.

Algorithm 5 shows the details of our algorithm. We make small changes to the algorithm in [QWZ23]. In Algorithm 5, γ denotes discounted factor, $\tau_t(i)$ denotes the sampling variance, τ_{max_bound} denotes the upper bound of the sampling variance, $\hat{\mu}_t(i)$ denotes the estimation of the expected reward of task i at round t, $\tilde{\mu}_1(i)$ denotes the discounted cumulative reward, $N_t(i)$ denotes the discounted number of plays of task i until round t, and s denotes the slope of τ_{max}. The key to the algorithm is to encourage exploration by increasing the sampling variance of unselected tasks and decreasing the sampling variance of selected tasks.

4.4.1 Model Architecture

Figure 4.2 illustrates the proposed model architecture. The encoder of a standard transformer is employed to process BERT features from the text modality, while the decoder handles VGGish features from the audio modality. The self-attention mechanism within the transformer facilitates the fusion of these two modalities. To ensure compatibility with the self-attention mechanism, which requires both features to have the same dimensionality, two fully connected layers are added on top of the transformer encoder to reduce the dimensionality of the BERT features. The MAB controller dynamically selects a task for training based on Algorithm 5.

4.4.2 Training Objective

The proposed MAB method selects a task from 6 tasks according to the MAB controller in each epoch of training to update the model. We choose the cross-entropy loss for classification tasks and the concordance correlation coefficient (CCC) loss for regression tasks. The CCC loss is defined as:

$$\mathcal{L}_{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where ρ denotes the Pearson correlation coefficient between the predicted values x and the ground truth y; μ_x and μ_y represent the means of x and y, respectively; and σ_x^2 and σ_y^2 are the variances of x and y, respectively.

4.5 Experimental Details

4.5.1 Model Configurations

We apply the pre-trained BERT model [1] to extract textual features and apply the pre-trained vggish model [2] to extract audio features. We fix the sequence length of text features to 60 and

[1] https://huggingface.co/bert-base-cased
[2] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

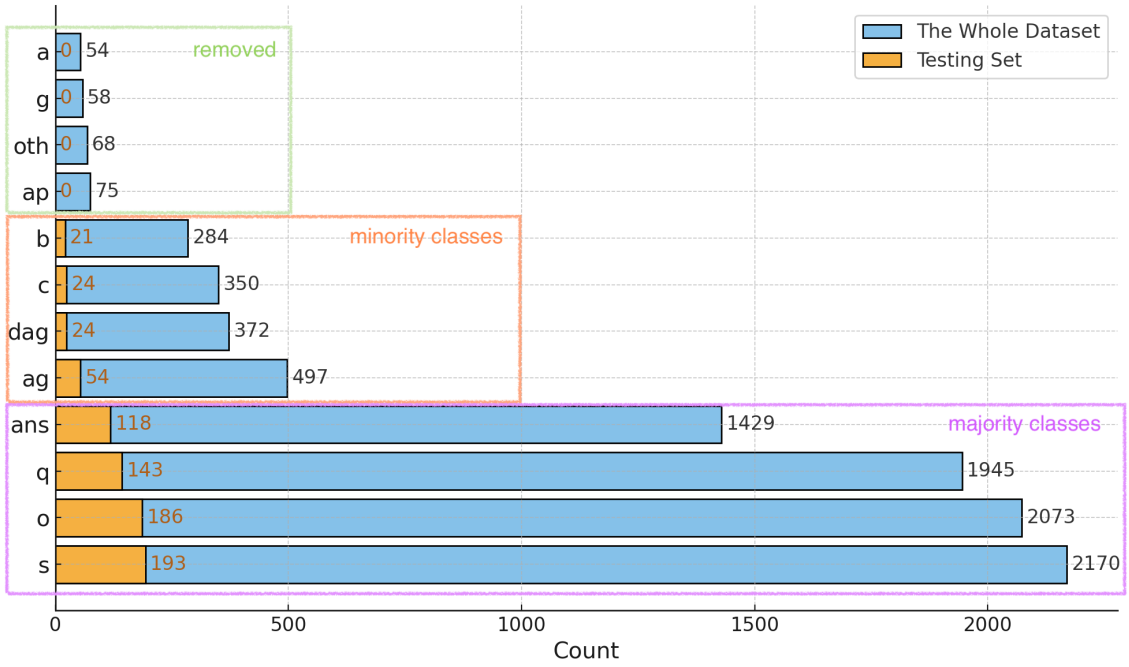


Figure 4.3: Class distribution for the whole dataset and the testing set we use.

the length of audio features to 20 by truncation or padding. The transformer part has 8 heads with 6 encoder and decoder layers.

We apply the Adam optimiser with a weight decay of 0.001. We train each model with 500 epochs and choose the model with the best validation loss. As in [GPB19], we initially set a slightly stronger prior for the primary task i^p, which $\hat{\mu}_1(i^p)$ and $\tilde{\mu}_1(i^p)$ are set to 0.3. $\hat{\mu}_1(i^a)$ and $\tilde{\mu}_1(i^a)$ for auxiliary tasks i^a are set to 0. $N_1(i)$, $\tau_1(i)$, τ_{max_bound}, and γ for all tasks are set to 0, 0.05, 0.5, and 0.9, respectively.

4.5.2 Datasets

We use the database proposed in [SGSB21] for all our experiments. This dataset is an extension of the original IEMOCAP dataset [BBL+08], adding 12 dialogue act (DA) tags including Greeting (g), Apology (ap), Command (c), Question (q), Answer (ans), Agreement (ag), Disagreement (dag), Statement-Opinion (o), Statement-Non-Opinion (s), Acknowledge (a), Backchannel (b), and Others (oth). The dataset has a total of 9376 audio samples. Figure 4.3 shows its class distribution.

The dataset used in Figure 4.4, Table 4.1, Figure 4.5, and Table 4.2 has 8 DA classes. We removed four DA classes ('ap', 'oth', 'g', 'a') with less than 0.8% of the total number of data samples. For emotion classification task, we replace the 'excitement' class in IEMOCAP with 'happiness' as in [XL15], and only consider classes 'anger', 'sadness', 'neutral', 'frustration', and 'happiness'. The final dataset has 7623 data samples in total. We did a random split with 70%, 20%, and 10% for training, validation, and testing sets, respectively.

The distribution of 8 DA classes in the testing set is shown in yellow in Figure 4.3. We call the 4 classes ('s', 'o', 'q', 'ans') with more than 117 data samples the majority class, and the 4 classes ('ag', 'dag', 'c', 'b') with less than 55 data samples the minority class. The dataset utilised in Table 4.3 has all 12 DA classes. To achieve a fair comparison, we utilise the same dataset and split as in [SGSB21], allocating 80% for training and 20% for testing.

4.5.3 Evaluation Metrics

The evaluation metrics used to assess model performance are Unweighted Average Recall (UAR), Macro-F1 (F1), and Accuracy. These metrics are calculated as follows:

$$\text{UAR} = \frac{1}{C}\sum_{k=1}^{C}\text{Recall}_k, \quad \text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}$$

$$\text{Macro-F1} = \frac{1}{C}\sum_{k=1}^{C}\text{F1}_k, \quad \text{F1}_k = \frac{2\cdot\text{TP}_k}{2\cdot\text{TP}_k + \text{FP}_k + \text{FN}_k}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$$

where C represents the total number of classes, k refers to the k-th class, and TP, FP, TN, and FN denotes the true/false positive, and true/false negative counts, respectively.

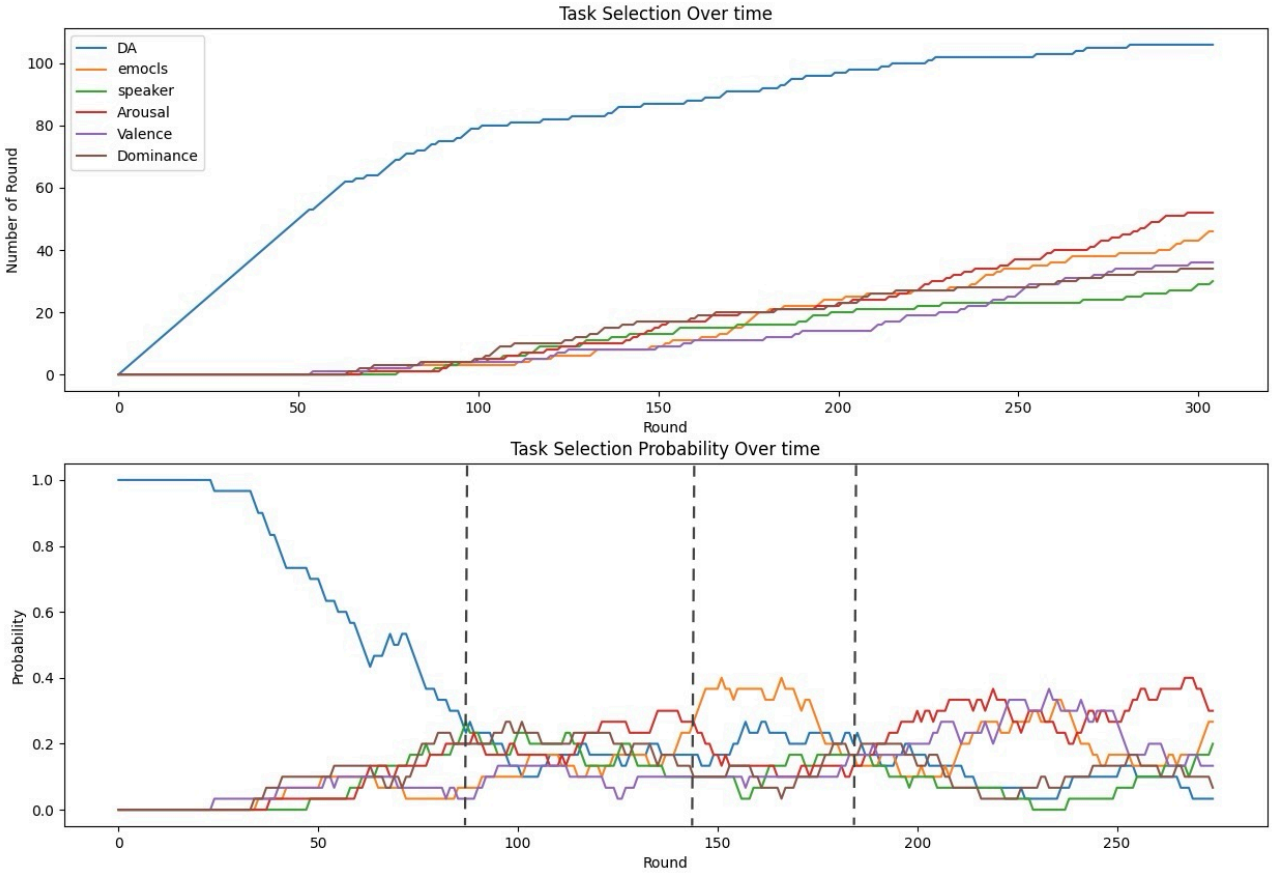


Figure 4.4: Task selection during the training process for the proposed MAB method.

4.5.4 Comparative Models

We compare our model against two baseline models (as in Table 4.1, Table 4.2, Figure 4.5, Figure 4.6) and the current state-of-the-art (SOTA) models (as in Table 4.3). The comparisons include:

1) ST: the single-task baseline model that trained solely on the DAC task;

2) MT: the multi-task baseline that trained on the same six tasks as our model but with a random task selection strategy during each training epoch;

3) SOAT Early Fusion: model form [SGSB21] utilizing an early fusion strategy;

4) SOAT Late Fusion: models in [SGSB21] utilizing an late fusion strategy.

Models	UAR	F1	Accuracy
ST	0.618±0.061	0.604±0.070	0.675±0.022
MT	0.522±0.052	0.499±0.054	0.608±0.019
Proposed	**0.667±0.013**	**0.655±0.013**	**0.683±0.010**

Table 4.1: Overall performance (8-class) on the testing set. Mean performance and standard deviations of 10 trials are reported.

4.6 Results and Conclusions

4.6.1 Task Selection

The top section of Figure 4.4 displays the cumulative counts of each task selected during training, while the bottom section illustrates the selection probability of each task throughout the training process. We compute the probability for task i using the number of times that task i is selected in a sliding window of length 30.

As shown by the gray line in the bottom section of Figure 4.4, the training process can be divided into four stages. In the first stage, the model first tends to choose the primary task that is most likely to bring the largest reward. As time goes on, the model finds that the reward of choosing only the primary task starts to decline, so it begins to explore other tasks and starts a comprehensive exploration in the second stage. In the third stage, the model finds that the emotion classification task is the most useful task, so there is a greater probability of selecting the emotion classification task. Finally, in the fourth stage, the model considers the arousal regression and the valence regression tasks as the first two useful tasks. The speaker classification and dominance regression tasks have a low selection probability during the entire training process, indicating that these two tasks are of low utility to the primary task.

4.6.2 Overall Performance

Table 4.1 shows the overall performance. The proposed method outperforms the two baseline methods on all metrics. The performance of the MT is far worse than that of the ST, which

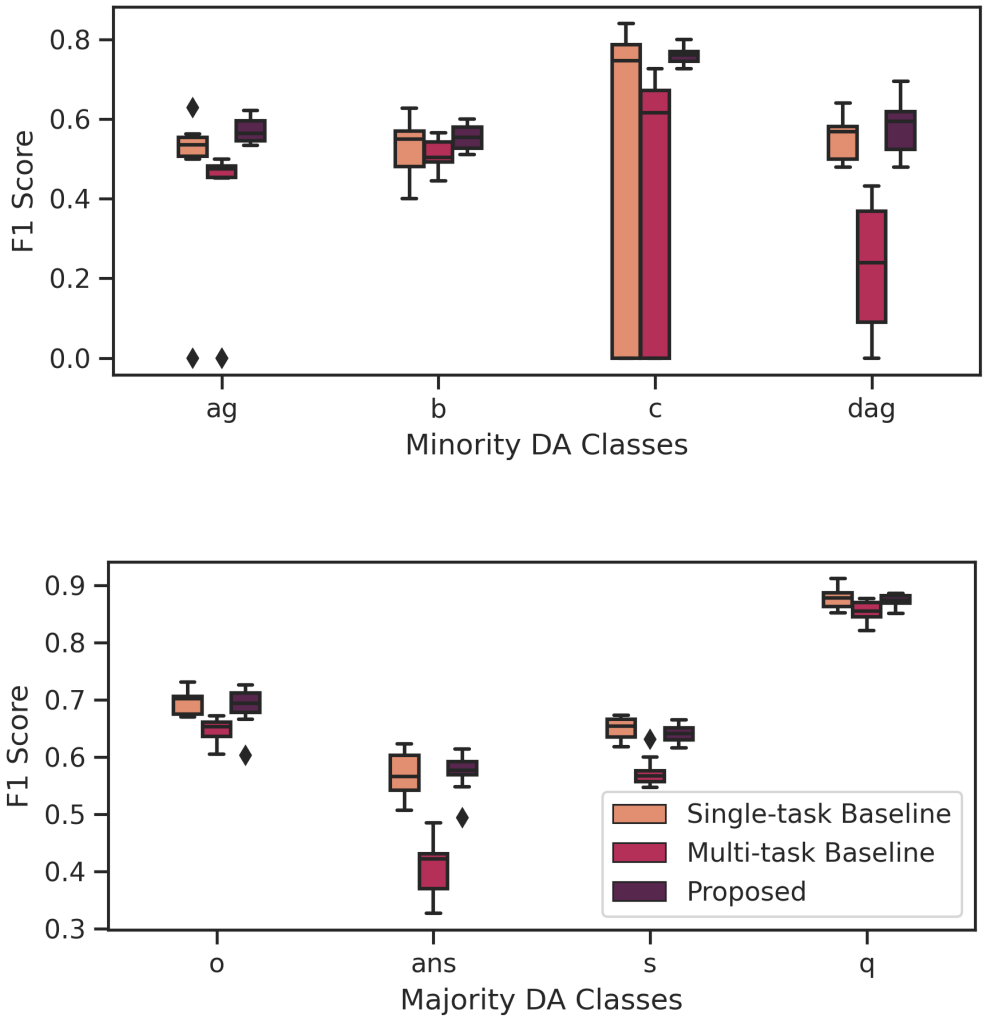


Figure 4.5: Boxplot of sub-class F1 on the testing set for 10 trials.

indicates that the random selection strategy can impede performance. The proposed method has a slight improvement in accuracy (metric not considering the class distribution) but has a significant improvement in F1 (Macro-F1) and UAR (metrics considering the class distribution) when compared to the two baselines. Our significance experiment (Welch's t-test [Wel47] at 5 % significance level) shows that this improvement of F1 and UAR is statistically significant. To find out the reason, we perform an analysis of sub-class performance.

4.6.3 Sub-class Performance

Figure 4.5 shows the box plot of the sub-class F1 scores. The F1 score of the MT method is significantly worse than the other two methods in almost all classes. The performance of the

Models	ag	c	dag	b
ST	0.489±0.176	0.473±0.408	0.550±**0.054**	0.530±0.068
MT	0.382±0.202	0.398±0.344	0.222±0.163	0.510±0.040
Proposed	**0.570±0.031**	**0.759±0.021**	**0.578**±0.068	**0.554±0.031**

Table 4.2: F1 for 4 minority classes on the testing set. Mean performance and standard deviations of 10 trials are reported.

proposed method is not significantly different from that of the ST method in four majority classes (the bottom part of Figure 4.5). However, for minority classes (the upper part of Figure 4.5), the proposed method obtained better results. Table 4.2 shows their sub-class F1 scores. For classes 'dag' and 'b', the proposed method has slightly better F1 scores than the ST method, while for classes 'ag' and 'c', the proposed method has much better results.

4.6.4 Stability of Proposed Method

We can observe an interesting phenomenon in class 'c'. Both Figure 4.5 and Table 4.2 demonstrate a huge instability of the ST method and MT method in the prediction of class 'c'. In fact, among the 10 ST models and 10 MT models recorded, 4 models each have an F1 score of 0. On the contrary, the proposed method has consistent and decent predictions.

The proposed method also has better stability for other classes besides class 'c'. Figure 4.5 and the standard deviation in Table 4.2 shows that the proposed method exhibits significantly improved stability in 3 out of 4 minority classes ('ag', 'c', 'b') compared to the ST, surpasses the stability of the MT across all 4 minority classes, and surpasses the two baseline methods in 3 out of 4 majority classes ('ans', 's', 'q'). Table 4.1 also shows that the two baseline methods have significantly larger standard deviations in F1 and UAR, indicating that the overall stability of the proposed method is better than that of the two baseline methods.

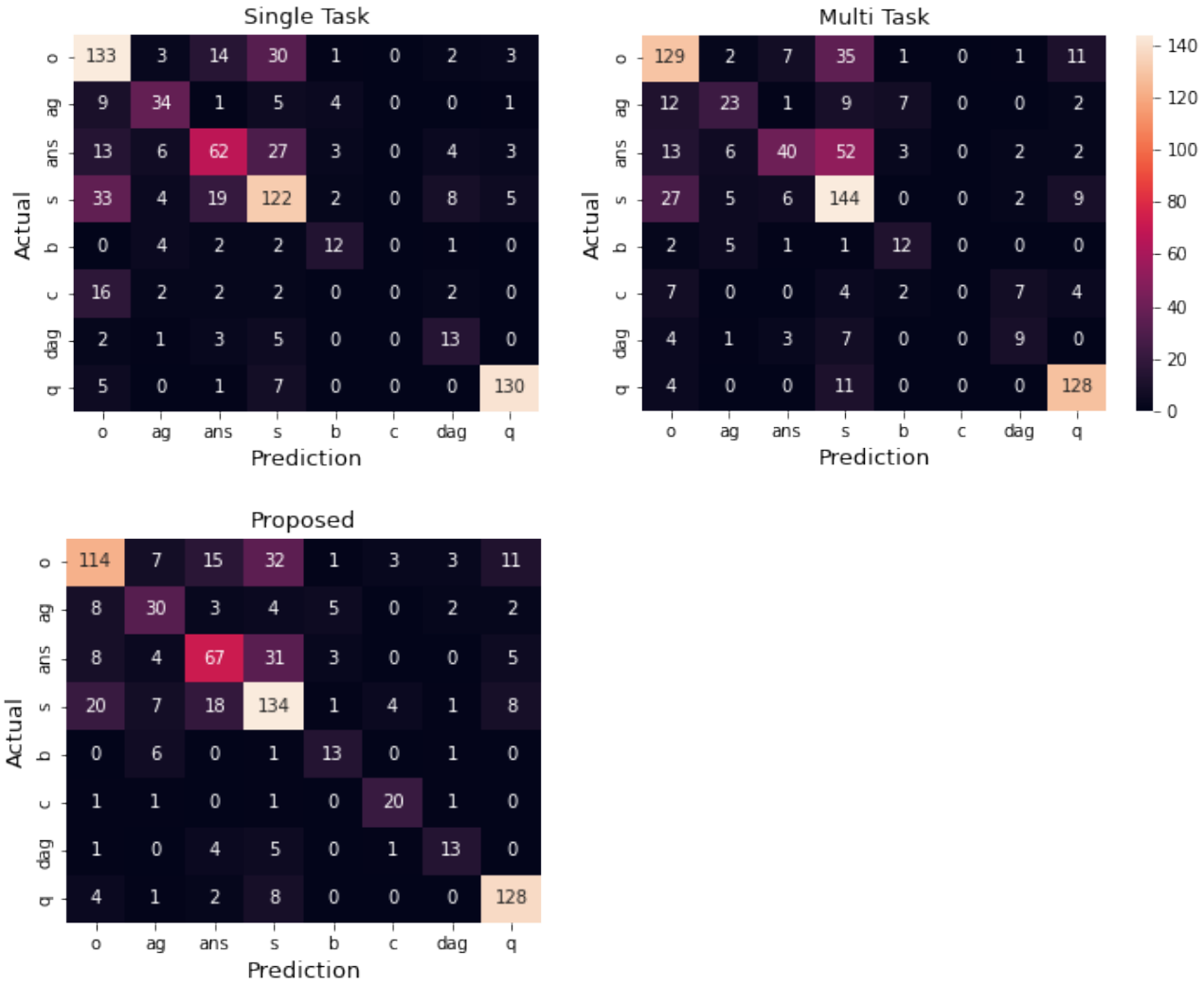


Figure 4.6: Case study for one trail: confusion matrices for the ST, MT baselines and our proposed model.

4.6.5 Misclassification Analysis

Figure 4.6 illustrates the confusion matrices for the ST, MT baseline models, and our proposed model in a single trial. The misclassification patterns differ across the models, with our proposed model generally demonstrating a more balanced and improved performance. Notably, for class 'c,' both the ST and MT baseline models failed to predict any instances as belonging to this category, while the proposed model demonstrated significantly improved performance in correctly classifying instances of this class.

Models	Accuracy	F1
SOTA early fusion [SGSB21]	0.6601	0.6385
SOTA late fusion [SGSB21]	**0.6963**	0.6786
Proposed	0.6951	**0.6858**

Table 4.3: Overall Performance (12-class) for the current SOTA and our proposed MAB method.

4.6.6 Comparison to SOTA

As shown in Table 4.3, although the proposed method obtains a matching accuracy with the SOTA [SGSB21], it has a considerable improvement on the F1 score. Note that the 12-class dataset used here has a more severe data imbalance problem, which shows the superiority of our proposed method.

4.6.7 Conclusions

This chapter studied the multi-task multi-modal DAC task and proposed a novel training strategy based on non-stationary MAB with discounted TS to deal with the task selection and task assignment challenges in MTL. Our results demonstrated that the proposed method can effectively identify task utility in different training stages and realise the end-to-end task assignment during training. The proposed method exhibited significant improvement in terms of UAR and F1 to both single-task and multi-task baselines. Experimental analysis of sub-class performance indicated that this improvement stemmed from the more stable performance of the minority classes. Further analysis illustrated that the overall stability of the proposed method is better than the baselines. The proposed method outperformed the current SOTA. Future work should also consider verifying the generalisability of the method, such as investigating primary tasks from diverse research fields.

4.7 Contributions

The main contributions of this chapter are as follows:

- We propose a novel training strategy based on non-stationary MAB with discounted TS to deal with the above two challenges of task selection and task assignment;
- We illustrate that the proposed method can effectively identify task utility in different training stages, avoid the negative transfer phenomena, and realise the end-to-end task assignment during training;
- We demonstrate that the performance improvement mainly comes from the significantly high stability of the minority classes;
- We compare the performance of the proposed method with the current state-of-the-art method [SGSB21] (SOTA) and show the superiority of the proposed method.

Chapter 5

Disentangled Emotion Features for Emotional Voice Conversion

Emotional features enrich speech with contextual information beyond literal words, enhancing intelligibility. However, accurately capturing and manipulating these emotional aspects in speech synthesis and conversion systems is challenging. Emotional Voice Conversion (EVC) is a task focused on transforming the emotional style of a source speech signal into a target style while preserving emotion-independent features, such as linguistic content and speaker identity. Previous EVC models often fail to disentangle emotional information from emotion-independent features that should be preserved, transforming them in a monolithic manner and resulting in less intelligible audio with severe distortions. In this chapter, we proposed a novel StarGAN-based EVC framework incorporating a two-stage training process that disentangles emotional features from emotion-independent attributes through the use of an autoencoder with dual encoders as the generator of the Generative Adversarial Network (GAN). The proposed model demonstrated favorable performance in both objective and subjective evaluations, particularly in terms of distortion, indicating its effectiveness of the disentangled emotional features in migrating distortions and enhancing intelligibility. Additionally, we validated the value of the generated audio for data augmentation in the Speech Emotion Recognition (SER) task. Results of data augmentation experiments for end-to-end SER indicated that the improvement

in audio quality obtained by the proposed model is not at the expense of its emotional conversion performance.

This chapter is based on [HCRS21] *An Improved StarGAN for Emotional Voice Conversion: Enhancing Voice Quality and Data Augmentation*, a conference paper accepted by Interspeech 2021 on August 30, 2021. It is important to note that the related work and comparative experiments presented in this chapter are limited to studies published prior to the submission date of the conference paper (March 2021). As the first author of this paper, I identified critical research gaps, proposed improvement ideas, designed the components of the improved StarGAN model, executed the entire experimental process, and authored the conference paper. The second author, Junjie Chen, contributed by discussing and refining the initial ideas and experimental design with me and provided insightful input during the revision of the paper. The third author, Dr. Georgios Rizos, provided the code and a technical report for the baseline StarGAN model to support this research. The last author, Prof. Björn Schuller, supervised the project throughout, offering helpful suggestions and guidance on the manuscript development.

5.1 Introduction

Emotional features provide contextual information that extends beyond literal words, contributing to enhanced speech intelligibility. However, effectively capturing and manipulating these emotional elements in speech synthesis and conversion systems remains a significant challenge. Emotional Voice Conversion aims to alter the emotional style of a source speech signal to a target style while preserving other emotion-independent information. A significant challenge in this task is the difficulty of obtaining high-quality parallel data due to the substantial labor costs involved [HZS17]. For instance, a typical dataset may include a source speech signal expressing a "happy" emotion for a specific sentence S by speaker A. However, if the task requires altering the emotion from "happy" to "sad", the dataset often lacks a corresponding recording of a sentence S spoken by speaker A with the "sad" emotion. Consequently, the training process for the model cannot rely on parallel data. Recent work proposed to utilize generative adversarial

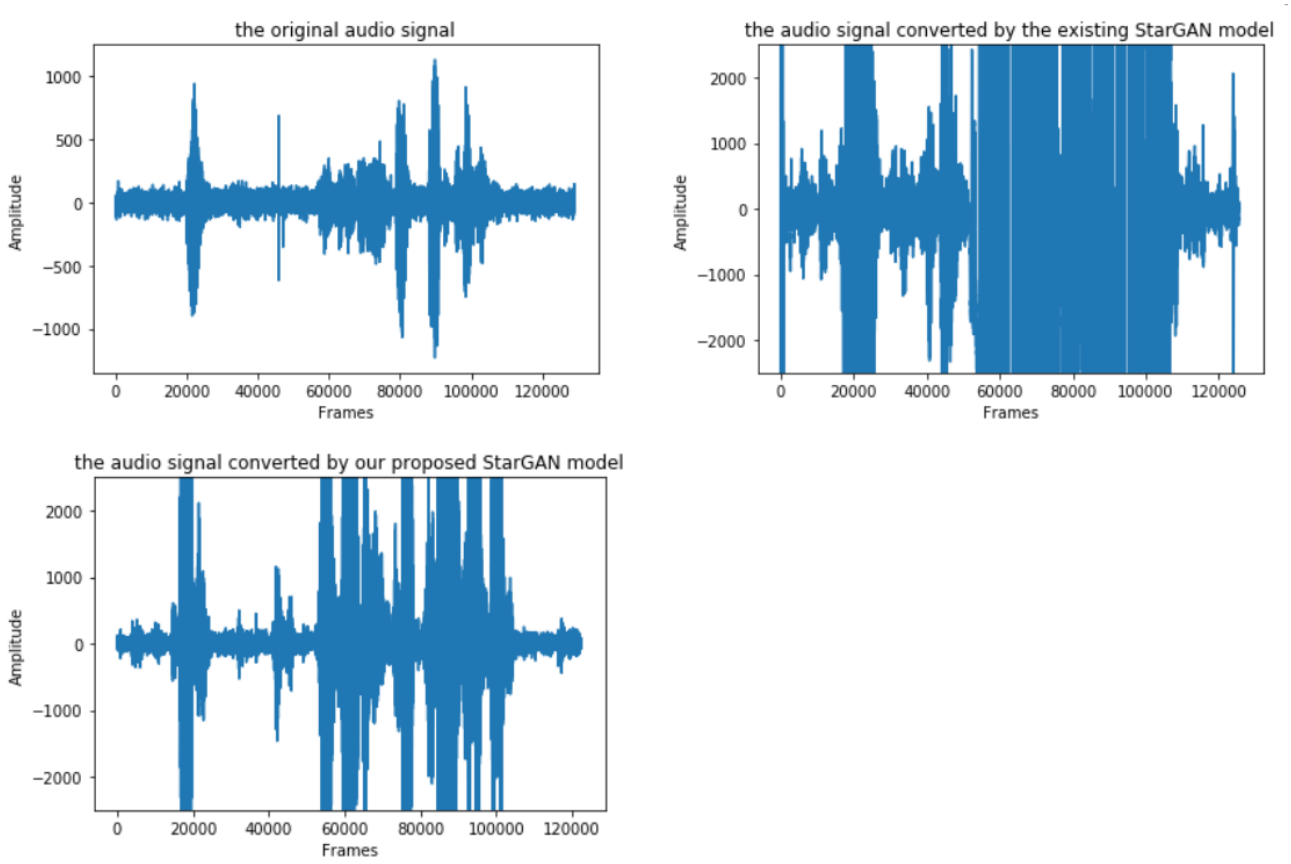


Figure 5.1: An example (Ses03F_script02_2_F043.wav in the IEMOCAP dataset) of waveforms for source and converted audio signals. Our improved StarGAN-EVC model introduces fewer artifacts in silent frames compared to the baseline StarGAN-EVC model [RBES20].

network (GAN) frameworks [GPAM+14, ZPIE17, CCK+18] to convert the spectrum features of a source audio signal to the features of the target [LCTA19, RBES20, ZSZL20, ZSLL20, SSV20], thus eliminating the need for parallel data. In addition, evidence shows that the GAN-generated audio signals, as a source of augmenting data, are valuable for improving the predictive performance of SER models [RBES20, BAS19], which not only indicates that the GAN-generated audio signals indeed carry effective emotional information, but also provides a promising data augmentation solution for the SER task [LAR+20, CPS+19, RS19, MBS+20].

However, when comparing the waveform of the source audio signal with that of the audio signal converted using the existing StarGAN-based EVC model [RBES20], we observe that certain silent frames in the original waveform are assigned disproportionately high amplitude values in the converted waveform. An example is shown in Figure 5.1. Since these silent frames represent silence between discrete words or sentences, excessive amplitude values should not be given to

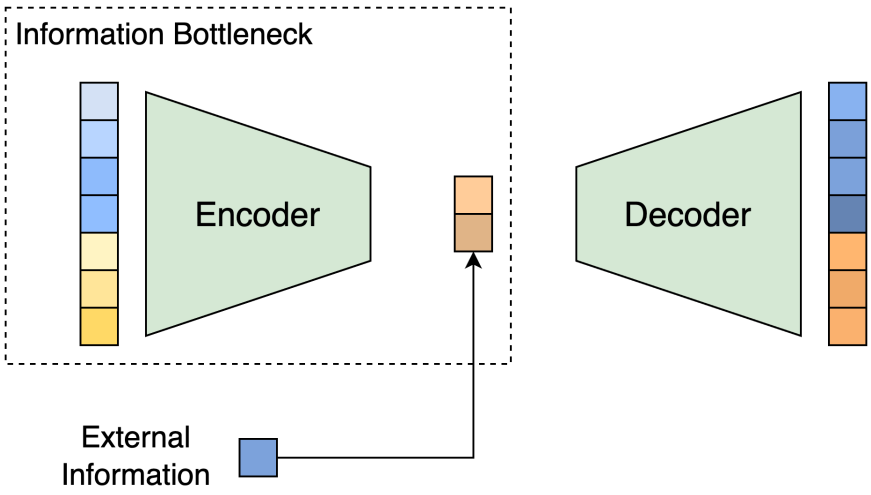


Figure 5.2: Information bottleneck in AE framework.

these frames in any conversion between emotions. This indicates that the existing StarGAN framework is prone to generating artifacts; we believe that this is because it transforms the input clip in a monolithic manner, instead of focusing on the emotional indices, as desired. This monolithic conversion manner tends to cause the distortion of the linguistic characteristics (e. g., cracked voices, or jitter) and damages the overall audio quality and speech intelligibility. Experimental results of existing EVC models also show that the mean opinion scores (MOS) for audio quality before and after conversion suffer from a significant drop [ZSLL20].

One approach to address this issue is to disentangle emotional features from emotion-independent features. Feature disentanglement methods based on autoencoder or variational autoencoder offer a potential solution (as shown in Figure 5.2) and have been successfully applied to the voice conversion (VC) task [HHW+16, LZS+18, HWL+19, QZC+19, QJHJM20, HLH+20]. In [HWL+19], by providing the speaker identity features as a condition to the decoder, the encoder learns to encode only the speaker-independent information in the process of minimizing the reconstruction loss. Experimental results from [HLH+20] show that the degree of disentanglement is positively correlated with the performance of the VC model and can be enhanced by both GANs and the speaker classifier. Inspired by these feature disentanglement methods, we propose to utilize an autoencoder with two encoders as the generator of the StarGAN as well as a two-stage training process for the model. The proposed model achieves favorable results in both the objective evaluation and the subjective evaluation in terms of distortion,

which reveals that the proposed model can effectively reduce distortion and enhance speech intelligibility. Furthermore, in data augmentation experiments for end-to-end SER, the proposed StarGAN model achieves an increase of 2 % in Micro-F1 and 5 % in Macro-F1 compared to the baseline StarGAN model, which indicates that the proposed model is more valuable for data augmentation.

5.2 Background

In this section, we first present the key feature disentanglement method that underpins the improved StarGAN-EVC, followed by an introduction to the StarGAN framework that forms its foundation.

5.2.1 Auto-encoder based Feature Disentanglement

The auto-encoder (AE) / variational auto-encoder (VAE) framework provides a possible solution for feature disentanglement and has been successfully applied to the VC and EVC tasks. This disentanglement mechanism follows the following principle [QZC+20]: if an information bottleneck exists in a neural network, the network will preferentially pass through information that cannot be provided elsewhere. As illustrated in Figure 5.2, the encoder in AE acts as the information bottleneck since the dimension of the encoder's output is usually much smaller than the dimension of its input. By providing the speaker identity embedding or the emotional embedding (blue blocks) as external information to the decoder, the encoder learns to encode only the speaker-independent or emotion-independent information (yellow blocks) in the process of minimizing the reconstruction loss. Experimental results from [HLH+20] show that the degree of disentanglement of this framework is positively correlated with the performance of the VC model.

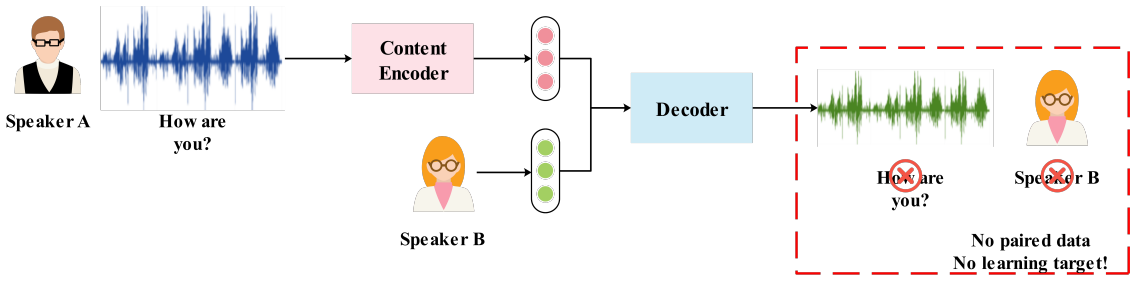


Figure 5.3: An example of the problem of having no learning target in the absence of paired training data.

5.2.2 StarGAN for Multi-domain Conversion

Directly applying the auto-encoder-based feature disentanglement framework on stylized voice conversion tasks (e. g., VC, EVC, and speaking style conversion) will cause the training-testing mismatch problem. The auto-encoder will simply learn to reconstruct the input audio signal throughout the training process without attempting to convert it. During the inference phase, the auto-encoder will try to convert the style for the first time. The conversion is an entirely new task for the model and has never been performed in the training process. This mismatch, to some extent, influences the performance of stylized voice conversion systems.

Some studies [CYLL18, LHL19] explore performing a second-stage training to address this mismatch problem. Those studies add a conversion stage to the auto-encoder-based feature disentanglement framework. However, in the absence of paired training data, directly adding a conversion stage to the training process will lead to the problem of having no learning target. For example, as shown in Figure 5.3, if we want to convert the voice of speaker A in a speech signal into the voice of speaker B, since we do not have the utterance of speaker B saying this sentence (no paired data), even if we add the conversion stage, the auto-encoder cannot carry out the reconstruction without a target. [LHL19] solves this by integrating a discriminator to determine whether the speech produced by the auto-encoder is real. This model at the second conversion stage follows GAN-based training. However, this method uses a single model for one-to-one domain conversion, which forces us to train multiple models and results in a high computational inefficiency when the multi-domain conversion is desired.

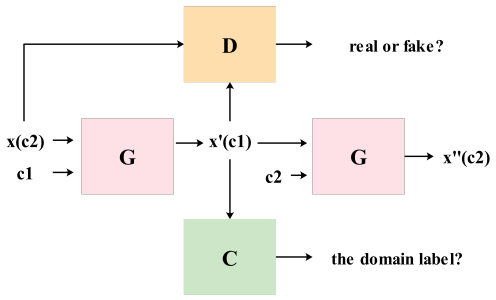


Figure 5.4: The architecture of StarGAN.

The architecture of StarGAN [CCK+18] allows multi-domain conversion with a single model. As shown in Figure 5.4, StarGAN has a generator G, discriminator D, and auxiliary classifier C. The generator G first takes a sample $x(c2)$ with a source domain label $c2$ and a target domain label $c1$ as input and outputs a converted version of x conditioned on $c1$, which is denoted as $x'(c1)$. The $x'(c1)$ and the source domain label $c2$ are then inputted back into the generator, which outputs a reconstructed version of $x''(c2)$. The input of discriminator D is either a real or fake sample x and its output is the probability of x being real. The classifier C takes a real or fake sample x as input and outputs a domain probability distribution of x. During training, The G tries to minimize the reconstruction loss and fool the D by generating a more realistic sample. The D tries to maximize the loss between the real sample and the fake sample. The C learns to predict the domain label and helps to enable the many-to-many emotional conversion.

5.2.3 Data Augmentation through Style Transfer

Data scarcity hinders the improvement of model performance in the field of SER, SI, ASR, and TTS, which is reflected not only in the lack of large-scale, well-annotated speech corpora but also in the imbalanced distribution over categories. Previous research seeks to leverage data augmentation strategies to address this problem. For example, in ASR, [HCC+14, GRAG09] generates more training samples by adding noise to clean audio, which can improve the robustness of ASR systems. [JH13] adapts Vocal Tract Length Normalization for data augmentation. This method is further extended to the large vocabulary continuous speech recognition

(LVCSR) in [CGK15]. [KPPK15] applies speed perturbation on raw audio for LVSCR tasks. [PCZ+19] proposes a simple but effective data augmentation method that consists of three kinds of deformations of the input log mel spectrogram. In SER, [AP17] shows augmenting the emotional dataset via speed perturbation leads to a significant improvement. [EFP+18] proposes a combination of oversampling and vocal tract length perturbation for the SER task.

Recent approaches have explored using style transfer (like EVC) models as a data augmentation strategy. These style transfer models are often based on GAN due to their ability to learn and mimic an input data distribution. In SER, [SGE18] investigates using vanilla GAN and conditional GAN to generate synthetic feature vectors. The experiments show that augmenting the training dataset with these synthetic feature vectors improves the SER performance. [CPS+19] adapts and improves a conditional GAN to generate synthetic spectrograms for the minority class in the dataset to address the data imbalance problem. [BNV19] further proposes a CycleGAN-based method that generates synthetic feature vectors through emotion style transfer to augment the data with target emotions. [LAR+20] proposes to combine the mixup data augmentation technique with a GAN to improve the generation of synthetic feature vectors and augment the training size of SER for performance improvement. [RBES20] proposes a StarGAN-based EVC model and shows the value of the generated samples for data augmentation in SER.

Compared with other data augmentation strategies, data augmentation through style transfer (like EVC) has two benefits. One benefit is that they enable more controllable and customizable data augmentation to address the data imbalance problem by synthesizing different speech styles to supplement the low-resource categories in datasets. Another benefit is that, since style transfer-based data augmentation strategies can increase the diversity of datasets by transferring components that are irrelevant to subsequent tasks, these strategies can reduce the interference of task-independent components on the model and promote the model to learn the real task-specific distribution. For example, let us imagine a scenario in which we have an emotional dataset where most of the audio samples for the category “happiness” come from male speakers. The model trained on this dataset will tend to associate those voice characteristics of males (such as low formant ranges) to the “happiness” category, which is undesired for a

generic emotion classifier. This wrong association will be effectively avoided by transferring the male voices from "happiness" samples into some female voices and expanding the training set.

5.3 Related Works

5.3.1 Feature Disentanglement

Inspired by style transfer and feature disentanglement technologies in computer vision, many studies in EVC and Expressive Speech Synthesis (ESS) tasks explore disentangling emotional features from speech to help with synthesizing expressive human-like speech. Style embedding is widely used in ESS. [SRBX+18] learns a fixed-length latent embedding of prosody which is almost independent of speaker identities. However, this method cannot transfer prosody independently of content, which means the latent embedding also contains content except for prosody. [WSZ+18] can transfer speaking styles independently of content by using global style tokens (GSTs). However, the experiments show that GSTs are not pure style embedding but contain rich speaker identity (timber) information. [ACP+20] proposed a VAE and Householder Flow-based model that can create expressive voices independent of the speaker identity (timber). However, the model's effectiveness is verified by the emotional strength conversion of only one emotion category ("exciting"), which needs further experiments on other emotions' modification. [HSTD20] notices that the style embedding transferred by ESS models contains also content information, which causes a "content leakage" problem. The same problem also occurs in EVC except for using the emotional embedding rather than the style embedding [ZSLL21, RBES20, EPL20]. [EPL20] reported the sound distortion introduced by EVC due to the impure emotional embedding transfer, which seriously damaged the content intelligibility of synthesized speech. [SRBX+18] shows that the speaker identity (timber) information is not completely preserved during prosody transfer. The MOS score for speaker similarity suffers a significant drop.

This evidence shows that either for EVC or ESS models, their disentanglement for emotion

or style is not complete. Features they transferred include not only emotional and stylistic information but also content and timber information that should be preserved. We attribute this to the overlap between emotional, timbral, and content-related information. These previous methods did not explicitly remove the interference of other irrelevant information during the disentanglement process.

5.3.2 Our Originality

Unlike the baseline StarGAN-based EVC model [RBES20], we explicitly separate emotion-independent features from emotional features during conversion. When providing the target emotion label to StarGAN, we use continuous emotional features extracted by a pre-trained emotion classifier rather than using one-hot vectors to represent each emotion. Similar continuous emotional features have also been used in [ZSLL20]. However, their model is based on VAW-GAN in the context of one-to-many emotional conversion, while our model is based on StarGAN and is able to complete the many-to-many emotional conversion. Besides, their model simply reconstructs the source spectral envelope without attempting to convert it during the whole training process. The authors only consider the emotion conversion when testing, which causes a mismatch between training and testing. Our model eliminates this mismatch by using a proposed two-stage training process.

5.4 Methodology

Since the harmonic spectral envelope (SP) and the fundamental frequency (F0) contour contain rich emotion-related features [SB13, YBL+04], we use the WORLD vocoder [MYO16] to decompose the raw audio signal into SP, F0, and aperiodicity parameters (AP) before converting. We then apply the proposed StarGAN framework to generate the SP with the target emotion. Next, we introduce the two training stages, the model architecture and the final conversion process.

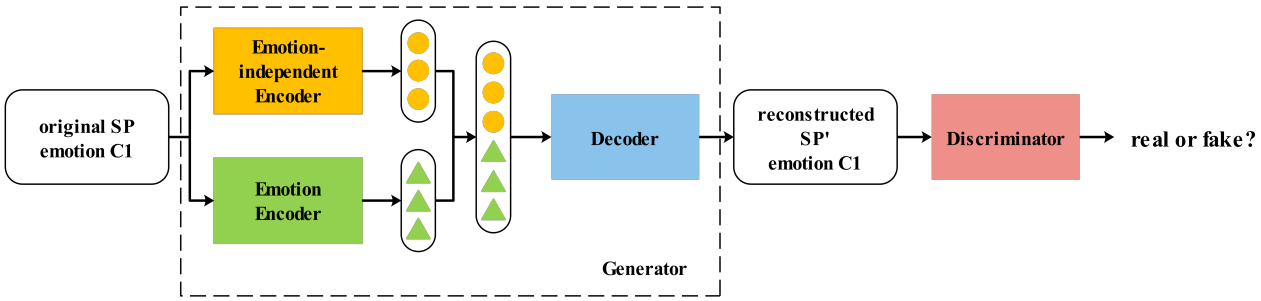


Figure 5.5: The model structure used in the first stage of training

5.4.1 Training

Training Stage1: Autoencoder training

In the first training stage, we propose to utilize an autoencoder with two encoders as the GAN generator G, which aims to learn an emotion-independent encoder capable of separating emotional features from emotion-independent features.

As shown in Figure 5.5, the autoencoder is trained by simply reconstructing the source spectral envelope SP_{C1} with emotion $C1$ without attempting to convert it, which means that the target emotion is not provided in this stage. The emotion encoder aims to encode emotional information, where its latent code $E(SP_{C1})$ is provided by a pre-trained emotion classifier (fixed during training stage 1) shown in Figure 5.6. By providing the emotion features of the source SP_{C1} as a condition, the emotion-independent encoder learns to eliminate emotional information from the input, thus making its latent code $I(SP_{C1})$ emotion independent. The decoder learns to reconstruct SP_{C1} using the concatenated latent codes, which can be formulated as:

$$SP'_{C1} = G(I(SP_{C1}), E(SP_{C1})). \tag{5.1}$$

To make the reconstructed SP'_{C1} more realistic, we train the autoencoder based on adversarial learning.

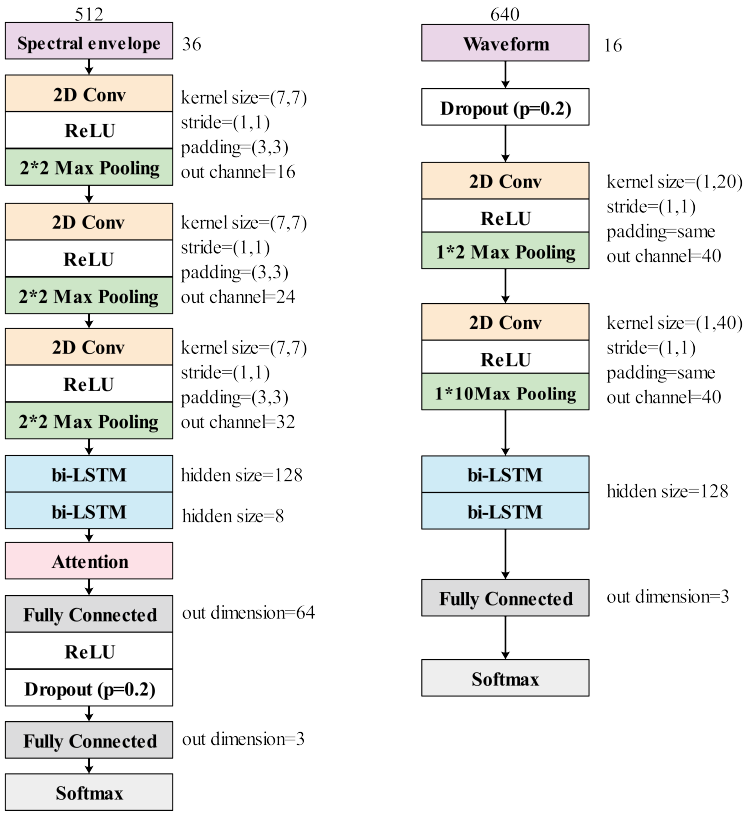


Figure 5.6: The structure of the pre-trained emotion classifier (left) used as the emotion encoder in the improved StarGAN and the SER model (right) used in the data augmentation experiments.

Training Stage2: StarGAN training

In the second training stage, the typical StarGAN training is implemented to eliminate the training-testing mismatch in [ZSLL20] and enable the many-to-many emotion conversion.

As shown in Figure 5.7, the emotion-independent encoder and the decoder are the generator of the StarGAN, denoted as G. We extract the emotion features of audio signals with the same emotion by using the pre-trained emotion classifier and average them as the feature representations of target emotion labels. We denote $\overline{E}_{C2}$ as the representation of target emotion $C2$. The generator G first reconstructs the SP'_{C2} from $\overline{E}_{C2}$ and $I(SP_{C1})$ and then reconstructs the SP''_{C1} from $\overline{E}_{C1}$ and $I(SP'_{C2})$:

$$SP'_{C2} = G(I(SP_{C1}), \overline{E}_{C2}), \tag{5.2}$$

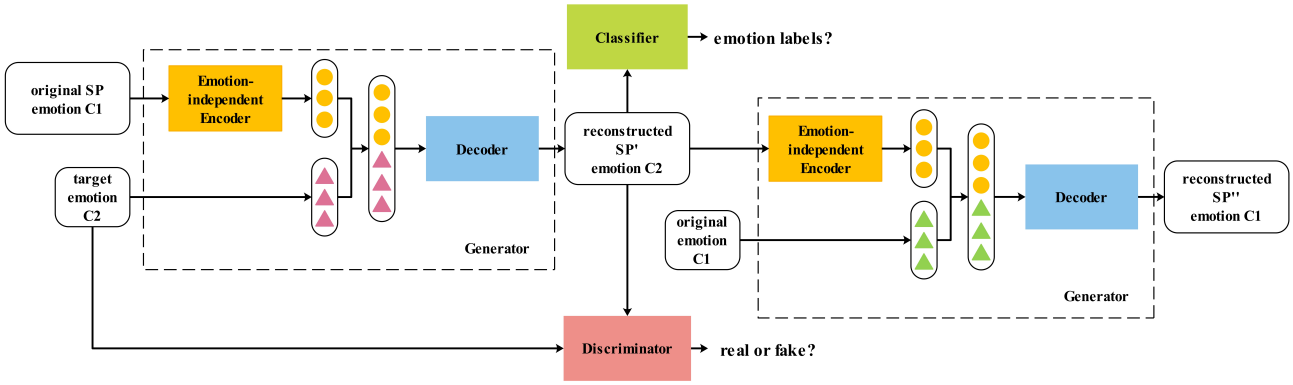


Figure 5.7: The model structure used in the second stage of training.

$$SP''_{C1} = G(I(SP'_{C2}), \overline{E}_{C1}). \tag{5.3}$$

The generator G tries to minimize the reconstruction loss and fool the discriminator D by generating a more realistic SP. The D tries to maximize the loss between the real SP and the fake SP. The domain classifier C learns to predict the emotion label and helps to enable the many-to-many emotion conversion.

5.4.2 Inference

We follow the relative logarithm Gaussian normalized transformation (LGNT) [LZY07] proposed by [RBES20] to convert the F0 contour of an audio signal from emotion $C1$ to the target emotion $C2$.

$$\begin{aligned} \log(F0') = ((\log(F0) - \mu(F0))\frac{\sigma(F0) + \Delta\sigma_{C_1,\ C_2}}{\sigma(F0)} \\ +\mu(F0) + \Delta\mu_{C_1,\ C_2}, \end{aligned} \tag{5.4}$$

where $\mu(F0)$ and $\sigma(F0)$ are the mean and variance of $\log(F0)$. $\Delta\mu_{C_1,\ C_2}$ and $\Delta\sigma_{C_1,\ C_2}$ are the average difference in the mean $\log(F0)$ and the average change in the variance of $\log(F0)$ between emotion $C1$ and $C2$.

At test time, with the spectral envelope converted by our improved StarGAN model and the F0 contour converted by the relative LGNT, we synthesize the audio signal with the target emotion via the WORLD vocoder [MYO16].

5.5 Experimental Details

5.5.1 Model Configurations

For more details of our improved StarGAN framework, we use the same discriminator and domain classifier structures as the baseline StarGAN-EVC model [RBES20]. The architecture of our emotion-independent encoder and decoder is the same as the generator in [RBES20, KKTH18]. The pre-trained classifier used in the first training stage is shown in the left part of Figure 5.6. The pre-trained classifier has the same structure as the domain classifier in the second stage. We extract 36 cepstral coefficients as an approximation to the whole spectral envelope for each training sample. In both the first and the second training stages, we set the batch size to 4 and use the Adam optimizer [KB14] with the learning rate (for G, D and C), $\beta 1$ and $\beta 2$ set to 1×10^{-4}, 0.5, and 0.999. We perform one generator update after five discriminator and domain classifier updates in the StarGAN training stage as in [GAA+17]. The first training stage lasts for 400 epochs, and we manually stop the second training stage when the loss reaches a plateau (approximately 380 epochs).

5.5.2 Datasets

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database [BBL+08] for all our experiments. This dataset has been recorded from ten actors in session 1 to session 5, including emotional scripts and improvised hypothetical scenarios. It contains approximately 12 hours of recordings with 9 emotions. The sampling rate of all recordings is 16 kHz. We utilize audio signals with three emotions, including angry, sad, and happy, from both scripted and improvised patterns, which contain 2 435 recordings in total.

Rating	Quality	Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying, but not objectionable
1	Bad	Very annoying and objectionable

Table 5.1: Explanation of MOS Scores.

5.5.3 Evaluation

To ascertain whether the separation of emotional features from emotion-independent features can indeed reduce the distortion of the generated audio signals and improve their overall quality, we perform an objective evaluation using the MCD and a subjective evaluation via the human preference test and the mean opinion scores (MOS) test in terms of distortion. To evaluate the value of the generated audio for data augmentation, we report the performance on the SER task.

Subjective Evaluation

We conduct two listening tests for our human perception experiments, the human preference test and the MOS test in audio distortion. 26 fluent English-speaking subjects participated in all experiments. Each of them was asked to evaluate 60 audio samples (30 pairs), which contain 30 audio samples generated by the three-class StarGAN-based EVC model [RBES20] (denoted as baseline StarGAN-EVC) and 30 audio samples generated by our proposed StarGAN-based EVC model (denoted as improved StarGAN-EVC). Both models were trained on the IEMOCAP dataset with the same train/validation/test split. Two audio samples of each pair were generated from one source sample in IEMOCAP to a randomly selected target emotion.

For the MOS test in terms of distortion, participants were shown a pair of samples for each question and were asked to rate each sample on the MOS scale shown in Table 5.1.

Objective Evaluation

We also perform an objective evaluation using MCD, which measures the distortion between two sequences of Mel-cepstral coefficients. It is a commonly used evaluation metric for assessing the quality of generated speech signals in VC [MSTS17, TWX+18, TWH+19, THW+20] and EVC. We calculate the MCD between the mel cepstra of reference audio signals and that of audio signals converted by both the baseline StarGAN-EVC and the improved StarGAN-EVC model for all six source-to-target emotion pairs.

Data Augmentation Experimental Design and Evaluation

We use an end-to-end SER model [ZWS19], which directly takes the waveform as its input, to perform the data augmentation experiments. The architecture is shown in the right part of Figure 5.6. We first standardized the raw audio signal by using the mean and standard deviation from the training set. The raw signal was then sampled at an interval of 40 ms with a step size of 10 ms. Given the 16 kHz sampling rate of raw signals, the waveform for each frame is of dimension 640. We randomly selected 16 continuous frames for each audio signal and got the fixed-size input waveform of dimension 16×640. Note that the SER model here is entirely different from the pre-trained SER model used in our improved StarGAN training. Also, note that the SER model here is slightly different from the model used in [RBES20] in terms of kernel sizes and filter numbers; that is why the data augmentation results we report for the baseline-StarGAN-EVC are different from the results shown in [RBES20]. Our SER model was trained using the Adam optimizer with the learning rate, $\beta 1$, and $\beta 2$ set to 1×10^{-5}, 0.5, and 0.999. The batch size was set to 4. We employ sessions 1-3 of the IEMOCAP dataset as the original training set, session 4 as the validation set, and session 5 as the testing set. We report the Micro-F1 and Macro-F1 scores to evaluate the performance of the SER model on the testing set. All results are averaged across 10 trials to reduce the effect of initialization biases.

5.5.4 Comparative Models

For the results of EVC performance in Table 5.2, Figure 5.8, and Table 5.3, the comparative models are:

1) baseline-StarGAN-EVC: the EVC model in [RBES20];

2) improved-StarGAN-EVC: the proposed EVC model.

For the results of data augmentation on the SER task in Table 5.4, the comparative models are:

3) sessions1-3: the SER model trained on the IEMOCAP dataset of sessions 1-3 (with no data augmentation);

4) sessions1-3+baseline-StarGAN: the SER model trained on IEMOCAP sessions 1-3 augmented with audio signals generated by the baseline-StarGAN-EVC model;

5) sessions1-3+improved-StarGAN: the SER model trained on IEMOCAP sessions 1-3 augmented with audio signals generated by our proposed improved-StarGAN-EVC model.

5.6 Results and Conclusions

5.6.1 EVC Performance

Subjective Evaluation Results

Table 5.2 illustrates the MOS test results in terms of distortion with 95 % confidence interval. We observe that the improved-StarGAN-EVC model outperforms the baseline-StarGAN-EVC model with a higher MOS score with no confidence interval overlap.

For the audio preference test, participants were asked to compare samples from each pair and choose the one they believe in having less perceived distortion (e. g., cracked voices or jitter).

Models	MOS
baseline-StarGAN-EVC	1.952 ± 0.121
improved-StarGAN-EVC	$\mathbf{2.355 \pm 0.084}$

Table 5.2: Results of the MOS test with 95 % confidence interval.

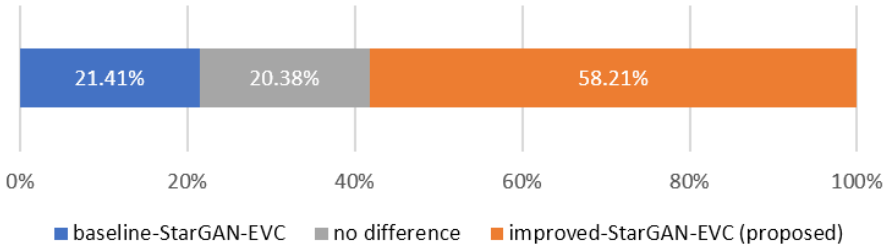


Figure 5.8: Results of audio preference test.

The audio preference test results are summarised in Figure 5.8. We observe that our proposed model obtained a preference of up to 58.21 %, which is almost three times the figure for the baseline-StarGAN-EVC model, indicating that humans clearly favored samples converted by our improved-StarGAN-EVC model.

objective Evaluation Results

Table 5.3 reports the MCD values of source-target emotion pairs and the overall MCD values for all converted audio signals regardless of emotion classes. We observe that our proposed improved-StarGAN-EVC model outperforms the baseline-StarGAN-EVC model in MCD for all terms, revealing that our proposed model can effectively reduce distortion. Besides, for the

MCD(dB)	baseline-StarGAN-EVC	improved-StarGAN-EVC
angry-happy	4.291	**4.046**
angry-sad	4.330	**4.265**
happy-angry	5.312	**4.719**
happy-sad	4.112	**3.879**
sad-angry	6.018	**4.826**
sad-happy	5.291	**3.916**
overall	4.777	**4.183**

Table 5.3: Results of the MCD for source-target emotion pairs and overall converted audio signals regardless of emotions.

Models	Micro-F1 (%)	Macro-F1 (%)
sessions1-3	59.87 (0.3889)	56.64 (0.3761)
sessions1-3+baseline-StarGAN	60.37 (0.7502)	55.41 (0.7933)
sessions1-3+improved-StarGAN	**62.62** (1.1956)	**60.34** (0.4701)

Table 5.4: Results of data augmentation experiments. Standard deviations for 10 trials are indicated in parentheses.

sad-happy and sad-angry emotion conversion, our model achieves a distortion reduction of up to 1.375 and 1.192, which indicates that the proposed feature separation method has a much better suppression effect on distortion when the targets are high-arousal emotions.

5.6.2 Data Augmentation Results

Although the audio distortion evaluation experiments show that our proposed feature separation method can effectively reduce distortion, we still need to determine whether or not this improvement in audio quality is at the expense of our model's emotional conversion performance.

Table 5.4 demonstrates the performance of SER models trained on training sets with different data augmentation. We observe that the model trained on sessions1-3+improved-StarGAN achieves higher Micro-F1 and Macro-F1 scores than the model trained only on sessions1-3, which indicates that the audio signals converted by our proposed model indeed carry effective emotional information and are valuable for data augmentation. We also note that the model trained on sessions1-3+improved-StarGAN achieves an absolute increase of 2 % in Micro-F1 and 5 % in Macro-F1 compared to the model trained on sessions1-3+baseline-StarGAN, which indicates that our model's improvement in audio quality is not at the expense of its emotional conversion performance; on the contrary, the audio signals generated by our proposed model are more consistent with the distribution of the original dataset and are more valuable for data augmentation.

5.6.3 Conclusions

We proposed an improved StarGAN-EVC framework with a two-stage training process that explicitly disentangles emotional features from emotion-independent features to reduce audio distortion and enhance emotional conversion performance. Experimental results from both objective and subjective evaluations demonstrated that the proposed model effectively reduces distortion and significantly enhances speech intelligibility, particularly for high-arousal emotions. Additionally, data augmentation experiments for end-to-end SER tasks validated that the improvement in audio quality achieved by the proposed model does not compromise its emotional conversion performance. These findings highlight the value of disentangled emotion features in enhancing the speech intelligibility of EVC models, as well as their potential for improving SER applications through high-quality data augmentation.

5.7 Contributions

The main contributions of this chapter are as follows.

- We propose a novel model structure as well as a two-stage training process based on StarGAN, which explicitly separates emotion-independent features from emotional features during emotional conversion;
- We perform both the objective evaluation using Mel-cepstral distortion (MCD) [Kub93] and the subjective evaluation in terms of distortion to ascertain whether this separation of features can indeed reduce the distortion of generated audio signals and enhance speech intelligibility;
- We replicate the data augmentation experiments of the baseline StarGAN-EVC model [RBES20] to showcase that our proposed model is more valuable for data augmentation.

Chapter 6

Achievements and Outlook

6.1 Summary of Achievements

This thesis aims to enhance speech intelligibility by enabling machines to better interpret speaker intentions in human speech and produce more comprehensible machine-generated speech, through the incorporation of paralinguistic features into speech processing systems. The main achievements of this thesis are summarized as follows.

- In response to the research questions RQ-1 and RQ-1a in Section 1.2, we proposed an unsupervised, flexible, and perceptually aligned approach for modeling prosody in speech. These modeled prosodic cues were integrated into a TTS system, achieving a prosody-aware model with strong generalizability, fine-grained prosody control, and natural prosody, enhancing intelligibility compared to existing TTS systems.
- In addressing the research question RQ-2 in Section 1.2, we proposed a novel training strategy that dynamically prioritizes valuable paralinguistic tasks during training while mitigating negative transfer, enabling a model that more effectively interprets the speaker's intentions.
- In tackling the research questions RQ-3 and RQ-3a in Section 1.2, we proposed an EVC model that incorporates disentangled emotional representations, reducing distortions in

converted speech while preserving emotional clarity and improving intelligibility. Additionally, we explored the use of converted speech for data augmentation, demonstrating its effectiveness in enhancing the performance of SER models.

6.2 Limitations and Future Perspectives

While this thesis demonstrates significant advancements in enhancing speech intelligibility through paralinguistic features, several limitations remain that open opportunities for further exploration.

6.2.1 For Chapter 3

Intonation Representation: Discrete Tokens vs. Continuous Representations

The learned intonation shape tokens in Chapter 3 provide a compact and label-free way to model phrase-final intonation. They are stable, easy to control, and effectively capture the most salient terminal contour types without requiring manual annotation. However, their discrete nature makes control relatively coarse, limiting fine adjustments within phrases, and they may be language-specific, as the same token inventory may not generalize well to tonal or typologically distant languages. In contrast, continuous, articulatory models such as qTA [POXT09, GZX+20] offer fine-grained, interpretable control through continuous parameters that represent target pitch height, slope, and rate of change, naturally accounting for cross-linguistic prosodic variation. Future work can explore a hybrid approach that combines the categorical clarity of shape tokens with the flexibility of continuous control, allowing both intuitive manipulation and precise modeling of prosody across languages.

Prominence and Rhythm Aspects of Prosody

ProsodyFM currently targets two controllable dimensions of prosody—phrasing and terminal intonation—but speech intelligibility is also shaped by prominence (perceived stress) and rhythm (isochrony, tempo variation), which we view as promising, complementary directions for future work.

For prominence, which is a pattern closely associated with phoneme-level realization, a practical next step is to mirror the intonation design in our ProsodyFM with phoneme-aligned prominence tokens. Each phoneme would be augmented with a discrete prominence token drawn from a small set of classes (e.g., "none/weak/strong"), giving the model a fine-grained and explicit handle on perceived stress that can be manipulated at inference time. The main challenge lies in obtaining reliable prominence labels for supervision. As suggested by prior work [RH07, WN07, Ros10], prominence in English can be derived automatically from continuous acoustic cues—such as F0, energy, duration, and spectral characteristics—and then mapped to a small, stable codebook of prominence levels, which can be used to train the model to realize and control prominence jointly with phrasing and intonation.

For rhythm, in contrast to intonation and prominence—which are tightly coupled to phoneme-level realization—is a more global property that governs how durations are distributed across an utterance or phrase. A promising future direction is therefore to introduce global rhythm control signals, analogous to the way we model speaker identity in ProsodyFM. Concretely, one could condition the duration modules on a learned rhythm embedding that encodes tempo and isochrony patterns at the utterance or phrase level. These global rhythm controls could be estimated from normalized duration statistics (such as inter-stress intervals or syllable timing patterns) during training and then manipulated at inference time to adjust rhythm at the utterance or phrase level.

Domain Adaptation

For real-world use, ProsodyFM must be robust across domains so that prosody modeling and control remain stable for audiobooks, conversations, and instruction-style speech. Achieving this robustness requires both adaptation and normalization. A promising first step is to factor out domain and style variability by introducing a separate domain/style indicator and training the model so that phrasing and intonation controls remain invariant while the indicator absorbs differences in speaking style and recording channel. When adapting to a new domain, the prosody control space would be kept fixed, and only the indicator embedding with a small set of adaptation layers would be tuned on the new data. In parallel, basic acoustic cues (F0, energy, duration) could be normalized per speaking style or per domain before inferring ProsodyFM controls, so that a given control signal corresponds to a comparable relative phrasing pattern or terminal intonation pattern across different speaking styles and recording conditions.

Cross-lingual Adaption to Mandarin

ProsodyFM currently focuses on English, a non-tonal language, which limits its applicability in multilingual settings. Integrating its prosodic modeling with cross-lingual TTS systems would significantly broaden its usefulness in real-world scenarios. For cross-lingual adaptation to tonal languages such as Mandarin, two main extensions are needed at the break and intonation levels. At the break level, our current intonational-phrase boundary detector relies on a pretrained PSST [RGT23] model that does not support Chinese, so we would replace it with a Mandarin-specific break-location detector: a high-accuracy Voice Activity Detection (VAD) module [YLL+24] would propose candidate break locations, and a learned classifier would filter these using both acoustic and text-side cues, serving as a drop-in replacement for PSST on Chinese speech. At the intonation level, because Mandarin expresses lexical tone in addition to sentence-level intonation, the terminal intonation patterns (questions, statements, hesitations, etc.) described in Chapter 3 will need to be modeled over a larger prosodic unit than just the last word of the intonational phrase in English [CJ11, XM12]. Concretely, we would shift conditioning from the "last word" to the last prosodic unit (e.g., a phrase or several characters),

adjust the granularity of the modeled intonation patterns accordingly, and constrain smoothing to the phrasal register only, avoiding any smoothing across syllable-level tone contours so as to preserve the micro-prosody of individual characters.

6.2.2 For Chapter 4

Task curricula vs. MAB

Curriculum learning is closely related to the proposed bandit-based task scheduler, as both seek to control the temporal ordering of tasks during training to stabilise and improve learning. In chapter 4, however, our primary goal was to investigate whether emotion-related auxiliary tasks can effectively support multi-modal DAC, so we started from a setting where task relatedness and difficulty were unknown and could not be reliably used to design a principled curriculum. We therefore cast multi-task learning as an online task-utility estimation problem and adopted a non-stationary bandit that reacts to observed improvements on DAC, rather than prescribing a fixed task order.

As future work, it would be interesting to explore curriculum learning as a complementary mechanism: curricula assume at least a partial ordering of tasks (e.g., by relatedness to DAC or by difficulty), and it is plausible that, especially for long-context dialogues where optimisation is more unstable, a well-designed curriculum could yield more stable and possibly larger gains than a purely reactive bandit schedule. A promising direction is to systematically compare fixed curricula, bandit-only schedules, and hybrid strategies that first use bandit statistics to infer an approximate task ordering and then apply bandit-style adaptation within or between curriculum stages, particularly on long-context, multi-modal DAC benchmarks.

Auxiliary-Task MTL in the Era of Speech Foundation Models

One aim of the chapter 4 is to investigate whether emotion-related auxiliary tasks are actually beneficial for DAC in a multi-modal setting, which motivates our explicit auxiliary-task MTL

formulation and the proposed bandit-based task scheduler. Looking ahead to practical DAC systems, and in light of the rapid progress in speech foundation models, explicit auxiliary-task MTL combined with a scheduler is unlikely to remain the only or even primary lever. Modern self-supervised speech encoders such as HuBERT [HBT+21], WavLM [CWC+22], and Whisper-style speech foundation models [RKX+23] already learn representations that jointly capture prosodic, phonetic, and semantic information, and they achieve strong performance on broad benchmarks like SUPERB with frozen backbones and lightweight task-specific heads. In such a regime, the marginal benefit of hand-designed auxiliary tasks may diminish as more of the relevant structure is absorbed during pre-training. Moreover, DAC is now part of standard spoken language understanding (SLU) benchmarks used to evaluate LALMs, such as SLUE-PERB [APC+24] and Dynamic-SUPERB Phase-2 [HCY+24], where current speech foundation models and LALMs already report strong DAC F1 scores—for example, Whisper-style OWSM v3.1 [PTC+24] and instruction-tuned systems such as UniverSLU [AFJ+24] perform well using only shallow heads or prompt-based interfaces.

At the same time, many speech-related tasks—especially in low-resource, highly imbalanced, or domain-specific settings—are unlikely to be fully solved "out of the box" by speech foundation encoders plus shallow heads or generic LALMs. In these regimes, one still has to decide how to adapt the backbone effectively, and this is where auxiliary-task MTL and our bandit scheduler remain relevant. Techniques such as adapter modules, prompt tuning, and representation probing primarily control how a strong backbone is specialised for each task, whereas our non-stationary bandit scheduler controls how training updates are allocated across tasks over time. These two axes are orthogonal. From a future perspective, our method is best viewed not as a competitor to strong speech/text backbones or LALMs, but as a complementary task-scheduling mechanism that can be layered on top of them, potentially enabling more sample-efficient and robust adaptation in challenging scenarios where foundation models alone are not yet sufficient.

Conversation-level Prosodic Cues

In chapter 4, lexical and emotional cues are modeled only at the utterance level, using audio/text embeddings and emotion-related auxiliary tasks, and conversation-level phenomena such as entrainment and turn-transition cues are not captured. However, prior work shows that the longer-range prosodic dynamics are closely linked to communicative functions like backchannels and floor management. For example, specific patterns — including pitch movements, intensity changes, and segmental lengthening—strongly influence backchannel likelihood [GH09, LGH11]. Similarly, turn-yield and turn-hold cues, such as final lengthening, pitch resetting, and pause timing, reliably signal floor transitions [EH05b, HELP09]. Future work should therefore integrate explicit conversation-level descriptors—such as cross-turn F0 and energy trajectories, overlap patterns, and entrainment measures—into the DAC framework. This can be achieved via dedicated conversation-level encoders or auxiliary tasks targeting turn transitions.

Domain Shift and Robustness

In chapter 4, the proposed emotion-assisted DAC model is trained on a single English corpus, without accounting for the interaction types (scripted, spontaneous), genre/platform (lab tasks, telephone, social media), or population differences (age, gender). Prior work on cross-domain DAC shows that even purely lexical models degrade under domain shift: [WLHW08] reports substantial drops when a cue-phrase–based DA tagger trained on one corpus is applied to another, while [VWvdG22] shows that transformer-based DAC models on social media require lexical normalisation, resampling of label distributions, and explicit use of dialog context to remain robust across platforms and domains. These results suggest that the gains obtained from emotion-related auxiliary tasks and bandit-based scheduling may not transfer uniformly across corpora, populations or interaction styles.

A prospective future plan to increase the generalizability and robustness of the current emotion-assisted DAC model is: (i) move from single-corpus training to cross-corpus, cross-domain fine-tuning, unifying label spaces over multiple DA datasets so that the model learns emotion and

DA patterns that generalise across settings [WLHW08]; (ii) exploit the finding that robustness improves when DAC is context-aware rather than utterance-only by encoding previous turns (text, prosody, speaker, and previously predicted DAs/emotions) [VWvdG22] and defining context-sensitive auxiliary tasks; and (iii) mitigate label imbalance and rare acts through class-balanced or smoothed sampling and by leveraging corpora where key dialogue acts (e.g., backchannels, questions, floor-taking moves) are more frequent [Kat24].

For evaluation, a standardized stress test should include training on one or more source corpora and testing on held-out target corpora, using fixed train/dev/test protocols so that results are comparable across studies. In addition, corpora should be annotated for demographic attributes, and evaluation should report not only overall macro-F1/UAR but also subgroup-wise and worst-group performance, making systematic drops for particular populations or domains explicit.

6.2.3 For Chapter 5

Limitations

The experiments in this chapter were carried out at the end of 2020 on a WORLD-based StarGAN EVC framework trained on a subset of IEMOCAP. From today's perspective (late 2025), both the baseline StarGAN-EVC and our improved StarGAN-EVC must be regarded as very weak systems: their subjective speech quality remains in the "poor" region of the MOS scale, with scores of 1.95 and 2.36 respectively (where a score around 2 corresponds to "poor, annoying but not objectionable" on the standard 5-point MOS scale). This low perceptual quality makes human listening tests for emotion classification difficult and unreliable. In contrast, modern EVC systems [PHY+25, QWL+25] routinely report MOS values around 4.0, i.e., close to natural speech quality.

Under these conditions, the experiments in this chapter should be interpreted as showing only that, on a very weak StarGAN-EVC baseline, explicit feature separation can reduce spectral distortion (MCD), slightly improve MOS, and generate speech signals that are useful for improv-

ing SER performance. The main contribution of this chapter therefore lies not in providing a practically competitive EVC system, but in its methodology and conceptual insight: it demonstrates that explicit disentanglement of style features is a promising direction for improving style transfer models. In practice, the most realistic use of this system is as a controlled generator for SER data augmentation rather than as a deployable EVC model. This idea has since been taken up and extended by later style transfer systems and data augmentation studies that explicitly cite our paper [HCRS21] when targeting synthetic data augmentation [IPL24] and higher-quality style transfer system [RWM+22].

There are two main technical reasons for the poor absolute quality of our system. First, StarGAN training is highly unstable and capacity-limited, which constrains the expressive power of the generator and makes it difficult to capture fine-grained acoustic details. Second, our feature disentanglement scheme likely suffers from information leakage. The inductive biases [LBL+19] in this scheme include an architectural bias in the form of an auto-encoder–style information bottleneck, and an invariance bias, where the final emotion embedding is obtained by averaging embeddings from multiple utterances with different speakers and linguistic content but the same emotion label, so that speaker- and content-specific components are encouraged to be washed out by averaging. However, the capacity of the information bottleneck is not carefully controlled, and we do not impose sufficiently strong inductive biases to explicitly eliminate speaker- and content-related information in the final emotion embeddings; as a result, speaker identity, phonetic content, and other nuisance factors likely remain partially entangled with emotion and can still be altered during conversion.

Future Perspective: Performance Improvement

From today's perspective (late 2025), several lines of work point to promising directions for improving the proposed EVC system. On the one hand, advanced generative paradigms such as diffusion and flow matching models [PLW+24, SYYC25, PHY+25] can replace the StarGAN-based backbone in this chapter; their likelihood- or regression-based training, instead of a purely adversarial min–max game, largely alleviates mode collapse and instability and should

provide better coverage of the target emotional speech distribution. On the other hand, stronger inductive biases can be introduced to enhance feature disentanglement. [WDY+21, YTZ+22, CXC+22, CLS+24] add a stronger invariance bias to the autoencoder framework by minimizing mutual information between style-related and style-independent features, thereby discouraging shared information and reducing leakage between the two latent spaces. [CYC+23] introduces a stronger architectural bias by using separate emotion and content encoders, an attention mechanism that highlights emotion-salient frames, and a loss that penalizes similarity between emotion and content features on those frames, thereby encouraging a more emotion-specific latent space.

Future Perspective: Zero-shot EVC vs. Fixed-class EVC

In this chapter, the proposed EVC framework is trained on a fixed set of discrete emotion classes. Recent advances in EVC, however, have demonstrated zero-shot settings that offer much greater flexibility and scalability. In zero-shot EVC [CLS+24, PHY+25, SYYC25] the target emotion style is specified dynamically, either by providing a reference utterance whose emotion is to be imitated or by supplying textual instructions that describe the desired emotion, so a single model can perform any-to-any conversion and naturally generalize to unseen emotion styles without retraining. By contrast, traditional fixed-class EVC realizes at best many-to-many conversion within a predefined emotion inventory; extending it to novel emotions would typically require collecting new labeled data and training (or at least fine-tuning) a separate model, which limits practicality when users demand fine-grained, personalized, or previously unseen emotional styles. Given these advantages in flexibility and generalization, zero-shot EVC represents a particularly promising direction for future work.

Future Perspective: Prosody-aware EVC

Prosody and emotion are tightly related but live at different levels: prosody is a relatively low-dimensional, fundamental aspect of speech, grounded in clear acoustic correlates such as F0, duration, speech rate, and intensity, while emotion is a higher-dimensional construct that

is largely realized through systematic configurations of these prosodic components. Empirical studies consistently show that high-arousal emotions (e.g., anger, joy) are associated with higher mean F0, larger F0 range, faster speech rate and higher intensity, whereas low-arousal emotions such as sadness exhibit lower F0, slower speech rate, longer pauses and reduced intensity [BS05, JL03, SCDC+01]. Emotion transfer is therefore intrinsically difficult—labeled emotional corpora are limited, labels are noisy and subjective, and the underlying affect space is high-dimensional—whereas transferring and controlling prosodic components is comparatively easier, because these components map directly to measurable physical quantities (pitch trajectories, segmental durations, energy contours), for which reliable automatic labels and even unsupervised modeling are feasible. In this view, converting the emotion of a speech signal can be framed as converting the underlying configuration of basic prosodic components. Progress in prosodic modeling—such as the explicit modeling of phrase breaks and terminal contours in chapter 3—can be directly propagated to EVC systems, and building strong prosodic models and interfaces for prosody control should be regarded as a necessary foundation for robust and scalable affect modeling in future EVC.

Future Perspective: A Balanced Evaluation Protocol for EVC

As discussed in the previous chapter of limitations, the proposed StarGAN-based EVC system still suffers from poor speech quality, so its most realistic use is as a controllable generator for SER data augmentation rather than as a deployable EVC model. As this field evolves rapidly, a balanced evaluation protocol for deployable EVC systems should therefore consider not only audio quality and the success of emotional transfer, but also the degree to which factors that ought to be preserved—most notably linguistic content and speaker identity—remain stable during conversion. Concretely, such an evaluation should cover at least four aspects: (i) speech quality, assessed with standard objective distortion measures (e.g., MCD) and subjective listening tests (e.g., MOS); (ii) emotion transfer, measured using appropriate objective similarity metrics [SYYC25] and subjective tests such as perceptual emotion identification [SYYC25]; (iii) content preservation, evaluated with ASR-based measures such as WER (or related metrics) and subjective tests such as perceptual content identification [CXC+22]; and (iv) speaker

preservation, assessed by objective speaker similarity scores [SYYC25] and subjective tests such as perceptual speaker similarity [YTZ+22]. Beyond these four aspects, a more advanced evaluation protocol can further incorporate overall intelligibility [UDL+25] and prosody naturalness/similarity [BCM+23, UDL+25], to capture how well the converted speech functions as fluent, context-appropriate communication.

6.2.4 Future Perspective: Unified Representations and Systems for Both Speech Understanding and Synthesis

In this thesis, three separate models are proposed for TTS, DAC, and EVC, each with its own tailored representations for prosody, dialogue act, and emotion. Looking ahead, a promising direction is to move from such task-specific systems toward unified large audio language models (LALMs) [DMO+24, DJL+25, GGK+25, WYH+25, XGH+25], in which a single model jointly supports synthesis (TTS/EVC) and understanding (including DAC) by using audio codec tokens as a shared representation. In the standard codec/encoder → adapter → LLM → decoder framework of LALMs, raw waveforms are first mapped by a neural audio codec (or continuous encoder) into a compact sequence of audio tokens that preserve both linguistic and paralinguistic cues. These tokens are then projected via a lightweight adapter into the embedding space of a large language model, which is trained on massive amounts of paired audio data for a broad spectrum of understanding tasks (DAC, SER, speaker identification, music classification, etc) and generation tasks (TTS, EVC, VC, audio question answering, music generation, etc). On the output side, the same class of audio tokens is predicted by the LLM for synthesis-oriented tasks and deterministically rendered back to waveform by the codec decoder. As a result, LALMs exhibit a rich suite of abilities—accurate audio understanding, high-fidelity speech synthesis and conversion, instruction following over mixed text–audio inputs, long-context conversational memory, and non-trivial reasoning over acoustic and linguistic information—all grounded in this shared audio token representation.

6.3 Ethical Considerations

The research topics explored in this thesis, including prosody-aware TTS, emotion-aided DAC, and EVC, hold significant potential to advance and enhance the efficiency of human-computer communication. However, the development, application, and implications of these proposed models raise several ethical considerations that must be carefully addressed.

Firstly, the development of the proposed models relies on datasets for training and evaluation, which may include sensitive or personal information, such as speech recordings and emotional annotations. This may raise ethical concerns regarding data privacy. To address these concerns, the datasets used in this thesis (VCTK [YVM19], LibriTTS [ZDC+19], EMOTyDA [SPSB20], IEMOCAP [BBL+08]) are all publicly available and obtained under official and legitimate licenses. Additionally, all personal information within these datasets is anonymized to ensure the protection of individual privacy.

However, anonymization alone may not provide sufficient protection as adversarial attacks on AI systems become more sophisticated. For example, emerging technologies such as re-identification techniques have made it possible for attackers to uncover hidden identity information from anonymized speech data. To mitigate such risks, future research should explore more powerful privacy protection technologies, such as differential privacy and federated learning. At the same time, strict ethical guidelines for dataset usage need to be developed and enforced. While this thesis uses publicly available and officially licensed datasets, future research involving new datasets must ensure that data contributors provide informed consent. This means participants must be clearly informed of how their data will be used, the scope of its usage, and any associated privacy risks.

Secondly, paralinguistic features, such as intonation, phrasing, and emotional expressions, exhibit significant variation across individuals with different cultural, linguistic, and gender backgrounds. This diversity poses challenges in ensuring fairness and inclusivity in the evaluation of TTS and EVC models, as these processes inevitably involve subjective human perception. To address potential issues of bias and fairness, this thesis incorporates fairness-focused subjective

evaluation metrics. These metrics are carefully designed to ensure that models are evaluated according to consistent and standardized criteria, minimizing the impact of subjective variability.

Beyond the preliminary step, future research may focus on developing more culturally and linguistically diverse evaluation methods. Researchers could employ interpretable machine learning techniques to identify and understand subtle, latent factors that contribute to biased judgments of speech. Furthermore, by incorporating findings on how individuals from diverse cultural and linguistic backgrounds perceive prosodic and paralinguistic cues, evaluation criteria could be designed to dynamically adapt over time, thereby reducing biases and improving the overall fairness of evaluation procedures.

Thirdly, while the proposed models offer significant benefits for human-machine interaction, they also pose potential risks of misuse. For example, a prosody-aware TTS system capable of synthesizing speech with human-like prosody could be exploited for malicious purposes, such as impersonation, fraud, or the creation of deepfake audio. To mitigate these risks, this thesis adopts a balanced approach: fundamental technical code is shared openly to maintain transparency, while access to specific model parameters is restricted exclusively for academic purposes. This is enforced through a certification process that explicitly prohibits any applications beyond academic research, ensuring the technology is employed responsibly and for the benefit of research and development.

Additionally, future work could focus on incorporating techniques like automated watermarking into generated speech, which would help detect fake or manipulated audio. Alongside these technical measures, researchers could collaborate closely with policymakers, industry leaders, and ethics committees to establish clear standards and guidelines that not only define the scope of the usage of these technologies but also outline effective safeguards against misuse. By integrating technical defenses and clear policies, the community can foster a responsible and trustworthy environment for the continued development of speech-based AI systems.

Bibliography

[ABB+91] Anne H. Anderson, Miles Bader, Ellen G. Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline C. Kowtko, Jan McAllister, Jim Miller, Cathy Sotillo, Henry S. Thompson, and Regina Weinert. *The HCRC Map Task Corpus*, volume 34. Language and Speech, 1991.

[ACP+20] Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba, and Roberto Barra-Chicote. Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6179–6183. IEEE, 2020.

[AFJ+24] Siddhant Arora, Hayato Futami, Jee-weon Jung, Yifan Peng, Roshan Sharma, Yosuke Kashiwagi, Emiru Tsunoo, Karen Livescu, and Shinji Watanabe. Universlu: Universal spoken language understanding for diverse tasks with natural language instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2754–2774, 2024.

[AMM+22] Syed Ammar Abbas, Thomas Merritt, Alexis Moinet, Sri Karlapati, Ewa Muszynska, Simon Slangen, Elia Gatti, and Thomas Drugman. Expressive, variable, and controllable duration modelling in TTS. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022*, pages 4546–4550. ISCA, 2022.

[AP17] Zakaria Aldeneh and Emily Mower Provost. Using regional saliency for speech

emotion recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2741–2745. IEEE, 2017.

[APC+24] Siddhant Arora, Ankita Pasad, Chung-Ming Chien, Jionghao Han, Roshan Sharma, Jee-weon Jung, Hira Dhamyal, William Chen, Suwon Shon, Hung-Yi Lee, et al. On the evaluation of speech foundation models for spoken language understanding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11923–11938, 2024.

[BAC+12] Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Fang, Koiti Hasida, Kyung-Tae Lee, Volha Petukhova, Andrei Popescu-Belis, and Laurent Romary. Iso 24617-2: A semantically-based standard for dialogue annotation. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 430–437, 2012.

[Bar06] Lisa Feldman Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46, 2006.

[Bar17] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

[BAS19] Alice Baird, Shahin Amiriparian, and Björn Schuller. Can deep generative audio be emotional? towards an approach for personalised emotional audio generation. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019.

[Bat24] Anton Batliner. Paralinguistics. https://gepf.falar.org/entries/63, 2024. Accessed: 2025-01-20.

[BBL+08] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[BCF+07] David I Beaver, Brady Clark, Edward Stanton Flemming, T Florian Jaeger, and Maria Wolters. When semantics meets phonetics: Acoustical studies of second-occurrence focus. *Language*, 83(2):245–276, 2007.

[BCM+23] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

[BDKG12] Mara Breen, Laura C Dilley, John Kraemer, and Edward Gibson. Inter-transcriber reliability for two systems of prosodic annotation: Tobi (tones and break indices) and rap (rhythm and pitch). *Corpus linguistics and linguistic theory*, 8(2):277–312, 2012.

[BGC+19] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International conference on machine learning*, pages 573–582. PMLR, 2019.

[BGH+11] Kristy Boyer, Joseph F Grafsgaard, Eun Young Ha, Robert Phillips, and James Lester. An affect-enriched dialogue act classification model for task-oriented dialogue. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1190–1199, 2011.

[BHA+17] Sarah A. Bibyk, Willemijn F. L. Heeren, Allison Albin, Walt Rankinen, Anthony P. Salverda, Christine Gunlogson, and Michael K. Tanenhaus. Boundary tone processing during online comprehension. Poster presented at the 30th Annual CUNY Conference on Human Sentence Processing, MIT, Cambridge, MA, March 2017.

[BHS20] Anton Batliner, Simone Hantke, and Björn Schuller. Ethics and good practice in computational paralinguistics. *IEEE Transactions on Affective Computing*, 13(3):1236–1253, 2020.

[BHTW18] Rianne van den Berg, Leonard Hasenclever, Jakub M Tomczak, and Max Welling. Sylvester normalizing flows for variational inference. *arXiv preprint arXiv:1803.05649*, 2018.

[BKMS22] Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 23(77):1–40, 2022.

[BNB+23] Anton Batliner, Michael Neumann, Felix Burkhardt, Alice Baird, Sarina Meyer, Ngoc Thang Vu, and Björn W Schuller. Ethical awareness in paralinguistics: A taxonomy of applications. *International Journal of Human–Computer Interaction*, 39(9):1904–1921, 2023.

[BNV19] Fang Bao, Michael Neumann, and Ngoc Thang Vu. Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In *INTERSPEECH*, pages 2828–2832, 2019.

[Boe01] Paul Boersma. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

[Bol98] Dwight Bolinger. Intonation in american english. *Intonation systems: A survey of twenty languages*, pages 45–55, 1998.

[BS05] Tanja Bänziger and Klaus R Scherer. The role of intonation in emotional expressions. *Speech communication*, 46(3-4):252–267, 2005.

[BZ20] Peter Birkholz and Xinyu Zhang. Accounting for microprosody in modeling intonation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8099–8103. IEEE, 2020.

[CA97] Mark G. Core and James F. Allen. Draft of DAMSL: Dialogue act markup in several layers. Unpublished manuscript, 1997.

[CAB+05] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maarten Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa

Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iskra Mclean, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: A pre-announcement. *Machine Learning for Multimodal Interaction*, 3869:28–39, 2005.

[Cah90] Janet E Cahn. The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8(1):1–1, 1990.

[Cal07] Sasha Calhoun. *Information structure and the prosodic structure of English: A probabilistic relationship*. PhD thesis, The University of Edinburgh, 2007.

[Can15] Francesco Cangemi. mausmooth. `https://ifl.phil-fak.uni-koeln.de/sites/linguistik/Phonetik/pdf-publications/2015/cangemi2015mausmooth.pdf`, 2015. Accessed: 2024-12-01.

[Car98] Rich Caruana. Multitask learning. In Sebastian Thrun and Lorien Y. Pratt, editors, *Learning to Learn*, pages 95–133. Springer, 1998.

[CC19] Eleanor Chodroff and Jennifer S. Cole. Testing the distinctiveness of intonational tunes: Evidence from imitative productions in american english. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019*, pages 1966–1970. ISCA, 2019.

[CCK+18] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[CGK15] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1469–1477, 2015.

[CH08] Arturo Camacho and John G Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3):1638–1652, 2008.

[CHBM24] Junjie Chen, Xiangheng He, Danushka Bollegala, and Yusuke Miyao. Unsupervised parsing by searching for frequent word sequences among sentences with equivalent predicate-argument structures. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3760–3772, 2024.

[CHJC+06] Ken Chen, Mark Hasegawa-Johnson, Aaron Cohen, Sarah Borys, Sung-Suk Kim, Jennifer Cole, and Jeung-Yoon Choi. Prosody dependent speech recognition on radio news corpus of american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):232–245, 2006.

[CHM22] Junjie Chen, Xiangheng He, and Yusuke Miyao. Modeling syntactic-semantic dependency correlations in semantic role labeling using mixture models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7959–7969, 2022.

[CHM24] Junjie Chen, Xiangheng He, and Yusuke Miyao. Language model based unsupervised dependency parsing with conditional mutual information and grammatical constraints. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6355–6366, 2024.

[CHMB25] Junjie Chen, Xiangheng He, Yusuke Miyao, and Danushka Bollegala. Improving unsupervised constituency parsing via maximizing semantic information. In *The Thirteenth International Conference on Learning Representations*, 2025.

[CJ11] Ping Jiang Aishu Chen and A Jiang. Representation of mandarin intonations: boundary tone revisited. In *Proc. 23rd North American Conference on Chinese Linguistics*, volume 1, pages 97–109, 2011.

[CLLW19] Yifang Chen, Chung-Wei Lee, Haipeng Luo, and Chen-Yu Wei. A new algorithm for non-stationary contextual bandits: Efficient, optimal and parameter-free. In *Conference on Learning Theory*, pages 696–726. PMLR, 2019.

[CLS+24] Hsing-Hang Chou, Yun-Shao Lin, Ching-Chin Sung, Yu Tsao, and Chi-Chun Lee. Zsdevc: Zero-shot diffusion-based emotional voice conversion with disentangled mechanism. *arXiv preprint arXiv:2409.03636*, 2024.

[CMS+23] Feng-Ju Chang, Thejaswi Muniyappa, Kanthashree Mysore Sathyendra, Kai Wei, Grant P Strimel, and Ross McGowan. Dialog act guided contextual adapter for personalized speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[CNM+25] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271, 2025.

[CPS+19] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan. Data augmentation using gans for speech emotion recognition. In *Interspeech*, pages 171–175, 2019.

[CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[Cry11] David Crystal. *A dictionary of linguistics and phonetics*. John Wiley & Sons, 2011.

[CST22] Jennifer Cole, Jeremy Steffman, and Sam Tilsen. Shape matters: Machine classification and listeners' perceptual discrimination of american english intonational tunes. In *Proceedings of the International Conference on Speech Prosody 2022*,

Proceedings of the International Conference on Speech Prosody 2022, pages 297–301. ISCA, May 2022.

[CWC+22] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[CWS+22] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

[CXC+22] Xunquan Chen, Xuexin Xu, Jinhui Chen, Zhizhong Zhang, Tetsuya Takiguchi, and Edwin R Hancock. Speaker-independent emotional voice conversion via disentangled representations. *IEEE Transactions on Multimedia*, 25:7480–7493, 2022.

[CYC+12] Sin-Horng Chen, Jyh-Her Yang, Chen-Yu Chiang, Ming-Chieh Liu, and Yih-Ru Wang. A new prosody-assisted mandarin asr system. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1669–1683, 2012.

[CYC+23] Yun Chen, Lingxiao Yang, Qi Chen, Jian-Huang Lai, and Xiaohua Xie. Attention-based interactive disentangling network for instance-level emotional voice conversion. In *Proc. Interspeech 2023*, pages 2068–2072, 2023.

[CYLL18] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. *arXiv preprint arXiv:1804.02812*, 2018.

[Dil05] Laura Christine Dilley. *The phonetics and phonology of tonal systems*. PhD thesis, Massachusetts Institute of Technology, 2005.

[Dil10] Laura C Dilley. Pitch range variation in english tonal contrasts: Continuous or categorical? *Phonetica*, 67(1-2):63–81, 2010.

[DJL+25] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

[DMO+24] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

[DWC+24] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[DZMS13] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 511–516. IEEE, 2013.

[EFP+18] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. CNN+LSTM architecture for speech emotion recognition with data augmentation. In Venkata Subramanian Viraraghavan and Suryakanth V. Gangashetty, editors, *2018 Workshop on Speech, Music and Mind, SMM 2018, Hyderabad, India, September 1, 2018*. ISCA, 2018.

[EH05a] Jens Edlund and Mattias Heldner. Exploring prosody in interaction control. *Phonetica*, 62(2-4):215–226, 2005.

[EH05b] Jens Edlund and Mattias Heldner. Exploring prosody in interaction control. *Phonetica*, 62(2–4):215–226, 2005.

[Ekm92] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[EPL20] Mohamed Elgaar, Jungbae Park, and Sang Wan Lee. Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational au-

toencoder. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7769–7773. IEEE, 2020.

[EWT+24] Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 682–689. IEEE, 2024.

[FGEB21] Mauajama Firdaus, Hitesh Golchha, Asif Ekbal, and Pushpak Bhattacharyya. A deep multi-task model for dialogue act classification, intent detection and slot filling. *Cognitive Computation*, 13:626–645, 2021.

[FPYT21] Kosuke Futamata, Byeongseon Park, Ryuichi Yamamoto, and Kentaro Tachibana. Phrase break prediction with bidirectional encoder representations in japanese text-to-speech synthesis. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*, pages 3126–3130. ISCA, 2021.

[Fra85] Clive Frankish. Modality-specific grouping effects in short-term memory. *Journal of Memory and Language*, 24(2):200–209, 1985.

[GAA+17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[GDM+24] Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11121–11125. IEEE, 2024.

[GGK+25] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

[GH09] Agustín Gravano and Julia Hirschberg. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of Interspeech 2009*, pages 1019–1022, Brighton, United Kingdom, 2009.

[GH11] Agustín Gravano and Julia Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634, 2011.

[GJD87] Cynthia Grover, Donald G Jamieson, and Michael B Dobrovolsky. Intonation in english, french and german: perception and production. *Language and Speech*, 30(3):277–295, 1987.

[GP10] Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1(1):68–99, 2010.

[GPAM+14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[GPB19] Han Guo, Ramakanth Pasunuru, and Mohit Bansal. Autosem: Automatic task selection and mixing in multi-task learning. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3520–3531. Association for Computational Linguistics, 2019.

[Gra18] David Graddol. Discourse specific pitch behaviour. In *Intonation in discourse*, pages 221–238. Routledge, 2018.

[GRAG09] Mark JF Gales, Anton Ragni, H AlDamarki, and C Gautier. Support vector machines for noise robust asr. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 205–210. IEEE, 2009.

[GZX+20] Yingming Gao, Xinyu Zhang, Yi Xu, Jinsong Zhang, and Peter Birkholz. An investigation of the target approximation model for tone modeling and recognition

in continuous mandarin speech. In *Proceedings of the annual conference of the International Speech Communication Association, INTERSPEECH*, volume 2020, pages 1913–1917. International Speech Communication Association (ISCA), 2020.

[HBGT15] Willemijn F. L. Heeren, Sarah A. Bibyk, Christine Gunlogson, and Michael K. Tanenhaus. Asking or telling—real-time processing of prosodically distinguished questions and statements. *Language and Speech*, 58(4):474–501, 2015.

[HBT+21] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[HCC+14] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[HCRS21] Xiangheng He, Junjie Chen, Georgios Rizos, and Björn W Schuller. An improved stargan for emotional voice conversion: Enhancing voice quality and data augmentation. In *Proc. Interspeech 2021*, pages 821–825, 2021.

[HCS24] Xiangheng He, Junjie Chen, and Björn W Schuller. Task selection and assignment for multi-modal multi-task dialogue act classification with non-stationary multi-armed bandits. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12091–12095. IEEE, 2024.

[HCY+24] Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *arXiv preprint arXiv:2411.05361*, 2024.

[HCZS25] Xiangheng He, Junjie Chen, Zixing Zhang, and Björn Schuller. Prosodyfm: Unsupervised phrasing and intonation control for intelligible speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24041–24049, 2025.

[HdL21] Daniel J Hirst and Céline de Looze. Measuring speech. fundamental frequency and pitch. *Cambridge Handbook of Phonetics*, 1:336–361, 2021.

[HELP09] Mattias Heldner, Jens Edlund, Kornel Laskowski, and Antoine Pelcé. Prosodic features in the vicinity of pauses, gaps and overlaps. In Martti Vainio, Reijo Aulanko, and Olli Aaltonen, editors, *Nordic Prosody: Proceedings of the Xth Conference, Helsinki 2008*, pages 95–106. Peter Lang, Frankfurt am Main, 2009.

[HG14] Kara Hawthorne and LouAnn Gerken. From pauses to clauses: Prosody facilitates learning of syntactic constituency. *Cognition*, 133(2):420–428, 2014.

[HHW+16] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–6. IEEE, 2016.

[HIM+14] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939, 2014.

[HLH+20] Wen-Chin Huang, Hao Luo, Hsin-Te Hwang, Chen-Chou Lo, Yu-Huai Peng, Yu Tsao, and Hsin-Min Wang. Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(4):468–479, 2020.

[HLL23] Ji-Sang Hwang, Sang-Hoon Lee, and Seong-Whan Lee. Pausespeech: Natural speech synthesis via pre-trained language model and pause-based prosody modeling. In *Asian Conference on Pattern Recognition*, pages 415–427. Springer, 2023.

[HRL+22] Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. *Advances in Neural Information Processing Systems*, 35:10970–10983, 2022.

[HRS08] Jinni Harrigan, Robert Rosenthal, and Klaus Scherer. *New handbook of methods in nonverbal behavior research*. Oxford University Press, 2008.

[HSN06] Ryuichiro Higashinaka, Katsuhito Sudoh, and Mikio Nakano. Incorporating discourse features into confidence scoring of intention recognition results in spoken dialogue systems. *Speech Communication*, 48(3-4):417–436, 2006.

[HSTD20] Ting-Yao Hu, Ashish Shrivastava, Oncel Tuzel, and Chandra Dhir. Unsupervised style and content separation by minimizing mutual information for speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3267–3271. IEEE, 2020.

[HSW+22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022.

[HTK+22] Xiangheng He, Andreas Triantafyllopoulos, Alexander Kathan, Manuel Milling, Tianhao Yan, Srividya Tirunellai Rajamani, Ludwig Küster, Mathias Harrer, Elena Heber, Inga Grossmann, et al. Depression diagnosis and forecast based on mobile phone sensor data. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4679–4682. IEEE, 2022.

[HVG91] Dik J Hermes and Joost C Van Gestel. The frequency scale of speech intonation. *The Journal of the Acoustical Society of America*, 90(1):97–102, 1991.

[HWL+19] Wen-Chin Huang, Yi-Chiao Wu, Chen-Chou Lo, Patrick Lumban Tobing, Tomoki Hayashi, Kazuhiro Kobayashi, Tomoki Toda, Yu Tsao, and Hsin-Min Wang. Investigation of f0 conditioning and fully convolutional networks in variational autoencoder based voice conversion. *arXiv preprint arXiv:1905.00615*, 2019.

[HZS17] Simone Hantke, Zixing Zhang, and Björn W Schuller. Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world. In *INTERSPEECH*, pages 3951–3955, 2017.

[IPL24] Karim M Ibrahim, Antony Perzo, and Simon Leglaive. Towards improving speech emotion recognition using synthetic data augmentation from emotion conversion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10636–10640. IEEE, 2024.

[JH13] Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, volume 117, page 21, 2013.

[JL03] Patrik N Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, 129(5):770, 2003.

[Kat24] Gloria Ashiya Katuka. *Investigating Automatic Dialogue Act Classification in Collaborative Learning through Federated Transfer Learning and Cross-Corpora Domain Adaptation.* PhD thesis, University of Florida, Gainesville, FL, USA, May 2024.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KGN16] Hamed Khanpour, Nishitha Guntakandla, and Rodney Nielsen. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2012–2021, 2016.

[KHK+22] Alexander Kathan, Mathias Harrer, Ludwig Küster, Andreas Triantafyllopoulos, Xiangheng He, Manuel Milling, Maurice Gerczuk, Tianhao Yan, Srividya Tirunellai Rajamani, Elena Heber, et al. Personalised depression forecasting using mobile sensor data and ecological momentary assessment. *Frontiers in digital health*, 4:964582, 2022.

[KKB20] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[KKKY20] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

[KKS21] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[KKTH18] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273. IEEE, 2018.

[KPPK15] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Sixteenth annual conference of the international speech communication association*, 2015.

[KTH+22] Alexander Kathan, Andreas Triantafyllopoulos, Xiangheng He, Manuel Milling, Tianhao Yan, Srividya Tirunellai Rajamani, Ludwig Küster, Mathias Harrer, Elena Heber, Inga Grossmann, et al. Journaling data for daily phq-2 depression prediction and forecasting. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2627–2630. IEEE, 2022.

[Kub93] Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128. IEEE, 1993.

[LAR+20] Siddique Latif, Muhammad Asim, Rajib Rana, Sara Khalifa, Raja Jurdak, and Björn W Schuller. Augmenting generative adversarial networks for speech emotion recognition. *arXiv preprint arXiv:2005.08447*, 2020.

[Lav94] John Laver. Voice quality and identity. *Annual Review of Applied Linguistics*, 15:12–23, 1994.

[LBL+19] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[LCB+23] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023.

[LCTA19] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1535–1548, 2019.

[Leo73] Laurence B. Leonard. The role of intonation in the recall of various linguistic stimuli. *Language and Speech*, 16(4):327–335, 1973.

[Lev93] Willem JM Levelt. *Speaking: From intention to articulation*. MIT press, 1993.

[LGH11] Rivka Levitan, Agustín Gravano, and Julia Hirschberg. Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*, pages 113–117, Portland, Oregon, USA, 2011. Association for Computational Linguistics.

[LHL19] Andy T. Liu, Po-Chun Hsu, and Hung-yi Lee. Unsupervised end-to-end learning of discrete linguistic units for voice conversion. In Gernot Kubin and Zdravko Kacic, editors, *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 1108–1112. ISCA, 2019.

[LHM22] Yinghao Aaron Li, Cong Han, and Nima Mesgarani. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *arXiv preprint arXiv:2205.15439*, 2022.

[LHR+24] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Lib75] Mark Yoffe Liberman. *The intonational system of English.* PhD thesis, Massachusetts Institute of Technology, 1975.

[LJD19] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

[LK19] Younggun Lee and Taesu Kim. Robust and fine-grained prosody control of end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5911–5915. IEEE, 2019.

[LLG19] Shengchao Liu, Yingyu Liang, and Anthony Gitter. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9977–9978, 2019.

[LMS00] D Robert Ladd, Ineke Mennen, and Astrid Schepman. Phonological conditioning of peak alignment in rising pitch accents in dutch. *The Journal of the Acoustical Society of America*, 107(5):2685–2696, 2000.

[LR18] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[LYH16] Giwoong Lee, Eunho Yang, and Sung Hwang. Asymmetric multi-task learning based on task relatedness and loss. In *International conference on machine learning*, pages 230–238. PMLR, 2016.

[LZS+18] Songxiang Liu, Jinghua Zhong, Lifa Sun, Xixin Wu, Xunying Liu, and Helen Meng. Voice conversion across arbitrary speakers based on a single target-speaker utterance. In *Interspeech*, pages 496–500, 2018.

[LZY07] Kun Liu, Jianping Zhang, and Yonghong Yan. High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin. In *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, volume 4, pages 410–414. IEEE, 2007.

[MBS+20] Silvan Mertes, Alice Baird, Dominik Schiller, Björn W Schuller, and Elisabeth André. An evolutionary-based generative approach for audio data augmentation. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2020.

[MD14] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 659–663. IEEE, 2014.

[MDBS12] Sven L Mattys, Matthew H Davis, Ann R Bradlow, and Sophie K Scott. Speech recognition in adverse conditions: A review. *Language and Cognitive processes*, 27(7-8):953–978, 2012.

[MKK09] Masanori Morise, Hideki Kawahara, and Haruhiro Katayose. Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

[MLYH21] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*, pages 7748–7759. PMLR, 2021.

[Mor17] Masanori Morise. Harvest: A high-performance fundamental frequency estimator from speech signals. In *18th Annual Conference of the International Speech Communication Association, Interspeech 2017*, pages 2321–2325. ISCA, 2017.

[MPSW23] Soumi Maiti, Yifan Peng, Takaaki Saeki, and Shinji Watanabe. Speechlmscore: Evaluating speech generation using speech language model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[MR03] Diego H. Milone and Antonio J. Rubio. Prosodic and accentual information for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 11(4):321–333, 2003.

[MSM+17] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017.

[MSTS17] Hiroyuki Miyoshi, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Voice conversion using sequence-to-sequence learning of context posterior probabilities. *arXiv preprint arXiv:1704.02360*, 2017.

[MTB+24] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11341–11345. IEEE, 2024.

[MYO16] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[N+97] Sieb Nooteboom et al. The prosody of speech: melody and rhythm. *The handbook of phonetic sciences*, 5:640–673, 1997.

[NHAC+22] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874. ACL, May 2022.

[NR03] Thierry Nazzi and Franck Ramus. Perception and acquisition of linguistic rhythm by infants. *Speech communication*, 41(1):233–243, 2003.

[OTO68] Daniel C. O'Connell, Elizabeth A. Turner, and Linda A. Onuska. Intonation, grammatical structure, and contextual association in immediate recall. *Journal of Verbal Learning and Verbal Behavior*, 7(1):110–116, 1968.

[PCZ+19] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In Gernot Kubin and Zdravko Kacic, editors, *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pages 2613–2617. ISCA, 2019.

[PHY+25] Yu Pan, Yanni Hu, Yuguang Yang, Jixun Yao, Jianhao Ye, Hongbin Zhou, Lei Ma, and Jianjun Zhao. Clapfm-evc: High-fidelity and flexible emotional voice conversion with dual control from natural language and speech. *arXiv preprint arXiv:2505.13805*, 2025.

[PLW+24] Navin Raj Prabhu, Bunlong Lay, Simon Welker, Nale Lehmann-Willenbrock, and Timo Gerkmann. Emoconv-diff: Diffusion-based speech emotion conversion for non-parallel and in-the-wild data. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11651–11655. IEEE, 2024.

[POXT09] Santitham Prom-On, Yi Xu, and Bundit Thipakorn. Modeling tone and intonation in mandarin and english as a process of target approximation. *The journal of the Acoustical Society of America*, 125(1):405–424, 2009.

[PPSZ20] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao. Non-autoregressive neural text-to-speech. In *International conference on machine learning*, pages 7586–7598. PMLR, 2020.

[PTC+24] Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, Yui Sudo, Muhammad Shakeel, Kwanghee Choi, Jiatong Shi, Xuankai Chang, et al. Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer. *arXiv preprint arXiv:2401.16658*, 2024.

[PTGK00] Carol R. Paris, Margaret H. Thomas, Richard D. Gilson, and J. Peter Kincaid. Linguistic cues and memory for synthetic and natural speech. *Human Factors*, 42(3):421–431, 2000.

[PVG+21] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning*, pages 8599–8608. PMLR, 2021.

[PWTL02] K Luan Phan, Tor Wager, Stephan F Taylor, and Israel Liberzon. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fmri. *Neuroimage*, 16(2):331–348, 2002.

[QJHJM20] Kaizhi Qian, Zeyu Jin, Mark Hasegawa-Johnson, and Gautham J Mysore. F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6284–6288. IEEE, 2020.

[QWL+25] Tianhua Qi, Shiyan Wang, Cheng Lu, Tengfei Song, Hao Yang, Zhanglin Wu, and Wenming Zheng. Promptevc: Controllable emotional voice conversion with natural language prompts. *arXiv preprint arXiv:2505.20678*, 2025.

[QWZ23] Han Qi, Yue Wang, and Li Zhu. Discounted thompson sampling for non-stationary bandit problems. *CoRR*, abs/2305.10718, 2023.

[QZC+19] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019.

[QZC+20] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pages 7836–7846. PMLR, 2020.

[RBES20] Georgios Rizos, Alice Baird, Max Elliott, and Björn Schuller. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3502–3506. IEEE, 2020.

[Red03] Laura Redi. Categorical effects in production of pitch contours in english. In *Proceedings of the 15th International Congress of the Phonetic Sciences*, pages 2921–2924, 2003.

[RGN+04] Burton S Rosner, Esther Grabe, Hannele BM Nicholson, Keith Owen, and Elinor L Keane. Prosody, memory load, and memory for speech. *Oxford University Working Papers in Linguistics*, 2004.

[RGT23] Nathan Roll, Calbert Graham, and Simon Todd. PSST! prosodic speech segmentation with transformers. In Jing Jiang, David Reitter, and Shumin Deng, editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 476–487, Singapore, December 2023. Association for Computational Linguistics.

[RH07] Andrew Rosenberg and Julia Hirschberg. Detecting pitch accent using pitch-corrected energy-based predictors. In *Interspeech*, pages 2777–2780, 2007.

[RHT+21] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *9th International*

Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[RK17] Vishnu Raj and Sheetal Kalyani. Taming non-stationary bandits: A bayesian approach. *CoRR*, abs/1707.09727, 2017.

[RKX+23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[Ros10] Andrew Rosenberg. Autobi-a tool for automatic tobi annotation. In *Interspeech*, pages 146–149. Makuhari, Chiba, Japan, 2010.

[RRT+19] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

[RS19] Georgios Rizos and Björn Schuller. Modelling sample informativeness for deep affective computing. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3482–3486. IEEE, 2019.

[RVC19] Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. *Advances in Neural Information Processing Systems*, 32, 2019.

[RWM+22] Dino Rattcliffe, You Wang, Alex Mansbridge, Penny Karanasou, Alexis Moinet, and Marius Cotescu. Cross-lingual style transfer with conditional prior vae and style loss. In *Proc. Interspeech 2022*, pages 4586–4590, 2022.

[SAMH22] Rose Sloan, Adaeze Adigwe, Sahana Mohandoss, and Julia Hirschberg. Incorporating prosodic events in text-to-speech synthesis. In *Proceedings of the Conference on Speech Prosody*, pages 287–291, 2022.

[SB13] Björn Schuller and Anton Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[SBP+92] Kim EA Silverman, Mary E Beckman, John F Pitrelli, Mari Ostendorf, Colin W Wightman, Patti Price, Janet B Pierrehumbert, and Julia Hirschberg. Tobi: A standard for labeling english prosody. In *ICSLP*, volume 2, pages 867–870, 1992.

[SC03] Marc Schroeder and Roddy Cowie. Paralinguistic information: The missing part in speech and speaker recognition. *Journal of Speech Communication*, 40(3):227–246, 2003.

[SC10] Jesse Snedeker and Elizabeth Casserly. Is it all relative? effects of prosodic boundaries on the comprehension and production of attachment ambiguities. *Language and Cognitive Processes*, 25(7-9):1234–1264, 2010.

[SCDC+01] Marc Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, and Stan CAM Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. In *Interspeech*, pages 87–90, 2001.

[SGE18] Saurabh Sahu, Rahul Gupta, and Carol Y. Espy-Wilson. On enhancing speech emotion recognition using generative adversarial networks. In B. Yegnanarayana, editor, *19th Annual Conference of the International Speech Communication Association, Interspeech 2018, Hyderabad, India, September 2-6, 2018*, pages 3693–3697. ISCA, 2018.

[SGSB21] Tulika Saha, Dhawal Gupta, Sriparna Saha, and Pushpak Bhattacharyya. Emotion aided dialogue act classification for task-independent conversations in a multimodal framework. *Cognitive Computation*, 13:277–289, 2021.

[Slo23] Rose Sloan. *Using Linguistic Features to Improve Prosody for Text-to-Speech*. Columbia University, 2023.

[SPSB20] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372, 2020.

[SRBX+18] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR, 2018.

[SRC+00] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.

[SSJ+98] Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492, 1998.

[SSV20] Ravi Shankar, Jacob Sager, and Archana Venkataraman. Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator. *arXiv preprint arXiv:2007.12932*, 2020.

[ST03] Jesse Snedeker and John Trueswell. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and language*, 48(1):103–130, 2003.

[SYYC25] Xiaosu Su, Bowen Yang, Xiaowei Yi, and Yun Cao. Diffemotionvc: A dual-granularity disentangled diffusion framework for any-to-any emotional voice conversion. *Proceedings of the Interspeech, Rotterdam, The Netherlands*, pages 17–21, 2025.

[Tay09] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

[THW+20] P Lumban Tobing, Tomoki Hayashi, Yi-Chiao Wu, Kazuhiro Kobayashi, and Tomoki Toda. Cyclic spectral modeling for unsupervised unit discovery into voice conversion with excitation and waveform modelingh. In *Proceedings of INTERSPEECH*, volume 2020, 2020.

[Tom23] Rosario Tomasello. Linguistic signs in action: The neuropragmatics of speech acts. *Brain and Language*, 236:105203, 2023.

[TPRG20] Francesco Trovo, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 2020.

[TR21] Jason Taylor and Korin Richmond. Confidence intervals for asr-based TTS evaluation. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021*, pages 2791–2795. ISCA, 2021.

[TSİ+23] Andreas Triantafyllopoulos, Björn W Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, et al. An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*, 111(10):1355–1381, 2023.

[TTN+17] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of selected topics in signal processing*, 11(8):1301–1309, 2017.

[TWH+19] Patrick Lumban Tobing, Yi-Chiao Wu, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda. Non-parallel voice conversion with cyclic variational autoencoder. *arXiv preprint arXiv:1907.10185*, 2019.

[TWX+18] Xiaohai Tian, Junchao Wang, Haihua Xu, Eng Siong Chng, and Haizhou Li. Average modeling approach to voice conversion with non-parallel data. In *Odyssey*, volume 2018, pages 227–232, 2018.

[UDL+25] Ismail Rasim Ulgen, Zongyang Du, Junchen Lu, Philipp Koehn, and Berrak Sisman. Objective evaluation of prosody and intelligibility in speech synthesis via conditional prediction of discrete tokens. *arXiv preprint arXiv:2509.20485*, 2025.

[VS08] Klára Vicsi and György Szaszák. Using prosody for the improvement of asr—sentence modality recognition. In *Proc. Interspeech*, Brisbane, Australia, 2008.

[VS10a] Klára Vicsi and György Szaszák. Using prosody to improve automatic speech recognition. *Speech Communication*, 52(5):413–426, 2010.

[VS10b] Klára Vicsi and György Szaszák. Using prosody to improve automatic speech recognition. *Speech Communication*, 52(5):413–426, 2010.

[VWvdG22] Marcus Vielsted, Nikolaj Wallenius, and Rob van der Goot. Increasing robustness for cross-domain dialogue act classification on social media data. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 180–193, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics.

[WDY+21] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. *arXiv preprint arXiv:2106.10132*, 2021.

[Wei16] Edda Weigand. *The dialogic principle revisited: Speech acts and mental states.* Springer, 2016.

[Wel47] Bernard L Welch. The generalization of 'student's'problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[Wig02] Colin W Wightman. Tobi or not tobi? In *Speech Prosody 2002, International Conference*, 2002.

[WKB13] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207, 2013.

[WLHW08] Nick Webb, Ting Liu, Mark Hepple, and Yorick Wilks. Cross-domain dialogue act tagging. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation*

(LREC'08), Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[WN07] Dagen Wang and Shrikanth Narayanan. An acoustic measure for word prominence in spontaneous speech. *IEEE transactions on audio, speech, and language processing*, 15(2):690–701, 2007.

[WP01] Marilyn Walker and Rebecca J Passonneau. Date: a dialogue act tagging scheme for evaluation of spoken dialogue systems. In *Proceedings of the first international conference on Human language technology research*, 2001.

[WPK+18] Jochen Wirtz, Paul G Patterson, Werner H Kunz, Thorsten Gruber, Vinh Nhat Lu, Stefanie Paluch, and Antje Martins. Brave new world: service robots in the frontline. *Journal of Service Management*, 29(5):907–931, 2018.

[WSZ+18] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International conference on machine learning*, pages 5180–5189. PMLR, 2018.

[WTG08] Duane G Watson, Michael K Tanenhaus, and Christine A Gunlogson. Interpreting pitch accents in online comprehension: H* vs. l+ h. *Cognitive science*, 32(7):1232–1244, 2008.

[WYH+25] Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

[XGH+25] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[XL15] Rui Xia and Yang Liu. A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on affective computing*, 8(1):3–14, 2015.

[XM12] Bo Robert Xu and Peggy Mok. Cross-linguistic perception of intonation by mandarin and cantonese listeners. In *Proceedings of the Sixth International Conference on Speech Prosody 2012*, pages 99–102, 2012.

[YBL+04] Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Zhigang Deng, Sungbok Lee, Shrikanth Narayanan, and Carlos Busso. An acoustic study of emotions expressed in speech. In *Eighth International Conference on Spoken Language Processing*, 2004.

[YCHT+17] Xuesong Yang, Yun-Nung Chen, Dilek Hakkani-Tür, Paul Crook, Xiujun Li, Jianfeng Gao, and Li Deng. End-to-end joint learning of natural language understanding and dialogue manager. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5690–5694. IEEE, 2017.

[YKS+23] Dong Yang, Tomoki Koriyama, Yuki Saito, Takaaki Saeki, Detai Xin, and Hiroshi Saruwatari. Duration-aware pause insertion using pre-trained language model for multi-speaker text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[YLH+19] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, et al. Durian: Duration informed attention network for multimodal synthesis. *arXiv preprint arXiv:1909.01700*, 2019.

[YLL+24] Qu Yang, Qianhui Liu, Nan Li, Meng Ge, Zeyang Song, and Haizhou Li. Svad: A robust, low-power, and light-weight voice activity detection with spiking neural networks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 221–225. IEEE, 2024.

[YTZ+22] SiCheng Yang, Methawee Tantrawenith, Haolin Zhuang, Zhiyong Wu, Aolan Sun, Jianzong Wang, Ning Cheng, Huaizhen Tang, Xintao Zhao, Jie Wang, et al. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. In *Proc. Interspeech 2022*, pages 2553–2557, 2022.

[YVM19] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). sound, 2019.

[ZCDS14] Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):115–126, 2014.

[ZDC+19] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. In *20th Annual Conference of the International Speech Communication Association, Interspeech 2019*, pages 1526–1530. ISCA, 2019.

[ZKG+25] Han Zhu, Wei Kang, Liyong Guo, Zengwei Yao, Fangjun Kuang, Weiji Zhuang, Zhaoqing Li, Zhifeng Han, Dong Zhang, Xin Zhang, et al. Zipvoice-dialog: Non-autoregressive spoken dialogue generation with flow matching. *arXiv preprint arXiv:2507.09318*, 2025.

[ZLY+21] Yuxiang Zou, Shichao Liu, Xiang Yin, Haopeng Lin, Chunfeng Wang, Haoyu Zhang, and Zejun Ma. Fine-grained prosody modeling in neural speech synthesis using tobi representation. In *Interspeech*, pages 3146–3150, 2021.

[ZPIE17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[ZSLL20] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. *arXiv preprint arXiv:2010.14794*, 2020.

[ZSLL21] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE, 2021.

[ZSZL20] Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li. Converting anyone's emotion: Towards speaker-independent emotional voice conversion. *arXiv preprint arXiv:2005.07025*, 2020.

[ZWS19] Zixing Zhang, Bingwen Wu, and Björn Schuller. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6705–6709. IEEE, 2019.