Doctoral Dissertation

Assessing the importance of social factors in dialogue processing

Liu Yuning

Supervisor　Masashi Unoki

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science
December, 2024

Abstract

In daily conversations, participants' behaviors and expressions are influenced by social factors such as roles and status. According to theories like communicative action and communication accommodation, individuals adjust their language and behavior to align with their partner's communication style. This adjustment fosters mutual understanding and enhances the success of the interaction. This natural tendency is both a cognitive process and a crucial social mechanism. Speech entrainment reflects a speaker's desire for social integration or identification with others. Speakers can modulate their entrainment to either reduce social distance by aligning with their partner or increase it through dis-entrainment.

Given the pervasive nature of entrainment in dialogues, this paper's core argument is that entrainment in conversations can reflect certain aspects of social distance. This social distance, in turn, is influenced by the topic and function of the conversation. Previous studies have also indicated that the topic and function in dialogues are part of social factors. Therefore, by quantifying entrainment metrics in conversations, we propose that it is possible to capture and reflect aspects of the social factors present in the interaction. These social factors can deepen our understanding of human interaction and provide objective insights. The main scientific question is how to leverage computational techniques to quantify social factors through entrainment. Unlike previous studies, this research, based on the Interactive Alignment Model (IAM), adopts a top-down theoretical framework and employs a linear model to measure entrainment across multiple acoustic

parameters. This approach, compared to traditional methods that calculate entrainment for individual features, better captures sociological factors in conversations. Specifically, this study focuses on exploring how participants employ different strategies to interact across varied scenarios. To test these hypotheses, three research questions are posed: (1) How can various social factors be quantified, and do these quantified factors exhibit different patterns across conversational scenarios? (2) Can these factors help classify or predict conversational scenarios? (3) Is the proposed method of measuring entrainment across multiple parameters superior to traditional methods that consider each feature separately?

To answer these questions, a corpus of Chinese conversations involving scenarios such as arguments, comforting, sharing happiness, and persuasion was developed. Additionally, the open-source IEMOCAP English corpus, containing emotionally rich conversations, was used to validate the findings.

The results demonstrate that speakers adjust their entrainment and dis-entrainment strategies depending on the conversational scenario. The proposed method bridges the gap between psychology and engineering, offering an objective framework for understanding human interaction, thus addressing the first research question. Introducing social factors into both corpora showed that scenario classification in the Chinese corpus achieved 62.3% accuracy, while predicting entrainment trajectories using acoustic features in the English corpus reached 49.0% accuracy. Integrating acoustic features with social factors improved accuracy by 5% and 2% in the respective corpora, answering the second research question. Finally, comparing the proposed method with traditional entrainment measures validated the third research question, demonstrating the superiority of the combined approach. This paper explores the role of entrainment in human interaction and presents

a computational framework to quantify sociological factors. These findings offer new insights into human-to-human interactions and are particularly promising for improving human-machine dialogue systems. By incorporating social factors, future human-machine systems could better perceive and respond to users' social cues, resulting in more natural and engaging interactions.

Keywords: social factors, conversational entrainment, social distance, Communication Accommodation, Speech act, Interactive Alignment Model

Acknowledgment

Firstly, I would like to express my heartfelt gratitude to my research supervisors, Prof. Jianwu Dang and Prof. Masashi Unoki, for their invaluable assistance and guidance throughout my research journey. Their expertise and support have been instrumental in helping me explore the fascinating field of cognitive science and successfully complete this paper.

I am also deeply grateful to my minor research supervisor, Prof. Shogo Okada and Prof. Aijun Li for their insightful feedback and contributions to my research. Their perspectives and expertise have added significant value to my work.

I extend my sincere appreciation to my research teams at His-lab in JAIST and the teams in Institute of Linguistics, Chinese Academy of Social Sciences. The stimulating discussions during our team meetings have played a crucial role in enhancing my research progress. I have cherished many wonderful memories with my teammates, and their camaraderie has been invaluable.

Lastly, none of this would have been possible without the unwavering support of my family and my boyfriend. They have been a constant source of inspiration and motivation, uplifting me in times of doubt. Their encouragement and guidance have propelled me forward, and I am forever grateful for their unwavering belief in me.

Acronym and Abbreviations

F0	Fundamental frequency
IAM	Interactive alignment model
LSA	Latent semantic analysis
PCA	Principal component analysis
SCC	Spearman's correlation coefficient
SVM	Support vector machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
BERT	Bidirectional Encoder Representations from Transformers
IEMOCAP	The Interactive Emotional Dyadic Motion Capture dataset
NHR	Harmonic-to-noise ratio
SHR	Sub-harmonic-to-harmonic ratio

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Research Background

In daily life, conversations happen constantly, but behind these seemingly simple exchanges lies a complex mechanism. According to the interactive alignment model (IAM) in Fig. 1.1 [1, 3], from the speaker's perspective, they formulate an utterance based on the dialogue situation. Initially, a high-level intention transforms into linguistic representations, starting with syntactic elements and progressing to phonological ones. Ultimately, these representations are translated into an articulatory program to produce low-level acoustic features. From the listener's standpoint, they decode the sound by converting it into successive levels of linguistic representation until they grasp the intention. Based on the theories of communicative action [4] and communication accommodation [5], dialogue is an act of cooperation, and it is important to respect others during the conversation. To build better relationships during conversation, people constantly adjust their language and behavior in communication to accommodate their conversation partner's social and cultural background, language ability, communication style, and other factors to reach consensus in the conversation and make it successful.

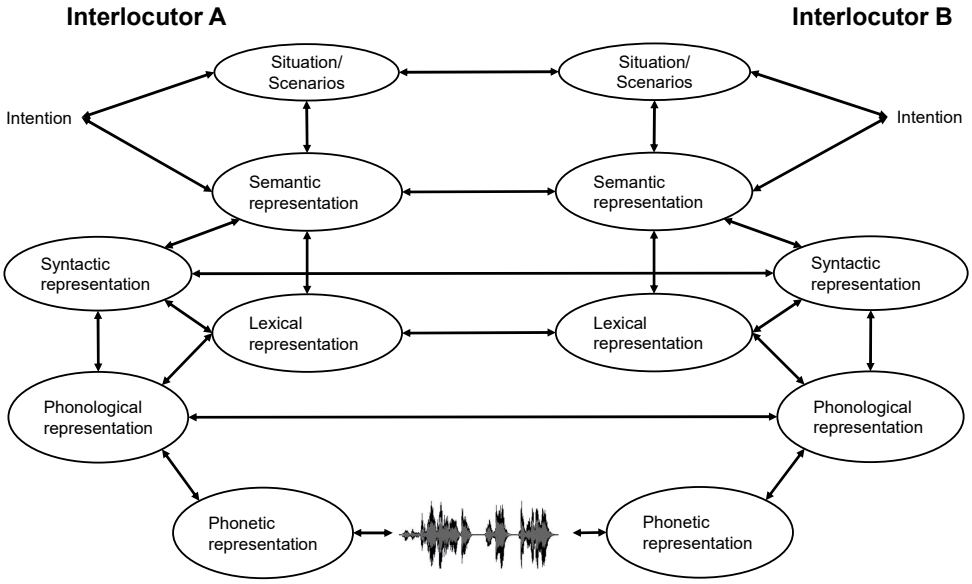

Figure 1.1: Schematic representation of the stages of comprehension and production processes according to the interactive alignment model [1]

Grace et al. [6]believed that the variety in one's utterances is influenced by several social and formality of the context. This occurs because the style of one's speech belongs to language variation which is concerned with the term of social factors. According to Holmes [7], these social factors include the participants, the setting, the topic, and the function. The participants is the person speak to and the kind of relationship you have with the person (interlocutor) determines your choice of words and also the intonation of voice. The setting means in what context the language exchanges occur, example: in the market, hall, office room in a home, university and so on. The topic is what is being talked about. And the function is the purpose of a conversation would influence the way a person speak. Sociolinguists think that there is a close relationship between language and the social environment. Each person or group has their own language style for expressing their intentions and purposes, which is influenced by the situation and the social environment. Consequently, the social environment and the language style used influence each other. These speech styles can be observed through in terms of these social factors. These social factors are crucial in human conversations, but they are often overlooked in current dialogue and human-machine interaction systems.

1.2 Conversational entrainment and social interaction

IAM theory holds that accurately conveying the interlocutor's intention based on the current situation is the most crucial aspect of dialogue. When intentions align, each level in this process should also align with each other. During this process, interlocutors seek to gain approval from their

conversation partner by becoming more similar to each other [2]. Therefore, Conversational entrainment is often regarded as the foundation for successful communication [8, 9]. However, other previous studies observed that interlocutors can simultaneously experience both "entrainment (alignment)" and "disentrainment (disalignment)" phenomena across various features. This means that they become more similar in some features while becoming increasingly dissimilar in others [10–12]. For example, a previous study investigated the correlation of entrainment on multiple acoustic and lexical features [13]. However, their results were quite divergent. Some features showed a positive relationship, indicating increased similarity between interlocutors, while other features exhibited a negative relationship, suggesting increased dissimilarity. This phenomenon is also consistent with reports from other previous studies indicating that participants reached a consensus on temporal-phonetic measures, exhibited entrainment on specific spectral-phonetic measures while diverging on others, and did not display entrainment on a syntactic measure [12].

This study posits that the presence of disentrainment in certain parameters does not contradict IAM theory. Rather, it suggests that dialogue is influenced by various social factors. These influences often manifest during the higher-level encoding processes of dialogue, such as topic, function, and social roles. Importantly, entrainment and disentrainment reflect not only linguistic adaptation but also underlying social dynamics, including social distance.

According to previous research [14, 15], entrainment can serve as a means of measuring social distance between interlocutors. Typically, stronger entrainment correlates with smaller social distance, resulting in more positive interaction outcomes. For instance, when interlocutors engage in comforting

or supportive conversations, entrainment is stronger, reflecting a desire to reduce social distance and increase social cohesion. Conversely, in conflictual or argumentative settings, interlocutors may consciously disentrain to increase social distance, signaling disagreement or discomfort. The degree of entrainment thus serves as a direct indicator of the social distance between speakers, and this distance can reflect the conversational topic and function.

Consequently, this study proposes that the contradictory entrainment phenomena observed in low-level features, even when high-level intentions align, may result from the failure to account for social factors during analysis. While alignment is often treated as a unitary concept [12], this perspective overlooks the influence of social factors such as participants' relationships, conversation topics, and purposes. Many previous studies have primarily focused on entrainment in low-level acoustic features, assuming that these features are independent of one another. However, this assumption overlooks the top-down influence of social factors on low-level behaviors, which can modulate the extent and type of entrainment observed.

Therefore, this study suggests that social factors—particularly those that affect the social distance between participants—influence low-level features in a top-down manner. To better assess the phenomenon of entrainment and the success of conversations, it is crucial to consider the relationship between these features from a top-down perspective, thereby enhancing our understanding of the mechanisms that underlie conversational entrainment.

1.3 Consider conversational entrainment from a high-level structure

Edlund et al.'s study has using proximity, convergence, and synchrony to quantify entrainment [16]. Proximity refers to whether features have similar mean values across interlocutors throughout the entire conversation, and convergence means whether the difference in feature values decreases across interlocutors over time. Whether the feature values of both interlocutors rise and fall together during the conversation is indicated as synchrony. In previous research, conversational entrainment (alignment) was used to describe the similarity of communicative behavior between interlocutors and often regarded as the foundation for successful communication [8, 9]. However, as previously mentioned, dialogue is influenced by numerous social factors, and due to the difficulty in quantifying these social factors, previous quantitative analyses of dialogic entrainment have mostly focused on independent low-level features without considering social factors. This may be a contributing factor to the findings in earlier studies where interlocutors can simultaneously experience both "entrainment" and "disentrainment" phenomena across various features.

In the process of conversation, a speaker's higher-level intentions are encoded through successive layers, ultimately resulting in the representations of speech signals. After this, the listener decodes the speech signals layer by layer to grasp the speaker's intentions. This research posits that, during this decoding process, people always consider the interaction and mutual influence among various low-level features, which are themselves influenced by social factors in the conversations. Therefore, it can be said that this form of entrainment, which considers the interaction between low-level features

to some extent, can reflect the social factors present in dialogue. This paper define this entrainment from a from a high-level structure view, which considers the interaction between low-level features, as "social factor".

1.4 Motivation and research goal

The philosophy guiding this research asserts that conversations in our lives are influenced by various social factors. High-level conversation entrainment is understood to reflect these social aspects to a certain extent, and these social factors can be utilized to classify and predict various conversational situations, thereby enhancing both human-human dialogues and human-machine interaction systems.

The motivation for this research comes from my interest in the entrainment phenomenon. In daily life, certain words and tones might be acceptable when conversing with classmates but inappropriate when addressing teachers or elders. This raises the question: Is it possible that different social factors in various contexts cause some words and tones to be suitable in Scenario A, leading to successful entrainment between interlocutors, but unsuitable in Scenario B, where another pattern of words and tones is needed to achieve entrainment in conversations?

If a method to describe these social factors in conversations could be found, it might help determine what kind of entrainment will occur in the current context. By exploring the mechanisms behind this phenomenon, entrainment, which considers various social factors, can be utilized to assess the state of a conversation. This approach can help us prevent unpleasant occurrences during dialogues and ultimately guide the conversation toward success. This would have significant implications for future human-computer

interaction systems. For instance, when customer service representatives are informed that they cannot establish good entrainment with a customer, they can try changing their conversation strategy by altering the words or tone used. Similarly, in human-machine interactions, the system can promptly switch topics when users lose interest in the current topic.

The final goal of this research is to explore the mechanisms of entrainment influenced by social factors in different conversational situations. To achieve this goal, three subgoals are set in this dissertation:

1. How to quantify various social factors, and do these quantified social factors exhibit different patterns in different conversational situations?

2. Can these social factors help classify or even predict conversational situations?

3. Is the proposed entrainment method, which considers relationships between different acoustic parameters in a top-down approach, superior to traditional methods that consider each feature separately?

1.5 Contribution and research novelty

This study makes several significant contributions to the field, the first is the development of a Linear Model for Measuring Social Aspects by proposing a novel linear model for quantifying social aspects within conversational data. By proposing a systematic approach to measure these social factors, the study advances our understanding of the nuanced influences of social dynamics on human interaction. Second is a methodological innovation in quantifying semantic information. This study proposes a method to quantify semantic information within conversations. By leveraging techniques such as BERT (Bidirectional Encoder Representations from Transformers), the research

provides a robust framework for extracting and analyzing lexical features, contributing to more comprehensive analyses of conversational dynamics.

These contributions offer valuable insights and methodological advancements for researchers investigating the intricate interplay between social factors and linguistic features in human communication.

Furthermore, this study introduces a linear model for quantifying social aspects as "social factors". The innovation lies in the development of a novel approach to capturing and quantifying social dynamics within conversational contexts. The derived social factors offer a promising avenue for achieving a more comprehensive understanding and classification of various conversation situations.

Finally, the significance of this study lies in its potential to utilize the derived "social factors" to prevent undesirable conversational outcomes and guide interactions toward more positive results. By predicting the trajectory of conversations based on the analysis of key social dynamics, it becomes possible to detect early signs of communication breakdowns, such as emotional misalignment or miscommunication, and intervene effectively to realign the interaction. It involves actively monitoring conversational dynamics to identify signals of potential misalignment or negative conversational patterns. For instance, a system could alert a customer service representative when communication starts to deviate from a positive path, prompting timely adjustments in tone, word choice, or conversational strategies to de-escalate the situation and avoid undesirable outcomes.

The development of real-time systems that track and analyze these social factors could significantly enhance the quality of human-computer interaction. Such systems could detect early communication issues and provide guidance to users on how to adjust their responses, whether in customer

service, mental health support, or conversational AI systems. This proactive approach to communication improvement, informed by the quantification of social dynamics, can lead to more efficient and effective interactions.

1.6 Dissertation outline

There are seven chapters in this dissertation. The remainder are organized as follows.

Chapter 2 introduces previous research on entrainment (alignment), focusing on the entrainment phenomena on various features in dialogues and how this phenomenon is quantified. I will also point out the shortcomings of these studies and aim to overcome these weaknesses in our research to enhance the investigation of entrainment mechanisms.

Chapter 3 explains the method of quantifying social factors and the calculation methods of the three entrainment metrics used in this study.

Chapter 4 elaborates on Experiment 1—Social factor for dialogue scenario classification. The questions of interest that this experiment aims to address are: (1) Do these quantified "social factors" exhibit different patterns in different current conversation situations? and (2) Can these "social factors" help classify the conversation situations?

Chapter 5 elaborates on Experiment 2—Social factors for conversation situation prediction. The questions of interest that this experiment aims to address are: (1) Do these quantified "social factors" exhibit different patterns in different future conversation situations? and (2) Can these "social factors" help predict the conversation situations?

Chapter 6 discusses the insights revealed by the current analysis and further examines whether entrainment at the semantic level is also influenced

by social factors.

Chapter 7 summarizes this paper, including the insights and contributions revealed in the current study.

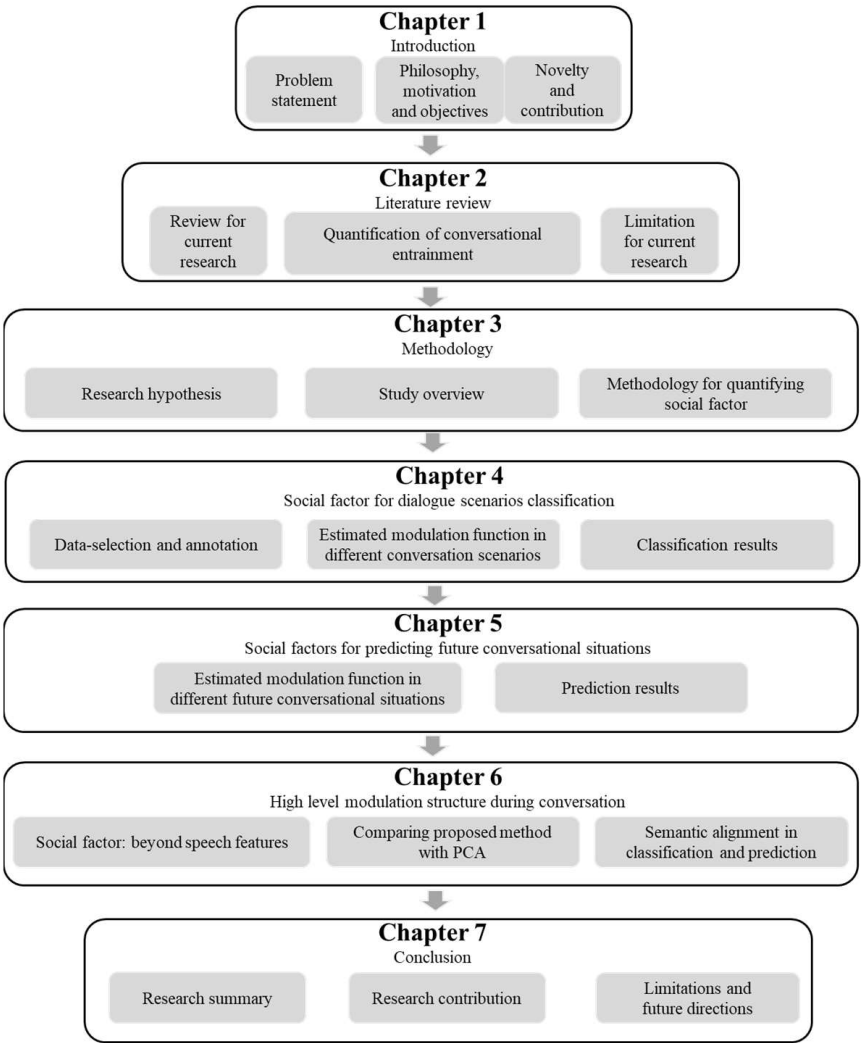

Figure 1.2: Thesis organization.

Chapter 2

Literature Review

Previous studies have extensively explored Conversational entrainment, demonstrating its potential for describing social factors. This chapter will focus on research related to conversational entrainment, highlighting how prior studies have quantified it and how it can be applied to human-computer interfaces.

2.1 What is entrainment?

Entrainment in human conversation refers to the tendency of speakers to synchronize or align their behaviors, such as speech patterns, gestures, and emotional expressions, with those of their interlocutors [17]. This alignment fosters rapport, understanding, and mutual engagement between speakers, ultimately enhancing the quality of communication [18–21].

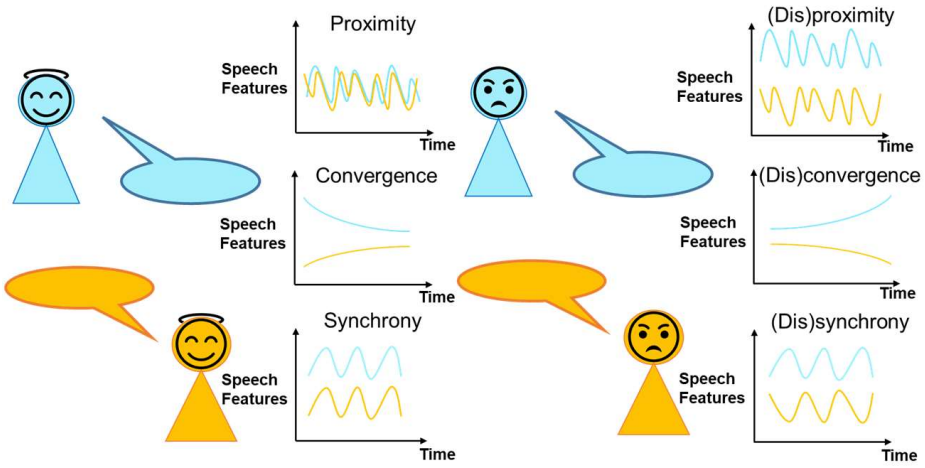

Figure 2.1: Schematic illustrations of Proximity, Convergence and Synchrony

To represent entrainment in conversations, researchers have proposed various metrics that capture different aspects of alignment between interlocutors. As show in Fig. 2.1, these metrics are proximity, which means whether the features have similar mean values across interlocutors throughout the entire conversation.Convergence, which means whether the difference in feature values decreases across interlocutors over time, and whether the feature values of both interlocutors rise and fall together during the conversation called synchrony.

By using these metrics, researchers can quantitatively analyze and measure the level of entrainment in conversations. This allows for a deeper understanding of the dynamics of social interaction and communication, as well as insights into the factors that influence interpersonal alignment and rapport.

2.2 Conversation entrainment occurs at multiple language levels

Garrod and Anderson used a designed computer maze game to explore how interlocutors use and interpret language within a restricted context [22]. They found that participants tended to use similar words and interpretation schemes to describe their positions in the cooperative maze game. Through an interactive card-sorting game, Branigan et al. found that participants not only tended to use similar vocabulary but also leaned towards employing the same grammatical structure as their partner had just used when describing a card to them [23]. The IAM posits that repeating the vocabulary and grammar used by the partner isn't merely a result of memory retrieval during speech. Instead, it's a mechanism through which conversationalists establish

a shared understanding, enhancing communication without focusing on each other's speaking habits. In previous research, this mechanism is referred to as conversation entrainment/alignment. Conversational entrainment was used to describe the similarity of communicative behavior between interlocutors.

Through extensive examination of conversational corpora, it has been observed that besides adjusting syntax and lexical features to match their partners [24–31], interlocutors also exhibit conversational entrainment in various low-level acoustic features, including their partner's fundamental frequency (F0) [15, 21, 32–34], vowel formants [35], voice onset time [36, 37], speech rate [38–41], pause or turn duration [42–44], pitch [13, 45], jitter [46], intensity [46, 47]. For instance, people may talk faster and louder when another participant in the argument raises their speaking rate and voice, or may speak more slowly and softly when comforting someone. This entrainment phenomenon has also been observed in nonverbal behaviors such as body movements [48–51] and facial expressions [52, 53] exhibited by both participants in the conversation. It's worth noting that this kind of alignment occurs not only in human-human dialogues but has also been confirmed to exist in human-machine interactions [54–56]. While conversational entrainment is a commonly observed phenomenon, the underlying mechanisms behind it remain unclear, particularly concerning the complex relationship between different speech features.

2.3 Quantifying conversational entrainment at linguistic levels

Although most studies have identified the existence of Conversational entrainment, they often remain at a qualitative level. Little research is available

that evaluates the properties of these entrainment measures. Without quantitative analysis, determining whether entrainment is universally present in most conversations becomes impossible, and assessing its sensitivity and stability also becomes challenging. What kind of distributions do these phenomena have for dialogue systems? Therefore, quantifying Conversational entrainment is a critical issue in this field.

Probabilistic measures

Church and Dubey et al. attempted to divide a long conversation into two parts. They used probabilistic measures to compute the likelihood of a single word or syntactic rule appearing in the latter half after its appearance in the former half. By counting the frequency of their co-occurrence, they aimed to determine if the probability of lexical and syntactic patterns was higher than the chance level, thereby confirming whether entrainment had occurred [57, 58].

The limitation of the probabilistic measure is that it requires a relatively large amount of text to conduct the computation because it relies on the observed frequency of words (or syntactic rules) to estimate the probability of co-occurrence. This means that to accurately determine if certain words or syntactic structures are appearing together more often than by chance, it need a substantial amount of conversational data. Without enough data, the measure may not be reliable or accurate, as rare co-occurrences might not be detected or might be misinterpreted due to insufficient sample size.

Document similarity measures

By Spearman's correlation coefficient (SCC) [59] or Latent Semantic Analysis (LSA) [60], it measures document similarity based on word frequency and co-occurrence or semantic similarity between documents [61, 62].

Repetition decay

Repetition decay, which is based on the concept of repetition effects, was notably demonstrated in the research by Branigan, Pickering, and Cleland [63]. Reitter et al. aims to quantify syntactic alignment by examining the decay rate of repetition probabilities of syntactic rules over time. Essentially, they observed how the likelihood of a syntactic rule being repeated decreased as the distance between the initial exposure and subsequent usage increased. They built a generalized linear model to analyze this repetition decay, using the distance between prime and target as a predictor [64]. The observation that the repetition rate of syntactic rules decreases as the distance increases suggests that there is a decrease in alignment strength over time. By using the regression coefficient of the predictor, they were able to estimate the strength of syntactic alignment. This approach provides a rigorous mathematical explanation for alignment phenomena, particularly from a probabilistic perspective. It helps distinguish alignment resulting from priming effects from random repetitions of linguistic elements.

However, there are limitations to this method. It cannot quantify alignment between individual pairs of texts, as it relies on analyzing repetition across sets of texts. Additionally, fitting a generalized linear model can be computationally intensive compared to other measures.

2.4 Quantifying conversational entrainment at acoustic levels

Unlike linguistic features, acoustic features inherently carry physical significance and can be represented using vectors. Therefore, the entrainment between acoustic features can be directly calculated using these vectors. As mentioned earlier, conversational entrainment is classified into three broad

categories: synchrony, proximity and convergence (see reference: [16, 65]).

As shown in Fig. 2.2a and Fig. 2.2c, synchrony means that the way people change their speech features, in terms of both direction and amount, is similar between different speakers. For example in Fig. 2.2c, the speaking rates of two speakers may be vastly different (i.e., low proximity). However, when one speaker increases their speaking rate, the other speaker may also increase theirs. Thus, despite low proximity between speakers, they are moving in parallel with each other, indicating high synchrony. Synchrony is typically measured using Pearson correlation coefficients [46, 66], and sometimes using cross-correlation [67–69], mutual information [70] or mean spectral coherence [71] between the acoustic features of interlocutors over a series of time points in the conversation [72]. A high degree of correlation reflects a high degree of similarity in the movement of speech features between interlocutors.

The concept of proximity is defined here as the distance of speech features between speakers in a conversation. For instance, if two speakers use similar speaking rates, the conversation is characterized by a high level of proximity (Fig. 2.2a and Fig. 2.2b). Conversely, if one speaker uses a fast speaking rate while the other uses a much slower rate, the proximity of the conversation is low (Fig. 2.2c and Fig. 2.2d). Typically, researchers rely on absolute difference of speech features to quantify proximity between speakers within the same turn. Sometimes, this calculation can also be done across different turns. For example, previous research examines differences between speakers in adjacent turns compared to non-adjacent turns [65, 73]. Convergence can be considered a subtype of proximity [2], as it describes whether the proximity between interlocutors increases or decreases over the course of a conversation.

Figure 2.2: Schematic of various combinations of high and low proximity and synchrony. (a) means the conversation is characterized by a high level of proximity and synchrony, (b) means the conversation is characterized by a high level of proximity and low level of synchrony, (c) means the conversation is characterized by a low level of proximity and high level of synchrony, and (d) means the conversation is characterized by a low level of proximity and low level of synchrony [2]

Although both synchrony and proximity are used to describe the similarity between interlocutors in a conversation, there are instances where, despite low proximity, the speakers are moving in parallel with each other, indicating high synchrony. Figure 2.2 (d) provides an example of this concept, two speakers may have highly nearly speech feature values (i.e., high proximity). However, despite this nearly, they may change their speech patterns differently during the conversation (i.e., low synchrony). This implies that a single measure alone cannot reliably determine the similarity between interlocutors; instead, consideration should be given to the overall changes in both characteristics.

2.5 Conversation entrainment as the basis for successful communication

The IAM posits that conversational entrainment is a mechanism through which interlocutors develop a shared understanding of the situation, thereby facilitating effective communication. Previous research define a general task that predicts conversational success from textual features. They found that the tendency to repeat language choices is linked to communicative success. This extends beyond sentence structure to lexical choices, with support vector machine (SVM) models [74] supporting the IAM at various representational levels. The results align with the view that speakers have a predisposition to adapt to each other, leading to task success. And syntactic adaptation correlates with task success early in dialogues more than lexical adaptation [75]. The study demonstrates the correlation between conversation entrainment and dialogue success, extending the concept of conversation entrainment from the field of cognitive psychology to a broader

range of applications.

Although conversation entrainment is an effective predictor of dialogue success, it has not been widely adopted in current dialogue systems. A unified, reliable, and simple framework is needed to standardize the calculation of entrainment metrics. For instance, in a recent study, researchers used customer service quality surveys as indicators of successful conversations between service representatives and customers. They assessed the success of these dialogues by computing the synchrony, proximity, and convergence of various acoustic feature metrics, including speech rate, jitter, shimmer, f0, and intensity max. However, their findings only marginally exceeded chance levels [76]. They also discovered that not all high synchrony values were associated with more successful conversations. The acoustic entrainment metrics that predicted dialogue success varied across different corpora. Only intensity-max convergence exhibited a positive correlation between conversational entrainment and predicted satisfaction in both corpora. These results suggest that, despite recent advancements, research on how entrainment predicts dialogue success is still in its early stages. First, it is challenging to ensure that an entrainment metric for a specific parameter is universally applicable across all corpora. Additionally, the relationship between dialogue success and the level of entrainment is not always positive; sometimes, lower entrainment levels may correlate with successful conversations [10–12].

These contradictory findings highlight the complex interactions between different speech features and the challenges in understanding entrainment. This concurrent entrainment and disentrainment across different features makes it challenging to evaluate the success of a conversation on the sole basis of the entrainment or disentrainment of a single feature. Contrary to conventional expectations, research has found that in certain cases, increasing

dissimilarity through disentrainment of specific features can be associated with positive social aspects in conversations [41]. This intriguing finding challenges the notion that closer alignment in conversational features always signifies a successful conversation. These findings underscore the limitation of evaluating conversation success on the sole basis of a single feature, highlighting the need for further exploration and deeper comprehension of the underlying mechanisms driving conversational entrainment. Therefore, finding a method to address these issues is crucial for leveraging conversation entrainment to predict dialogue success.

2.6 From a structure view to consider the conversation entrainment

Given these issues, prior research suggests discussing entrainment phenomena among different acoustic parameters from a structured perspective. Researchers try to investigate whether entrainment in one feature correlates with entrainment in another feature among interlocutors.If the intrinsic relationships between parameters can be identified, it becomes possible to infer the entrainment of other features based on the entrainment of a single feature. This approach could explain why entrainment is sometimes inversely related to dialogue success and might lead to the development of universally applicable entrainment metrics across various corpora. However, the results were disappointing.

Rahimi et al. examined the entrainment correlations between four acoustic features and five lexical features. Their results varied significantly: some correlations were not significant, while among those that reached significance, some exhibited positive correlations, while others showed negative correla-

tions [13]. Cohen Priva and Sanker conducted a similar study, investigating the convergence correlations between several acoustic speech features within the same conversation. This study also found that there were no cross-feature correlations between the levels of alignment [77]. Similarly, Weise and Levitan also check the relation between acoustic entrainment and lexical entrainment [78]. In this research, they used a Principal Component Analysis (PCA) [79] structure to consider the relationship of them. However, no relationship was found between the degree of alignment across different features. The authors concluded that alignment is not a singular, uniform behavior occurring simultaneously across levels and features. Instead, it comprises various loosely connected behaviors, possibly stemming independently from different cognitive mechanisms or serving distinct communicative or social purposes.

While their results may not be thrilling, the authors' research represents a crucial step forward. First, the lack of found relationships between features could be attributed to their use of only a few parameters to investigate entrainment phenomena. Additionally, in Weise and Levitan's study, they relied solely on PCA to integrate relationships between these features. PCA is a common data analysis method that does not impose strong prior assumptions or incorporate prior knowledge to investigate relationships between different speech features.

2.7 Limitation in previous research

Conversation entrainment, as a feature that can measure social factors in dialogue, holds significant potential for future conversational and natural language processing systems. However, despite extensive research over

the years, its application in human-computer interaction systems remains limited.

Theoretical limitation

Firstly, conversation entrainment has predominantly been treated as a phenomenon in cognitive psychology and linguistics, without sufficient quantitative research. This lack of quantitative analysis hinders its parameterization and subsequent application in human-computer interaction systems. Secondly, as previous studies have reported, there is currently no unified and effective theoretical framework for quantitatively representing conversation entrainment. Investigating conversation entrainment based on only a few parameters is incomplete and often leads to contradictory results that are not universally applicable across different corpora. Thirdly, although previous studies have attempted to explore the relationships between different features' entrainment, these studies often lack alignment with existing theoretical frameworks, such as using PCA without any prior knowledge to investigate the intrinsic relationship between different features. Therefore, it is essential to incorporate existing theories to guide our investigation of entrainment phenomena.

Technological limitation

Firstly, the methods for calculating entrainment of linguistic features and acoustic features lack granularity consistency. Acoustic features, which can be directly extracted from speech, allow for finer granularity. Consequently, the calculation of acoustic entrainment can be done both within and between speaking turns, even capturing pitch variations before and after each word. On the other hand, lexical feature entrainment quantification relies on analyzing repetition across sets of texts. This type of entrainment may span several speaking turns, making it difficult to capture fine-grained changes within

a single utterance. This mismatch in granularity might explain why it is challenging to find correlations between the two. Secondly, previous research has employed the concept of conversational entrainment to characterize the synchrony of communicative behaviors between interlocutors and attempted to utilize conversational entrainment to distinguish between successful and unsuccessful dialogues. However, it is important to acknowledge the limitations of binary classifications in evaluating dialogue success. This approach may oversimplify the evaluation process. Dialogue outcomes are influenced by numerous subtle factors, such as the context of the conversation and the emotional states of the speakers. For instance, conversations involving arguments or speakers expressing anger or sadness are more likely to be classified as unsuccessful. Simply categorizing all instances of conflict and sadness as unsuccessful may overlook the nuanced factors that influence the dialogue process due to overly broad classification. Moreover, exploring multi-class classifications allows conversational entrainment to have broader applications in human-machine interaction systems.

To address these limitations, it is essential to explore multi-class classifications, allowing conversational entrainment to have broader and more accurate applications, particularly in human-machine interaction systems. By recognizing the diverse outcomes that a conversation can have, this approach can better capture the complexities of human communication.

Inspired by the Interactive Alignment Model (IAM) and previous research, this study hypothesizes that social aspects exert a consistent influence on conversational scenarios of the same style. To quantify this influence, we define the "social factor" to measure the quantitative impact of social dynamics on conversational entrainment. As noted in related work, different speech features may exhibit contradictory behavior in the same conversation.

However, previous studies have often relied on single features to quantify entrainment, which provides an incomplete representation of the interaction [76, 78].

To address this issue, the method proposed in this study aims to derive social factors across different conversational styles by integrating multiple speech features. This approach captures the complex interplay among these features, enabling a more comprehensive representation of conversational entrainment. By considering how multiple features interact, this method effectively addresses the contradictory entrainment phenomena that can occur simultaneously across different acoustic features. This results in a measure that not only reflects the influence of social factors but also accommodates the variations and inconsistencies observed in different entrainment patterns.

Our hypothesis is that social factors have a consistent influence on conversational entrainment within the same type of conversational scenario. By capturing the underlying patterns of entrainment, even when contradictory behaviors occur across different features, the social factor provides a more holistic understanding of the conversation. This method not only allows for better classification of conversational scenarios but also enables the prediction of the future trajectory of the interaction, whether it evolves into a conversation that may create discomfort or one that fosters positive engagement and joy between interlocutors.

2.8 Summary

This chapter reviewed relevant prior research on conversational entrainment, beginning with an introduction to the phenomenon as it occurs across multiple levels of speech. The chapter then discussed how previous studies

have quantified these conversational entrainment phenomena and explored their potential applications in human-machine interaction systems. Finally, this chapter identified the limitations of earlier research from both theoretical and technological perspectives.

First, the chapter described how conversational entrainment manifests in various speech dimensions, including lexical, syntactic, and acoustic levels. It highlighted the importance of understanding these layers to grasp the full scope of conversational entrainment.

Next, it examined the methods used by researchers to quantify conversational entrainment, such as measuring synchrony, proximity, and convergence of speech features like speech rate, jitter, shimmer, f0, and intensity max. It emphasized the challenges and inconsistencies in these quantification methods, noting that a unified and effective theoretical framework is still lacking.

Furthermore, the chapter explored how conversational entrainment has been linked to human-machine interaction systems, assessing its potential to predict dialogue success. However, it noted that the application of these findings to interactive systems remains limited due to the absence of a standardized approach.

From a theoretical standpoint, the chapter pointed out that conversational entrainment has often been treated as a broad phenomenon in cognitive psychology and linguistics without sufficient quantification. This lack of quantification has hindered its parameterization and practical application in interactive systems.

Technologically, the review highlighted the granularity mismatch in measuring linguistic and acoustic feature entrainment. Acoustic features, being directly extractable from speech, allow for finer-grained analysis compared

to lexical features, which rely on text-based repetition analysis and span multiple turns. This granularity mismatch may contribute to the difficulty in identifying correlations between different feature alignments.

Additionally, the chapter emphasized the limitations of binary classifications in evaluating dialogue success. It argued for a more nuanced approach that considers the complexity of dialogue dynamics, such as the context and emotional states of the speakers, and suggested using scenario classification and future emotion prediction tasks instead.

Overall, this chapter underscores the need for a more refined and comprehensive understanding of conversational entrainment, proposing that future research should aim to address the identified theoretical and technological gaps to enhance its applicability in human-machine interaction systems.

Chapter 3

Method and Materials

3.1 Research hypothesis

The philosophy guiding this research asserts that conversations in our lives are influenced by various social factors. High-level conversation entrainment is perceived to reflect these social aspects to a certain extent in this research. It is suggested that these social factors can be leveraged to classify and predict various conversational situations, thereby enhancing both human-human dialogues and human-machine interaction systems.

This study posits that interlocutors use their intentions to guide conversations in various scenarios, resulting in distinct modulations of their speech features. Specifically, this research hypothesize that in different conversational contexts, such as arguing, comforting, convincing, and sharing happiness, speakers adjust their acoustic and linguistic features in ways that reflect their communicative goals.

Building on the IAM theory, it is proposed that the modulation of speech features is not random but systematically influenced by the conversational intent. For instance, during an argument, speakers may increase their speaking speed and volume, leading to faster and louder speech patterns. Conversely, in a comforting scenario, the speech might become slower and

softer, aiming to soothe the listener.

To test this hypothesis, this research will analyze dialogues from a Chinese TV drama that depicts daily life, focusing on the four identified scenarios: arguing, comforting, convincing, and sharing happiness. These scenarios are chosen because they encapsulate a wide range of emotional interactions and reflect the complex dynamics of human communication. By examining these common conversational contexts, the aim is to understand the underlying mechanisms and the role of social factors in shaping these interactions.

Additionally, the research will incorporate the IEMOCAP corpus, which is a widely recognized dataset in emotion research, to enhance the robustness and applicability of our findings. The IEMOCAP corpus includes emotional labels such as Angry, Sad, Neutral, and Happy, providing a comprehensive framework for analyzing emotional expressions and their impact on dialogue scenarios. It is hypothesized that by using these labels, future conversational situations can be predicted, elucidating how different social factors influence the entrainment process across various emotional contexts.

This research further hypothesize that the social factors quantitied by three metrics of entrainment (proximity, convergence, and synchrony) can effectively differentiate and predict these conversational scenarios. Proximity refers to the average distance between the speech features of interlocutors, convergence indicates the degree to which these features become more similar over time, and synchrony measures the simultaneous coordination of speech features between speakers. By quantifying these social factors, the aim is to demonstrate that they can serve as reliable indicators of the underlying conversational intent and scenario.

In summary, our research hypothesis is that the intentional modulation of speech features by interlocutors in different conversational scenarios can be

effectively captured and differentiated using the social factors of proximity, convergence, and synchrony. This, in turn, can enhance our understanding of the entrainment process and improve the prediction of future conversation situations based on emotional and contextual cues.

3.2 Construction of the "Golden Marriage" Corpus

To evaluate the proposed method's effectiveness in quantifying conversational entrainment, we first constructed a custom Chinese corpus derived from the dialogue of the Chinese TV drama Golden Marriage. This corpus was designed to closely approximate real spoken interactions in various social scenarios, adhering to clear guidelines for data selection, transcription, and annotation. The decision to extract content from Golden Marriage was driven by its realistic portrayal of family dynamics, which allowed us to focus on the topic and function aspects of social factors, aligning with the study's objectives.

The "Golden Marriage" TV series, jointly produced by the Beijing Television Art Center and Century Star Run Company in 2007, is a 50-episode drama that chronicles the life of a Beijing couple over the course of 50 years, from 1956 to 2005. Each episode represents one year in the couple's life, covering significant personal and social events. The show has been praised for its grounded depiction of daily family life in China, making it one of the most realistic portrayals of Chinese family dynamics in recent decades.

This study chose "Golden Marriage" as the source for our corpus due to its rich conversational content, which covers a wide range of emotional and

social situations. By focusing on four key scenarios—arguing, convincing, comforting, and sharing happiness—the corpus allows us to explore conversational entrainment across different thematic and functional contexts.

3.2.1 Data Collection and Selection

To ensure the selected dialogues were suitable for detailed acoustic and interactional analysis, we applied three primary criteria for data selection:

Audio Quality: The first criterion was the clarity of the audio. Many scenes in "Golden Marriage" include background music or environmental sounds, making it essential to choose dialogues where the speakers' voices were clear and not masked by noise. Annotators manually reviewed each scene and selected those with the highest speech quality, ensuring the data could be processed effectively for feature extraction.

Dialogue Length: Each dialogue had to contain at least six turns, with a minimum of three turns per speaker. This criterion ensured that interactions were substantial enough to capture meaningful conversational dynamics. Dialogues with fewer than six turns were excluded to maintain data consistency.

Content and Scenario Relevance: The corpus focuses on four conversational scenarios—arguing, convincing, comforting, and sharing happiness—to reflect diverse communicative purposes. These scenarios were selected because they highlight distinct topics and functions, providing clear, observable interactional patterns. This focus allowed us to systematically examine how the topic and function of a conversation influence entrainment dynamics.

The selection process was carried out by five professional annotators who each watched the entire TV series from beginning to end. The annotators were tasked with independently identifying dialogue segments that they believed fell into one of the four conversational scenarios: arguing,

convincing, comforting, or sharing happiness. Each annotator initially made their own judgments, selecting scenes they felt best represented these specific interaction types.

After the initial selection, the annotators came together to compare their choices and reach a consensus. For a dialogue segment to be classified under one of the four scenarios, at least three of the five annotators needed to agree on its categorization. For instance, if Annotator A judged a particular segment to belong to the arguing category, and at least two other annotators also categorized it as arguing, then that segment would be officially included in the corpus as an arguing dialogue.

This process ensured that the final selected dialogues reflected a shared understanding among multiple annotators, rather than being based on a single subjective judgment. By incorporating this method of majority voting, we aimed to increase the reliability of the classification and ensure that the chosen dialogues accurately represent the intended conversational scenarios. The dialogues are believed to capture real-life conversational behaviors, providing a robust dataset for analyzing how entrainment patterns vary across different interaction types.

3.2.2 Transcription and Multi-Level Annotation

Following selection, the chosen dialogues were manually transcribed into orthographic Chinese text, adhering to strict guidelines to ensure accuracy. Each dialogue was annotated using the Praat software toolkit [80], incorporating multi-level annotations to capture both segmental information (e.g., speaker turns) and higher-level features (e.g., conversational scenarios). This approach is consistent with other established corpora, such as CASIA-CASSIL, which also employ multi-layered annotation frameworks.

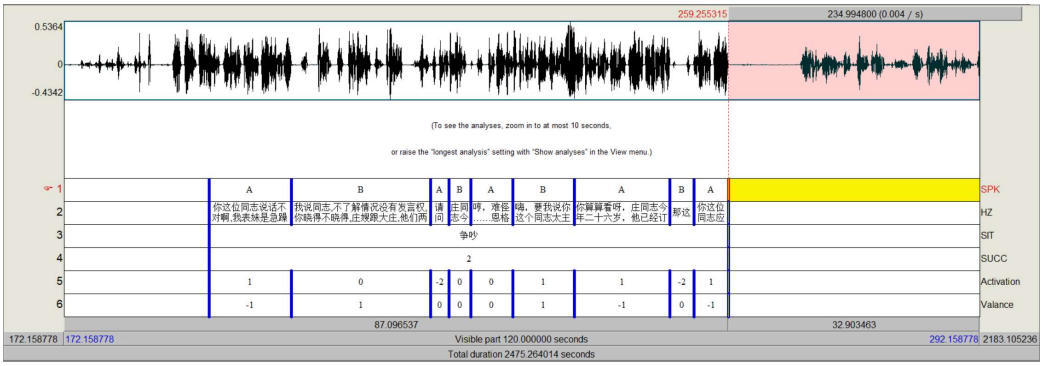

Figure 3.1: An example of an annotated dialogue using Praat

Turn-Level Annotation: Each dialogue was segmented into individual turns, marking when one speaker stopped and the other began. This segmentation allowed us to analyze how conversational entrainment evolved within each interaction.

Speaker Roles: Annotators labeled each speaker as either the dominant or follower, providing insight into how power dynamics and role-switching influenced entrainment patterns.

Scenario: Each dialogue was labeled according to the specific scenario it represented (e.g., arguing or comforting), offering a nuanced understanding of how topic and function shape interactional dynamics.

The audio data was also processed into 16-bit PCM WAV files to enable the extraction of speech features such as fundamental frequency (F0), energy, and harmonic ratios. These features were critical for analyzing the acoustic properties of the dialogue and how they related to conversational entrainment.

3.2.3 Corpus Statistics and Features

The final "Golden Marriage" corpus consists of 193 dialogues, categorized into the four aforementioned scenarios. The average length of a dialogue is 75 seconds, with approximately 11.75 turns per dialogue. The distribution of dialogues by scenario is shown in Table 6.1.

The transcription was manually corrected for accuracy, and each dialogue was segmented into turns. Z-score normalization was applied to ensure consistency across dialogues and speakers, enabling a more reliable analysis of the speech features.

3.2.4 Corpus Significance

The "Golden Marriage" corpus provides a rich dataset for studying conversational entrainment in realistic family settings. By focusing on the topic and function of conversations—such as arguing, convincing, and comforting—we are able to capture the core interactional dynamics that shape conversational entrainment. The corpus offers insights into how these social factors influence the alignment of speech patterns in emotionally charged interactions.

In addition, the realism of "Golden Marriage", often praised for its portrayal of daily life in China, ensures that the dialogues closely reflect natural conversation. This makes the corpus a valuable resource for studying conversational dynamics and developing models of human-machine interaction that respond to real-life social cues.

In future research, this corpus could be extended to explore how participants and setting further influence conversational entrainment. The inclusion of these additional social factors could enable a more comprehensive understanding of how conversational dynamics are shaped across different

Table 3.1: Statistics of "Golden Marriage" corpus

	Number of dialogue	Duration/per dialogue (second)	Number of turns/per dialogue	Total duration (minute)
Arguing	64	76.7	13	81.8
Comforting	67	82.5	11	91.1
Convincing	41	79.6	11	54.4
Sharing happiness	21	61.0	12	21.3

social contexts.

3.3 Interactive Emotional Dyadic Motion Capture dataset

After conducting experiments on our self-constructed Chinese corpus, this study going to explore whether similar patterns could be observed in languages other than Chinese. Therefore, we chose the Interactive Emotional Dyadic Motion Capture dataset (IEMOCAP), which is in English, to conduct further experiments.

Interactive Emotional Dyadic Motion Capture dataset [81], which focuses on capturing emotional expressions within conversations. Despite being a designed laboratory recording dataset, the IEMOCAP dataset employed two distinct methodologies: scripted sessions resembling plays and improvised hypothetical scenarios. In the second approach, actors were granted significant freedom in expressing their emotions. This approach enables the data in the IEMOCAP corpus to, to some extent, mirror genuine spoken interaction dynamics.

The objective of this study with the "Golden Marriage" corpus was to use the social factor to differentiate conversational entrainment patterns and classify conversation scenarios accordingly, and in the IEMOCAP, the classification task involved predicting the future trajectory of the conversation scenarios based on past conversational entrainment patterns. To accomplish this, the approach involved deriving the social factor using the historical information of dialogue turns from both interlocutors. By using historical information-based social factor, the goal was to predict trajectory of the conversational entrainment. The trajectory of conversational entrainment,

whether it is positive or negative, is predicted by the emotional labels of the upcoming turn. This prediction can effectively guide us in preventing unpleasant occurrences between interlocutors during conversations. It can also serve as a valuable guideline for the creation of more effective and user-friendly human-machine dialogue systems. Such systems can facilitate seamless communication and collaboration between humans and machines, improving interaction experiences.

3.4 Overview of proposed method

This study simplify the interactive alignment model into a linear model, as shown in Fig. 3.2. In this linear model, the input is the intention based on the current dialogue situations/scenarios(e.g., whether you want to comfort someone or share happiness with someone), and this input is encoded, encapsulating various social factors present in the ongoing conversation. The output consists of low-level acoustic features generated while expressing the speaker's intention.

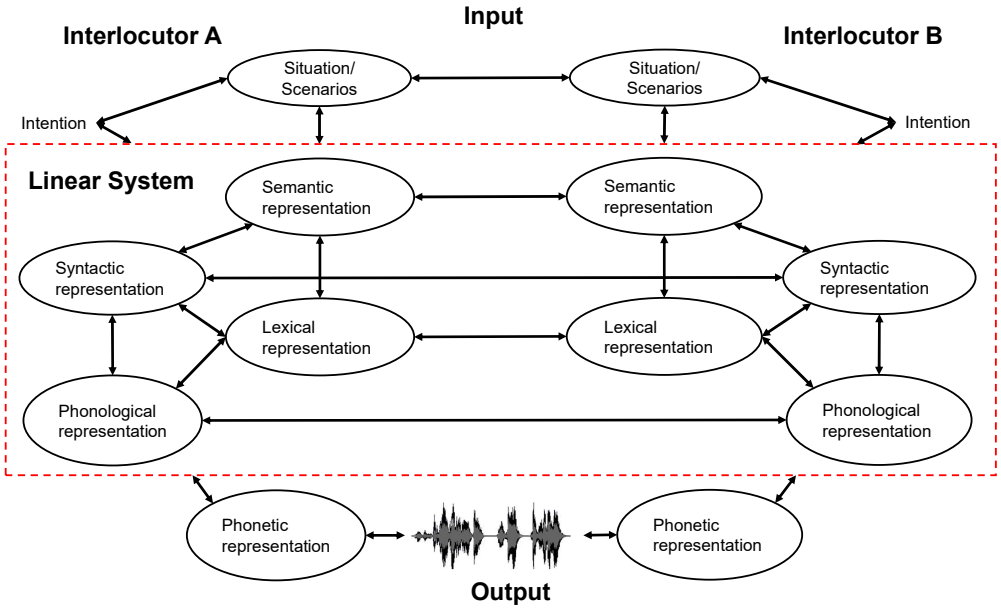

Figure 3.2: A simple linear model to describe the interactive alignment model

As shown in Fig. 3.3, the inverse process can be viewed as the modulation of the interlocutors' speech features by different dialogue strategies. The output modulated speech features can reflect high-level conversation scenarios. The entrainment between the modulated speech features is defined as the social factor. According to our hypothesis, three metrics of the social factor (proximity, convergence, synchrony) can effectively differentiate the dialogue style and conversation scenario of the ongoing dialogue. The definitions of these metrics are given in Section 3.6.

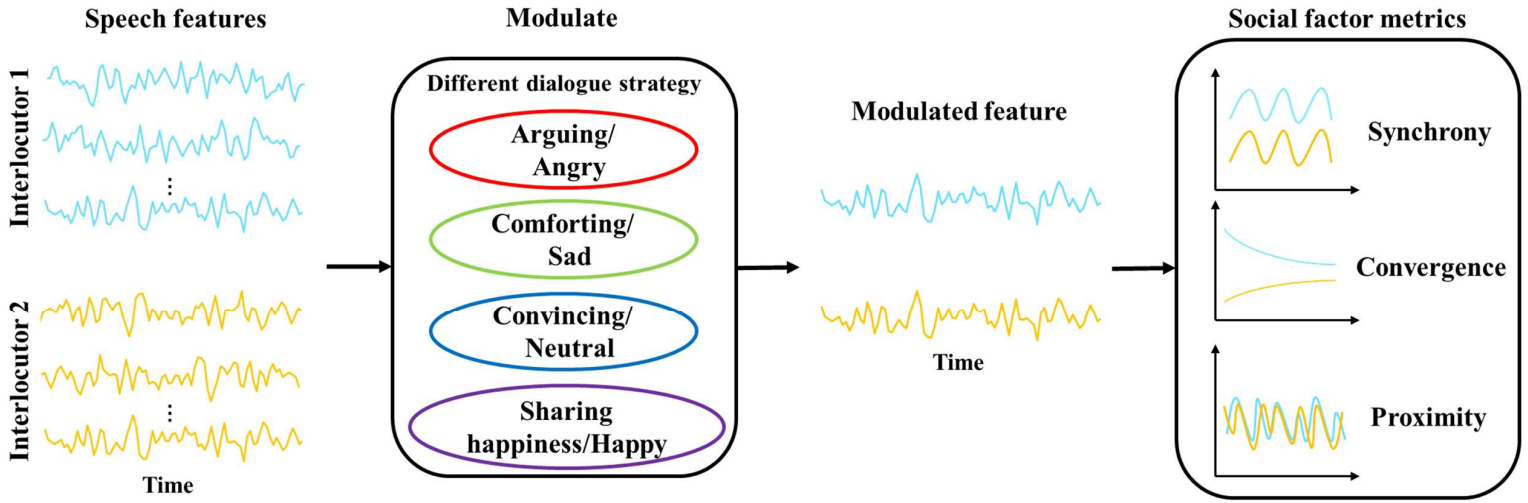

Figure 3.3: Modulation of interlocutors' speech features by different dialogue strategies related to social factors

This study first aim to infer the modulation function under different conversation scenarios by a linear model. Then evaluated the effectiveness of the social factor computed from the modulated speech features in classifying and predicting conversation situations.

3.5 Methodology for quantifying social factor

The linear model describes how different dialogue strategies modulate various low-level speech features. The objective with this model is to discover a vector $\boldsymbol{\omega} \in \mathbb{R}^{\boldsymbol{n}}$ that multiplies the speech features $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{t}}$, where $\boldsymbol{n}$ is the dimension of speech features and $\boldsymbol{t}$ is the length of the acoustic series. This multiplication aims to maximize the contrast between two different conversation scenarios by optimizing the weighted sum of speech features.

For example, consider the modulated acoustic feature $\boldsymbol{X_A} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{t}}$ from an arguing scenario, and the modulated feature $\boldsymbol{X_S} \in \mathbb{R}^{\boldsymbol{n} \times \boldsymbol{t}}$ from a sharing happiness scenario as the reference time series. The goal is to find the $\boldsymbol{\omega}$ that maximizes the following ratio $\boldsymbol{\lambda}$,

$$\boldsymbol{\lambda} = \frac{\|\boldsymbol{\omega^T X_A}\|^{\boldsymbol{2}}}{\|\boldsymbol{\omega^T X_S}\|^{\boldsymbol{2}}}. \tag{3.1}$$

Where $\|\cdot\|^{\boldsymbol{2}}$ represents the squared magnitude of the vector. Thus, $\boldsymbol{\lambda}$ is the ratio of the magnitude of the arguing speech features modulated through the $\boldsymbol{\omega}$, to the magnitude of the sharing happiness speech features modulated through the same $\boldsymbol{\omega}$. By expressing the squared magnitudes as dot products, the equation can be rewritten as:

$$\boldsymbol{\lambda} = \frac{\boldsymbol{\omega^T}(\boldsymbol{X_A X_A^T})\boldsymbol{\omega}}{\boldsymbol{\omega^T}(\boldsymbol{X_S X_S^T})\boldsymbol{\omega}}. \tag{3.2}$$

It is observed that $\boldsymbol{A} = \boldsymbol{X_A X_A^T}$ and $\boldsymbol{S} = \boldsymbol{X_S X_S^T}$ are two covariance matrices of the original speech features. Rewriting Eq. (3.2) using covariance matrices leads to the forms $\boldsymbol{\omega^T A \omega}$ and $\boldsymbol{\omega^T S \omega}$, which are known as the quadratic forms:

$$\boldsymbol{\lambda = \frac{\omega^T A \omega}{\omega^T S \omega}.} \tag{3.3}$$

Accordingly, the objective function is to find the vector $\boldsymbol{\omega}$ that maximizes $\boldsymbol{\lambda}$.

$$\boldsymbol{\arg\max_{\omega} \frac{\omega^T A \omega}{\omega^T S \omega}.} \tag{3.4}$$

The objective function in Eq. (3.4) is used to find the optimal vector $\boldsymbol{\omega}$ to maximize the variance ratio between $\boldsymbol{A}$ and $\boldsymbol{S}$, the goal is to use the modulation function to capture how speech features vary across different conversational contexts. By applying this function, we can derive a representation of the speech signals that reflect the distinct characteristics of each conversation scenario.

When the covariance ratio between different scenarios is maximized, the modulated data can clearly distinguish between these conversational situations. In other words, the modulation function allows us to optimize the separation between different conversation types by emphasizing the most distinctive features of the speech signals. This is the key role of the modulation function: to transform the speech features in a way that highlights the differences between scenarios, enabling clearer classification and analysis of the interactions.

However, it is possible to expand the objective function to include additional vectors $\boldsymbol{\omega_2}$ up through $\boldsymbol{\omega_M}$, where $\boldsymbol{M}$ is less than the dimension of speech features $\boldsymbol{n}$. Each $\boldsymbol{\omega_i}$ is subject to the constraint that it maximizes (or

minimizes) $\boldsymbol{\lambda_i}$ while being uncorrelated with the previous components. This extension, similar to principal component analysis (PCA), allows for a better description of the modulation function by incorporating more components. Considering additional components allows for the capture of a greater amount of variation in the modulation function, potentially enhancing its ability to modulate speech features. Thus, the inclusion of multiple $\boldsymbol{\omega_i}$ in the objective function aims to explore a richer representation of the modulation function and enhance its performance in capturing the desired characteristics of conversational entrainment. Therefore, Eq. (3.3) can be rewritten as:

$$\boldsymbol{\Lambda} = \frac{\boldsymbol{W^T A W}}{\boldsymbol{W^T S W}}, \tag{3.5}$$

where $\boldsymbol{W}$ is a modulation function, and each diagonal element of $\boldsymbol{\Lambda}$ is the multivariate ratio in the direction of the corresponding column in $\boldsymbol{W}$.

The normalization of the denominator can be achieved by setting $\boldsymbol{W^T S W} = \mathbf{1}$, thereby converting the maximization of $\boldsymbol{\Lambda}$ into a quadratic programming problem with an equality constraint. The problem can be formulated as:

$$\begin{aligned} \boldsymbol{obj.}\ & \arg\max_{\boldsymbol{W}} \boldsymbol{W^T A W}, \\ & \boldsymbol{s.t.}\ \boldsymbol{W^T S W} = \mathbf{1}. \end{aligned} \tag{3.6}$$

A lagrange multiplier can be used to solve this quadratic programming problem [82,83]. Thus, the generalized eigendecomposition of the covariance matrices $\boldsymbol{A}$ and $\boldsymbol{S}$ provides the solution to Eq. (3.5). The eigenvalue of the generalized eigenvalue problem represents the value of $\boldsymbol{\Lambda}$, and the corresponding eigenvector is the optimal modulation function $\boldsymbol{W}$. In principle, regardless of whether the value of $\boldsymbol{\Lambda}$ is maximized or minimized, the

corresponding $\boldsymbol{W}$ can effectively distinguish between arguing scenarios and sharing happiness scenarios. Hence, the rearrangement of $\boldsymbol{W}$ was conducted based on the value of $\boldsymbol{\Lambda}$,

$$\boldsymbol{W} = [\boldsymbol{\omega_1}, \boldsymbol{\omega_2}, ..., \boldsymbol{\omega_{2i-1}}, \boldsymbol{\omega_{2i}}, ..., \boldsymbol{\omega_{M-1}}, \boldsymbol{\omega_M}]. \tag{3.7}$$

In Eq. (3.7), $\boldsymbol{\omega_{2i-1}}$ represents the modulation function corresponding to the $\boldsymbol{i-th}$ largest eigenvalue, and $\boldsymbol{\omega_{2i}}$ represents that corresponding to the $\boldsymbol{i-th}$ smallest eigenvalue.

3.6 Calculation of social factor

As mentioned before, in this study, "social factor" are defined as the social distance quantified by entrainment. Through the calculations of the interaction between low-level features in the previous section, modulated speech features were obtained that incorporate these interactions. Next, the social factor will be calculated using the three entrainment metrics referenced in previous research: proximity, convergence, and synchrony. These metrics for conversational entrainment have been widely used in previous studies [2, 76].

Originally, proximity, convergence, and synchrony were defined as independent metrics, each capturing a different aspect of the interaction between interlocutors. Specifically:

Proximity measures the distance between the features of two speakers (e.g., how close or far their speech rates or pitch are on average). Synchrony refers to how the features of two speakers change in parallel over time (e.g., if both speakers increase their pitch or speech rate simultaneously). Convergence tracks whether the difference between these features decreases over time, indicating that the speakers are becoming more aligned.

While these metrics are theoretically independent, they may exhibit correlations in practice when measured on real conversational data. For example, high synchrony in speech patterns may coincide with greater proximity as speakers' features converge. In this study, the interaction between these three metrics is considered using modulated speech features, allowing for a more comprehensive representation of social factors in conversation.

Additionally, our study primarily focused on global entrainment by treating the entire dialogue or individual turns as the unit of analysis. Specifically calculated entrainment based on the interaction between interlocutors A and B at the level of turns rather than dividing the conversation into smaller segments, such as the beginning, middle, or end. This approach allowed us to capture the overall alignment between speakers throughout the entire conversation, rather than focusing on the local entrainment or how it might fluctuate at different stages of the conversation.

3.6.1 Proximity

$\boldsymbol{f^A}$ and $\boldsymbol{f^B}$ are defined as the processed data from the dominant speaker and follower speakers, respectively. The proximity between $\boldsymbol{f^A}$ and $\boldsymbol{f^B}$ $\left(\boldsymbol{prox^{A,B}}\right)$ can be measured as the negated absolute difference of the mean values of $\boldsymbol{f^A}$ and $\boldsymbol{f^B}$ [76], that is,

$$-\left|\bar{\boldsymbol{f^A}} - \bar{\boldsymbol{f^B}}\right| \tag{3.8}$$

where $\bar{\boldsymbol{f^A}}$ and $\bar{\boldsymbol{f^B}}$ stand for the mean value of $\boldsymbol{f^A}$ and $\boldsymbol{f^B}$, respectively. When $\left(\boldsymbol{prox^{A,B}}\right)$ is close to zero, $\boldsymbol{f^A}$ and $\boldsymbol{f^B}$ are on average close to each other, while when it is far from zero, they are distant.

3.6.2 Convergence

Convergence $\left(\boldsymbol{conv}^{\boldsymbol{A,B}}\right)$ between $\boldsymbol{f}^{\boldsymbol{A}}$ and $\boldsymbol{f}^{\boldsymbol{B}}$ can be measured as the Pearson correlation coefficient between $-\left|\bar{\boldsymbol{f}}^{\boldsymbol{A}} - \bar{\boldsymbol{f}}^{\boldsymbol{B}}\right|$and time $\boldsymbol{t}$, which can be respectively calculated as Eq. (3.9) and (3.10) [76]

$$\boldsymbol{D(t)} = -\left|\boldsymbol{f}^{\boldsymbol{A}} - \boldsymbol{f}^{\boldsymbol{B}}\right| \tag{3.9}$$

$$\boldsymbol{conv}^{\boldsymbol{A,B}} = \frac{\int_{\boldsymbol{t_{st}}}^{\boldsymbol{t_{end}}} (\boldsymbol{D(t)} - \bar{\boldsymbol{D}}) \cdot (\boldsymbol{t} - \bar{\boldsymbol{t}}) \boldsymbol{dt}}{\sqrt{\int_{\boldsymbol{t_{st}}}^{\boldsymbol{t_{end}}} (\boldsymbol{D(t)} - \bar{\boldsymbol{D}})^{\boldsymbol{2}} \boldsymbol{dt} \cdot \int_{\boldsymbol{t_{st}}}^{\boldsymbol{t_{end}}} (\boldsymbol{t} - \bar{\boldsymbol{t}})^{\boldsymbol{2}} \boldsymbol{dt}}} \tag{3.10}$$

where $\bar{\boldsymbol{D}}$ and $\bar{\boldsymbol{t}}$ denote the mean values of $\boldsymbol{D(t)}$ and $\boldsymbol{t}$, respectively. Positive/negative values of this metric indicate that $\boldsymbol{f}^{\boldsymbol{A}}$ and $\boldsymbol{f}^{\boldsymbol{B}}$ become closer to/further apart from each other as the conversation proceeding.

3.6.3 Synchrony

Synchrony between $\boldsymbol{f}^{\boldsymbol{A}}$ and $\boldsymbol{f}^{\boldsymbol{B}}$ $\left(\boldsymbol{sync}^{\boldsymbol{A,B}}\right)$ can be measured as the Pearson correlation coefficient between $\boldsymbol{f}^{\boldsymbol{A}}$ and $\boldsymbol{f}^{\boldsymbol{B}}$. The calculation of $\left(\boldsymbol{sync}^{\boldsymbol{A,B}}\right)$ is presented in Eq. (3.11) [76]

$$\boldsymbol{sync}^{\boldsymbol{A,B}} = \frac{\int_{\boldsymbol{t_{st}}}^{\boldsymbol{t_{end}}} \left(\boldsymbol{F}^{\boldsymbol{A}}(\boldsymbol{t})\right) \cdot \left(\boldsymbol{F}^{\boldsymbol{B}}(\boldsymbol{t})\right) \boldsymbol{dt}}{\sqrt{\int_{\boldsymbol{t_{st}}}^{\boldsymbol{t_{end}}} \left(\boldsymbol{F}^{\boldsymbol{A}}(\boldsymbol{t})\right)^{\boldsymbol{2}} \boldsymbol{dt} \cdot \int_{\boldsymbol{t_{st}}}^{\boldsymbol{t_{end}}} \left(\boldsymbol{F}^{\boldsymbol{B}}(\boldsymbol{t})\right)^{\boldsymbol{2}} \boldsymbol{dt}}} \tag{3.11}$$

where $\boldsymbol{F}^{\boldsymbol{A}}(\boldsymbol{t}) = \left(\boldsymbol{f}^{\boldsymbol{A}}(\boldsymbol{t}) - \bar{\boldsymbol{f}}^{\boldsymbol{A}}\right)$, $\boldsymbol{F}^{\boldsymbol{B}}(\boldsymbol{t}) = \left(\boldsymbol{f}^{\boldsymbol{B}}(\boldsymbol{t}) - \bar{\boldsymbol{f}}^{\boldsymbol{B}}\right)$. Positive values of $\left(\boldsymbol{sync}^{\boldsymbol{A,B}}\right)$ indicate that $\boldsymbol{f}^{\boldsymbol{A}}$ and $\boldsymbol{f}^{\boldsymbol{B}}$ behave in synchrony with each other, while negative values indicate the opposite directions.

Traditionally, these measures were determined separately on different speech features. However, what sets our method apart from previous research is that these measures are determined on the basis of the modulated

speech features, taking into account the interaction among multiple speech features. Therefore, our method can better characterize the social aspects in conversation.

3.7 Features extraction and data preprocessing

3.7.1 Acoustic features extraction

According to previous research [11, 76], the VOICESAUCE toolkit [84] was employed to extract the following speech features: fundamental frequency (F0), energy, harmonic-to-noise ratio (HNR), and subharmonic-to-harmonic ratio (SHR). These studies also examined voice quality, as measured through jitter, shimmer and found these acoustic features to potentially serve in different interactive strategies [85–87]. Therefore, the jitter and shimmer features were also extracted. Additionally, considering that envelope information can reveal the rhythmic features of speech [88], an envelope extraction was performed using a window length of 25 ms and a frameshift of 1 ms for each dialogue. This study also incorporated parameters reflecting the speakers' perception process: loudness, sharpness, roughness, and fluctuation, these parameters are extracted using MATLAB functions (The MathWorks, Inc., 2023) to provide insights into psychological auditory perceptual features. These attributes play a pivotal role in understanding how individuals perceive and interpret sound, thus influencing their subjective listening experiences. Both acoustic and perceptual features are assessed at the turn level within dialogues. Each feature is determined at the turn level in a dialogue, and z-score normalization is applied to ensure that the data have an average value

of zero.

3.7.2 Linguistic (semantic) features extraction

According to Fig. 3.5, to parameterize the semantics of each turn in the dialogue, a "semantically similarity" was calculated between the same speaker using their two adjacent turns [89]. For quantitatively measuring this kind of semantic distance, The approach utilized the BERT model [90] to represent each turn as a fixed-length vector (768 dimensions in this case). Each element of the vector encoded the semantics of the original turn. Subsequently, the "semantic similarity" of each turn was calculated by comparing Pearson's correlation between the 768-dimensional vector of the current turn with that of the next turn. The correlation provides a semantic similarity measure for each turn. This similarity is calculated for the same speaker using his/her two adjacent turns, turn by turn.

According to the calculation of entrainment, Fig. 3.4 shows examples of conversations with high (a) and low convergence (b), and with high (c) and low synchrony (d). For high convergence, two semantic curves of the speakers gradually become more and more similar during a dialogue, while the high synchrony shows that speakers are consistently behaving in a similar way.

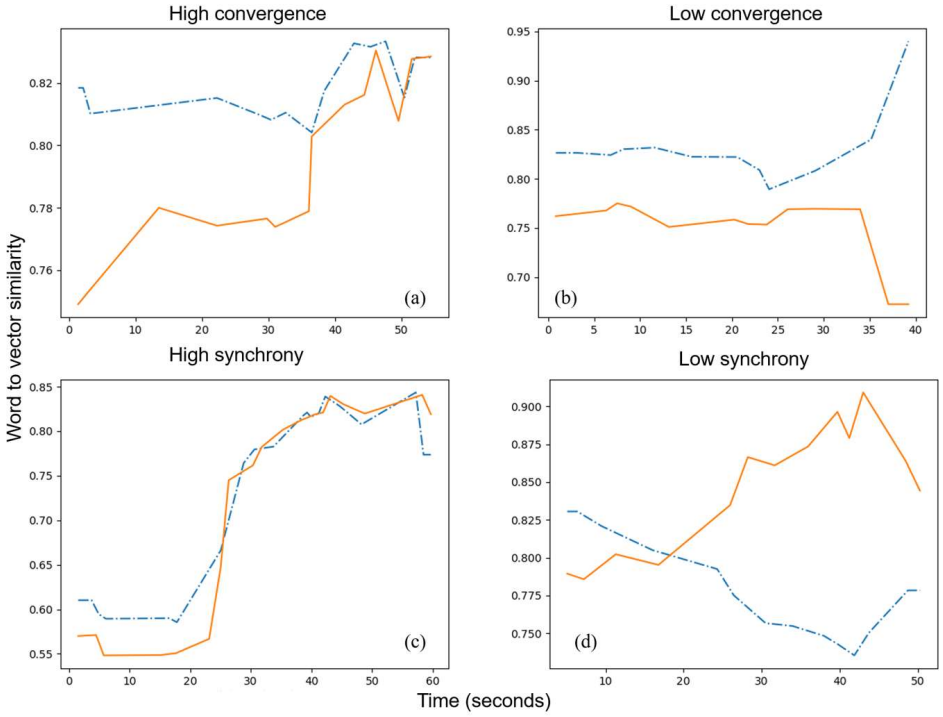

Figure 3.4: Sample conversations with different values of the semantic entrainment metrics: (a) shows high convergence of semantic entrainment, (b) shows low convergence of semantic entrainment, (c) shows high synchrony of semantic entrainment, and (d) shows low synchrony of semantic entrainment.

To ensure data alignment for calculating social factor between two interlocutors, preprocessing of the speech feature data is necessary [43, 91]. In dialogues, interlocutors A and B may have a different number of turns, resulting in different time and sample point counts for the extracted speech features from their voices. To address this misalignment, resampling techniques are utilized to adjust the number of sampling points, ensuring alignment between interlocutors A and B on their speech features.

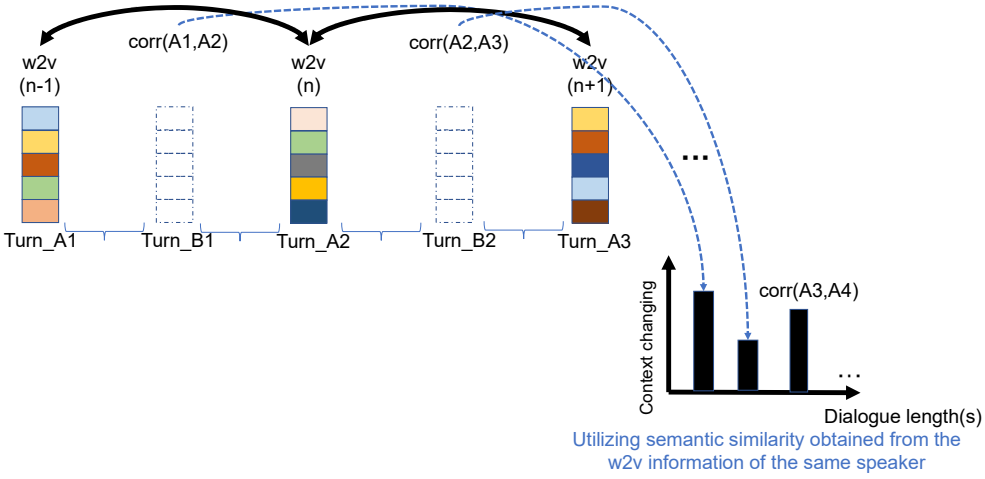

Figure 3.5: Method to extract semantic features.

3.8 Summary

This chapter outlines the methodology employed to investigate how social factors influence conversational entrainment. The research hypothesis posits that interlocutors modulate their speech features according to their communicative goals, influenced by social factors. By examining various conversational contexts such as arguing, comforting, convincing, and sharing happiness, the study aims to capture these modulations and predict future conversational situations.

The methodology integrates Communication Accommodation Theory and Interactive Alignment Model to create a linear model that encodes social factors from dialogue situations. This model focuses on three key metrics—proximity, convergence, and synchrony—to quantify the social factor, reflecting the alignment between interlocutors' speech features.

Data from two corpora, the "Golden Marriage" corpus and the IEMOCAP dataset, are utilized to evaluate the proposed method. The "Golden Marriage" corpus provides real-life inspired scenarios, while the IEMOCAP dataset offers a range of emotional expressions. By analyzing these data sets, the study demonstrates that the social factor, derived from the interaction of multiple speech features, effectively differentiates and predicts conversational scenarios.

Acoustic and semantic features are extracted using tools like VOICESAUCE and BERT, respectively. The social factor is then calculated based on these features, emphasizing the importance of considering both acoustic and semantic elements in conversational analysis.

Overall, this methodology aims to enhance the understanding of conversational dynamics and improve the prediction of future conversational

situations, thereby contributing to more effective human-human and human-machine interactions.

Chapter 4

Social factor for dialogue scenarios classification

The process of social factor derivation conversation-scenario classification is shown in Fig. 4.1. A subset of dialogues from the corpus was used as the training set to estimate the modulation function that can distinguish between a negative (arguing) scenario and positive (sharing happiness) scenario. Subsequently, the modulation function was applied to unseen test data to calculate the social factor. To assess the performance of the social factor in classifying different conversation scenarios, a supervised support vector machine classifier was utilized for classification tasks, and its classification accuracy was measured. The role of an SVM is to find an optimal hyperplane that maximally separates different classes in the feature space. An SVM is known for its ability to handle high-dimensional data and has been widely used in various research fields [92, 93].

As mentioned in Section 3.5, $\boldsymbol{W}$ is estimated from a 2-class conversation scenario. If extending the modulation function to a multi-class conversation scenario is desired, combining several modulation functions becomes necessary. For example, predicting the modulation function in a four conversation scenarios (such as arguing, comforting, convincing, and sharing happiness), can be represented as:

$$\hat{\boldsymbol{W}} = [\boldsymbol{W}_{\mathrm{A}}^{\boldsymbol{T}}, \boldsymbol{W}_{\mathrm{Comf}}^{\boldsymbol{T}}, \boldsymbol{W}_{\mathrm{Conv}}^{\boldsymbol{T}}, \boldsymbol{W}_{\mathrm{S}}^{\boldsymbol{T}}]. \tag{4.1}$$

In this case, the modulation function consists of multiple vectors, each representing a specific conversation scenario. For example, $\boldsymbol{W}_{\mathrm{A}}^{\boldsymbol{T}}$ denotes the modulation function that distinguishes the arguing scenario from other scenarios. Combining these individual modulation functions captures the modulatory effects of different dialogue strategies on various conversation scenarios. The resulting modulation function $\hat{\boldsymbol{W}}$ enables us to effectively modulate and integrate the speech features in a multi-class setting, providing a comprehensive representation of conversational entrainment.

Through this analysis, the aim was to validate the effectiveness and versatility of the proposed method in capturing important aspects of conversations and its potential applicability to various conversational tasks.

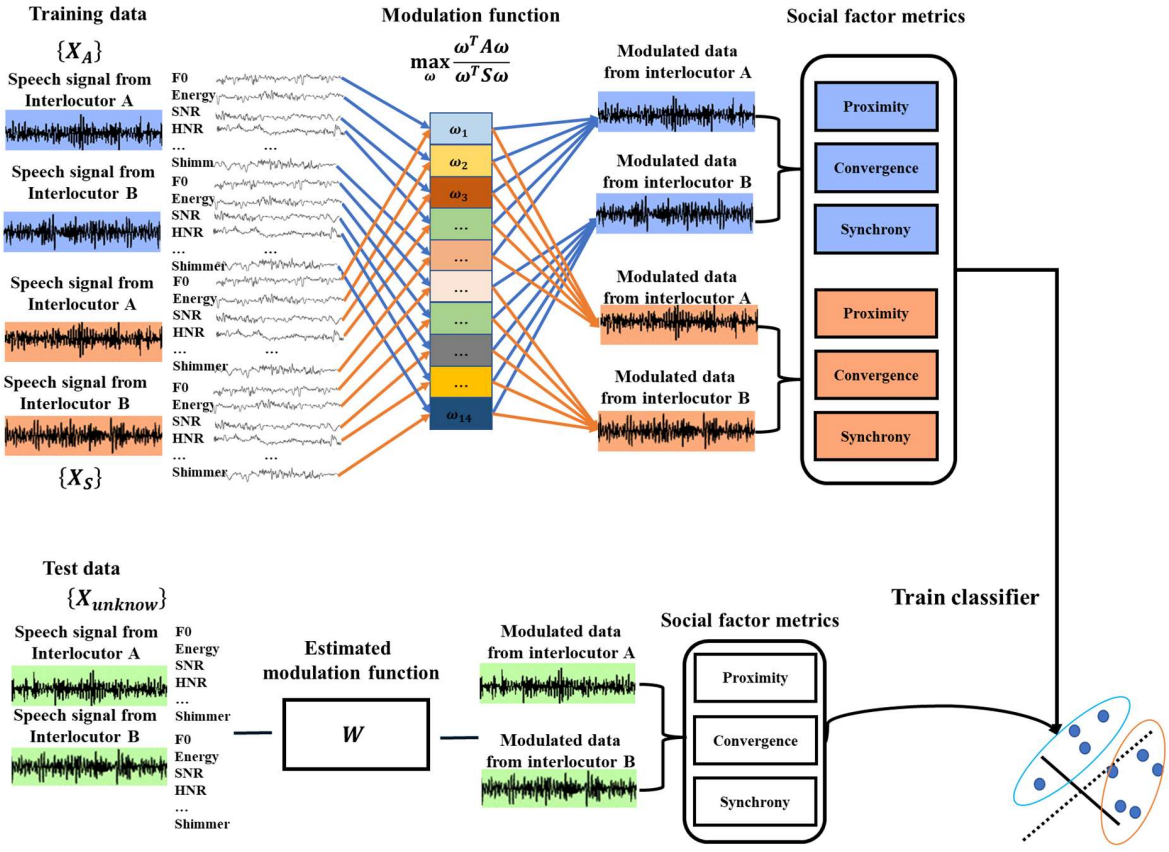

Figure 4.1: Procedure for modulation function estimation and social factor classification.

As mentioned in Section 3.2.3, 193 dialogues of the four scenarios were selected from "Golden Marriage" TV drama. The average length of a dialogue was about 75 s, and on average, a dialogue contained 11.75 turns as show in Table 6.1.

4.1 Estimated modulation function in different conversation scenarios

Based on the linear model introduced in Section 3.5, the modulator functions are estimated. According to our hypothesis, different conversation scenarios have distinct modulator functions, while similar conversation scenarios generate similar modulators. Hence, even modulator functions estimated from different training datasets exhibit commonalities in the same conversation scenarios. To investigate this, 100 random estimations of the modulator functions were conducted using different data samples. Each estimation involved randomly selecting 50% of the data from the "Golden Marriage" corpus to estimate the modulator function.

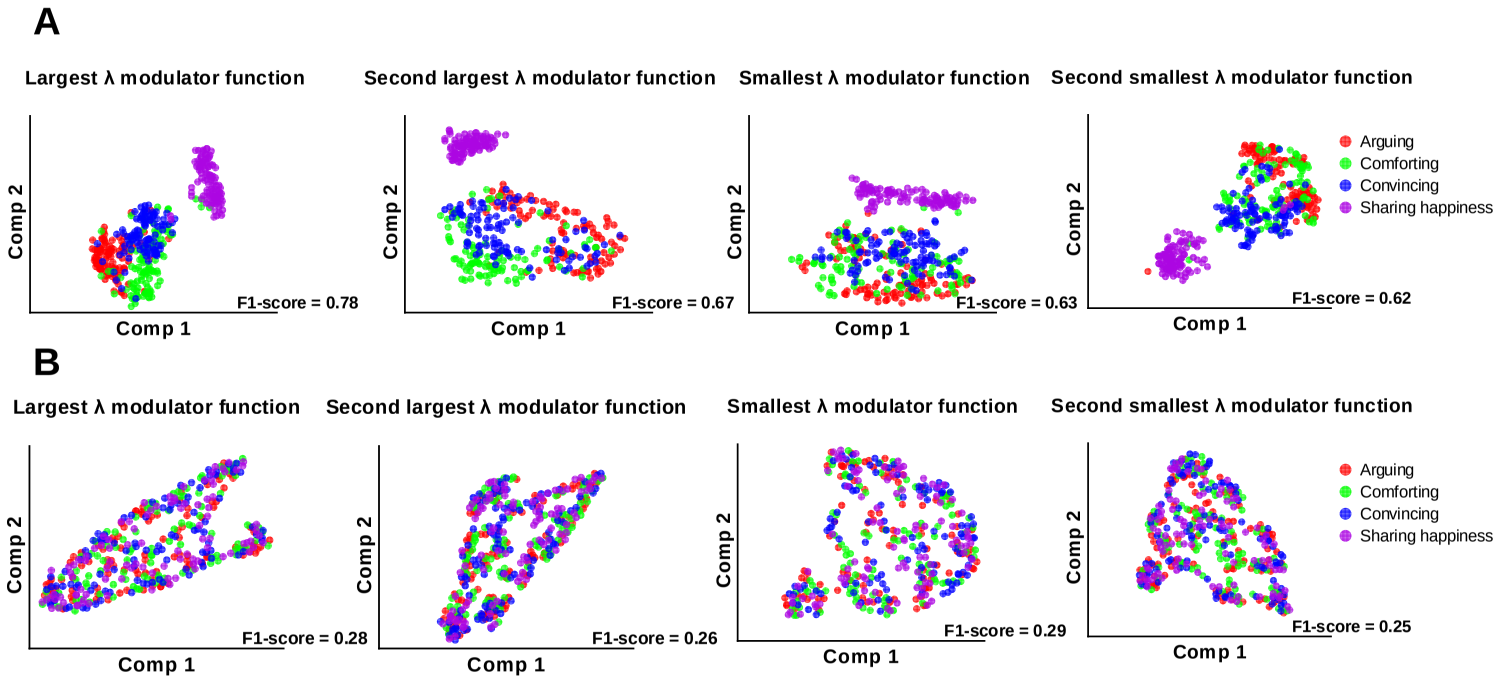

Figure 4.2: Modulator functions in different conversation scenarios. (A) illustrate the modulator functions estimated for different conversation scenarios and (B) show the modulator functions resulting from the permutation test.

To provide a visual representation of the results, t-SNE was used to visualize the modulators in a two-dimensional space [94]. Compared to other dimensionality reduction methods, which focus on preserving global variance in a linear manner, T-SNE is better at capturing non-linear relationships between features. In conversational data, these relationships are essential as they reflect the dynamic and subtle interactions between interlocutors. T-SNE allows us to visualize distinct clusters of conversational behaviors that emerge from these complex relationships, making it more suitable for this type of analysis.

Additionally, a permutation test [95]was conducted to support the findings as a contrast condition. In this permutation test, the labels in the corpus were randomly shuffled, and the shuffled-label data was used to estimate the modulator functions.

The modulator functions estimated for different conversation scenarios are illustrated in Fig. 4.2A. The modulator functions resulting from the permutation test are shown in Fig. 4.2B, from left to right, the figures depict the modulator function corresponding to the first and second largest/smallest $\boldsymbol{\lambda}$ values. Each point in the figure represents the result of one estimation of the modulator function for each conversation scenario. There are 100 points for each conversation scenario. The colors of the different points represent different conversation scenarios. Subsequently, the results of the 2-D t-SNE projection were clustered using the k-means algorithm [96]. The cluster number was set to 4, and 1,000 repetitions of k-means with random initial states were carried out.the clusters represent different conversational scenarios influenced by various social factors. The F1-score shown on the T-SNE visualization helps evaluate the clustering performance. The closer the F1-score is to 1, the better the clustering result, indicating that the

conversation data points within each cluster are more accurately grouped based on their social and speech features. In this study, after applying T-SNE, we calculated the F1-score to assess the quality of the clustering. A higher F1-score suggests that the social factors driving the clustering are well-represented, meaning that similar conversation scenarios are grouped together more effectively.

In contrast, when we shuffled the labels to create a random distribution, the F1-score dropped significantly. This drop demonstrates that the original clustering is not random but reflects meaningful patterns in the data. Therefore, the combination of T-SNE and the F1-score provides strong evidence that the clusters are valid and reflect distinct conversational dynamics. These results indicate that our linear model can capture certain social rules during conversations. When the labels are shuffled, it becomes impossible to observe these patterns from the shuffled data. This observation supports the hypothesis that interlocutor speech features are modulated by different dialogue styles. It also indicates the successful estimation of these modulator functions using the linear model.

4.2 Accuracy in classifying conversation scenarios

The estimated modulation functions were applied to extract modulated speech features from the test dataset, and the social factor was derived accordingly. The accuracy of classifying conversation scenarios using social factor is illustrated in Fig. 4.3.

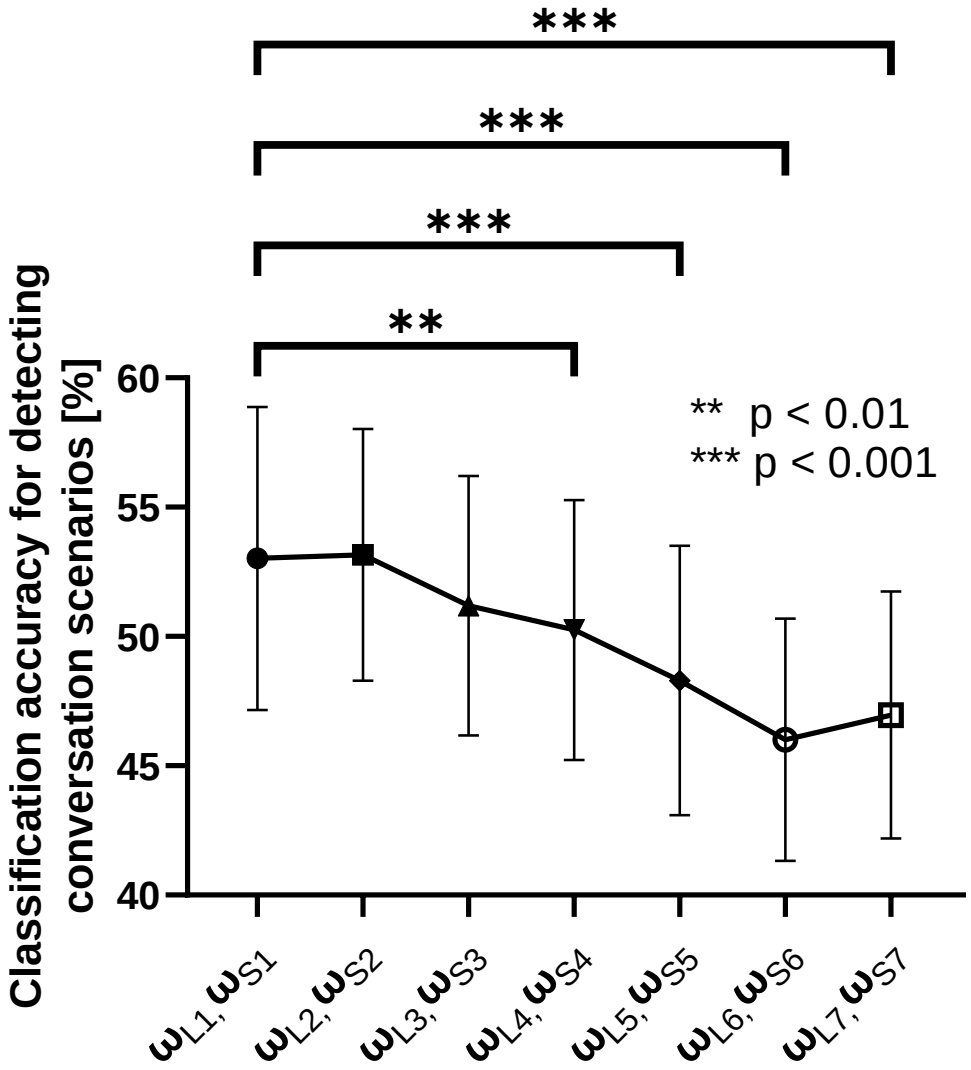

Figure 4.3: Combination of social factor from different modulation functions used for conversation-scenario classification

The "Golden Marriage" dataset was augmented by segmenting the conversations into three parts: the initial third, middle third, and final third of each conversation. The labels for each segment remained the same as the original conversation-scenario labels. To ensure result stability, 100 random training and testing iterations were conducted. Given the significant variability in the number of conversation-scenario labels in the dataset, consistency was maintained by ensuring an equal number of testing samples for each label in every classification iteration. To tackle the imbalance in the number of conversations across different labels, approximately 50% of the total data was constituted by randomly selecting 15 conversations from the label with the fewest instances. These selected conversations were used as the testing set. The remaining 50% of conversations were used as the training set, ensuring that each label had equal representation for a balanced evaluation of the model's performance across different conversation scenarios. Along the x-axis, the classification results present the combination of the social factor extracted from different modulation functions. From left to right, $\boldsymbol{\omega_{L1}}$ and $\boldsymbol{\omega_{S1}}$ represent the combination of the social factor extracted from the modulation functions corresponding to the largest and smallest $\boldsymbol{\lambda}$, respectively. This is followed by the combination of the social factor extracted from the modulation functions corresponding to the the largest and smallest and second largest and smallest $\boldsymbol{\lambda}$. The rightmost $\boldsymbol{\omega_{L7}}$ and $\boldsymbol{\omega_{S7}}$ represent the combination of the social factor extracted from all modulation functions. The social factor from the largest second largest and smallest and second smallest modulation functions demonstrated better classification results. However, when combining more social factor (after the fourth largest and smallest modulation functions), there was a significant decrease in the accuracy of classifying conversation scenarios ($\boldsymbol{F = 30.6, p < 0.001}$). These social

factor from the fourth largest and smallest modulation functions appeared to have little modulation effect on the speech information during the conversation process. They exhibited social aspects that were unrelated to the types of conversation scenarios and could be present in any conversation scenario.

For the task of four-class dialogue scenario classification, using conventional original speech features, the proposed social factor, and the combination of social factor and conventional original speech features separately, the highest classification accuracy of 71.0% when using.

4.3 Summary

In this chapter, the process of deriving social factors for classifying conversation scenarios using the "Golden Marriage" corpus was outlined. A subset of dialogues was utilized to train a modulation function capable of distinguishing between different conversation scenarios, such as arguing and sharing happiness. This modulation function was then applied to unseen test data to calculate the social factor. A supervised support vector machine classifier was used to assess the performance of the social factor in classifying various scenarios, demonstrating its ability to handle high-dimensional data effectively.

The approach was extended to multi-class conversation scenarios by combining multiple modulation functions, enabling a comprehensive representation of conversational entrainment. The data-selection process for the "Golden Marriage" corpus involved criteria such as video quality, dialogue length, and content, ensuring the collection of high-quality dialogues representing four specific scenarios: arguing, comforting, convincing, and sharing happiness. Annotators manually transcribed and annotated each dialogue,

resulting in a corpus of 193 dialogues with detailed multi-layer annotations.

Our analysis included estimating modulation functions for different conversation scenarios and visualizing these functions using t-SNE. The results showed distinct clusters for different scenarios, supporting our hypothesis that interlocutor speech features are modulated by different dialogue styles. Furthermore, the accuracy of classifying conversation scenarios using social factors was evaluated, revealing that the combination of social factors from specific modulation functions contributed to higher classification accuracy.

Overall, this chapter, based on the "Golden Marriage" corpus, utilizes the proposed linear method grounded in IAM theory to integrate different acoustic parameters in a top-down approach to study entrainment phenomena. The modulation function of this linear model demonstrates that different conversational strategies exist across various scenarios. Interlocutors use these strategies to choose whether to entrain or dis-entrain with each other, which answers Research Question 1:

- Do these quantified "social factors" exhibit different patterns in different conversational situations?

  The answer is yes—these quantified "social factors" do exhibit different patterns in different conversational situations.

Additionally, this research believe that the entrainment metrics, which comprehensively account for multiple acoustic features, can effectively help us distinguish between different conversational scenarios. This answers Research Question 2 and also proves that our method successfully bridges psychological theory with engineering approaches:

- Can these "social factors" help classify conversational situations?

  The answer is yes—these "social factors" can help classify conversa-

tional situations.

Chapter 5

Social factors for predicting future conversational situations

The IEMOCAP database is a widely used corpus for studying emotional expressions in human language and nonverbal behavior. It was created by researchers at the University of Southern California. The corpus consists of recorded movie dialogues from 10 different actors, amounting to over 12 hours of data. These dialogues were recorded in an interactive manner, with two actors performing based on given emotional scripts. High-quality microphones and cameras were used during the recording process to capture both audio and video data simultaneously.

The dialogues cover various emotional states, including anger, happiness, neutrality and sadness. These emotions are annotated using rating scales, providing quantitative information about the expressed emotional states in each sentence. Each dialogue is annotated with voice features, basic emotion labels, and manually marked emotion transition points. Based on our hypothesis, the influence of social factor leads to consistent conversational entrainment patterns within similar conversation scenarios. These patterns would results in entrainment or disentrainment between interlocutors. Consequently, the future trajectory of interlocutors in the same dialogue scenario can be predicted using these patterns. In this corpus, four emotion labels

were used to denote the trajectory of speakers in specific dialogue scenarios, and the social factor was used to predict these labels.

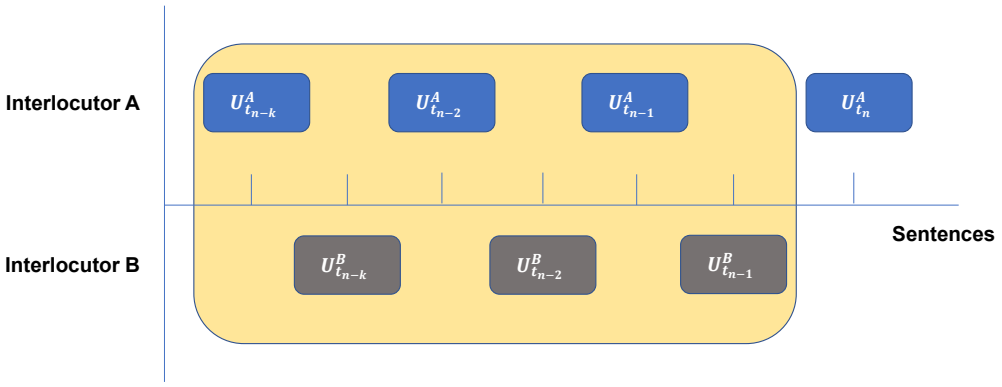

Figure 5.1: Definition of trajectory of conversational entrainment

The approach for labling the experimental data is illustrated in Fig. 5.1 [97]. The figure presents the definition of the context sequence in a conversation. There are two interlocutors in the given conversation, $\boldsymbol{U_A}$ represents the utterances of speaker A, $\boldsymbol{U_B}$ represents the utterances of speaker B, and $\boldsymbol{t}$ denotes the number of turns in the conversation. When taking one speaker alone into account, for instance, speaker A, the context information for speaker A can be denoted as $(\boldsymbol{U^A_{t_{n-k}}, \ldots, U^A_{t_{n-2}}, U^A_{t_{n-1}}})$. When taking both interlocutors into account the interactive context information would become $(\boldsymbol{U^A_{t_{n-k}}, U^B_{t_{n-k}}, \ldots, U^A_{t_{n-2}}, U^B_{t_{n-2}}, U^A_{t_{n-1}}, U^B_{t_{n-1}}})$. On the basis of the sentences labeled as 'happy' at the $\boldsymbol{U^A_{t_n}}$, the trajectory of the conversational entrainment of $\boldsymbol{U^A_{t_{n-k}}}$ and $\boldsymbol{U^B_{t_{n-k}}}$ associated with positive (happy) scenarios can be predicted. To maintain consistency with the corpora used in this study in terms of specifications and duration, this study set t=**4**, meaning that we used a total of eight turns—four from interlocutor A and four from interlocutor B—as the input data. The statistical results of these conversations are listed in Table 5.1. In this corpus, the trajectory of conversational entrainment involves four patterns, conflict (angry), unhappiness (sad), happiness (happy), or neutral.

5.1 Estimated modulation functions for different future conversational situations

Similar to the approach outlined in Section 4.1, the estimation of modulation functions for predicting different future conversational situations in the IEMOCAP database was conducted. Figure 5.2A illustrates the estimated modulation functions for these future situations, and Fig. 5.2B shows the modulation functions resulting from the permutation test. Each figure

Table 5.1: Statistics of IEMOCAP database (history information:8 turn)

	Number of dialogue	Duration/per dialogue (second)	Number of turns/per dialogue	Total duration (minute)
Angry	**909**	**3.68**	**8**	**55.7**
Sad	**645**	**4.01**	**8**	**43.8**
Neutral	**1143**	**3.84**	**8**	**67.8**
Happy	**3812**	/	**8**	**240.5**

represents the modulation function corresponding to the largest/smalles and second largest/smallest $\boldsymbol{\lambda}$ values. Each point represents the result of one estimation of the modulation function for each conversational entrainment patterns. The colors of the points represent a different future situations during the conversation. The k-means algorithm was also applied to cluster the modulation functions obtained from the 2-D t-SNE projection, repeating the clustering process 1,000 times. The modulation functions also displayed distinct clusters for different future situation of conversations. The average F1-scores for the 1,000 repetitions exceeded 0.79, significantly surpassing the chance level observed in the permutation test.

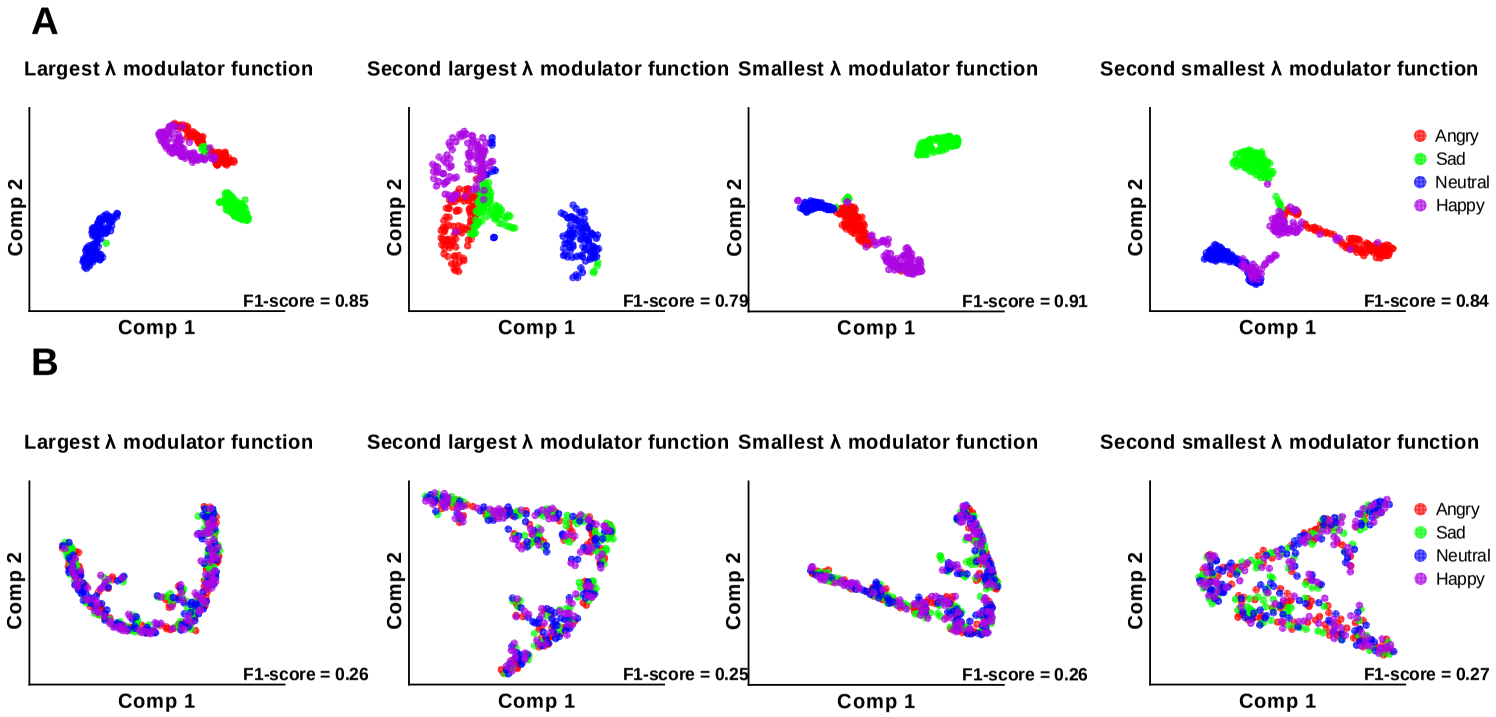

Figure 5.2: modulation functions for future conversational situations. (A) illustrate the modulator functions estimated for different future conversational situations and (B) show the modulator functions resulting from the permutation test.

The classification accuracy of predicting the future conversational situations using the social factor is illustrated in Fig. 5.3. Similar to the "Golden Marriage" corpus, 100 random training and testing iterations were conducted to ensure result stability. In each iteration, 300 conversations from different labels were randomly selected as the testing set, while the remaining conversations were used as the training set. Additionally, an ANOVA [98] was conducted to examine which modulation functions' social factor contributed to higher classification accuracy ($\boldsymbol{F = 299.9, p < 0.001}$). The social factor from the five largest and smallest modulation functions demonstrated better classification results for predicting the future conversational situations (average classification accuracy is 48.8%).

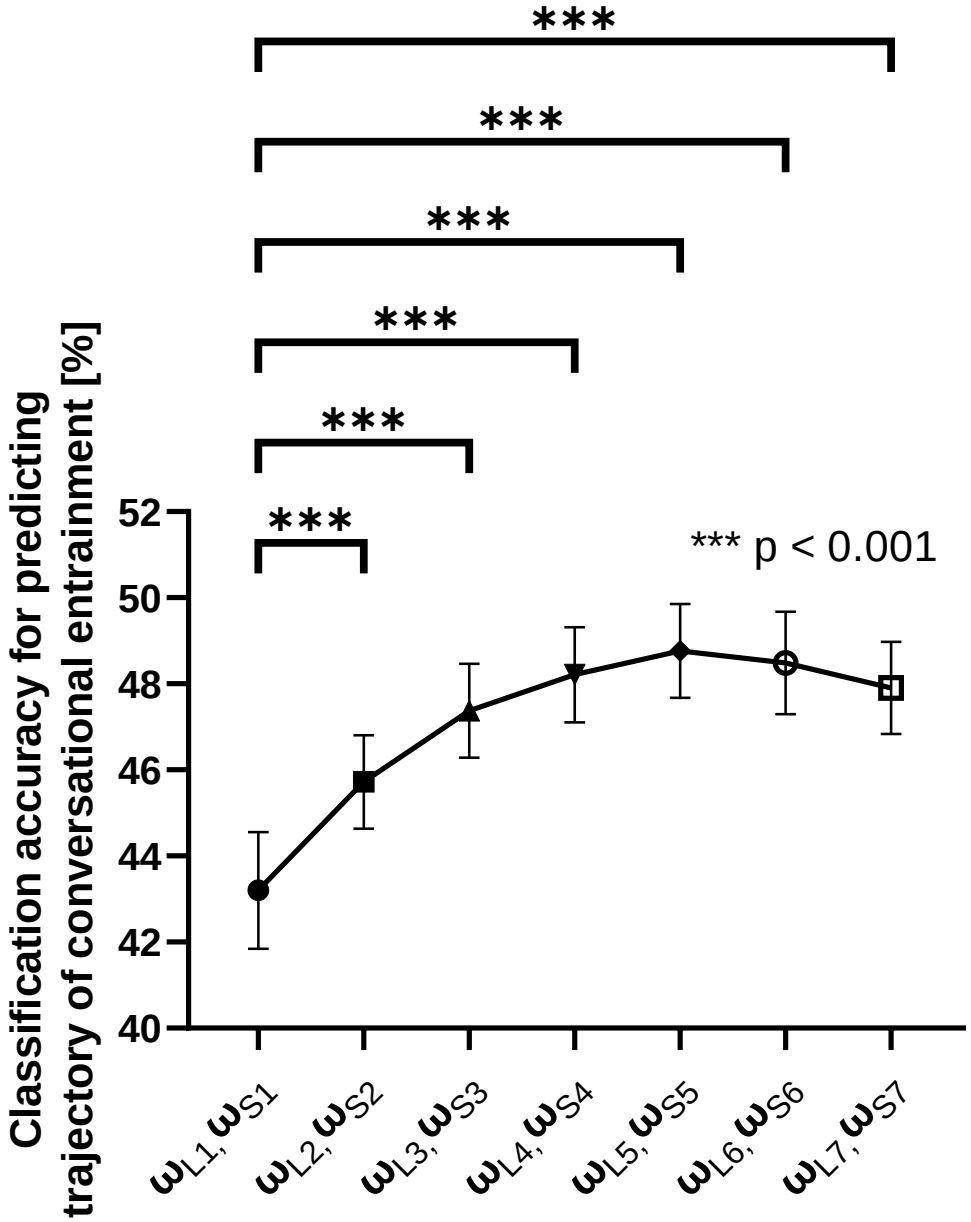

Figure 5.3: Combination of social factor from different modulation functions used for the future conversational entrainment patterns

For predicting the four conversation scenarios, the highest four-class classification accuracy of 56.0% was achieved when using the combination of conventional original speech features and the proposed social factor together.

5.2 Summary

In this chapter, the process of predicting the trajectory of conversational entrainment using social factors derived from the IEMOCAP database was described. The IEMOCAP database is a comprehensive corpus that includes over 12 hours of recorded movie dialogues from 10 different actors, annotated with various emotional states such as anger, happiness, neutrality, and sadness.

The hypothesis posited that social factors prompt consistent conversational entrainment patterns within similar conversation scenarios, enabling the prediction of future trajectories of interlocutors in the same dialogue. Four emotion labels—anger, sadness, neutrality, and happiness—were employed to denote the future situation of speakers in specific dialogue scenarios.

The methodology involved estimating modulation functions for different future situation of conversations. The estimated modulation functions were visualized using t-SNE, and distinct clusters for different conversational situations were observed. This clustering was further supported by a k-means algorithm, which showed that the modulation functions could effectively differentiate between various future situation of conversations.

Additionally, the accuracy of predicting future conversational situations using social factors was evaluated. Across 100 random training and testing iterations, it was observed that the social factors derived from the five largest and smallest modulation functions yielded the best classification results,

averaging 48.77% accuracy. An ANOVA test confirmed the significant contribution of these modulation functions to the classification accuracy.

Overall, this chapter, based on the IEMOCAP corpus, demonstrates through the proposed linear modulation function that different emotional scenarios in English also exhibit distinct strategies. Moreover, these strategies can be used to predict future emotional (or conversational) trends, proving that our proposed method is not only effective for the Chinese corpus but also works well for the English corpus. This suggests that our method has the potential to be applicable in cross-linguistic systems in the future. These findings address the following research questions:

- Do these quantified "social factors" exhibit different patterns in different conversational situations?

  The answer is yes—these quantified "social factors" exhibit different patterns in different conversational situations.
- Can these "social factors" help predict future conversational situations?

  The answer is yes—these "social factors" can help predict future conversational situations.

Chapter 6

Discussion

6.1 High level modulation structure during conversation

In previous studies, researchers shared the same hypothesis as ours, which suggests that there is an interaction among different features in conversational entrainment. This hypothesis motivated them to search for a high-level structure that could explain conversational entrainment by considering this type of interaction that spans multiple speech features. They aimed to move beyond basic entrainment measures of individual speech features and focus on a more comprehensive analysis of conversational entrainment. To explore the underlying patterns of this interaction, researchers employed various methods, including examining linear correlations between pairs of conversational entrainment across different speech features, conducting clustering analysis, and using PCA [77, 78]. However, their findings contradicted the initial expectations, suggesting that entrainment across different features exhibits loose connections and should be considered independently. While these results were disappointing, they serve as a valuable reminder to explore alternative methods and techniques for a more accurate analysis and modeling when investigating conversational entrainment.

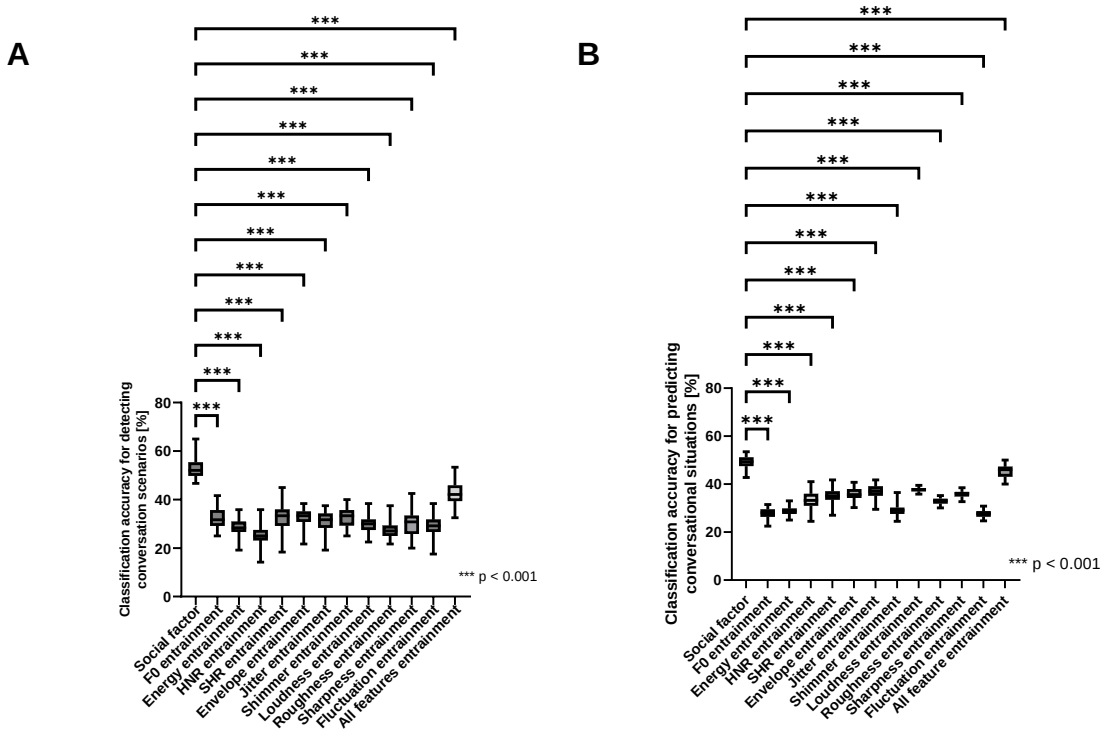

Figure 6.1: Comparison between social factor and independent entrainment metrics. (A) shows the comparison between social factor and independent entrainment metrics for the "Golden Marriage" corpus, and (B) shows the comparison for the IEMOCAP database.

In this study, a high-level modulation structure was introduced to capture the interaction across multiple speech features. This approach is grounded in the Communication Accommodation Theory, which hypothesizes that the speech features of interlocutors are modulated by social factors. By applying a linear model, these high-level social factors were successfully captured and represented, providing a more comprehensive measure of the interaction dynamics in conversations. The results of this study provide strong evidence supporting the validity of the hypothesis and the effectiveness of social factors in predicting conversational outcomes.

Previous findings have suggested that entrainment across different speech features often exhibits loose connections and should be considered independently. To test this, we conducted experiments to compare the performance of models using social factors derived from our proposed method with models using independent speech feature entrainment metrics, which do not consider interactions between multiple features.

The independent speech feature entrainment metrics used in the comparison were calculated based on methods from previous research, where each feature's entrainment was analyzed in isolation, without accounting for the interaction between features. The comparison between social factor and independent entrainment metrics of "Golden Marriage" corpus and IEMOCAP database is illustrated in Figs. 6.1 A and 6.1 B, respectively. Along the x-axis, from left to right, the figure displays the classification accuracy of our proposed method, the method using independent speech feature's conversational entrainment metrics, and the method using the combination of all these independent speech feature's conversational entrainment metrics. An ANOVA demonstrated that the classification accuracy of our proposed social factor was the highest among all metrics in

both "Golden Marriage" corpus ($\boldsymbol{F = 144.9, p < 0.001}$) and IEMOCAP database ($\boldsymbol{F = 225.6, p < 0.001}$). Post-hoc tests for the ANOVA revealed that the prediction accuracy significantly improved with our method, even compared with the combination of all individual conversational entrainment metrics in both "Golden Marriage" corpus ($\boldsymbol{F = 5.53, p < 0.001}$) and IEMOCAP database ($\boldsymbol{F = 23.59, p < 0.001}$).

Overall, this section, based on both the "Golden Marriage" and IEMOCAP corpora, highlights the added value of using a modulation function to integrate multiple acoustic dimensions into social factors. This approach significantly enhances the accuracy and robustness of classification and prediction tasks related to conversational dynamics. Furthermore, it answers Research Question 3:

- Is the proposed entrainment method, which considers relationships between different acoustic parameters in a top-down approach, superior to traditional methods that consider each feature separately?

  The answer is yes—the proposed method, which integrates multiple acoustic parameters in a top-down approach, is superior to traditional methods that consider each feature in isolation.

6.2 social factor: beyond speech features

In this study, social factor derived from speech features are believed to capture the influence of social aspects on conversations. These social factor are considered to contain information beyond the original speech features. To evaluate their effectiveness, speech features were combined with social factor in a conversation-scenario classification task, as shown in Fig. 6.2.

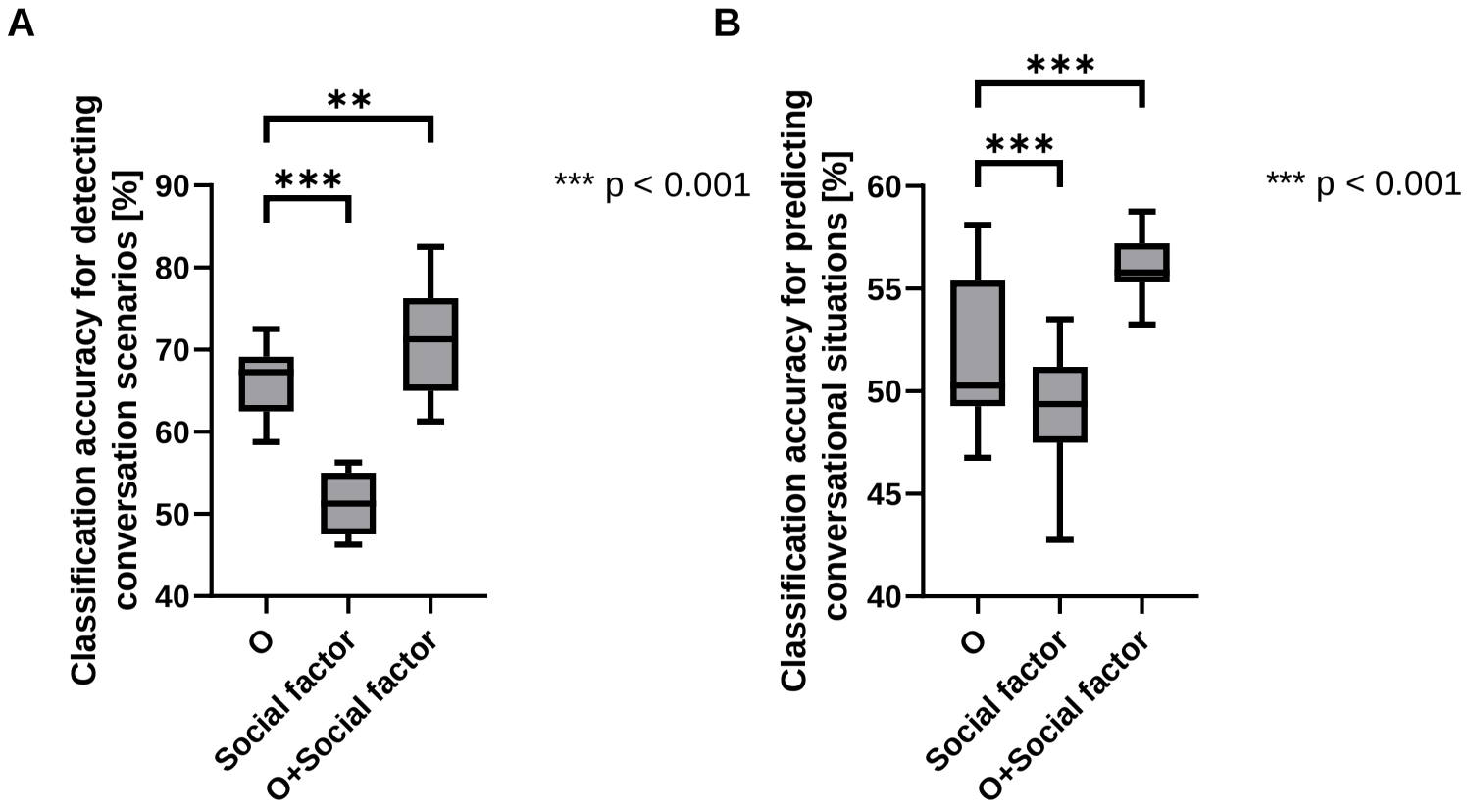

Figure 6.2: Classification results using different speech feature/social factor combinations. (A) shows the results of the "Golden Marriage" corpus, and (B) shows the results of the IEMOCAP database.

The x-axis represents different feature combinations: the conventional original speech features (O), the propsed social factor (Social factor), and the combination of social factor and conventional original speech features (O + Social factor). On the Chinese corpus, O achieved an accuracy of 62.3% in classifying four common conversation scenarios, whereas Social factor achieved an accuracy of 52.9%. On the IEMOCAP database, the accuracy of predicting the future conversational situations using O was 52.0%, while Social factor achieved an accuracy of 48.8%. An overall increase in accuracy of 6.5% on the "Golden Marriage" corpus and 4.0% on the IEMOCAP database was observed with O + Social factor. An ANOVA was conducted to compare the classification results of these feature combinations in detecting conversation scenarios and predicting the future conversational situations. It indicated that speech features and social factor had equal effectiveness in classifying conversation scenarios and predicting conversational situations for both the "Golden Marriage" corpus ($\boldsymbol{p = 0.58}$) and IEMOCAP database ($\boldsymbol{p = 0.24}$). The results indicate that the combination of social factor and speech features achieves the best performance in both the "Golden Marriage" corpus (Fig. 6.2 A) and IEMOCAP database (Fig. 6.2 B), with accuracies of 68.8% and 56.0%, respectively.

However, when comparing the classification results of independent features' entrainments with O, no significant differences were observed in the "Golden Marriage" corpus ($\boldsymbol{p = 0.48}$) and IEMOCAP database ($\boldsymbol{p = 0.52}$). This suggests that the additional features derived from the combined entrainments did not significantly improve classification performance compared with using O. This may indicate that independent features' entrainments did not effectively capture the social aspect information that goes beyond O.

6.3 Comparing proposed method with PCA

In previous studies, researchers attempted to use PCA to determine the interactions between different speech features and represent social factor [99]. PCA is a well-known statistical method used for analyzing multivariate time series data. It is used to carry out an orthogonal transformation of a set of observed variables into a set of uncorrelated variables called principal components. The first component captures the maximum variance of the observed data, and each subsequent component explains the maximum possible variance while being orthogonal to the previous components. To compare the performances of our method and PCA, PCA was applied to derive social factors, and the accuracy of conversation-scenario classification was compared using the PCA-derived social factors and those derived with our proposed method. Figures 6.3A and 6.3B illustrate these accuracy of PCA on the "Golden Marriage" corpus and IEMOCAP database. A Student's t-Test revealed that our proposed method achieved higher classification accuracy compared with PCA (both $\boldsymbol{p < 0.001}$). These results indicate that our proposed method better derives social aspects than PCA.

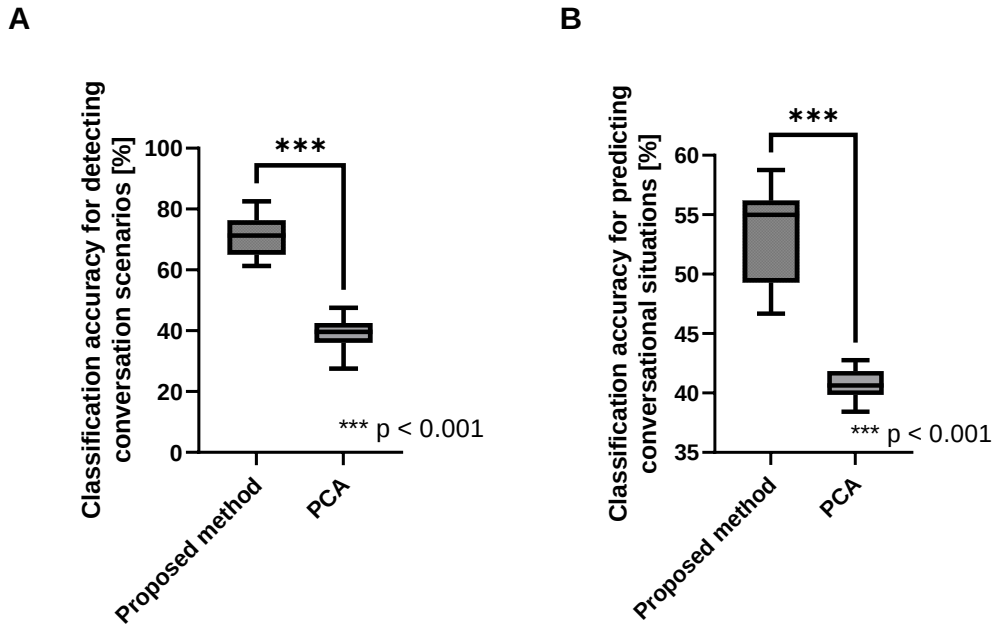

Figure 6.3: Comparison between propsed method and PCA. (A) shows the comparison results for the "Golden Marriage" corpus, and (B) shows the comparison results for the IEMOCAP database.

PCA has limitations in deriving social factor. First, it is descriptive rather than inferential, as it focuses on finding a set of principal components that maximize the variance of the observed data without considering specific reasoning and causal relationships. This means that PCA provides a representation of the data without necessarily explaining the underlying mechanisms and factors driving social interactions. social factor often involve complex interpersonal dynamics and behavioral patterns that require more reasoning and understanding to accurately determine their influences. Second, PCA is used to extract principal components by maximizing the variance of the data. However, in the context of social factor, where maximizing correlation may be more important than maximizing variance,

PCA may fall short. social factor often involve interpersonal relationships and mutual influences, where maximizing correlation between features may be more critical than maximizing variance. Therefore, using PCA alone may not accurately determine the importance and impact of social factor.

In contrast, our proposed method offers an effective hypothesis to describe the relationship between speech features and social factor. Through this method, the correlation between social factors within the same type of conversation scenario can be maximized. Thus, our method is more suitable for deriving and capturing these social factor compared with PCA. Overall, our proposed method takes into account the complex interactions between speech features and social factor, providing a more comprehensive and effective representation.

6.4 Alignment between speech features and lexical features

In previous studies [13, 78, 100], researchers measured various features at both the lexical and speech levels and examined the correlation in alignment between these different features. They specifically investigated whether the degree of alignment between interlocutors on one feature was related to their degree of alignment on the other. This approach aimed to explore the potential interconnectedness and relationships between different linguistic features and their contribution to conversational entrainment. However, it is worth noting that these studies have rarely found significant relationships between speech and lexical features, suggesting limited associations between these two types of features.

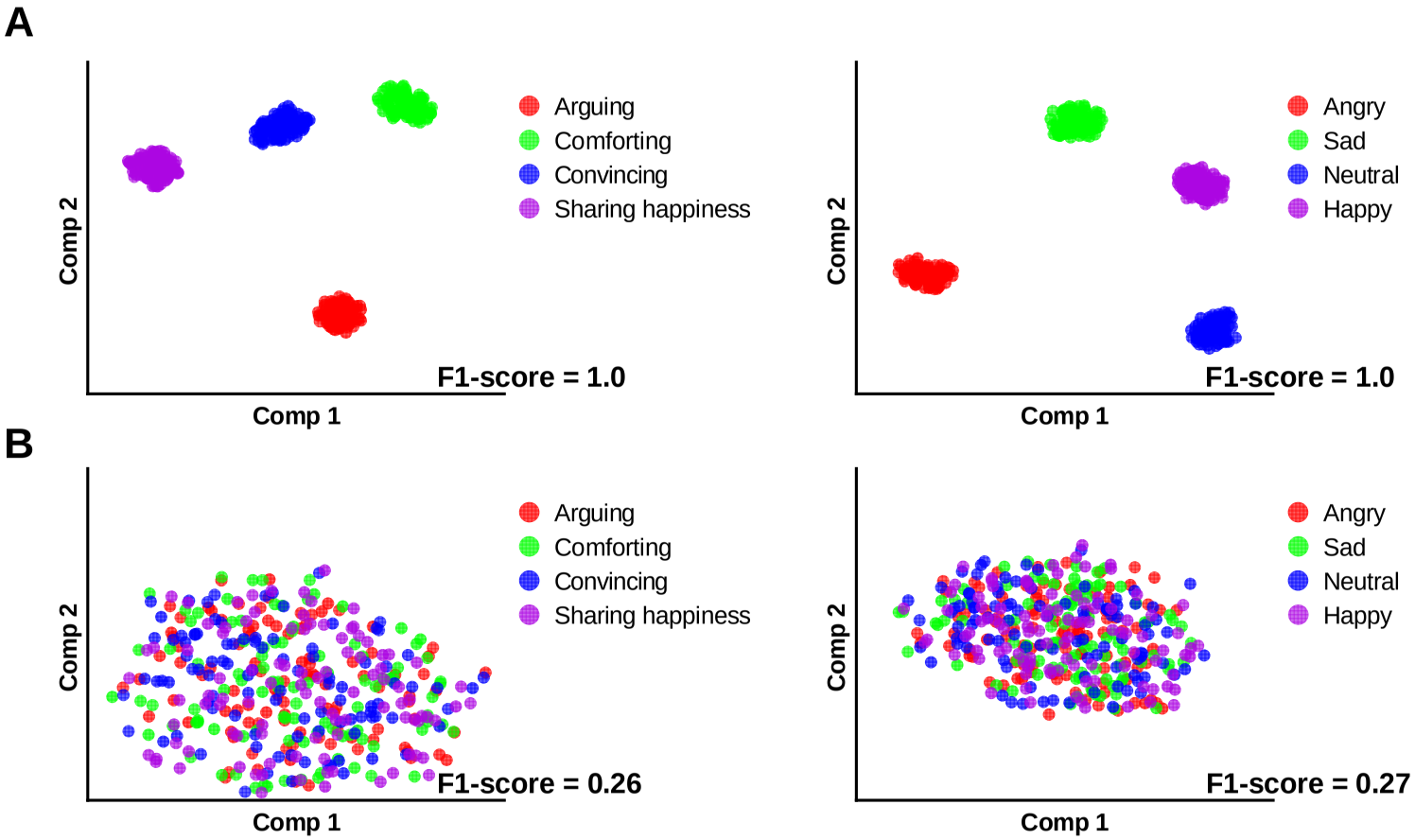

Figure 6.4: Visualization of semantic information in different conversation scenarios and future conversational situations. (A) illustrates the semantic information in different conversation scenarios and future conversational situations. (B) shows the semantic information resulting from the permutation test

In the study, BERT was employed to extract semantic information from conversations at the turn-level [90], serving as a representation of lexical information. During the training of the modulation functions, consideration was given to both the speech features and the semantic information of each turn, taking into account the different interlocutors involved in the dialogue. The extracted semantic information of interlocutors A and B was then transformed using t-SNE into a 2-D space for visualization. This process was repeated 100 times to ensure the robustness and reliability of the results. The visualization results of the semantic information are presented in Fig. 6.4A, where each point represents the reduced representation of the averaged semantic information in each random iteration. Thus, for each type of conversation-scenario and future conversational entrainment trajectories, there are a total of 100 points. The color of each point corresponds to different conversation scenarios and different future conversational entrainment trajectories.

The analysis of semantic information reveals distinct clustering for different conversation scenarios and future conversational entrainment trajectories, similar to the modulation functions based on speech features. However, when conducting a permutation test by shuffling the labels in Fig. 6.4B, the clear classification pattern disappears. This consistency observed in the distribution of both speech and lexical features' modulation functions suggests alignment between different levels of language features in terms of conversation scenario. In different scenarios, distinct strategies are exhibited in terms of acoustic rhythm and word choice, resulting in noticeable differentiations. This indicates that different social aspects guide our communication with others in different conversational contexts.

These findings highlight the importance of considering both speech and

lexical features and their alignment at different levels of language in understanding the dynamics of social interaction in conversations. Therefore, the BERT were utilized to extract word vectors from conversations at the turn level [90], serving as a representation of semantic features. To characterize semantic changes between different turns in the conversation, a semantic similarity method from previous research was utilized to compute the Pearson correlation of word vectors across different turns. [101]. This method enables the depiction of consistency or discrepancy in semantic content between consecutive turns. Significant changes in correlation between turns may occur during pivotal moments such as arguments or other unexpected events within the dialogue.

As previously mentioned, prior research has typically considered the entrainment of various speech features independently. However, during the dialogue process, low-level acoustic features are perceived by both interlocutors. On the basis of this psychological perception, interlocutors organize semantic information to provide active feedback to each other, ensuring smooth communication. Therefore, utilizing our approach, the integration of low-level acoustic features with pragmatic features, which consider the interaction among acoustic and semantic features, may yield improved results for the social factor. Table 6.1 shows the average accuracies of classifying dialogue scenarios and predicting the future conversational situations using social factors which consider pragmatic features. Here, 'A' indicates that the social factor is calculated by only acoustic features, while 'A + W' is derived from acoustic and semantic features, emphasizing their interplay in pragmatics. This underscores the importance of analyzing not only the acoustic features and their psychophysical perception in conversation but also integrating semantic information for a comprehensive understanding of

Table 6.1: Classification results by considering pragmatic features. 'A': combination of acoustic features and social factors derived from acoustic features. 'A + W': amalgamation of acoustic and semantic features combined social factors derived from all of them to consider pragmatic features. 'O' represents the original acoustic features. 'O + E' denotes a combination of traditional conversational entrainment and original acoustic features. 'O + S' means the integration of the proposed social factor and original acoustic features.

Dataset	**Features**	**A**	**A+W**
Golden Marriage	$\boldsymbol{O}$	62.3%	70.0%
	$\boldsymbol{O + E}$	60.5%	64.0%
	$\boldsymbol{O + S}$	67.0%	71.9%
IEMOCAP	$\boldsymbol{O}$	51.3%	52.0%
	$\boldsymbol{O + E}$	51.5%	52.7%
	$\boldsymbol{O + S}$	52.9%	54.2%

their pragmatic functions in human interaction. As shown, with the inclusion of more interlocutors' features to consider pragmatic aspects, the predictive accuracy of dialogue scenarios gradually improves. Upon combining all three types of features, the integrated social factor proposed outperforms the results based solely on acoustic features by 9.6% and 5.2% in the two corpora, respectively.

According to our hypothesis, different conversation scenarios are linked with distinct modulator functions, while similar conversation scenarios produce similar modulators. Consequently, even modulator functions estimated from different training datasets should exhibit commonalities in corresponding conversation scenarios.To investigate this, 40 random estimations of the modulator function were conducted using varying data samples. Each estimation involved randomly selecting 30% of the data from the corpus to

derive the modulator function. It was found that the modulator functions from different scenarios exhibit distinct clustering, indicating that conversational participants utilize different strategies to adapt to various scenarios, aligning with the hypothesis. Furthermore, it is worth noting that with the addition of parameters reflecting pragmatic functions, the differentiation of modulator functions across different scenarios becomes more pronounced. This suggests that integrating pragmatic features allows for a more precise description of the strategies users implement during conversations. These findings underscore the efficacy of the proposed method in capturing social information and describe the pragmatic aspects across different conversation scenarios.

6.5 Summary

This chapter explores the high-level modulation structure during conversation, focusing on the interactions among multiple speech features and their role in conversational entrainment. Previous studies hypothesized such interactions but found weak connections between different speech features, suggesting the need for independent analysis. A method based on Communication Accommodation Theory was introduced, hypothesizing that social factors influence speech features and can be captured using a linear model. The results supported this hypothesis, showing that social factors derived from speech features improve classification accuracy in conversation scenarios and prediction of conversational situations compared to independent speech feature metrics.

The study demonstrated that combining social factors with conventional speech features yields the highest classification accuracy, indicating the

complementary nature of these features. This method outperformed PCA in deriving social aspects, as PCA's focus on variance maximization is less suited for capturing the complex correlations inherent in social interactions.

The alignment between speech and lexical features was also investigated, using BERT to extract semantic information and visualize it in different conversation scenarios. The results revealed distinct clustering, suggesting that both speech and lexical features align with conversation scenarios, supporting the need to consider multiple linguistic levels in social interaction analysis.

Finally, the importance of pragmatic features was highlighted by integrating acoustic and semantic information. This integration enhanced the predictive accuracy of dialogue scenarios, demonstrating the necessity of considering pragmatic aspects for a comprehensive understanding of conversational dynamics. The findings suggest that different conversation scenarios correspond to distinct modulator functions, and incorporating pragmatic features refines the differentiation of these functions, providing a robust method for capturing social information in conversations.

Chapter 7

Conclusion

7.1 Summary

The final objective of this research is to investigate the mechanisms of entrainment influenced by social factors across diverse conversational scenarios. To accomplish this objective, several sub-goals were outlined within this dissertation:

- How to quantify these various social factors?
- Does the quantified "social factors" exhibit different patterns in different conversational situations?
- Can these "social factors" aid in classifying or predicting conversational situations?

Through our research efforts, the aforementioned questions are addressed as follows:

- A1: A linear methodology for quantifying the "social factor" is introduced.
- A2: Yes. The quantified "social factor" demonstrates distinct patterns across varying current and future conversational scenarios.
- A3: Yes. The identified "social factor" exhibits potential in both classifying and predicting conversation situations.

7.2 Contributions

Social factor play a crucial role in human-to-human conversations, yet they are often overlooked in current dialogue systems and human-machine interaction systems. Previous studies have attempted to measure these social aspects using conversational entrainments on the basis of different speech features. However, these studies often failed to consider the interactions between these features, which can lead to conflicting interpretations of social aspects. It is crucial to comprehensively consider the conversational entrainment patterns of these speech features. Our proposed method integrates multiple speech features and captures their intricate interplay, quantifying the social aspects as the "social factor". This approach takes into account the complexities of social aspects in conversations and offers a new perspective to overcome the challenges posed by these complexities. Through the empirical analyses of both Chinese and English corpora, the efficacy of social factors in current conversation-scenario classification tasks and future conversational situation prediction tasks of conversations was demonstrated. These results indicate that social factor plays the same importance as that of the acoustic features, and the combination of social factor and acoustic features improved prediction accuracy, which denotes that social factor and acoustic features have distinguishable information that can compensate for each other.

The implications of our proposed method extend to the development of effective and natural human-machine dialogue systems. Understanding and implementing the mechanisms underlying socially influenced interactions are crucial in the field of human-machine interaction. By incorporating these mechanisms into human-machine systems, machines can learn to perceive and respond to users' cues, resulting in more engaging and interactive experiences.

For example, an advanced dialogue system equipped with an understanding of social dynamics can adapt its responses to re-engage and retain users' interest during conversations where they may feel bored or restless. The ability to observe and respond to users' social cues contributes to the creation of more satisfying interactions. By comprehending the intricate dynamics of socially influenced human-human interactions, researchers and engineers can enhance the development of interactive technologies that better meet the needs and expectations of users. This, in turn, can lead to the creation of more effective and user-friendly human-machine dialogue systems that facilitate the communication and collaboration between humans and machines. While some researchers have used conversational entrainment to predict the success of a conversation [76, 102], there have been few studies applying conversational entrainment to predict more complex tasks related to human-machine interaction. By extending conversational entrainment to social factor, our research highlights the significance of considering these social factor that encompass social aspects in human-machine interaction systems. This should be taken into account for future research.

In conclusion, the contributions of this study include proposing a novel method to quantify social factor in conversations, validating the effectiveness of these social factor, providing new insights into dialogue systems and human-machine interaction, and expanding the perspective of dialogue research. These contributions offer valuable insights into the development of more intelligent and socially-aware dialogue systems in the future.

This study proposed a novel concept of the "social factor" to describe the entrainment phenomenon at the high-level feature. The modulation functions for different conversation scenarios and future conversational entrainment patterns were estimated using two distinct language corpora: the "Golden

Marriage" corpus and the IEMOCAP database. Our results indicate that the estimated modulation functions exhibited distinct clusters for different conversation scenarios and future conversational entrainment patterns, highlighting the effectiveness of our method. The classification accuracy of conversation scenarios and future conversational situation patterns was further evaluated using the derived social factor. The results indicate that the social factor outperformed independent features' entrainments in both tasks. And the proposed social factor exhibit comparable effectiveness to traditional acoustic features in this study. Moreover, the combination of speech features with social factor revealed that social factor encompass valuable information beyond traditional speech features alone. These results are consistent across both the "Golden Marriage" corpus and IEMOCAP database, which underscores the importance of considering social factor in future dialogue systems and human-machine interaction systems.

The findings of this study contribute to our understanding of social dynamics in conversations and have implications for the development of human-machine dialogue systems. By incorporating social factor into dialogue systems, machines can better perceive and respond to users' social cues, leading to more engaging and interactive interactions.

7.3 Remaining works

This study recognizes the importance of examining both global and local entrainment dynamics. While the current focus has been on global entrainment, capturing overall alignment between speakers throughout the entire conversation, we acknowledge that entrainment is a dynamic process that may vary during different phases of a dialogue.

Future work will aim to explore these fine-grained temporal changes by investigating local entrainment—how synchronization strengthens or weakens within specific segments of the conversation. By analyzing these more detailed shifts in alignment, we can gain a deeper understanding of how entrainment evolves over time and how various factors, such as the context or content of the conversation, may influence these dynamics. This exploration will help uncover whether specific phases of a dialogue (e.g., the beginning, middle, or end) exhibit distinct patterns of alignment.

In summary, while the current study focuses on global entrainment, future research will delve into the dynamics of local entrainment to provide a more nuanced understanding of conversational alignment as it develops and changes over the whole dialogue.

While the formulation of the three entrainment metrics in this study captures global synchrony across the entire conversation, previous research in psychology has highlighted the importance of delay-based synchrony, where one speaker's behavior influences the other with a slight lag. This delay reflects the natural time required for interlocutors to process and respond to each other's speech cues. Incorporating such temporal dynamics can provide a more nuanced understanding of conversational alignment, particularly in real-time settings where speech patterns evolve continuously.

Although delay-based synchrony was not the focus of this study, future research could benefit from integrating dynamic synchrony measures that account for varying delays. This approach would allow for a more detailed analysis of turn-taking dynamics and how synchrony develops at different stages of a conversation, offering deeper insights into the timing and flow of interaction.

It is also important to note that only F0, energy, HNR, SHR, envelope,

jitter, shimmer and loudness, sharpness, roughness and fluctuation were considered as the speech features. For future research, it would be valuable to explore the inclusion of additional speech features like boundary tone, the silence and gap in conversation and so on. Furthermore, our exploration of social factor in this study focused on the global level. Future investigations could incorporate more complex convolutional models to capture the dynamic changes of social factor at the local level. Future research can also explore the application of social factor in various domains, such as dialogue generation, sentiment analysis, and user modeling. Investigating the generalizability of social factor across different languages and cultures would provide valuable insights into cross-cultural communication and understanding.

References

[1] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[2] C. J. Wynn and S. A. Borrie, "Classifying conversational entrainment of speech behavior: An expanded framework and review," *Journal of Phonetics*, vol. 94, p. 101173, 2022.

[3] M. J. Pickering and S. Garrod, "An integrated theory of language production and comprehension," *Behavioral and brain sciences*, vol. 36, no. 4, pp. 329–347, 2013.

[4] J. Habermas, *Moral consciousness and communicative action*. MIT press, 1990.

[5] M. Dragojevic, J. Gasiorek, and H. Giles, "Communication accommodation theory," *The international encyclopedia of interpersonal communication*, pp. 1–21, 2015.

[6] G. I. K. Sari, "Social factor and social dimension in terms of speech style," *Language Horizon*, vol. 7, no. 1, 2019.

[7] R. Wardhaugh and J. M. Fuller, *An introduction to sociolinguistics*. John Wiley & Sons, 2021.

[8] M. J. Pickering and S. Garrod, "Alignment as the basis for successful communication," *Research on Language and Computation*, vol. 4, pp. 203–228, 2006.

[9] D. Reitter and J. D. Moore, "Alignment and task success in spoken dialogue," *Journal of Memory and Language*, vol. 76, pp. 29–46, 2014.

[10] Š. Beňuš, "Conversational entrainment in the use of discourse markers," in *Recent Advances of Neural Network Models and Applications: Proceedings of the 23rd Workshop of the Italian Neural Networks Society (SIREN), May 23-25, Vietri sul Mare, Salerno, Italy*. Springer, 2014, pp. 345–352.

[11] R. Levitan, *Acoustic-prosodic entrainment in human-human and human-computer dialogue*. Columbia University, 2014.

[12] R. Ostrand and E. Chodroff, "It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue," *Journal of phonetics*, vol. 88, p. 101074, 2021.

[13] Z. Rahimi, A. Kumar, D. J. Litman, S. Paletz, and M. Yu, "Entrainment in multi-party spoken dialogues at multiple linguistic levels." in *Interspeech*, 2017, pp. 1696–1700.

[14] Š. Beňuš, "Social aspects of entrainment in spoken interaction," *Cognitive Computation*, vol. 6, pp. 802–813, 2014.

[15] S. W. Gregory Jr and S. Webster, "A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions." *Journal of personality and social psychology*, vol. 70, no. 6, p. 1231, 1996.

[16] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," in *Proc. Interspeech 2009*, 2009, pp. 2779–2782.

[17] R. L. Street Jr, "Evaluation of noncontent speech accommodation," *Language & Communication*, vol. 2, no. 1, pp. 13–31, 1982.

[18] H. Giles, N. Coupland, and I. Coupland, "1. accommodation theory: Communication, context, and," *Contexts of accommodation: Developments in applied sociolinguistics*, vol. 1, 1991.

[19] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker, "Language style matching predicts relationship initiation and stability," *Psychological science*, vol. 22, no. 1, pp. 39–44, 2011.

[20] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, no. 6, pp. 419–426, 2013.

[21] S. A. Borrie, T. S. Barrett, M. M. Willi, and V. Berisha, "Syncing up for a good conversation: A clinically meaningful methodology for capturing conversational entrainment in the speech domain," *Journal of Speech, Language, and Hearing Research*, vol. 62, no. 2, pp. 283–296, 2019.

[22] S. Garrod and A. Anderson, "Saying what you mean in dialogue: A study in conceptual and semantic co-ordination," *Cognition*, vol. 27, no. 2, pp. 181–218, 1987.

[23] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic co-ordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13–B25, 2000.

[24] H. P. Branigan, M. J. Pickering, J. F. McLean, and A. A. Cleland, "Syntactic alignment and participant role in dialogue," *Cognition*, vol. 104, no. 2, pp. 163–197, 2007.

[25] A. A. Cleland and M. J. Pickering, "The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure," *Journal of Memory and Language*, vol. 49, no. 2, pp. 214–230, 2003.

[26] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation." *Journal of experimental psychology: Learning, memory, and cognition*, vol. 22, no. 6, p. 1482, 1996.

[27] R. Ostrand and V. S. Ferreira, "Repeat after us: Syntactic alignment is not partner-specific," *Journal of memory and language*, vol. 108, p. 104037, 2019.

[28] H. Friedberg, D. Litman, and S. B. Paletz, "Lexical entrainment and success in student engineering groups," in *2012 ieee spoken language technology workshop (slt)*. IEEE, 2012, pp. 404–409.

[29] A. M. Rosenthal-von der Pütten, L. Wiering, and N. Krämer, "Great minds think alike. experimental study on lexical alignment in human-agent interaction," *i-com*, vol. 12, no. 1, pp. 32–38, 2013.

[30] J. Kejriwal and Štefan Beňuš, "Relationship between auditory and semantic entrainment using Deep Neural Networks (DNN)," in *Proc. INTERSPEECH 2023*, 2023, pp. 2623–2627.

[31] I. Ivanova, W. S. Horton, B. Swets, D. Kleinman, and V. S. Ferreira, "Structural alignment in dialogue and monologue (and what attention

may have to do with it)," *Journal of Memory and Language*, vol. 110, p. 104052, 2020.

[32] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[33] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 11–19.

[34] U. D. Reichel, Š. Beňuš, and K. Mády, "Entrainment profiles: Comparison by gender, role, and feature set," *Speech Communication*, vol. 100, pp. 46–57, 2018.

[35] J. S. Pardo, K. Jordan, R. Mallari, C. Scanlon, and E. Lewandowski, "Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures," *Journal of Memory and Language*, vol. 69, no. 3, pp. 183–195, 2013.

[36] K. Nielsen, "Specificity and abstractness of vot imitation," *Journal of Phonetics*, vol. 39, no. 2, pp. 132–142, 2011.

[37] K. Shockley, L. Sabadini, and C. A. Fowler, "Imitation in shadowing words," *Perception & psychophysics*, vol. 66, pp. 422–429, 2004.

[38] L. S. Casasanto, K. Jasmin, and D. Casasanto, "Virtually accommodating: Speech rate accommodation to a virtual interlocutor." in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 32, no. 32, 2010.

[39] B. G. Schultz, I. O'BRIEN, N. Phillips, D. H. McFarland, D. Titone, and C. Palmer, "Speech rates converge in scripted turn-taking conversations," *Applied Psycholinguistics*, vol. 37, no. 5, pp. 1201–1220, 2016.

[40] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Acoustic-prosodic entrainment in slovak, spanish, english and chinese: A cross-linguistic comparison," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 325–334.

[41] J. M. Pérez, R. H. Gálvez, and A. Gravano, "Disentrainment may be a positive thing: A novel measure of unsigned acoustic-prosodic synchrony, and its relation to speaker engagement." in *INTERSPEECH*, 2016, pp. 1270–1274.

[42] J. D. Matarazzo, M. Weitman, G. Saslow, and A. N. Wiens, "Interviewer influence on durations of interviewee speech," *Journal of Verbal Learning and Verbal Behavior*, vol. 1, no. 6, pp. 451–458, 1963.

[43] J. Edlund, J. B. Hirschberg, and M. Heldner, "Pause and gap length in face-to-face interaction," 2009.

[44] C. De Looze, C. Oertel, S. Rauzy, and N. Campbell, "Measuring dynamics of mimicry by means of prosodic cues in conversational speech," in *ICPhS 2011*, 2011.

[45] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Language and speech*, vol. 55, no. 2, pp. 231–248, 2012.

[46] S. A. Borrie, N. Lubold, and H. Pon-Barry, "Disordered speech disrupts conversational entrainment: A study of acoustic-prosodic entrainment and communicative success in populations with communication challenges," *Frontiers in psychology*, vol. 6, p. 151543, 2015.

[47] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability." *Journal of Personality and Social Psychology*, vol. 32, no. 5, p. 790, 1975.

[48] A. Paxton and R. Dale, "Interpersonal movement synchrony responds to high-and low-level conversational constraints," *Frontiers in psychology*, vol. 8, p. 1135, 2017.

[49] J. Holler and K. Wilkin, "Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue," *Journal of Nonverbal Behavior*, vol. 35, pp. 133–153, 2011.

[50] M. M. Louwerse, R. Dale, E. G. Bard, and P. Jeuniaux, "Behavior matching in multimodal communication is synchronized," *Cognitive science*, vol. 36, no. 8, pp. 1404–1426, 2012.

[51] K. Shockley, M.-V. Santana, and C. A. Fowler, "Mutual interpersonal postural constraints are involved in cooperative conversation." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, no. 2, p. 326, 2003.

[52] H. Drimalla, N. Landwehr, U. Hess, and I. Dziobek, "From face to face: the contribution of facial mimicry to cognitive and emotional empathy," *Cognition and Emotion*, 2019.

[53] D. N. McIntosh, "Spontaneous facial mimicry, liking and emotional contagion," *Polish Psychological Bulletin*, vol. 37, no. 1, p. 31, 2006.

[54] S. E. Brennan, "Lexical entrainment in spontaneous dialog," *Proceedings of ISSD*, vol. 96, pp. 41–44, 1996.

[55] H. P. Branigan, M. J. Pickering, J. Pearson, J. F. McLean, and A. Brown, "The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers," *Cognition*, vol. 121, no. 1, pp. 41–57, 2011.

[56] N. Suzuki and Y. Katagiri, "Prosodic alignment in human–computer interaction," *Connection Science*, vol. 19, no. 2, pp. 131–141, 2007.

[57] K. Church, "Empirical estimates of adaptation: The chance of two noriegas is closer to p/2 than p2," in *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, 2000.

[58] A. Dubey, P. Sturt, and F. Keller, "Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, 2005, pp. 827–834.

[59] J. Hauke and T. Kossowski, "Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data," *Quaestiones geographicae*, vol. 30, no. 2, pp. 87–93, 2011.

[60] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[61] D. Huffaker, J. Jorgensen, F. Iacobelli, P. Tepper, and J. Cassell, "Computational measures for language similarity across time in online communities," in *Proceedings of the analyzing conversations in text and speech*, 2006, pp. 15–22.

[62] Y. Wang, D. Reitter, and J. Yen, "Linguistic adaptation in conversation threads: Analyzing alignment in online health communities," in *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics*, 2014, pp. 55–62.

[63] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic priming in written production: Evidence for rapid decay," *Psychonomic Bulletin & Review*, vol. 6, pp. 635–640, 1999.

[64] D. Reitter, F. Keller, and J. D. Moore, "Computational modelling of structural priming in dialogue," in *Proceedings of the human language technology conference of the naacl, companion volume: Short papers*, 2006, pp. 121–124.

[65] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. Interspeech 2011*, 2011, pp. 3081–3084.

[66] P. Sedgwick, "Pearson's correlation coefficient," *Bmj*, vol. 345, 2012.

[67] P. Bourke, "Cross correlation," *Cross Correlation", Auto Correlation—2D Pattern Identification*, vol. 596, 1996.

[68] D. H. Abney, A. Paxton, R. Dale, and C. T. Kello, "Complexity matching in dyadic conversation." *Journal of Experimental Psychology: General*, vol. 143, no. 6, p. 2304, 2014.

[69] J. S. Pardo, I. C. Jay, and R. M. Krauss, "Conversational role influences speech imitation," *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2254–2264, 2010.

[70] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[71] G. L. Murphy and D. L. Medin, "The role of theories in conceptual coherence." *Psychological review*, vol. 92, no. 3, p. 289, 1985.

[72] C.-C. Lee, M. Black, A. Katsamanis, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. Interspeech 2010*, 2010, pp. 793–796.

[73] N. Lubold and H. Pon-Barry, "Acoustic-prosodic entrainment and rapport in collaborative learning dialogues," in *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, 2014, pp. 5–12.

[74] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[75] "Alignment and task success in spoken dialogue," *Journal of Memory and Language*, vol. 76, pp. 29–46, 2014.

[76] R. H. Gálvez, L. Gauder, J. Luque, and A. Gravano, "A unifying framework for modeling acoustic/prosodic entrainment: definition and evaluation on two large corpora," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2020, pp. 215–224.

[77] U. C. Priva and C. Sanker, "Distinct behaviors in convergence across measures," in *Proceedings of the 40th Annual Meeting of*

the Cognitive Science Society, CogSci 2018, Madison, WI, USA, July 25-28, 2018, C. Kalish, M. A. Rau, X. J. Zhu, and T. T. Rogers, Eds. cognitivesciencesociety.org, 2018. [Online]. Available: https://mindmodeling.org/cogsci2018/papers/0294/index.html

[78] A. Weise and R. Levitan, "Looking for structure in lexical and acoustic-prosodic entrainment behaviors," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 297–302.

[79] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[80] P. Boersma, "Praat: doing phonetics by computer," *http://www. praat. org/*, 2006.

[81] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[82] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[83] B. Ghojogh, F. Karray, and M. Crowley, "Eigenvalue and generalized eigenvalue problems: Tutorial," *arXiv preprint arXiv:1903.11240*, 2019.

[84] S. I. Levitan, J. Xiang, and J. Hirschberg, "Acoustic-prosodic and lexical entrainment in deceptive dialogue," in *Proc. 9th International Conference on Speech Prosody*, 2018, pp. 532–536.

[85] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[86] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *2015 AAAI spring symposium series*, 2015.

[87] M. Vashkevich, A. Petrovsky, and Y. Rushkevich, "Bulbar als detection based on analysis of voice perturbation and vibrato," in *2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 2019, pp. 267–272.

[88] T. Wu, L. Zhao, and Q. Zhang, "Research on speech synthesis technology based on rhythm embedding," in *Journal of Physics: Conference Series*, vol. 1693, no. 1. IOP Publishing, 2020, p. 012127.

[89] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[90] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[91] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, 2014.

[92] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[93] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, 2020.

[94] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[95] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance." *Journal of machine learning research*, vol. 11, no. 6, 2010.

[96] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE access*, vol. 8, pp. 80 716–80 727, 2020.

[97] X. Shi, S. Li, and J. Dang, "Dimensional emotion prediction based on interactive context in conversation." in *INTERSPEECH*, 2020, pp. 4193–4197.

[98] L. St, S. Wold *et al.*, "Analysis of variance (anova)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.

[99] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to

affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[100] R. Ostrand and E. Chodroff, "It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue," *Journal of Phonetics*, vol. 88, p. 101074, 2021.

[101] Y. Liu, A. Li, J. Dang, and D. Zhou, "Semantic and acoustic-prosodic entrainment of dialogues in service scenarios," in *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 71–74.

[102] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, ser. HLT '11. USA: Association for Computational Linguistics, 2011, p. 113–117.

Publications

Journal

[1] Liu, Yuning, Unoki, M. Linear Model Approach for Investigating the Comprehensive Entrainment in Conversation[J]. Journal of Signal Processing, 2024, 28(4): 133-136.

[2] Liu, Yuning, Zhou, D., Unoki, M., Dang, J., Li, A & Okada, S. Investigation of social factor in conversational entrainments. IEEE access, 2024. (under review)

International Conference

[3] Liu, Yuning, Li, A., Dang, J.,& Zhou, D. Semantic and acoustic-prosodic entrainment of dialogues in service scenarios. In: Companion Publication of the 2021 International Conference on Multimodal Interaction. 2021. p. 71-74.

[4] Liu, Yuning, Zhou, D., Unoki, M., Dang, J., & Li, A. Dialogue scenario classification based on social factors. In: 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2022. p. 379-383.

[5] Liu, Yuning, Unoki, M. Linear Model Approach for Investigating the Comprehensive Entrainment in Conversation. International Workshop

on Nonlinear Circuits, Communications and Signal Processing. 2024, February.

Domestic Conference

[6] Liu, Y., Li, A., Dang, J., & Zhou, D. Investigation of Semantic and Acoustic-Prosodic Entrainment in Service Scenarios. Phonetics Division of the Chinese Language Society. 2021, July.

[7] Liu, Y., Unoki, M. Emotion Prediction based on Conversation Entrainments. JHES2023. 2023, September.

[8] Liu, Y., Unoki, M. Conversation Scenario Classification Based on Conversation Entrainment. Acoustical Society of Japan. 2024, March.