2023-05-01

On The Predictability Of Appropriate Prosody Of Dialog Markers Directly From The Local Context

Anindita Nath
University of Texas at El Paso

ON THE PREDICTABILITY OF APPROPRIATE PROSODY OF DIALOG MARKERS DIRECTLY FROM THE LOCAL CONTEXT

ANINDITA NATH

Doctoral Program in Computer Science

APPROVED:

Nigel G. Ward, Ph.D., Chair

David G. Novick, Ph.D.

Olac Fuentes, Ph.D.

Carla Contemori, Ph.D.

Heike Lehnert-LeHouillier, Ph.D (NMSU)

Stephen Crites, Ph.D.
Dean of the Graduate School

To my

MOTHER who let me have my wings

&

my late FATHER who ensured I was brave enough to dream of having them

ON THE PREDICTABILITY OF APPROPRIATE PROSODY OF DIALOG MARKERS DIRECTLY FROM THE LOCAL CONTEXT

by

ANINDITA NATH, MBA, MCA

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2023

Acknowledgements

I begin by expressing my deepest gratitude to my esteemed advisor, Dr. Nigel G. Ward, who guided me at every step of my Ph.D. journey. This endeavor would not have reached its final destination without his invaluable supervision, relentless support, and tutelage throughout the course of this program.

I would also like to express my appreciation to my dissertation committee members — Dr. David G. Novick and Dr. Olac Fuentes, both of the Computer Science department, Dr. Carla Contemori of the Linguistics Department, and Dr. Heike Lehnert-LeHouillier of the New Mexico State University—for their mentorship. Special thanks to Dr. David G. Novick for his treasured support, which was influential in shaping my experiment methods and critiquing my results. Additionally, I sincerely thank Dr. Heike Lehnert-LeHouillier and her team at New Mexico State University for enabling the experiments in Chapter 8 using their corpus of collected autistic data.

My gratitude extends to the Computer Science Department at UTEP for funding my travel to Lisbon, Portugal, to present my publication at the ISCA conference, *Speech Prosody, 2022*.

Many thanks to the CS faculty—who had been largely influential in sharpening my technical skills—and the staff at UTEP for all their hard work and dedication in providing me with the means to complete my degree and prepare me as a computer scientist.

I would be remiss in not mentioning my friends, and colleagues for the cherished time we spent together in the formal and social settings that made me feel welcomed in this alien nation since the day I landed here.

Finally, words will fall short in expressing my gratitude to my mother and late father. Without their tremendous understanding, life-long support, and encouragement, it would be impossible for me to even dream of pursuing, let alone completing, my doctoral degree. This one is dedicated to you both!

Abstract

Today's state-of-the-art spoken dialog systems lack context-appropriate prosody in their responses, often making them sound unnatural. Better modeling of this contextual dependency would enable natural prosodic responsiveness. Accordingly, this dissertation explores the extent to which the prosody of a dialog marker can be predicted directly from the prosody of its local context. The prediction performance was evaluated in terms of the similarity between the predicted and the observed prosodic features as measured by the reduction of root mean square error from the baseline. This prediction task was accomplished for multiple combinations of various sets of context features and different machine learning algorithms. Simple machine-learning models, without any knowledge of pragmatic intent or phonetic structure, could predict prosody, to a certain extent, for each of the most common twelve types of dialog markers in a corpus of unstructured American English dialogs. A simple feed-forward multi-layered artificial neural networks model performed best, with an overall average reduction in prediction error of 42%. This proposed prosody prediction approach has value also for a task-oriented dialog domain.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Variation in Prosody for Dialog Markers

Humans participating in dyadic conversations influence each other's engagement, emotions, and behaviors (Burgoon et al., 1995) and prosody plays an important part in this process. Humans use their knowledge of prosody to infer the intended meaning of a spoken word from its dialog context, to distinguish between different contextual uses of the same word, and to respond with context-appropriate prosody.

The following dialog snippets illustrate how the same word (in this case, a dialog marker) can vary in its prosody according to the dialog context. These examples—part of my exploratory qualitative study (Chapter 3)—were obtained from the *ISG Billing Support Corpus* (Ward et al., 2005), which comprises recorded telephonic conversations between a human customer and a human or virtual agent for accomplishing tasks like paying credit card bills and enquiring about past transactions or current credit balances. The participants conducted the same task with a spoken dialog system (i.e., the virtual agent) and the human agent, not necessarily in that order.

Context 1

Human Agent: What is your bank routing number?

Customer: It's 879668321.

Human Agent: Okay. What is your bank checking account number?

The word *okay* here conveys that the agent acknowledges the information provided by the customer and also seeks more information on the **same sub-topic**. The most observable prosodic characteristics of this *okay* are a late pitch peak in the first syllable, a downslope

and lower pitch in the second syllable, and high harmonicity[1] .

Context 2

Customer: I need to make a payment.

Human Agent: Okay. What is your account number?

In this context, *okay* conveys that the agent acknowledges the customer's wish but needs to violate her likely expectation for the next action and seek more information on a **different sub-topic**. *Okay* here is loud, breathy, and relatively short in duration.

These examples were from dialogs between two neurotypical humans. For comparison, I also listened to human-machine conversations in similar dialog contexts. This revealed that the *okay*s spoken by the virtual agent (alternatively, the spoken dialog system) did not vary at all: they sounded flat and monotone, with no variation in their prosodic patterns that could help distinguish among their contextual uses.

Some humans, such as those with autistic spectrum disorder, may also lack context sensitivity (Scholten et al., 2015) and often exhibit atypical prosody in conversations (Kanner, 1943; Asperger and Frith, 1991; Nakai et al., 2014; Fusaroli et al., 2017).

To summarize this section, humans typically convey their intent to their speaking partners by appropriately shaping the prosody of their responses. To do this, they generally consider the context in which the dialog occurs. However, such context-appropriate prosodic behavior is still absent in spoken dialog systems.

1.2 Motivation

As noted above, current spoken dialog systems are often incapable of exhibiting human-like responsive behavior, specifically with respect to adaptive prosody. Consequently, their conversations with users are less human-like, that is, less natural. Hence, there is a need for models that enable the selection of better prosody in system responses to make them appropriate to the pragmatic intentions or the local dialog context. Improved responsiveness

[1]high harmonics to noise ratio, reflecting a near "singing" voice

could improve perceived naturalness and ultimately enable systems:

- to retain the attention of the user for a longer period,
- to increase user engagement and rapport in interactive games,
- to improve user satisfaction in commercial applications, as previously suggested by Acosta and Ward (2011); Lubold and Pon-Barry (2014); Li et al. (2019); Gálvez et al. (2020); Choi and Agichtein (2020).

To detect atypical prosodic patterns in autistic populations—especially in children—prior research has compared their patterns with the corresponding prosodic patterns in neurotypical counterparts (Dahlgren et al., 2018). Automatic detection of atypical prosody—by automatically learning its differences from the neurotypical prosodic behavior— could help in the early diagnosis of this disorder (Chi et al., 2022). However, these approaches do not consider the typical context dependency of response prosody, which if incorporated may improve these automatic comparisons.

With these possible improvements as an ultimate goal, this research explores a novel approach that would:

- enable a spoken dialog system to select prosody in its responses that is more appropriate to the context and thus, improves naturalness in its responsive behavior, and
- enable direct and improved comparisons between atypical and neurotypical prosody and ultimately, improve the automatic detection of atypical prosodic patterns in autistic dialogs.

To be more specific, in this dissertation I claim that it is possible to predict the appropriate prosody of dialog markers directly from their local dialog context prosody. I choose to focus only on dialog markers instead of all words and utterances because they:

1. occur frequently,

2. are often more semantically independent of local lexical context than most words,

3. have a core procedural and not conceptual meaning (Fraser, 1999)

4. are important in dialog since they:

    - can be explicit indicators of the local discourse structure (Fraser, 2009),
    - can be powerfully indicative of the intent of an utterance (Fraser, 2009).
    - serve many important functions, including managing turn-taking, marking topic structure, and expressing stance (Louwerse and Mitchell, 2003; Fraser, 1999)

   and, finally

5. often stands alone, prosodically. That is, their prosody is usually their own and is less often affected by the larger prosodic patterns that govern many word sequences.

Hence, dialog markers are a reasonably good initial exploration point to test the predictivity of local context prosody in spoken dialogs.

1.3 Dissertation Preview

In this dissertation, I will review the literature (Chapter 2). I will discuss what is considered natural responsiveness in interactions in general, what progress has been made over the years toward achieving this responsiveness in spoken dialog systems (specifically with respect to prosodic behavior), what key challenges remain to be solved, and how I propose to solve a few of them. Then I will discuss what is meant by atypical prosody as exhibited by humans with autistic spectrum disorder, current approaches to differentiate this from neurotypical prosody, the likely benefits of automatic detection of atypical prosody, and one of the key challenges in doing so that this dissertation aims to address. Next, I will review prior research on dialog markers and their associated prosody, some of the key limitations of these past works, and how the novelties of my proposed approach are likely to address them. Chapter

2 concludes with the main question I address in this dissertation: how well can the dialog marker prosody be predicted directly from the local context prosody?

Chapter 3 describes an initial qualitative study on the context-prosody mappings of a single dialog marker, *okay*, in an in-house task-oriented dialog corpus.

Next, in Chapter 4, I will discuss the methodological and evaluation details and then the initial results for predicting dialog marker prosody directly and only from the prosody of its local context. Chapter 5 investigates the factors that affect the prediction performance of this proposed approach and Chapter 6 discusses performance improvement through different machine learning algorithms.

Subsequent chapters describe the methods that test and report whether the proposed prosody prediction approach can be generalized for task-oriented dialogs (Chapter 7) and for dialogs involving neurotypical and autistic children (Chapter 8).

Finally, Chapter 9 summarizes my research findings and discusses their implications and possible future directions.

Chapter 2

Related Research

2.1 Spoken Dialog Systems and Responsive Prosody

This section briefly reviews prior research on responsiveness in human-human interactions and its benefits. Then it discusses in detail the approaches toward achieving this natural responsiveness in spoken dialog systems, specifically concerning their prosodic behavior.

2.1.1 Natural Prosodic Responsiveness

Human interactions involve verbal and non-verbal features (Mandal, 2014), including *prosody*. Prosody refers to the suprasegmental acoustic features of speech (Lehiste, 1970), including pitch, intensity, and duration, that act as non-verbal signals capable of communicating information above and beyond what is explicitly stated verbally (Cutler et al., 1997; Ferreira, 2006). Since intended meaning cannot be predicted entirely from word forms, for example, when the same word is used in different contexts, human listeners must rely on these extralinguistic cues to uncover the word's meaning in the context. (Shintel et al., 2006; Nygaard et al., 2009; Roettger and Rimland, 2020). It is also observed that most humans can adapt their responses to these inferred interlocutor intents or the communicative context by appropriately adjusting these prosodic signals (Tzeng et al., 2019; Xie et al., 2021). Such adaptable, responsive behavior demonstrated by typical human speakers, without conscious awareness, while interacting with a fellow interlocutor, is what some speech researchers term *natural* or *human-like interaction* (Edlund et al., 2008).

This is of practical importance because it is not only *what* one says but also *how* one

says it that is important for effective communication. Appropriate prosody in the speaker's responses plays an important role in various real-life situations. For example, features extracted from the vocal tone and prosody can quantify the non-linguistic communication channel between the interviewer and interviewee, which can also help predict the outcomes of job interviews accurately (Soman and Madan, 2009). Appropriate prosody can strengthen social bonds and can be used to generate supportive responses in human-human interactions (Mauchand and Pell, 2021). Prosodic cues also help the listener determine whether the speaker is being sarcastic (negative irony) or teasing (positive irony), that is, understand the speaker's stance towards the listener (Mauchand et al., 2020).

Natural responsiveness is also related to the phenomenon of *acoustic-prosodic entrainment*. Human dialog partners sometimes become more similar to each other in their prosodic behavior—pitch, loudness, or speaking rate—during a conversation. This entrainment could directly correlate with rapport (Lubold and Pon-Barry, 2014), primarily measured from the perceptual perspective. This is supported in the aggregate results from self-reported rapport as perceived by each participant: speakers who entrain tend to be more in rapport with each other. Prosodic entrainment has even been shown to lead to greater success in student learning in an intelligent dialog tutoring system (Thomason et al., 2013). Also, it has been seen that entrainment of the robot's prosodic features to that of the child user's speech in a game that recognized only two action words, *go* and *jump* (Sadoughi et al., 2017), engendered greater engagement than the version of the game that did not entrain.

2.1.2 Towards Prosodically Responsive Spoken Dialog Systems

This section briefly surveys progress towards incorporating responsiveness in spoken dialog systems.

To have a virtual assistant or a chat companion system with adequate intelligence has seemed elusive and has existed only in Sci-Fi movies for a long time. But with the evolution of speech-based technologies, today's spoken dialog systems have come a long way from accomplishing simple tasks, such as the provision of air travel (Hempel, 2010), to be used

in more complex scenarios, for example, in-car applications (Geutner et al., 2002), personal assistants such as Siri, Google Now or Microsoft's Cortana (Janarthanam et al., 2013), smart homes (Krebber et al., 2004), and interaction with robots (Foster et al., 2013). During the last two decades, spoken dialog systems have been increasingly used in providing services such as interviews (Ghanem et al., 2005), counseling (Hubal and Day, 2006), chronic symptoms monitoring (Black et al., 2005; Migneault et al., 2006), medication prescription assistance and adherence (Bickmore and Giorgino, 2006), changing dietary behavior (Delichatsios et al., 2001), promoting physical activity (Farzanfar et al., 2005), helping cigarette smokers quit (Ramelson et al., 1999) and speech therapy (Saz et al., 2009). Recent technological advances that enabled end-to-end and transformer-based neural speech synthesis achieved enormous task success in conversational AI, for example, in *Alexa* (Ram et al., 2018).

Despite the progress in spoken dialog systems, accommodating *human-like* or *natural responsiveness* remains one of the top challenges (Ward and DeVault, 2016), especially while designing an open-ended dialogue system (Huang et al., 2020). Neural methods have helped develop context-aware and expressive dialogue systems in open-domain and task-oriented genres but either mostly in text-based bots or have only improved linguistic information (Ni et al., 2022). However, incorporating human-like prosodic behavior in system responses remains to be accomplished (Ni et al., 2022). The above systems can accomplish their intended task(s) but fail to exhibit natural prosodic behavior to the extent needed to improve engagement and rapport with users and increase user satisfaction. One way systems could achieve this is by better approximating the adaptive behavior observed in human-human interactions.

Research on improving the responsiveness of spoken dialog systems to make them more human-like or natural has boomed in recent years. For example, in the context of job interviews, a socially adaptive virtual recruiter (Youssef et al., 2015), that adapted its behavior according to social constructs (attitudes, relationships, etc.) depending on user's behavior was perceived by the users as a more credible agent than a scripted one.

Acoustic-prosodic entrainment in human-human dialogs was modeled for a spoken dialog

system that served as a learning companion to increase its effectiveness (Lubold et al., 2015). In particular, adapting by shifting the text-to-speech output's pitch contour towards the user's mean pitch resulted in the highest measures of rapport and naturalness. Overall, the system demonstrated greater rapport with the student users, enabling them to have a higher learning rate in a collaborative learning domain.

In one of the early works on improving adaptive behavior in the e-learning domain (Forbes-Riley and Litman, 2004), it was found that it is useful to understand the students' affective states through the correct interpretation of the non-verbal cues communicated by them. The students' negative, neutral, and positive emotions could automatically be predicted at each student turn in a dialog to an accuracy of 84%, utilizing contextual acoustic-prosodic and other linguistic information from the speech signal. This predictive ability was used to build an intelligent tutoring spoken dialog system that automatically predicted and adapted to student uncertainty. Students learned more if the tutor responded to their explicit questions and this pedagogically relevant state (Litman, 2013). A Wizard-Of-Oz spoken dialog system (Forbes-Riley and Litman, 2012) — that adapted to two user states, uncertainty, and disengagement, in real time — showed that increased adaptivity reduced uncertainty levels and disengagement of their users as well as improved their learning. Based on these findings, a fully-automated affect-adaptive spoken tutoring system was designed and evaluated (Litman and Forbes-Riley, 2014). The results showed that adapting to affect is better than not adapting at all, but there was no significant difference between multiple-state (user-uncertainty and user-disengagement) and single-state (only user-uncertainty) adaptivity. Also, only males showed improved learning due to adaptivity, indicating that such systems' utility may vary by gender. The system, however, could only be adaptive in terms of the linguistic content of the responses.

In an attempt to improve the prosodic behavior of systems, Acosta and Ward (2011) showed that the emotional prosody of interlocutor responses could be improved utilizing the prosody associated with the emotional categories identified in the immediately preceding user utterance. Until then, either emotional categories of user speech were recognized, or

system-side emotions were expressed with the help of prosody. This work combined both of these functionalities for the first time. The emotional state of each speaker utterance was recognized by its associated prosodic features, and then this information was used to determine the emotional coloring for the next interlocutor's utterance. A spoken dialog system, Gracie, which used this coloring technique, achieved greater rapport with its users than baselines with no or non-contingent coloring. This was one of the early successful attempts to render the system's responses with human-like prosody.

More recently, Li et al. (2019) combined a prosody-based emotional valence recognition model and a text-based sentiment analysis model to improve the emotion processing and reaction module of a spoken dialog system. Using linear regression, prosody was used to predict the labeled valence in an utterance. The prosodic features used were energy, creakiness, pitch lowness, pitch highness, narrow pitch range, wide pitch range, and speaking rate, each computed over four time periods preceding the end point of each utterance: –1600 ms to –1100 ms, – 1100 ms to –600 ms, –600 ms to –100 ms, and –100 ms to 0 ms. Their novel method of combining valence from speech and sentiment from text achieved better emotion recognition than the state-of-art models. Encouraged by these results, the authors proposed a "reactive emotion expression method", where the system's emotion category and level were predicted using the parameters predicted by the emotion recognition. This proposed system generated more authentic and effective emotions that resulted in more natural and human-like conversations, compared to conventional dialogue systems that provided robotic and unnatural reactions since they did not consider emotion, as reported by ten participants in a subjective experiment. This also shows that even the simplest machine learning algorithms can effectively achieve natural responsiveness if used with appropriately engineered features.

The above systems or approaches only adapted their linguistic content to the context or were adaptive to only a handful of user states or emotional dimensions. Recently, incorporating end-to-end-speech synthesis (Shen et al., 2018) and transfer learning (Éva Székely et al., 2019) synthesized more expressive and natural speech for public speaking and casual conversation. However, these approaches for speech synthesis did not consider the dialog

context. Hence, the synthesized speech often did not sync with the dialog flow.

Several studies introduced dialog-related information to speech synthesis: tokens representing emphasis (Tsiakoulis et al., 2014), dialog acts (Hojo et al., 2020), and dialog history up to ten turns (Guo et al., 2021), but they employed only linguistic features. In one of the recent attempts to model the typical contextual dependency of prosody observed in human-human dialogs, Yamazaki et al. (2021) employed neural methods that modeled the differential F0 in a dialog context, along with other linguistic information, to synthesize speech with appropriate F0 contours. The synthesized F0 was closer to the recorded F0 for a Japanese corpus of spontaneous multi-modal conversations, also demonstrated through a subjective evaluation of perceived appropriateness. Although this work is a step in the right direction towards improving prosodic responsiveness in human-machine interaction, pitch contours represent only a part of the voice characteristics. To make the machines sound more natural sound, it is necessary to approximate the human-like adaptive prosodic behavior for all or most of the prosodic characteristics of a voice signal.

2.1.3 Key Limitations of Responsiveness Research

Though previous research has shown progress toward building responsive spoken dialog systems, key challenges—some of which are enumerated below—remain before commercial dialog systems can be made highly naturally responsive, specifically with respect to prosodic behavior.

1. Previous spoken dialog systems were responsive to only a small finite set of user states—uncertainty and disengagement (Forbes-Riley and Litman, 2012)—or emotional dimensions (activation, evaluation, and power) (Acosta and Ward, 2011).

2. Prior research (Acosta and Ward, 2011) used limited context information, only from one speaker and only from the immediate past utterance.

3. Previous responsive models were built to function only in a specific dialog domain such

as the persuasive dialog domain (Acosta and Ward, 2011) or the e-tutoring application domain (Litman and Forbes-Riley, 2014).

4. Even state-of-the-art methods of speech synthesis (Yamazaki et al., 2021) used both lexical and prosody context information to generate spoken responses in a TTS with context-appropriate prosody but with respect to only F0 contours.

This research will address the above challenges, respectively, by:

1. developing a continuous model that would predict the prosody for all aspects of each target word.

2. including prosodic context information from both the dialog partners and from a wider time period: wider recent past and future context.

3. investigating the possibility of building a more generalized cross-domain responsive model.

4. by predicting appropriate prosody for more than just one feature and that too, only using the local context prosody.

2.2 Autistic Spectrum Disorder and Atypical Prosody

While this does not relate directly to the main question, this section reviews research on atypical prosody patterns observed in humans with autism spectrum disorder, detection of these patterns via comparisons with neurotypical prosody, automatic detection of autistic prosody through machine learning approaches, and how our novel approach can improve or add value to these methods.

2.2.1 Atypical Prosody in Humans

Atypical prosody has been investigated in many childhood disorders to date, for example, in language impairment (SLI), hearing impairment, Down syndrome (DS), childhood fluency

disorder, Williams syndrome (WS) and autistic spectrum disorder (ASD). The specific characteristics of atypical prosody are diverse and complex (Peppé, 2018). They may include misplaced stress, articulatory irregularities, disrupted speech rhythm, monotonous speech, sudden rise in volume, unusual intonation patterns, and disordered phrasing.

Autism spectrum disorder(ASD) has been associated with atypical prosody, specifically, with respect to pitch, duration, and intensity, ever since (Kanner, 1943) and Asperger and Frith (1991) published the first systematic studies of ASD. Common autistic prosodic behavior includes monotonous voice, atypical pitch with no pitch variation, harsh or hoarse voice, and either too loud or too quiet, especially in autistic children (Baltaxe et al., 1984; Sheinkopf et al., 2000; Kaland et al., 2013).

Atypical prosody could be the underlying cause of various functional and social challenges that autistic people face. Impaired prosodic skills were significantly linked with executive dysfunction traits: divided attention, working memory/sequencing, set-switching, and inhibition (Filipe et al., 2018). Zampella et al. (2020) demonstrated in their study that children with ASD exhibited less interactional synchrony than their typically developing peers, which may be associated with impaired social functioning in ASD. Severe deficits in social communication, usually characterizing autism spectrum disorder, could often be caused due to their lesser capability of processing and recognizing emotional prosody (Rosenblau et al., 2017).

2.2.2 Towards Modeling Autistic Prosody

Research shows that the traditional ways of detecting atypical prosody in autistic speech are by comparing the same with conversations involving neurotypical humans using either perception or acoustic analysis (Nadig and Shaw, 2011; Nakai et al., 2014). Dahlgren et al. (2018) used both perceptual and acoustic analyses to study the production of prosody in children with autism spectrum disorder and to additionally investigate whether prosodic characteristics of the voice could be used as clinical markers for autism spectrum disorder. Eleven children each, from a group diagnosed with autism spectrum disorder and the other with typical development, were recorded while they told a story that was elicited with ex-

pression. Perceptual analysis was carried out using a standardized Swedish clinical voice evaluation procedure. For the acoustic analysis, features considered were: average fundamental frequency (F0), F0 range, F0 variation, speech rate, and the number of words produced per utterance. The perception analysis revealed no difference between the two groups, whereas the acoustic analysis showed an increased number of words per utterance, atypical fluency, and speech rate in autistic children. Disfluency in producing discourse markers *um* when compared to typically developed peers proved to be an important pragmatic marker in children with autism spectrum disorder (Irvine et al., 2016). Though used extensively in differentiating the two groups, this traditional process is time-consuming, lengthy, and labor-intensive, sometimes also involving trained physicians.

Recent research has, therefore, shifted its focus to the automatic detection of atypical prosody using machine learning, which is both time and cost-effective and may facilitate early detection of autistic prosody. Chi et al. (2022) demonstrated that the random forest classifier achieved 70% accuracy, the fine-tuned wav2vec 2.0 model achieved 77% accuracy, and the convolutional neural network achieved 79% accuracy when classifying children's audio as either ASD or NT in an experiment involving cellphone-recorded child speech audios curated from the *Guess What?* mobile game.

2.2.3 Key Challenge in Modeling Autistic Prosody

Current approaches on auto-classification of autistic prosody are mostly a black box and do not facilitate factor-based studies. Also, these comparative studies do not consider the typical dependency of a word or utterance's prosody on the prosody of its local context, observed in conversations involving neurotypical humans, which may be lacking in autistic people. Scholten et al. (2015) demonstrated that autistic people, especially children, were less sensitive to contextual prosody than their neurotypical counterparts as they could not detect irony in spoken words, mostly cued by the local context prosody. Hence, comparing the two groups' context-responsive behavior could be the key factor in improving the automatic detection of atypical autistic prosodic patterns. This dissertation explores how far the local

context dependency information is directly useful in automatically distinguishing autistic prosodic behavior from neurotypical one.

2.3 Dialog Markers and Prosody

My strategy for tackling the challenges in the aforementioned prior research (c.f. Section 2.1.3 and Section 2.2.3) will be to focus on an important and accessible special case: dialog markers.

2.3.1 Dialog Markers Properties

Dialog markers have been much studied for their important role in speech. The research works cited in this section establish this importance, describe some functions of their prosody and indicate how this connection can be exploited to accomplish various speech-related tasks.

In this dissertation, I use the term dialog markers simply to refer to discourse markers used in a dialog. Discourse markers have been defined as "sequentially dependent elements which bracket units of talk" (Schiffrin, 1987). They mark transition points in communication and may facilitate the construction of a mental representation of the semantic and pragmatic organization of the dialog (Louwerse and Mitchell, 2003). They are used extensively in spontaneous speech and function as indicators of the structure of dialog, such as *now*, which marks the beginning of a new topic, *but*, which indicates contrasting information, and so on. They enable cohesion in a conversation (Louwerse and Mitchell, 2003) by signaling how an upcoming utterance relates to the prior dialog (Fraser, 1999). They help the listener to develop an expectation of the pragmatic intent of the upcoming utterance (Byron and Heeman, 1997). A study of dialog markers in English implied that it is essential for language learners to be aware of these markers and their pragmatic functions (Zarei, 2013; Torabi Asr and Demberg, 2013).

2.3.2 Towards Automatically Identifying Dialog Markers' Prosody

Dialog markers are often associated with their own characteristic prosody. Literature showed the presence of a linear systematic relationship (Shriberg and Lickley, 1993) between the clause-internal filled-pauses (such as *uh* and *um*) *f0* values and their corresponding past prosodic context, represented by the closest preceding *f0* peaks.

Hirschberg and Litman (1993) proposed an intonation model, based on pitch accent and phrasing, to disambiguate dialog markers (or as the authors called them, *cue phrases*). The authors tagged the transcript of the corpus for two different usages of tokens of the word *now*: according to whether it represented a sentential or a dialog use. This was mainly done utilizing simple human perception. The intonational phrase containing each token plus the preceding and succeeding intonational phrases were then digitized and pitch-tracked. Dialog and sentential uses of the tokens were then compared along several acoustic dimensions, accent type being one of them. Empirical results showed that the discourse uses of the cue phrases were either deaccented or bore an L* pitch accent whereas the sentential uses bore an H* or a complex pitch accent. Building on these findings, Litman (1996) used machine learning to classify the cue phrases according to their discourse or sentential functions, achieving higher accuracy than the state-of-the-art manual classification models. Also in Slovak (Mareková and Benus, 2020), context prosodic cues helped in disambiguating the functional meanings of the word *no*, analogous to *okay* in English.

Not only in English but also in the Japanese language, discourse markers are characterized by their prosodic features. Kawamori et al. (1998) demonstrated through the analysis of a task-oriented Japanese corpus that certain words which were until then considered redundantly used in an interaction actually fulfilled the characteristics of discourse markers in this language. These were categorized as responsiveness (back-channels in English): *hai* that could mean several things like *yes* or *ok* in English and fillers: *aaa* in Japanese which could be *mm* in English. Also, the work demonstrated that these categories could easily be distinguished by their associated prosodic characteristics like pitch, vowel length, and phonetic forms.

One dialog marker, namely *okay*, has been shown to have as many as ten pragmatic functions (Table 2.1) in a collaborative task domain in American English (Gravano et al., 2007b). These discourse functions were disambiguated with the help of contextual cues as was shown by a perception study. In an empirical study, Novick and Sutton (1994) distinguished between three following broad classes of acknowledgments, based on exchange structure in dialog utterances:

1. Other → ackn: acknowledgment forms the second phase in an utterance pair that follows the other speaker's utterance.

2. Self → other → ackn: Self speaker initiates an exchange, Another speaker (eventually) completes the exchange, and Self then utters an acknowledgment, and

3. Self + ackn: Self includes an acknowledgment in an utterance outside of an utterance pair.

They recognized thirteen different speech-act patterns that were present within these classes which accounted for the specific uses of acknowledgments in a task-oriented speech corpus. In another work, pitch change in a single word served as a prosodic cue that helped distinguish between different uses of the word *right*—namely affirmative answer and acknowledgment, pronounced with a falling intonation, or direction, pronounced with rising intonation—in a task-oriented dialog corpus of spontaneous speech (Ward and Novick, 1995).

In another work, (Lai, 2009), showed how prosody, specifically pitch range, can distinguish whether the dialog marker *really* is used for questioning or expressing surprise. (Freeman et al., 2015), investigated the role and importance of prosody in categorizing and characterizing different stance-related dialog functions, namely, stance strength and polarity of *yeah*. It was seen that prosody helped in communicating meaning as six stance categories – agreement, no stance, backchannel, opinion, reluctance, and convincing – could be distinguished through a combination of intensity contour and duration cues, and the pitch was particularly useful for distinguishing the strength of stance. Gravano et al. (2007a)

Table 2.1: Discourse functions of *okay*, as described by Gravano et al. (2007b)

AI: Acknowledgment/agreement.
A2: Backchannel.
C: Cue beginning discourse segment.
E: Cue ending discourse segment.
P: Pivot beginning (A1+C).
F: Pivot ending (A1 + E).
N: Literal modifier.
B: Back from a task.
K: Check.
S: Stall.

presented results of a perception study showing that contextual and acoustic cues were useful in the disambiguation of various dialog-pragmatic functions of the word *okay* such as agreement/acknowledgment, backchannel, and cue to discourse beginning. Further, they showed that acoustic features capturing the pitch excursion at the right edge of *okay* feature prominently in disambiguation, whether other contextual cues were present or not.

Correlations between dialog markers and their characteristic prosody were exploited to predict dialog relations in English spoken monologs (Kleinhans et al., 2017). This co-dependency was used to distinguish between various dialog functions of the Swedish markers *men* (*but* or *and* in English) and *sa* (*so* in English), like signaling the beginning of a new topic, return to a previous topic, and different kinds of dialog moves (Hansson, 1999). The prosody of dialog markers was also useful to classify them in multiple dialog domains of the European Portuguese Corpus (Cabarrão et al., 2015).

Yet another study (Beach, 2020) explored this correlation of dialog markers with prosody, which revealed that *okay*s that were more prosodically marked, with more extreme pitch, loudness, duration, timing, and overall vocal quality, were used by speakers to display a wide range of orientations (e.g., when disagreeing, displaying aggravation, treating others'

actions as odd or bizarre, exuding happiness and excitement) rather than simply to achieve acknowledgment, acceptance, or assessment of the prior speaker's actions.

More recently, a preliminary analysis by Figueroa et al. (2022) revealed that designing an annotation scheme for communicative feedback functions required both the lexical forms of the words and their associated prosodic characteristics to be taken into account to distinguish the functions. This work annotated a total of 1627 short single-worded feedback tokens from the Switchboard corpus—that were preceded and followed by at least 5 seconds of silences—for each of the ten different communicative functions enumerated in Table 2.2. Each of the ten different functions had its own prosodic characteristics in terms of duration, mean pitch, pitch slope, and pitch range. Specifically, their statistical analysis showed that in terms of duration, functions *Sympathy*, *Strong Surprise* and *Disapproval* had significantly longer duration compared to the other functions. Also, *Agree* had a significantly longer duration than *Continue*. In terms of mean pitch, functions *Non-understanding* and *Mild Surprise* had significantly higher mean pitch than *Strong Surprise*, *No*, *Yes*, *Disapproval*, *Agreement*, and *Sympathy*. Also, *Disagree* and *Strong Surprise* had a significantly higher mean pitch than *Continue* and *Agreement*. In terms of pitch slope, *Non-Understanding* was the only function with a rising slope (*0.919*), that was significantly different from others, all of which had downward, negative pitch slopes. In terms of pitch range, functions *Mild Surprise*, *Non-understanding*, and *Strong Surprise* had a significantly wider pitch range than *Agree*, *Continue* and *No*. The authors further observed that uses of some lexical tokens like *yeah*, *um*, and *mhm* were particularly ambiguous and their corresponding feedback functions could frequently only be identified based on their prosodic characteristics.

In an exploratory study, Wallbridge et al. (2021) demonstrated the value of the non-lexical channel of speech through experimental results that showed that providing the non-lexical context in addition to its lexical counterpart helped to increase participants' ability to discriminate actual responses from the sampled ones. This discriminative task presented the participants with 1 true correct (T0, T1) sample, where T0 was the dialogue context (either text-only or audio+text) from the preceding turns and T1 was a potential response, along

Table 2.2: Major feedback functions with descriptions ((Figueroa et al., 2022))

Feedback Functions	Description
Continue	Continue speaking. I hear you and I'm listening but not necessarily agreeing/disagreeing.
Non-Understanding	I'm uncertain if I understood/heard what you said.
Agree	I agree with what you said.
Disagree	I doubt what you said is true. I disagree with what you said.
Yes	I am giving a positive response/answer to your yes/no question.
No	I am giving a negative response/answer to your yes/no question.
Sympathy	I'm expressing sympathy/pity/sorrow/concern/compassion to a negative statement.
Disapproval	I am showing disapproval/disgust.
Mild Surprise	I am showing mild surprise, showing slight interest.
Strong Surprise	I am showing strong surprise; I am impressed.

with 3 lexically-equivalent (T0, T1́) samples where T1́ was extracted from elsewhere in the corpus. Each of 67 native English-speaking participants completed a 20-question Qualtrics survey, where they were asked to rate how likely each sample was to be the true one on a scale from 1 ('Very Unlikely') to 4 ('Very Likely'). It was seen that when participants had access to the audio context of the immediately preceding turn along with its lexical counterpart, accuracy, and cross-entropy performance were significantly better, as people could more effectively discriminate between different prosodic realizations of the response, than when presented with only the lexical context. Thus, the results of this non-behavioral task indicate that the local prosodic non-lexical context strongly relates to the prosodic form of the next utterance.

In another recent study, Raso et al. (2022) used prosody to not only define the general function of being a discourse marker but also to distinguish between the specific functions performed by different kinds of markers. Mean intensity, f0, and the f0 slope up to the stressed syllable were the prosodic features that enabled the classification of 3 labeled discourse marker categories—*ALL* which includes markers establishing "social cohesion among the interlocutors" or disambiguating "who is the addressee of the utterance, using titles,

epithets, and proper names", *CNT* which includes markers that "push the listener to do something or to stop doing something", and *INP* that includes markers "taking the turn or beginning the utterance"— with a global 74% accuracy. Further qualitative analyses revealed that:

- discourse markers tagged as *EXP (Expressive)* —those that convey some surprise or emotional support—were characterized by a rising f0 shape on the stressed syllable, and
- discourse markers tagged as *EVD (Evidentiator)*—those that highlight what was said and also, secure the other speaker's attention—had slightly rising f0 shape, low intensity, and short duration.

The above research works revealed the variation of dialog markers' prosody according to their uses or functions in a dialog context, but they mostly explored the lexical context rather than the prosodic context. The use of prosodic context is potentially valuable, as shown by the above perceptual analysis establishing the connection of the dialog marker's prosody with its prosodic context. However, this association with context prosody has not been exploited further or modeled directly.

Although prior research on improving the responsiveness of spoken dialog systems used state-of-the-art speech synthesis models (c.f. Section 2.1), they largely ignored the utility of dialog markers, so well recognized in discourse structure and functions. Smith et al. (2022) presented a set of design principles on leveraging this prosodic information associated with discourse markers uses such as backchanneling, turn-taking, and so on, hypothesizing that the inclusion of this "conversational intelligence" would improve the system's usability and decrease a user's cognitive load. Though a subjective evaluation result of a mock system demonstrated some truth in this hypothesis, it has not yet been experimentally verified. Before incorporating the context-dependent prosodic behavior of dialog markers into a speech synthesis model, the first step is to model these mappings.

2.3.3 Key Limitations of Dialog Marker Research

While the prior research suggests how dialog markers may be used to improve responsive behavior in dialog systems, several shortcomings remain. The prior research:

1. mostly used lexical context to distinguish among various dialog marker classes or functions.

2. mostly treated one or, at the most, two dialog markers.

3. used prosody to disambiguate a few different categories of uses or functions of dialog marker, from two in (Hirschberg and Litman, 1993) to ten in (Gravano et al., 2007a) or very recently three categories in (Figueroa et al., 2022).

4. did not investigate whether these models are applicable across domains.

In this dissertation, I aim to address these limitations, respectively, by:

1. investigating the predictive ability of specifically the context prosody.

2. modeling the context-prosody mappings for more than a few dialog markers, specifically for the twelve most common ones found in American English.

3. predicting the prosody of each individual dialog marker instance.

4. testing the proposed context-dependent prosody prediction across multiple domains.

In sum, though there has been extensive research on improving the natural responsive behavior of the spoken dialog systems, there is still much room for improving the state of the art, especially with respect to the prosody in responses, which are often not yet appropriate to the dialog context. Prior research has also established that dialog markers' prosody depends on or varies with local context prosody but this prosodic information has not been put to any direct use.

To fill in the major research gaps mentioned above, I aim to investigate in this dissertation:

How well can a dialog marker prosody be predicted directly from its local context prosody?

Additionally, can this predictability succeed for more than:

- two dialog marker types,
- one prosodic feature, and
- a single dialog domain?

In addition, I explore the related question of whether this proposed prediction approach is able to differentiate between neurotypical and autistic prosody automatically.

Chapter 3

Exploratory Qualitative Study

3.1 Purpose

This study[1] aimed to discover simple mappings from dialog context to prosody for one of the most commonly found dialog markers in American English, namely *okay*. I could identify such mappings in one dialog corpus, the *ISG Billing Support Corpus* (Ward et al., 2005). This corpus includes approximately 90 minutes of role-playing audio, in which human customers interacted with a human agent and then, also repeated the same task with a virtual agent over the phone, though not always in that order. Tasks included paying credit card bills, enquiring about recent transactions, and other credit card-related tasks.

This exploration was based on the premise that various dialog markers can be categorized according to their associated pragmatic functions or use in a dialog context, and thus it expands on the study by Novick and Sutton (1994).

The main objective behind this study was two-fold:

1. Observe first-hand, albeit on a small scale, the details of one example of context-prosody mapping for dialog markers to supplement what was gleaned from previous work (Section 2.3).

2. Identify the most important prosodic features in such mappings, to determine what to include later in my proposed model.

[1]This chapter is based on Nath (2020)

3.2 Methodology

To explore mappings from context to prosody for the dialog marker *okay*, my approach comprised the following steps:

1. Listen to the dialogs.

2. Mark all points of occurrences of the dialog marker *okay*.

3. Group them into various context categories[2] based on their associated pragmatic functions. To accomplish this, a two-step bottom-up approach was followed.

   (a) Corresponding to each instance of *okay* in the corpus, the exact time-stamps were noted along with the type of its speaker, that is, whether it was the human agent, the human customer, or the virtual conversational agent. Transcripts of the utterances immediately preceding and following an *okay* were saved.

   (b) The above information was then used to categorize the pragmatic function of each *okay* in a given context and then to group them based on their common functions. For example, the contextual uses of *okay* similar to those of *Context-1* and *Context-2*, described in the Introduction, were categorized as *AST (AcknowledgementSameTopic)* and *ATS (AcknowledgementTopicShift)*, respectively.

4. Use Praat to visualize the prosodic features, namely, pitch, intensity, and duration, present in these categories. Praat graphs[3] generated for each occurrence of *okay* were studied to find the common prosodic characteristics for each context category of *okay*.

5. Listen to the audio again and, based entirely on human perception, note any additional subtle prosody that characterizes each context category and distinguishes it from the

[2]Audios available at https://github.com/anath2110/ISG-Credit-Card-Billing-Dialog-Corpus/tree/master/OkayWithPastFutureContext

[3]Graphs available at https://github.com/anath2110/ISG-Credit-Card-Billing-Dialog-Corpus/tree/master/OkayPraatFigures

others. For example, *okay* spoken by the human customer as an acknowledgment of the human agent's speech and also to indicate the end of the conversation was perceived to be more cheerful than the other *okays*.

I repeated the steps 1, 2 and 3a for each of the 123 *okay*s in the corpus.

3.3 Results

The major context categories and their common prosodic characteristics are listed below:

- Acknowledgment leading to a new topic, i.e., initiating Topic Shift (ATS):
  For example,
  Agent: What is your credit card account number?
  Customer: (provides the number).
  Agent: Okay, what can I help you with?

  This contextual use of *okay* was often characterized by high intensity (that is, was quite loud), clarity, pitch rise (more if there was a greater topic shift), breathiness, and was mostly short in duration.

- Acknowledgment, then, seeking more information on the Same Topic (AST):
  For example,
  Customer says his bank routing number.
  Agent: Okay, What is your bank account number?

  Common prosodic characteristics for this context category were more harmonicity than in other contexts, late pitch peak when prefiguring a question, and more downslope when the acknowledgment was that of heavier information, such as bank name or user intent vs. account number.

- Acknowledgment, then Reply (AR): For example,
  Agent: Will you like me to find out your last check number?

Customer: Yes.

Agent: Okay. Your last check number is (says the number).

*Okay*s, in these contexts, were observed to have a pitch drop or rise, a breathy first syllable, sometimes devoiced, co-articulated with the next word, and was not followed by any pause. They were short and quiet.

- Acknowledgement only (ACK):

  For example,

  Agent: Your payment is processed.

  Customer: (satisfied) Okay.

  *Okay*s here were short, clipped, or fading out slowly, loud at /k/ and early in the second syllable, and mostly the first syllable was higher in pitch than the second.

- Acknowledgement, then, Ending Conversation (AEC):

  For example:

  Agent: Last date of payment is midnight tomorrow.

  Customer: Okay. That's all I needed to know. Bye.

  This contextual use of *okay* did not have a single specific prosodic pattern but was perceived, in different scenarios, to be either:

  - cheerful, where /o/and the diphthong /eɪ/ were high pitched, or
  - loud and low-pitched at the beginning with gradually rising pitch, or
  - just flat pitched and abruptly ending.

- End of Conversation (EC):

  For example:

  Customer: That's all. Bye.

  Agent: Okay. Bye.

In such contexts, *okay*s were shortest of all, very quiet, almost inaudible, devoiced, and breathy.

- Transition to Different Speaker (TDS):

  For example,

  Customer: I just don't know what I had given..

  Agent: Okay.

  Customer: I don't remember if you could give me that information..

  In these contexts, *okay*s were mostly lengthened, overlapped with the interlocutor's speech, quite loud with the higher pitch at the end, and ended abruptly.

From the above observations, it could be inferred that the most common prosodic features that helped distinguish each context of *okay* were volume, duration, pitch range, pitch height, harmonicity, breathiness, creakiness, and peak disalignment.

3.4 The Salience of Context-Inappropriate Prosody

In the absence of context-appropriate prosody in response, the speaker's intended meaning or the local dialog goal is not successfully communicated to the interlocutor. To investigate whether this claim holds in a practical scenario, I conducted an informal study with ten research students from different areas who attended a poster session at an international conference where I presented the findings of the initial phase of this exploration. I asked each participant to listen to the original recording and an edited version of the same dialog snippet, but where the context had been purposely mismatched with the prosody. More specifically, I interchanged an *okay* spoken in the context *AcknowledgementSameTopic* (*AST*), in one of the dialogs with that spoken in the context *End of Conversation, (EC)* in another dialog. All of the participants confirmed that the out-of-context prosody was distinct in the edited audio[4]. The outcome of this informal experiment, thus, demonstrated—to a certain extent—the significant contribution of context-appropriate prosody to communicate the speaker's

intended meaning in a dialog.

3.5 Discussion

In a nutshell, this exploratory study:

- finds that various contextual uses of the same dialog marker can be disambiguated by their associated prosody,
- provides a list of prominent prosodic features that enable this disambiguation, and
- finally, confirms that the use of appropriate prosody contributes to communicating the speaker's intended meaning of a word in a dialog context.

This qualitative analysis, however, is limited in terms of scalability and reproducibility since it mostly involved human perception and, therefore, consumed an enormous amount of effort and time.

Moving forward, to test the original research hypothesis — more specifically, to test the extent to which the prosody of local dialog context is directly informative to predict the prosody of a dialog marker— it is necessary to learn these context-prosody mappings automatically. Based on the aforementioned findings of this exploration, I now have a key set of prosodic features that could be used, in some combination, to represent an instance of a dialog marker and its local context. I decided to employ supervised machine learning to build the proposed predictive model. The detailed methodology and evaluation results are described in the following chapter.

[4]Audio available at https://github.com/anath2110/ISG-Credit-Card-Billing-Dialog-Corpus/tree/master/OkayEditedMismatched

Chapter 4

Proposed Prosody Prediction

This chapter details Nath and Ward (2022)'s demonstration that appropriate prosody for a dialog marker can be approximately predicted directly from the prosodic information obtained from its context of use in a dialog.

4.1 Data Set

As the aim is to develop a predictive prosodic model generally suitable for open-domain conversations, the Switchboard (Godfrey et al., 1992) corpus of American English telephone conversations is well suited to the task. Also, in open-ended spontaneous conversations, we are more likely to find a greater frequency and more variation in the contexts of use of the dialog markers than in task-oriented dialogs since the nature of the genre— topic or goal of conversations, the intent of the speakers and opinion of the participants— influences discourse marker use (Tübben and Landert, 2022; Verdonik et al., 2008).

After excluding recordings with poor audio quality or artifacts that bothered the pitch tracker, 1900+ conversations were considered involving 400+ speakers. All audio spans bearing labels from the list in Table 4.1 were considered, according to the Picone transcriptions (Deshmukh et al., 1998). These were the most common dialog markers, those with more than 1000 instances each in the corpus. No subjective labels (Calhoun et al., 2010) or any additional checks were done, so cases, where the word was not actually being used as a dialog marker, were not excluded.

Table 4.1: Number of instances of each dialog marker used from the Switchboard corpus.

	token count
huh	1417
now	5910
oh	14053
okay	3915
really	11009
right	12448
uh	52230
uh-huh	12155
um	16392
well	16701
yeah	33768
yes	3393
Total	183391

4.2 Prosodic Features Predicted

Ultimately, I would like to predict every detail of the prosody of each dialog marker token: the value for every feature at every frame of the token. However, for now, only four features are predicted namely:

- loudness or volume, as measured by its acoustic correlate *log energy*,
- pitch, as measured by its acoustic correlate *fundamental frequency* or *f0*, estimated by the pitch tracker *fxrapt* (Talkin and Kleijn, 1995) from the MATLAB's speech processing tool, *Voicebox* (Brookes, 2019),
- *cepstral flux*, as a measure of lengthening and reduction. This computes frame-to-frame cepstral difference to capture the rate of change of cepstrum in a dialog segment using James Lyons's Cepstral Coefficient's implementation, and
- the *harmonic ratio* (Kim et al., 2005), which is a proxy for harmonicity and, indirectly, other properties of voicing, including the absence of creakiness, breathiness, and devoicing.

Table 4.2: Predicted features

mean over the token	max over the token
log energy	log energy
cepstral flux	cepstral flux
pitch	pitch
harmonic ratio	harmonic ratio

The relevance of pitch, energy, and timing is well known; harmonicity was also included since it appears to help differentiate among roles for many dialog markers, as seen earlier in Section 3.3.

For each of these four features, two experiments were conducted. One was to predict the average of the feature values over the entire token. The other was to predict the maximum feature value in the entire token, as both contribute to what is perceived. Table 4.2 summarizes this information. The pitch frames with undefined values were excluded from the computations of mean and max.

4.3 Prosodic Features used for Prediction

As the primary goal is to explore, I used neither a maximal set of features nor a minimal one. Rather a set of 72 prosody features were chosen that were diverse, convenient, reliable, and broadly covered the local context. Contextual features were used for both speakers: the one who produced the dialog marker and the interlocutor. Together these features cover the time from 3.2 to 0 seconds before the start of the dialog marker and the time from 0 to 3.2 seconds after its end (Figure 4.1). Future information was also considered because it was observed, as described in Chapter 3, that often the prosody of the dialog marker is suitable not only based on what came before but also for what is upcoming, either by the same speaker or by the interlocutor, to the extent that the prosody of a dialog marker can guide the interlocutor's future behavior. Specifically, for each speaker, 36 features were computed as shown in Table 4.3: 9 base features, each computed over 4 time spans. All were computed using the Midlevel Prosodic Features Toolkit (Ward, 2015—2022). The four

Table 4.3: Features used for prediction from 3.2 sec context). Times are in milliseconds relative to the start (s) and end (e) of the dialog marker whose prosody is being predicted

	Past Windows, from the start (s) of the token	Future Windows, from the end (e) of the token
log energy lengthening peak disalignment	s–3200 to s–800, and s–800 to s	e to e+800, and e+800 to e+3200
creakiness pitch lowness pitch highness narrow pitch wide pitch speaking rate	s–1600 to s–200, and s–200 to s	e to e+200, and e+200 to e+1600

pitch-related features enabled everywhere-meaningful computation of pitch information, even over windows with few or no pitch points (Ward, 2019). The "peak disalignment" feature measures the displacement between energy peaks and pitch peaks (Ward, 2018). For this data, this feature generally measures the late (delayed) peak occurring in stressed syllables. The specific time windows were chosen based on some initial intuitions about the rate of local prosodic change relative to broader movements and were not subsequently optimized or revisited. Together these features capture much information about the local prosody—from the immediate and wider recent context—and the local turn-taking state.

Both the predicted features, i.e., the dialog marker token features, and the predictors, i.e., the local context features, were z-normalized per speaker as follows:

$$z = (x - \mu)/\sigma$$

where x is the computed feature value for a single frame in a conversation, μ is the mean over the channel, and σ is the standard deviation of the values per speaker in the conversation.

A z-normalized value represents the number of standard deviations that the original

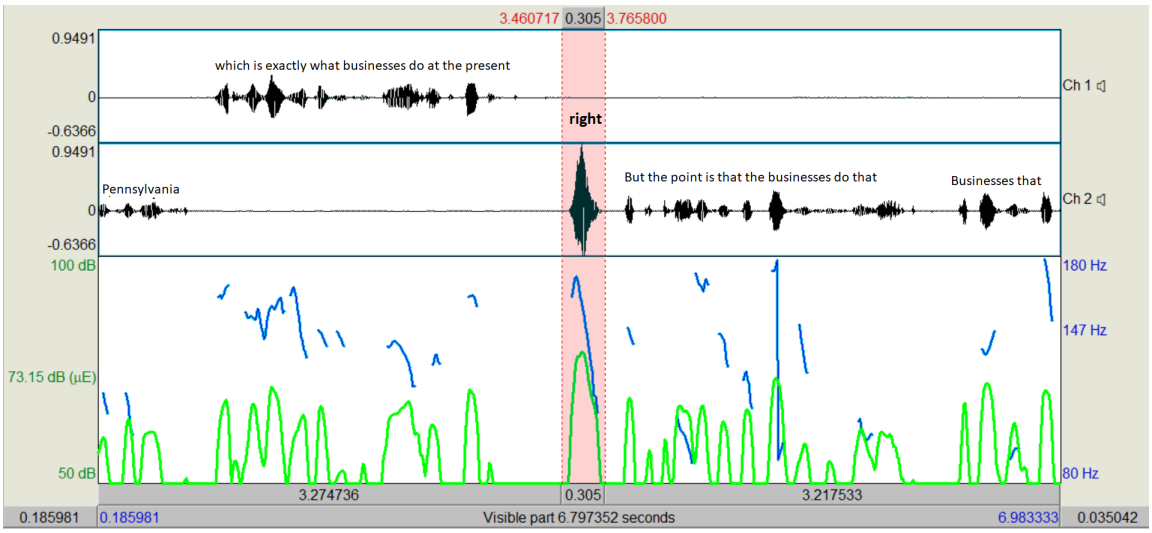

Figure 4.1: Diagrammatic representation of 3.2 sec context audio segment, each from the past and future of the dialog marker token, *right*: Pitch (blue) and Volume (green) contours generated by *Praat*

value is from the mean. The mean of a z-normalized dataset is 0, and the standard deviation is 1. This computation was done using MATLAB's in-built *normalize* function. This z-normalization was done to reduce the effects of intrinsic speaker differences. Also, when a z-normalized dataset is used to fit a machine learning model, any outlier is likely to have a lesser influence on the model fit.

4.4 Evaluation Approach

An intra-corpus evaluation approach (*Evaluation* phase in Fig. 4.2), was followed. Each model, one for each of the twelve dialog marker types, was evaluated with a disjoint train-test split of 70:30, chosen such that the test set contained no dialogs seen in the training set. However, no speaker separation was done, meaning the same speaker could speak in both the train and test sets, possibly more than once.

The performance of each model was evaluated in terms of similarity between the predicted and observed prosody. Owing to the lack of any universally accepted measure for determining prosodic similarity, this was approximated by one of the most popular met-

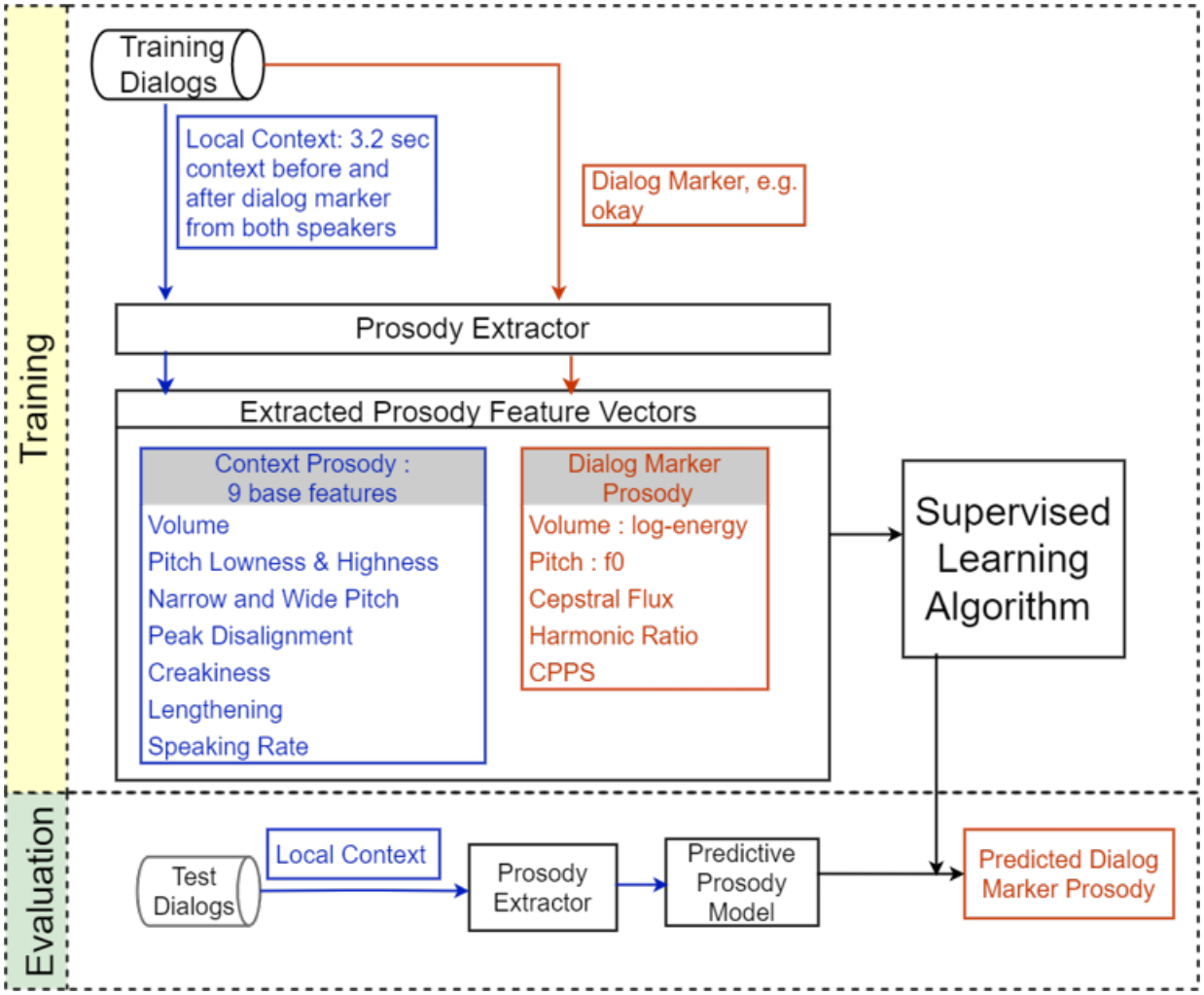

Figure 4.2: Proposed prediction model: An overview

rics used in regression tasks, namely, root mean squared error (RMSE). The utility of local context information was measured by the percent reduction in the RMSE values for model predictions compared to the baseline. The baseline, here, is simply predicting the average over all instances of the particular dialog marker type under investigation. For example, the baseline to compare any predicted feature value for an instance of *yeah* would simply be the global average of its value across all instances of *yeah*.

4.5 Linear Regression Model

4.5.1 Training the Model

The training phase of Fig. 4.2 presents an overview of the process of training the proposed predictive prosody model. Unlike prior research that either primarily relied on human perception (c.f. Section 2.3) or followed a time-consuming approach of generating hand-crafted rules (Acosta and Ward, 2011), in this work, I follow an automated approach for predicting context-dependent prosody.

My first approach involved one of the simplest machine learning algorithms: multi-variate linear regression. This algorithm used the default least-square sum function to select the best-fit line. Using a relatively simple approach enables easy backward reasoning to analyze the underlying features better and further improve the solution. This enabled me to easily examine how the context features affected the dialog markers' prosody.

A separate model was developed for each dialog marker category, as it is not expected that the same rules to work well for all dialog marker types: for both *huh* and *okay*, for example (cf. Section 5.2).

As explained earlier, each model predicted two types of features for each dialog marker token, namely mean and max. The first represents the average or the mean, and the latter represents the maximum of the features over all frames for a single dialog marker token.

4.5.2 Evaluation Results

Table 4.4: Prediction errors with the baseline and with the linear regression model, and error reduction rate for predicting mean (respectively maximum) features. Prediction errors are the average RMSE values obtained for each of the 12 dialog markers. le is log energy, cf is cepstral flux, p is pitch, and hr is harmonic ratio.

	predicting mean features				predicting max features			
	le	cf	p	hr	le	cf	p	hr
Baseline Average RMSE	0.61	0.82	0.67	0.66	0.74	1.38	1.95	1.18
Model Average RMSE	0.44	0.64	0.52	0.46	0.49	0.92	1.34	0.74
Reduction, %	28.7	22.4	23.6	29.6	33.4	31.9	31.1	37.2

The above model is evaluated as described in Section 4.4. Table 4.4 shows the quality of the baseline and linear regression model predictions. The errors are lower with the model, with reductions ranging from 22% to 37%, showing that the local prosodic context is informative. The benefit is statistically significant for all four predicted features and in both cases of predicting mean and max features (matched pairs t-tests, $p << 0.01$). It is also

Table 4.5: Prediction error reduction rate[1]per dialog marker type using linear regression: Percent reduction in root mean squared error for predicting mean (respectively maximum) feature values.

	predicting mean features					predicting max features				
	le	cf	p	hr	Avg.	le	cf	p	hr	Avg.
huh	29	43	26	34	33	22	46	27	15	28
now	19	21	30	25	24	23	21	44	39	32
oh	17	27	15	24	21	35	37	25	37	34
okay	18	30	7	6	15	19	31	18	40	27
really	46	16	-1	15	19	37	39	35	48	40
right	37	2	7	25	18	31	14	19	39	26
uh	43	21	56	30	38	52	48	33	32	41
uh-huh	12	20	34	50	29	18	11	31	39	25
um	31	22	29	32	28	56	50	36	27	42
well	25	24	23	33	26	32	24	40	33	32
yeah	38	25	20	44	32	49	38	40	54	45
yes	30	19	37	37	31	28	22	26	45	30
Average	29	22	24	30	26	33	32	31	37	33

[1]Note: Average values, in this and in any subsequent table in this dissertation that reports percent

seen that the reductions were greater for the maximum features than for the mean features, although this difference may be largely due to outliers. Also, it is seen that the mean and maximum harmonic ratio were most easy to predict: their predictions had the highest error reduction rate, 30%, and 37% (Table 4.5), respectively, from the baseline. Predicting mean pitch was hardest for the response marker type, *really*.

4.6 Failure Analysis

The above evaluation results support the hypothesis: local prosodic context alone is sufficiently informative to approximately predict the dialog markers' prosody

This section probes into the regularities that the above model learned and discusses the strengths and limitations of prediction from prosodic context alone.

- For most dialog markers, the most predictive features were the peak disalignment features. For most windows, these features had correlations of 0.20 or higher with high pitch, high volume, and high harmonicity. This is likely because peak disalignment often marks times of shared laughter, questions, and other high-engagement dialog acts (Ward, 2018), and these generally call for enthusiastic dialog markers.

- There was also a tendency to matching: more specifically, when the immediate past context exhibits higher volume or pitch, the prosody of the dialog marker often does too, for example, when acknowledging new information.

- Some specific dialog markers had additional unique tendencies. For example, for the word *now*, high pitch correlated with a high pitch by the same speaker over the next few windows, likely due to its forward-looking role, as in introducing new subtopics.

- Most of the strong correlations were with contextual behavior by the person who produced the dialog marker, but there were also interlocutor effects. For example, the

reduction of prediction errors for a model from its respective baseline, were computed before rounding off the values.

word *okay* tended to be lower in pitch when the interlocutor's cepstral flux was low in the context. This happened likely due to the use of lengthening and reduction to mark a low density of new information or seeking only weak feedback.

For insight into why the model sometimes performed well and sometimes poorly, I started by considering Table 4.5. I noted that:

1. The relatively high predictability for *uh* and *huh* was likely because they often have no independent prosody or meaning beyond their roles in the local context.

2. Low predictability was observed for mean features of *really* and for both mean and max features of *right*, which are sometimes dialog markers, but sometimes just adverbs and adjectives, in which roles they likely have different prosodic tendencies.

3. The prosody of *okay* was also hard to predict, perhaps because it often is deployed to convey a specific meaning or function rather than just fitting passively in the context. This can also be verified from the initial exploratory results (cf. Section 3.2), where the prosodic pattern of *okay* was more pronounced in scenarios where it had distinctive pragmatic functions like acknowledgment and topic shift or acknowledgment and seeking confirmation than in scenarios where it was used to acknowledge the speaker's information or to end the conversation.

To further understand where this model succeeded and failed, its performance was examined on specific tokens: for each dialog marker type, the five for which the predictions were least accurate and the five for which they were most accurate. This was done subjectively, relying on my perceptions and those of a native speaker researcher and qualitative inductive methods.

Factors that were common when the model failed included:

1. background noise in the audio segment. (Feature computations in the midlevel toolkit are not robust to noise.)

2. long monologues (a dialog activity type uncommon in Switchboard, and therefore likely rare in the training data).

3. one speaker with an unusual accent or perhaps a speech impediment,

4. incorrect annotations, though very rare in occurrence, for example, where the label was *um*, but the sound was more like *hmm*. (The model for *um*, of course, had not been trained to predict the prosody of *hmm* tokens.)

5. sequences of dialog markers, such as *well, yeah* and *oh, okay*. (The prosody of markers in a sequence is apparently different from those in isolation, the more common case in the training data.)

6. *okay* at the end of the conversation, where it was short and breathy as part of the closing, which is also what was observed in my exploratory study, Chapter 3, and

7. *huh* when produced as a repair question or strong exclamation.

The cases where the model's predictions were most accurate include:

1. typical backchannel uses of *yeah*,

2. times the speaker and interlocutor shared happy or excited agreement, for example, *You're pretty Texan, yes . . . [interlocutor laughter]*, and

3. sympathetic productions of *really* in the context of talk about troubles or problems, as in *that can really be a problem.*

Overall, it seems that the model tends to perform well when the local dialog context is of a type that is common in the Switchboard genre.

Chapter 5

Improving Prediction, Part 1: Investigating Contributing Factors

The aim of this chapter is to investigate the factors contributing to the prediction performance of the proposed prosody prediction model and thereby, improve the same. Specifically, the same linear regression algorithm is trained with:

1. either only the past or the future context features
2. either the speaker's or the interlocutor's context features
3. features from all the dialog markers' instead of only the one whose prosody is being predicted
4. fine-grained context features
5. features from a wider (10 sec) context
6. features whose values are scaled to the range (0–1),
7. a feature set that additionally includes the CPPS feature, and
8. features combined from both speech and text modalities

5.1 Type of Context

This section tests the prediction model's performance via some ablation experiments, that is, removing specific components to understand their contribution to the overall model.

Table 5.1: Summary of results for predicting mean features using linear regression with either only the past or the future context.

	past context only				future context only			
	le	cf	p	hr	le	cf	p	hr
Baseline Average RMSE	0.61	0.82	0.67	0.66	0.61	0.82	0.67	0.66
Model Average RMSE	0.58	0.80	0.60	0.61	0.61	0.78	0.58	0.66
Reduction, %	6.00	2.90	10.2	7.70	0.00	4.88	13.4	0.00

5.1.1 Ablation Study #1: Future & Past vs. Future Only vs. Past Only

To investigate the utility of local dialog context for many realistic scenarios where the future context is not usually available, I measured the quality of predictions made using only the past context features.

A linear regression model was trained on the local context features from both the dialog speakers, but now only on those prior to the dialog marker token: specifically on 3.2 sec of past context only. Thus, any future context information was excluded entirely.

Using the past context features alone provided some benefit with the reduction in prediction error that ranged from 2.9% to 10.2% (Table 5.1), but was much less than those obtained when also using the future context (c.f. Table 4.4).

To explore whether using the future context alone has any benefits, a linear regression model was trained on only the future context features from both the dialog speakers. Mixed performance results (Table 5.1) were produced. Predicting mean cepstral flux and mean pitch has improved performance more than when using the past context alone, while predicting mean log energy and harmonic ratio produces no benefit at all.

We may safely conclude, however, that using both the context (c.f Table 4.4), enables more appropriate prosody prediction than using either of the past or future context alone. This further implies that at its current state, this context-dependent approach of predictive prosody modeling is not directly applicable to real-time prosody predictions.

5.1.2 Ablation Study #2: Both Speaker vs. Self Only vs. Interlocutor Only

In this section, I describe yet another ablation study. I compared the performance of linear regression models that used only one of the dialog speaker's context information: either the same speaker who produced the dialog marker or the interlocutor.

Linear regression using only the interlocutor context produced better prediction results than that using only the self-speaker context (Table 5.2) for predicting the mean of three out of four response features: cepstral flux, pitch, and harmonic ratio. The harmonic ratio was the easiest to predict when using only the interlocutor's context. With respect to the dialog marker type, *yes* and *really* had the best overall prediction results. Perception analysis of some of the dialog marker tokens that had the best-predicted mean harmonic ratio revealed that:

- the harmonicity of *really* was predicted best when it was low or when the speaker was less harmonic in the context, that is either: i) when they were breathier using *really* as an adjective in a self-directed speech segment (Ward et al., 2022), or ii) when they sounded nasal to attach more importance and confidence in their response to a context where *really* was used as an interjection, marking the receipt of surprising new information as in *oh really*.

- the low harmonicity of *yes* was predicted best when the speaker's voice was creaky in the context, demonstrating their confidence and authority.

It can be seen that neither of the individual contexts generated predictions as well as using both speakers' contexts (c.f. Table 4.4), for any of the features.

5.2 Type of Training: Generic vs Type-Specific

Here, a global training approach was adopted where the model was trained on the entire set of dialog markers and tested for its ability to predict the token features of each particular type.

Table 5.2: Summary of results for predicting mean features using linear regression with only one speaker's context from both the past and the future.

	self-speaker context only				interlocutor context only			
	le	cf	p	hr	le	cf	p	hr
Baseline Average RMSE	0.61	0.82	0.67	0.66	0.61	0.82	0.67	0.66
Model Average RMSE	0.56	0.74	0.54	0.63	0.57	0.73	0.49	0.54
Reduction, %	8.20	9.76	19.40	4.55	6.56	10.98	26.87	18.18

Table 5.3: Summary of results for predicting mean features for any dialog marker type using linear regression trained on features of all the marker types

	predicting mean features			
	le	cf	p	hr
Baseline Average RMSE	0.61	0.82	0.67	0.66
Model Average RMSE	0.46	0.70	0.51	0.53
Reduction, %	24.5	14.8	23.8	20.1

For example, the model that predicted the mean features of each token of *yeah* was trained on the features from all of the twelve dialog marker types. This was done to investigate if there is any inter-categorical effect in the prosody of the dialog markers, that is, for example, whether the prosody of the local context that affected the prosody in the dialog marker token *yeah* would also equally affect the prosody of the token *okay*, if used in a similar context.

As seen in Table 5.4, this global model performed modestly worse than those trained on the specific dialog markers (c.f. Table 4.4): the prediction errors were more for the former. This result is easy to understand as it reflects the prosody of local context for one marker type, say, *yeah* is not as informative while predicting the prosody of another type of response, for example, *yes*. Table 5.4 lists the individual RMSE values for predicting each marker type's mean of token features. Also, as seen in Table 5.5, predicting mean cepstral flux was rather more difficult with the lowest average prediction error reduction of 15%. Also, prediction for response marker types, *um* was the worst.

Table 5.4: Prediction errors per dialog marker type using the generic model: Root mean squared error for predicting a token's mean feature values of any type using linear regression model trained on context feature sets of all marker types.

	predicting mean features				
	le	cf	p	hr	Avg.
huh	0.44	0.77	0.68	0.46	0.59
now	0.44	0.61	0.41	0.46	0.48
oh	0.48	0.95	0.68	0.56	0.67
okay	0.48	0.85	0.48	0.45	0.57
really	0.54	0.75	0.54	0.60	0.61
right	0.47	0.57	0.40	0.47	0.48
uh	0.45	0.72	0.47	0.51	0.54
uh-huh	0.44	0.49	0.37	0.66	0.49
um	0.40	0.59	0.42	0.44	0.46
well	0.51	0.82	0.67	0.59	0.65
yeah	0.48	0.64	0.48	0.51	0.53
yes	0.43	0.67	0.51	0.62	0.56
Average	0.46	0.70	0.51	0.53	0.55

Table 5.5: Percent reduction of prediction errors per dialog marker type using the generic model: Percent reduction of root mean squared error, from the baseline of predicting a token's mean feature values of any type, using linear regression trained on features of all marker types.

	predicting mean features				
	le	cf	p	hr	Avg.
huh	39	15	16	28	25
now	19	27	34	21	25
oh	14	11	15	11	13
okay	24	-3.7	29	8.2	14
really	10	8.5	20	13	13
right	16	29	29	24	24
uh	20	30	4	30	21
uh-huh	17	25	30	4.3	19
um	9	-28	28	28	9.1
well	29	28	26	27	28
yeah	38	11	19	22	22
yes	41	-6.3	34	21	22
Average	25	15	24	20	21

5.3 Alternate Context Feature Sets

5.3.1 Fine-Grained Context Features

The original hypothesis of this research being supported to a certain extent, I try using an alternative set of features to represent the context used to train the models, an approach hypothesized to improve the prediction quality.

Table 5.6: Set of fine-grained context (3.2 sec) prosody features used for prediction. Times are in milliseconds relative to the start (s) and end (e) of the dialog marker whose prosody is being predicted

	Past Windows, from the start (s) of the token	Future Windows, from the end (e) of the token
log energy lengthening peak disalignment	s-3200 to s-2400, s-2400 to s-1600, s-1600 to s-800 and s-800 to s	e to e+800, e+800 to e+1600, e+1600 to e+2400 and e+2400 to e+3200
creakiness pitch lowness pitch highness narrow pitch wide pitch speaking rate	s-3200 to s-1600, s-1600 to s-800, and s-800 to s	e to e+800, e+800 to e+1600, and e+1600 to e+3200

Table 5.7: Summary of prediction results using linear regression trained on the set of fine-grained context prosody features (Table 5.6).

	predicting mean features			
	le	cf	p	hr
Baseline Average RMSE	0.61	0.82	0.67	0.66
Model Average RMSE	0.45	0.66	0.50	0.49
Reduction, %	26.5	20.0	24.7	26.1

Table 5.8: Root mean squared error for predicting mean feature values using linear regression trained on the set of fine-grained context features (Table 5.6)

	predicting mean features				
	le	cf	p	hr	Avg.
huh	0.45	0.72	0.67	0.48	0.58
now	0.42	0.62	0.43	0.48	0.49
oh	0.47	0.84	0.63	0.52	0.62
okay	0.48	0.74	0.49	0.46	0.54
really	0.40	0.62	0.55	0.48	0.51
right	0.47	0.57	0.41	0.49	0.48
uh	0.47	0.73	0.43	0.54	0.54
uh-huh	0.44	0.49	0.42	0.45	0.45
um	0.42	0.56	0.38	0.45	0.45
well	0.47	0.76	0.61	0.54	0.59
yeah	0.48	0.63	0.49	0.52	0.53
yes	0.43	0.62	0.51	0.44	0.50
Average	0.45	0.66	0.50	0.49	0.52

More windows were used for computing the local context features. The 9 base features remained the same, namely, log energy, lengthening, creakiness, peak disalignment, pitch lowness, pitch highness, narrow pitch, wide pitch and speaking rate, but they were computed over more windows, as shown in Table 5.6 than before (c.f. Table 4.3). A total of 120 context features, 60 per speaker, were computed.

Linear regression that used this new set of local context features reduced the RMSE of predicting mean pitch, to 24.7 % (Table 5.7), which is slightly better than when trained with the original set of context features (Table 4.4).

However, training with this extended set of context features had a lesser positive effect on predicting the remaining three response features. Prediction errors by the model in terms of RMSE values and the error reduction rate from the baseline for predicting mean log-energy, cepstral flux, pitch, and harmonic ratio for each dialog marker type are listed in Tables 5.8 and 5.9, respectively. Interestingly, as is seen in Table 5.9, response token of type *um* shows a negative error reduction, -21%, for predicting mean cepstral flux and only a slight positive improvement, 4% for predicting mean log energy. This could imply that the prosody of *um* is more influenced by immediate local context.

Table 5.9: Percent reduction in the root mean squared error for predicting mean features using linear regression trained on the set of fine-grained context features (Table 5.6)

	predicting mean features				
	le	cf	p	hr	Avg.
huh	37	21	17	25	25
now	22	26	30	17	24
oh	16	21	21	17	19
okay	25	9	27	7	17
really	33	24	20	30	27
right	17	29	27	22	23
uh	16	29	12	26	21
uh-huh	16	23	21	34	24
um	4	-21	33	25	10
well	34	33	33	34	33
yeah	37	13	15	21	21
yes	41	1	34	43	30
Average	27	20	25	26	24

5.3.2 Wider Context

Table 5.10: Set of Context features used for prediction from a wider context (10 sec). Times are in milliseconds relative to start (s) and end (e) of the dialog marker whose prosody is being predicted

	Past Windows, from the start (s) of the token	Future Windows, from the end (e) of the token
log energy lengthening	s-10000 to -8000,	e to e+2000,
peak disalignment creakiness	s-8000 to -6000,	e+2000 to e+4000
pitch lowness pitch highness	s-6000 to -4000	e+4000 to e+6000,
narrow pitch wide pitch speaking rate	s-4000 to s-2000, and s-2000 to s	e+6000 to e+8000, and e+8000 to e+10000

In this section, I experiment with another set of context features (Table 5.10). This time the local context region includes 10 sec of context each to the past and the future of the dialog

marker token from both the dialog speakers from the dialog marker token. Linear regression trained on this wider context is hypothesized to improve the prediction performance.

Table 5.11: Summary of prediction results using linear regression trained on prosody features from a wider context 10 sec (instead of 3.2 sec) each from the past and the future as well as from both the speakers.

	predicting mean features			
	le	cf	p	hr
Baseline Average RMSE	0.61	0.82	0.67	0.66
Model Average RMSE	0.43	0.64	0.49	0.46
Reduction, %	29.8	22.6	26.5	30.7

Table 5.12: Root mean squared error for predicting mean feature values using linear regression trained on 10 sec local context features both from the past and the future of both the speakers (Table 5.10).

	predicting mean features				
	le	cf	p	hr	Avg.
huh	0.49	0.47	0.60	0.41	0.49
now	0.45	0.65	0.38	0.47	0.49
oh	0.35	0.97	0.74	0.43	0.62
okay	0.50	0.74	0.64	0.36	0.56
really	0.41	0.64	0.52	0.48	0.51
right	0.46	0.58	0.39	0.46	0.47
uh	0.25	0.75	0.11	0.52	0.41
uh-huh	0.46	0.52	0.35	0.36	0.42
um	0.28	0.36	0.35	0.45	0.36
well	0.50	0.77	0.56	0.54	0.59
yeah	0.44	0.63	0.54	0.38	0.50
yes	0.58	0.58	0.72	0.63	0.63
Average	0.43	0.64	0.49	0.46	0.50

As seen in Table 5.11, this model performed moderately but consistently better than the linear regression trained on only 3.2 sec of context (c.f Table 4.4) or with the extended set of fine-grained context feature set (c.f Table 5.7), for predicting mean of each of the response features: log energy, cepstral flux, pitch, and harmonic ratio.

Table 5.13: Percent reduction in the root mean squared error for predicting mean feature values using linear regression trained on 10 sec of local context features both from the past and the future of both the speakers (Table 5.10).

	predicting mean features				
	le	cf	p	hr	Avg.
huh	31	48	25	36	35
now	17	23	38	20	24
oh	38	10	8	32	22
okay	21	10	7	27	16
really	31	21	24	30	27
right	18	27	30	26	25
uh	55	27	78	29	47
uh-huh	13	20	34	48	29
um	36	22	38	26	30
well	31	33	38	34	34
yeah	43	12	8	42	26
yes	21	8	6	19	13
Average	30	22	28	31	27

5.3.3 Adding CPPS feature

In this section, Cepstral Peak Prominence Smoothed (CPPS) is used both as a predictor as well as a predicted/response feature.

Hillenbrand et al. (1994) measures CPP or Cepstral Peak Prominence as the difference in amplitude between the cepstral peak and the corresponding predicted magnitude for the quefrency at the cepstral peak. It is calculated using an inverse Fast Fourier Transform (FFT) of the log power spectrum of a voice signal. A CPP variation is called Cepstral Peak Prominence-Smoothed (CPPS). It has different calculation algorithms, such as smoothing the cepstrum before extracting the peak and using 1024 frames every 2 ms instead of 10 ms (P.S. and Pebbili, 2020), which makes CPPS more robust and with no influence of artifacts. The literature demonstrates CPPS as an acoustic measure of overall voice quality (Maryn et al., 2009), negatively correlating with breathiness in voice (Ward et al., 2022). It integrates measures of several features describing the aperiodicity and waveform of the acoustic voice signal (Fraile and Godino-Llorente, 2014). It is considered a reliable cue of overall dysphonia (Wolfe and Martin, 1997; Heman-Ackah et al., 2002) and also has been used in clinical voice

Table 5.14: Set of context prosody features used as predictors in Table 4.3 extended to also include the feature Cepstral Peak Prominence Smoothed (CPPS). Times are in milliseconds relative to the start (s) and end (e) of the dialog marker whose prosody is being predicted

	Past Windows, from the start (s) of the token	Future Windows, from the end (e) of the token
log energy lengthening peak disalignment CPPS	s–3200 to s–800, and s–800 to s	e to e+800, and e+800 to e+3200
creakiness pitch lowness pitch highness narrow pitch wide pitch speaking rate	s–1600 to s–200, and s–200 to s	e to e+200, and e+200 to e+1600

Table 5.15: Summary of results for predicting prosody features including mean CPPS using linear regression trained on the set of predictors that also include CPPS (Table 5.14)

	predicting mean features				
	le	cf	p	hr	cpps
Baseline Average RMSE	0.61	0.82	0.67	0.66	0.60
Model Average RMSE	0.42	0.63	0.50	0.46	0.38
Reduction, %	31.8	24.1	24.5	30.6	35.9

evaluation (Murton et al., 2020). Hence, this feature could be extremely useful in predicting appropriate dialog marker prosody.

Here, Cepstral Peak Prominence Smoothed is computed using the code developed by Marcin Włodarczak in 2020 and available in the mid-level toolkit (Ward, 2015—2022)). The local context (3.2 sec) feature set (Table Table 4.3) is extended to include the CPPS from both the past and future context of both the speakers (Table 5.14). Linear regression trained

with this feature set is applied to predict mean CPPS in addition to predicting mean log energy, cepstral flux, pitch, and harmonic ratio.

The average prediction performance is seen to be improved, Table 5.15, for each of the response features (log energy, cepstral flux, pitch, and harmonic ratio), over what was obtained with the original set of context features (c.f Table 4.4). It is also seen that adding a breathiness feature (CPPS) is usefully informative for specific tokens and features, notably those that predicted least well with the original set of context features (c.f. Table 4.5).

As a predicted feature, CPPS was the easiest to predict overall than the other response features. The best average predictions were obtained for the dialog marker *really* and the worst for *yes*.

Further listening to 20 of these audio segments of *yes* that were predicted with large prediction errors revealed that:

- Of those with predicted CPPS much higher than the true value, several:
  - were emphatic with a stressed vowel,
  - had a rising pitch, and
  - and had no perceived breathiness.
- Of those with predicted CPPS much lower than the true value, several:
  - were creaky and lengthened, and
  - sometimes ended abruptly.

Overall, it can be said that the prediction for CPPS failed miserably when the tokens were rather idiosyncratic uses in the corpus that did not conform to the local dialog flow.

Table 5.16: Percent reduction in the root mean squared error for predicting mean feature values using linear regression trained local context features that also include CPPS both from the past and the future of both the speakers (refer Table 5.14).

	predicting mean features					
	le	cf	p	hr	cpps	Avg.
huh	27	26	25	35	36	29
now	33	45	27	35	41	36
oh	33	24	25	35	36	30
okay	20	22	30	45	38	31
really	45	22	57	31	42	39
right	39	26	22	45	39	34
uh	19	30	17	26	35	25
uh-huh	21	23	32	27	35	28
um	31	20	39	42	40	34
well	39	5	10	27	30	22
yeah	49	18	5	19	32	24
yes	21	33	10	11	25	20
Average	32	24	24	31	36	29

5.3.4 Optimal Feature Set

As was illustrated in the previous sections, linear regression using 10 sec of context features outperformed the one using 3.2 sec of context features. Also, linear regression using 3.2 sec context performed better with CPPS as one of the predictors than without it. Hence, I decided to create an optimal predictor set with the best possible combination of feature sets so far: 10 sec context that additionally includes CPPS as the predictor (Table 5.17).

As expected, linear regression trained on the above set of features produced a higher quality of predictions for each predicted feature: mean log energy, cepstral flux, pitch and harmonic ratio (Table 5.18) than the one trained on only 3.2 sec of context (c.f Table 4.4) or on 10 sec of context features with no CPPS (c.f. Table 5.12) or even using CPPS in a 3.2 sec context (Table5.15). This also reduced the overall average prediction error by 37% which is the best result so far. Thus, it is demonstrated that an appropriate set of engineered features can ensure better prediction even with the simplest of machine learning algorithms.

In another experiment, a step-wise linear regression algorithm was employed to garner

Table 5.17: Set of context features used for prediction from 10 sec context also including CPPS as a predictor. Times are in milliseconds relative to start (s) and end (e) of the dialog marker whose prosody is being predicted

	Past Windows, from the start (s) of the token	Future Windows, from the end (e) of the token
log energy lengthening peak disalignment creakiness pitch lowness pitch highness narrow pitch wide pitch speaking rate CPPS	s-10000 to -8000, s-8000 to -6000, s-6000 to -4000 s-4000 to s-2000, and s-2000 to s	e to e+2000, e+2000 to e+4000 e+4000 to e+6000, e+6000 to e+8000, and e+8000 to e+10000

Table 5.18: Summary of results for predicting prosody features including mean CPPS using linear regression trained on a set of features that also include CPPS as a predictor (Table 5.17)

	predicting mean features				
	le	cf	p	hr	CPPS
Baseline Average RMSE	0.61	0.82	0.67	0.66	0.60
Model Average RMSE	0.39	0.60	0.42	0.40	0.34
Reduction, %	37.1	27.3	37.2	39.1	42.6

more information on the more contributing predictors using the same feature set as above. At first, linear regression using MATLAB's *fitlm* method was used to create a model using the entire feature set. Then the *step* method, using a 5-fold cross-validation approach, recursively removes a predictor from the set if the increase in the adjusted R-squared value of the model is less than -0.01 or, in other words, if the goodness of the fit is reduced by more than 1%. The prediction results were the same as in Table 5.18. Also, none of the predictors was excluded from the model at any point. This could possibly mean that each predictor contributed with some positive value to this prediction task, at least for this model.

5.4 Scaled Features

In another experiment, features were scaled, using the min-max scaling formula shown below:

$$x - x_{min}/(x_{max} - x_{min})$$

where x_{max} and x_{min} are the maximum and the minimum values of the feature x across each channel of each conversation, respectively.

This technique shifted the feature values for each speaker and re-scaled them so that they end up ranging between 0 and 1.

Scaling the feature values had a slight positive effect when compared to baseline, Table 5.19, for predicting mean log-energy, cepstral flux, pitch, and harmonic ratio. Predicting mean log-energy showed the least improvement from the baseline. Table 5.20 shows the model's prediction errors and Table 5.21 shows the percent reduction of these errors from the baseline for each dialog marker type. Such poor prediction quality that demonstrates the inferiority of this normalization technique is not surprising. This scaling of data increases the possibility of outliers having a high negative impact on the model's performance.

Table 5.19: Summary of prediction results with linear regression trained on 3.2 sec of local context features, scaled within [0-1] instead of z-normalization.

	predicting mean features			
	le	cf	p	hr
Baseline Average RMSE	0.13	0.06	0.105	0.11
Model Average RMSE	0.12	0.06	0.09	0.10
Reduction, %	3.23	5.59	5.65	5.72

5.5 Multi-modality: Text plus Speech

Having already supported my original research claim that the prosody of the local context is directly informative to predict the prosody of the target dialog marker token to a certain

Table 5.20: Root mean squared error for predicting mean feature values using linear regression trained on 3.2 sec of local context features that are scaled within [0-1] instead of z-normalization.

	predicting mean features				
	le	cf	p	hr	Avg.
huh	0.11	0.07	0.13	0.11	0.11
now	0.13	0.05	0.09	0.10	0.09
oh	0.13	0.06	0.13	0.12	0.11
okay	0.13	0.07	0.10	0.10	0.10
really	0.11	0.05	0.11	0.10	0.10
right	0.12	0.06	0.09	0.10	0.09
uh	0.13	0.06	0.07	0.10	0.09
uh-huh	0.12	0.05	0.10	0.11	0.10
um	0.12	0.06	0.08	0.10	0.09
well	0.12	0.06	0.11	0.11	0.10
yeah	0.13	0.06	0.10	0.11	0.10
yes	0.13	0.06	0.09	0.09	0.09
Average	0.12	0.06	0.09	0.10	0.09

Table 5.21: Percent reduction in the root mean squared error from the baseline for predicting mean feature values using linear regression trained on 3.2 sec of local context features, scaled within [0-1] instead of z-normalization.

	predicting mean features				
	le	cf	p	hr	Avg.
huh	5.2	-1.5	0.8	4.3	2.2
now	-2.8	12	10	9.1	7.0
oh	1.5	-3.4	-0.8	4.0	0.3
okay	4.5	2.8	4.8	4.8	4.2
really	4.3	12	11	11	9.6
right	4.8	1.6	2.3	4.0	3.2
uh	3.0	12	5.4	1.9	5.5
uh-huh	-1.7	7.4	1.9	3.5	2.8
um	4.1	0.0	9.7	9.5	5.8
well	6.3	9.1	10	5.8	7.8
yeah	6.0	3.2	4.8	5.1	4.8
yes	3.0	12	9.1	5.3	7.3
Average	3.2	5.6	5.7	5.7	5.0

extent, here, I investigate whether adding the lexical information of the context has any effect on the prosody prediction results.

Following are the hypotheses tested here.

1. Adding lexical information from the context in addition to prosody improves prediction performance.

2. Predictive ability of the context prosody alone is more than the lexical context.

3. Individual lexical context information is more useful than the average of the entire lexical context.

4. Using more lexical context information improves prediction.

The lexical context was represented by an embedding vector obtained from a pre-trained word-embedding file, *glove.6B.50d*. These pre-trained word embeddings were generated using *GloVe* (Pennington et al., 2014), an unsupervised learning algorithm that produced linear representations of the word vector space and has since been used in various text mining tasks. *glove.6B.50d* has 6 billion tokens, 400K vocabulary, is uncased, and has 50 dimensions. Thus, a given word was represented by a 50-dimensional vector of numbers.

Before computing the lexical embeddings, each word was first:

- expanded: contractions, if any, were expanded to their full forms, for example, *there's* $\rightarrow$ there is, *they've* $\rightarrow$ they have, *I'm* $\rightarrow$ I am, and so on and then,

- lemmatized, with the use of a vocabulary and morphological analysis of words that normally aims to remove inflectional endings only and to return the base or dictionary form of a word (i.e. the *lemma*). Here, this is done using the *normalizeWords* function of MATLAB with the parameter *Style* set as *lemma*.

Among many other possible ways to define the local lexical context, here this included three words each from the past and the future of the target dialog marker token from both the speakers, that is twelve words in total. In situations when both speakers speak at the same time, some of these context words might overlap with each other or with the target token, entirely or partially. Nevertheless, even in such cases, each of the context words was separately added to the set. To test the first hypothesis, prosodic context (3.2 sec to the past and the

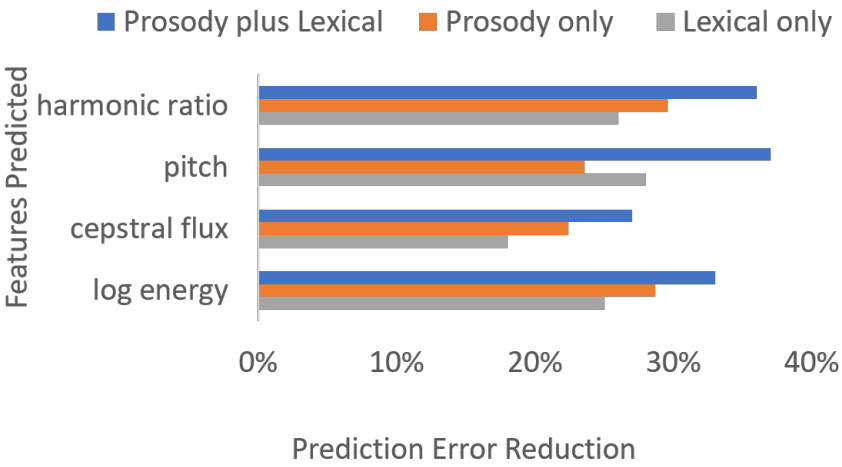

Figure 5.1: Prediction performance comparison using i) prosody plus lexical context, ii) only context prosody, and iii) only lexical context

future of the target token and from both speakers) was represented by the 72-dimensional feature vector (c.f. Table 4.3) and the lexical context by 50-dimensional lexical embeddings averaged over all the context words. Therefore, each predictor was a 122-dimensional vector (72-dimensional prosody vector + 50-dimensional average lexical embedding vector). The results, (Figure 5.1) supported the first hypothesis since linear regression trained on both prosodic and lexical contexts performs better for predicting the mean of each feature when compared to the model trained on only prosodic context.

Using only the lexical information (excluding prosody entirely) to represent the context worsens the prediction results (Figure 5.1) when compared to using only prosodic context, other than for predicting mean pitch, which is, surprisingly, improved. Thus, the second hypothesis is also supported for three out of four predicted features.

Table 5.22: Prediction performance comparison when using i) average lexical embeddings vs. ii) individual lexical embeddings.

	average lexical embeddings				individual lexical embeddings			
	le	cf	p	hr	le	cf	p	hr
Baseline RMSE	0.61	0.82	0.67	0.66	0.61	0.82	0.67	0.66
Model RMSE	0.46	0.68	0.48	0.49	0.46	0.67	0.48	0.49
Reduction, %	25.5	18.0	28.0	25.9	24.9	19.2	27.8	25.3

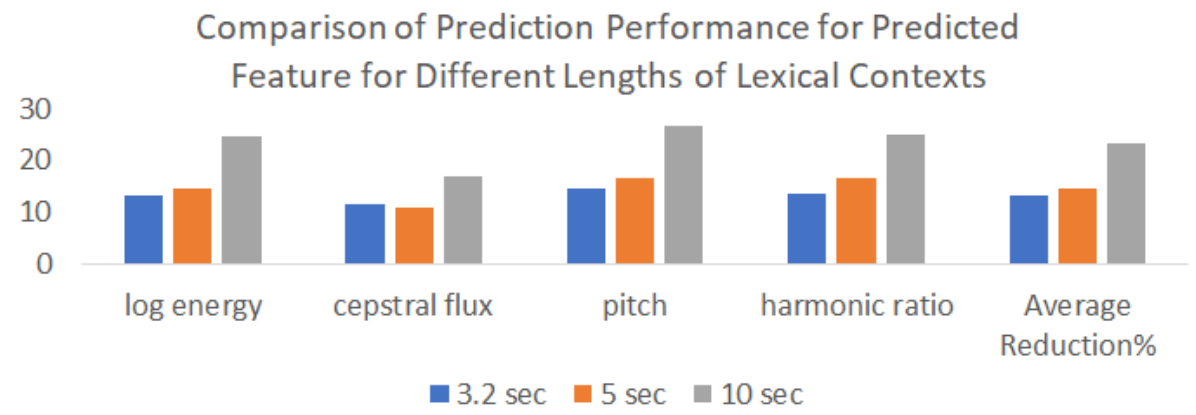

Figure 5.2: Prediction performance comparison for each mean feature predicted using only lexical information from i) 3.2 sec context, ii) 5 sec context, and iii) 10 sec context

To test the third hypothesis, instead of computing the average, the individual lexical embedding vectors corresponding to each of the twelve context words were used, thus, the lexical context was here a 200-dimensional vector. However, no significant difference was found in the prediction results (Table 5.22) when compared to using the average lexical embeddings.

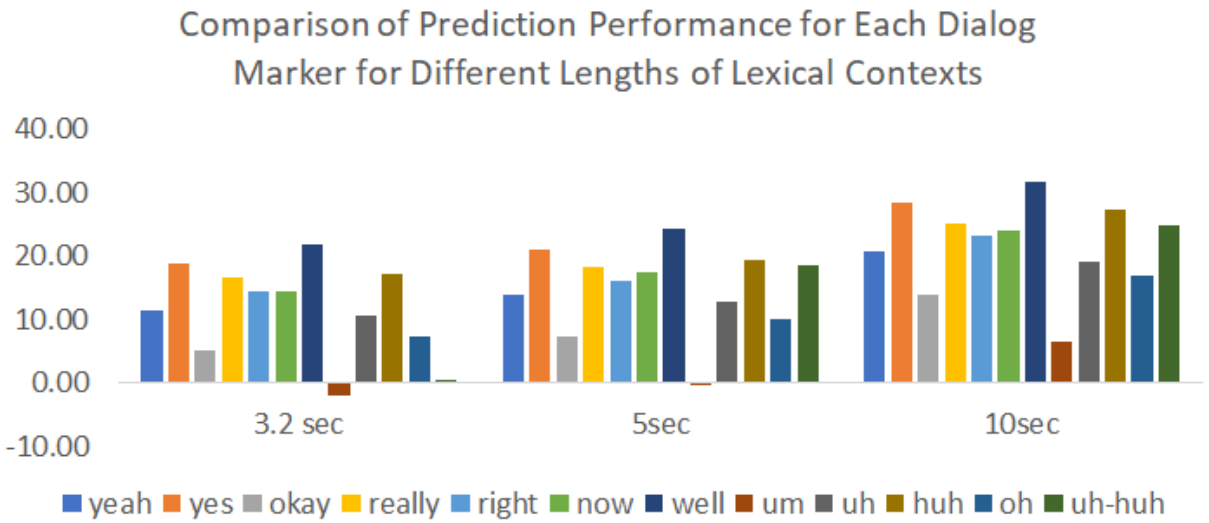

Figure 5.3: Prediction performance comparison for predicting prosody for each dialog marker when using only lexical information from i) 3.2 sec context, ii) 5 sec context, and iii) 10 sec context

To verify the final hypothesis of this section, linear regression trained on 3.2 sec and

5 sec, and 10 sec of lexical context was used to predict the mean prosodic features. It is seen (Figure 5.2) that the more lexical information is used, the overall average prediction is improved.

Incidentally, among the dialog markers (Figure 5.3), *um* had the worst prediction results across the three conditions. This could imply that the *um*'s prosody is probably more influenced by the context prosody than by the lexical context information, which is understandable since it is primarily used as a backchannel or filler without any semantic meaning.

To conclude, the addition of text modality, that is to say, lexical information, to prosody, does improve prediction performance. Also, the better predictive ability of prosodic information over its lexical counterpart was established.

Chapter 6

Improving Prediction, Part 2: Different Learning Algorithms

This chapter describes the performance of a few other machine learning algorithms that were used with the aim of improving predictive performance over the linear regression model.

6.1 Auto Best-Fit Model

Here, I investigate the value of an auto-optimized regression model. More specifically, the *fitlinear* function in MATLAB, with its *OptimizeHyperparameters* parameter set as *auto* was used to build the predictive model. This function automatically chooses an optimized set of hyper-parameters, selecting either the least-square sum or the SVM as the best regression-fitting approach that minimizes the loss at the end of each training iteration. This model was trained using the same set of context features described in Table 4.3.

Comparing the performance of linear regression that uses only the least-square sum method to generate the best-fit line with this auto best-fit approach provided no benefit whatsoever in prediction quality. The error reduction rate was exactly the same in both cases (Table 4.5 and c.f. Table 6.1). Apparently, the algorithm chose the least-square sum method as the best-fit approach each time since this probably minimized the training error. This could mean that the underlying relationship between the predictor and the predicted data was mostly linear in nature.

Table 6.1: Percent reduction of prediction errors per dialog marker type using the auto-optimized linear regression model that chooses the better learner between the least square sum and SVM.

	predicting mean features					predicting max features				
	le	cf	p	hr	Avg.	le	cf	p	hr	Avg.
huh	29	43	26	34	33	22	46	27	15	28
now	19	21	30	25	24	23	21	44	39	32
oh	17	27	15	24	21	35	37	25	37	34
okay	18	30	7	6	15	19	31	18	40	27
really	46	16	-1	15	19	37	39	35	48	40
right	37	2	7	25	18	31	14	19	39	26
uh	43	21	56	30	38	52	48	33	32	41
uh-huh	12	20	34	50	29	18	11	31	39	25
um	31	22	29	32	28	56	50	36	27	42
well	25	24	23	33	26	32	24	40	33	32
yeah	38	25	20	44	32	49	38	40	54	45
yes	30	19	37	37	31	28	22	26	45	30
Average	29	22	24	30	26	33	32	31	37	33

6.2 Random Forest Model

Continuing the search for a better machine learning algorithm with the aim of improving the predictive performance, the random forests algorithm was tried.

The random forests algorithm employs a type of ensemble machine learning algorithm called Bootstrap Aggregation or bagging. The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean of prediction of the individual trees. It is often known to produce better results with large datasets than linear regression and is faster than non-linear SVM. Based on this, it was hypothesized that the random forests model would outperform the other models in terms of its predictive performance.

A random forest model trained on the same feature set as in Table 4.3, which was run for 30 iterations for each dialog marker type, showed an overall better performance in predicting the mean features. MATLAB's *fitrensemble* function was used that returned an ensemble regression model of boosting a default of 100 decision trees using either the *LSBoost* (least square boosting) or *Bag* (bagging) approach and other hyper-parameters— *NumLearningCycles*, *LearnRate* and *MaxNumSplits*— that were automatically optimized

Table 6.2: Summary of results for predicting mean features of a dialog marker token using random forests algorithm.

	predicting mean features			
	le	cf	p	hr
Baseline Average RMSE	0.61	0.82	0.67	0.66
Model Average RMSE	0.42	0.64	0.47	0.47
Reduction, %	32.0	22.8	29.2	28.8

using *Bayesian Optimization*. The best value that minimized the 5-fold cross-validation loss for the ensemble was returned as the final result after 30 iterations. In general, if there is any missing data in the predictors, i.e., the value of the optimal split predictor for an observation is missing, this algorithm uses *surrogate splits* that sends the observation to the left or right child node using the best surrogate predictor. This is done to maintain the quality of predictions. However, since I already ensured that no NaNs or missing values in the computed features would be used for training, there was hardly any chance of such a predicament.

Table 6.3: Percent reduction in root mean squared error from the baseline for predicting mean feature values using the random forests model.

	predicting mean features				
	le	cf	p	hr	Avg.
huh	42	28	16	23	27
now	30	27	34	22	29
oh	20	26	28	19	23
okay	27	11	28	7	18
really	39	24	23	33	30
right	20	30	30	23	26
uh	27	34	17	31	27
uh-huh	25	29	36	39	32
um	18	-16	39	30	17
well	39	33	38	37	37
yeah	39	14	20	23	24
yes	44	4	37	44	32
Average	32	23	29	29	28

6.2.1 Results using 3.2 sec of context

As seen while comparing Table 4.4 with Table 6.2, this random forests model reduced the prediction error from the baseline for most of all the mean features predicted, namely, log energy, cepstral flux, and pitch. The highest reduction was 29.2% for predicting mean pitch, which is more than what was achieved by the linear regression model (Table 4.4).

From the reduction in prediction error rate, Table 6.3, it is evident that there is substantial improvement in predicting mean token features for each dialog marker type by this model, with the overall average prediction error reduction of 28% or the best of 32% for mean log energy, over linear regression (c.f. Table 4.5).

6.2.2 Results using 10 sec of context

Table 6.4: Summary of results for predicting mean features of a dialog marker token using random forests algorithm trained with 10 sec context feature set including CPPS as a predictor (Table 5.17).

	predicting mean features				
	le	cf	p	hr	CPPS
Baseline Average RMSE	0.61	0.82	0.67	0.66	0.60
Model Average RMSE	0.38	0.57	0.41	0.42	0.31
Reduction, %	37.9	30.3	38.9	36.4	47.6

Inspired by the success of the linear regression model's prediction performance when trained with 10 sec of context that also had CPPS as one of the predictors, the random forests model —with other hyper-parameters remaining the same as the previous model— was now trained with the feature set in Table 5.17. Features predicted also included mean CPPS along with mean log energy, cepstral flux, pitch, and harmonic ratio.

On average, predictions were improved for each of the dialog marker types as well as for each predicted feature (Table 6.4 and Table 6.5) when compared to the model using only 3.2 sec of context without the CPPS predictor (c.f Table 6.2). Interestingly, predictions for mean cepstral flux for the dialog marker *um*, which was the hardest to predict in the previous model (c.f. Table 6.3), was quite improved to achieve 6% error reduction (Table 6.5). The

overall average prediction error reduction for this model was 38% which is also better, albeit marginally, than what linear regression achieved with the same feature set (c.f. Table 5.18).

Table 6.5: Percent reduction in root mean squared error from the baseline for predicting mean feature values using random forests algorithm with the feature set in Table 5.17.

	predicting mean features					
	le	cf	p	hr	CPPS	Avg.
huh	44	36	34	36	40	38
now	35	38	41	38	38	38
oh	27	28	29	28	36	30
okay	30	13	33	10	32	23
really	41	34	43	41	41	40
right	30	29	33	29	44	33
uh	39	36	35	41	37	38
uh-huh	36	35	42	39	48	40
um	24	6	45	35	44	31
well	45	41	50	43	53	46
yeah	43	24	31	38	45	36
yes	49	26	47	47	49	44
Average	38	30	39	36	48	38

6.3 K Nearest Neighbors

With the aim of even further improving the prediction results, the KNN approach was used.

Table 6.6: Prediction results summary for the kNN (k=3) model using 3.2 sec context feature set without the CPPS predictor.

	predicting mean features			
	le	cf	p	hr
Baseline Average RMSE	0.61	0.82	0.67	0.66
Model Average RMSE	0.54	0.72	0.51	0.58
Reduction, %	11.9	12.2	24.3	11.9

6.3.1 Results using 3.2 sec of context

kNN is a non-parametric method that approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood. A kNN model, using MATLAB's *fitcknn* function with the number of neighbors, k=3, and the default Euclidean distance as the metric to decide the closest neighbors, was trained with the original set (Table 4.3) of prosody features computed over 3.2 sec of context.

Contrary to expectations—given the success of this approach in many supervised predictive modeling tasks— the results were quite disappointing, as seen in Table 6.6, since only for predicting mean pitch were the results close to the linear regression model (c.f Table 4.4) while the rest had much lower prediction error reduction rate.

If we look into more detailed prediction results for each dialog marker type in Table 6.7, *yes* had the overall highest reduction in average prediction error while *well* had the worst predictions.

Table 6.7: Percent reduction of root mean squared error per dialog marker type for the kNN model using 3.2 sec context without the CPPS predictor.

	predicting mean features				
	le	cf	p	hr	Avg.
huh	1.4	1.1	1.2	1.6	1.3
now	11	17	47	14	22
oh	11	16	38	11	19
okay	18	15	25	16	18
really	30	32	52	12	31
right	8.9	33	48	11	25
uh	1.8	1.9	2.0	1.4	1.8
uh-huh	1.9	0.0	1.9	1.4	1.3
um	4.5	2.2	26	8.2	10
well	1.4	0.0	0.0	1.2	0.7
yeah	0.0	1.4	1.7	1.5	1.2
yes	49	33	48	59	47
Average	12	12	24	12	15

Following are some of my inferences:

- The prediction results of the kNN model indicated some sort of *clustering* in the dia-

log markers' prosodic mapping to their context: the context categories into which the dialog markers can be grouped together are not only distinguishable by their corresponding pragmatic functions but also the associated prosody.

- Dialog markers *uh* and *well* —usually used as fillers—, and *uh-huh* and *huh* — usually used as back-channels— had an overall lower reduction in average prediction error, implying that they were particularly hard to predict.

- Dialog marker *yes* — usually a response — and *really*— usually an adjective but sometimes also an interjection— were easier to predict. One common factor was that the speakers usually continued to hold the floor for some time in the local future after the target token, and thus, likely provided sufficient context prosodic information to the model to enable easy prediction of the marker's prosody.

- The kNN model predicted the prosody of *okay* well above the overall average, unlike the linear regression model (c.f. Table 4.5) which predicted its prosody with a much lower prediction error reduction than the other dialog markers. This result aligns with my exploratory study results (c.f. Chapter 3) that *okay*'s prosody could be significantly distinguished in different pragmatic contexts. For example, *okay* said in an *Acknowledgement* context category is prosodically different from that in an *End of Conversation* context. kNN, essentially used as a clustering algorithm, is probably able to easily predict the prosody of *okay* because of their inherent clustered prosodic behavior.

6.3.2 Results using 10 sec of context

Looking to further improve the prediction performance, kNN (k=5) was now trained on the optimal predictor set (Table 5.17) that has contributed to the best prediction results with each learning algorithm used so far.

Prediction was improved for all features except mean pitch (Table 6.8) which remained

Table 6.8: Summary of results for predicting mean features using kNN(k=5) trained on 10 sec context feature set with CPPS as an additional predictor (Table 5.17).

	predicting mean features				
	le	cf	p	hr	CPPS
Baseline Average RMSE	0.61	0.82	0.67	0.66	0.60
Model Average RMSE	0.51	0.67	0.51	0.57	0.42
Reduction, %	16.6	18.2	24.1	14.0	30.3

Table 6.9: Percent reduction in root mean squared error from the baseline for predicting mean feature values using kNN(k=5) algorithm trained on 10 sec context feature set, also including CPPS as a predictor (Table 5.17).

	predicting mean features					
	le	cf	p	hr	CPPS	Avg.
huh	12	19	13	1	13	11
now	16	24	36	16	35	25
oh	12	21	24	24	22	21
okay	19	17	27	18	30	22
really	33	33	37	14	35	30
right	14	35	39	12	37	27
uh	16	2	18	4	12	10
uh-huh	12	15	14	18	16	15
um	6	4	31	9	27	15
well	13	12	10	1	16	10
yeah	14	3	17	15	18	13
yes	29	35	30	37	35	33
Average	17	18	24	14	30	21

the same as before (c.f. Table 6.6). The overall average prediction error was reduced to 21% (Table 6.9) and hence, improved upon the previous model by 6%. However, the predictions were still not of the level achieved by random forests or even linear regression algorithms with the same feature set. Moderate prediction performance of kNN (even with an increased number of neighbors and increased context information) implies that in this unstructured corpus of dialogs, the diverse nature of dialog marker prosody, though dependent on context, could not be commonly categorized in general. However, it is to be noted that the predictions for *okay* were better than the overall average even this time (Table 6.9), further supporting my inference regarding the clustered nature of *okay*'s prosody as also noted in the previous

section.

6.4 Artificial Neural Networks

The success of advanced neural methods in synthesizing context-aware expressive speech has seen quite a success in recent times (c.f. Section 2.1). I decided to start with a simple version of an artificial neural network, intending to further improve the prediction quality.

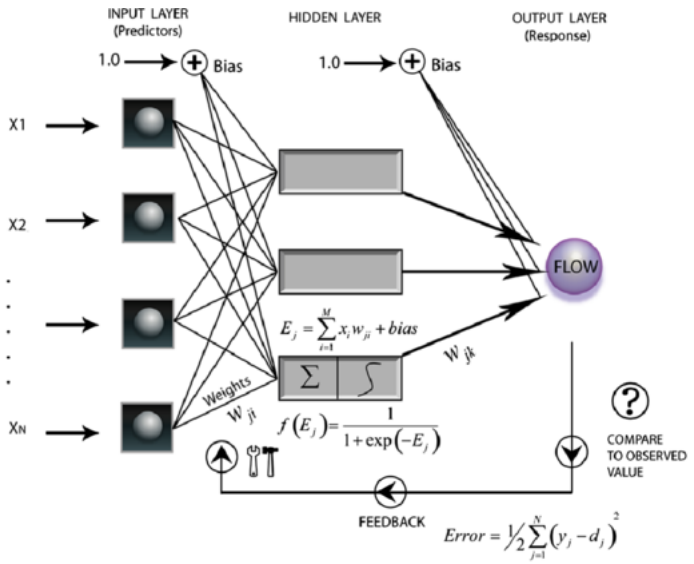

Figure 6.1: Conceptual Diagram of feedforward Artificial Neural Network (Demirel et al., 2009)

An artificial neural network for regression task was created using MATLAB's *fitrnet* method (conceptual diagram in Figure 6.1). This method trains a feedforward, fully connected neural network for regression. The first fully connected layer of the neural network has a connection from the input (predictor data), and each subsequent layer has a connection from the previous layer. Each fully connected layer multiplies the input by a weight matrix and then adds a bias vector. An activation function follows each fully connected layer, excluding the last. The final fully connected layer produces the network's output, namely predicted response values. This network was trained with the argument *OptimizeHyperparameters* set to "auto", and for reproducibility, the *AcquisitionFunctionName* argument set

to "expected-improvement-plus" in a *HyperparameterOptimizationOptions* structure. These arguments cause the *fitrnet* to search for optimized values for the following set of hyperparameters, that would produce a model with the lowest 5-fold cross-validation error at the end of 100 epochs, using *Bayesian optimization*:

- Activation functions for fully connected layers over: {'relu', 'tanh', 'sigmoid', 'none'}.
- Regularization Term Strength over continuous values in the range $([1e-5, 1e5]/NumObservations)$, where the value is chosen uniformly in the log-transformed range. The objective function for minimization is composed of the mean squared error (MSE) loss function and the ridge (L2) penalty term.
- LayerBiasesInitializer—type of biases for initial fully connected layer— over the two values {'zeros', 'ones'}.
- LayerWeightsInitializer—function to initialize fully connected layer weights— over the two methods {'glorot' (Glorot and Bengio, 2010), 'he'(He et al., 2015)}.
- LayerSizes: *fitrnet* optimizes over 1, 2, and 3 fully connected layers, excluding the final one. Also, *fitrnet* optimizes each fully connected layer separately over 1 through 300 neurons in the layer, sampled on a logarithmic scale.

Table 6.10: Summary of results for predicting mean features of a dialog marker token using an artificial neural network model trained on 10 sec context feature set with CPPS as an additional predictor.

	predicting mean features				
	le	cf	p	hr	CPPS
Baseline Average RMSE	0.85	41.0	45.7	0.11	1.19
Model Average RMSE	0.52	25.0	26.0	0.07	0.60
Reduction, %	39.3	39.0	43.1	36.4	49.7

At each epoch, the objective function minimized was *log(1 + cross-validation loss))*. Also, a cross-validation check with early stopping was used to ensure no overfitting in the training

data. Also note that by default, *fitrnet* uses the limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) (Nocedal and Wright, 2006) as its loss function minimization technique. This is a non-parametric algorithm that is not dependent on the learning rate and is suited to finding a global minimum.

Table 6.11: Percent reduction in root mean squared error from the baseline for predicting mean feature values using feed-forward neural network algorithm trained on 10 sec context feature set, also including CPPS as a predictor (Table 5.17).

	predicting mean features					
	le	cf	p	hr	CPPS	Avg.
huh	25	31	35	31	36	32
now	59	56	53	54	57	56
oh	42	33	39	36	51	40
okay	33	34	44	33	35	36
really	25	31	35	31	34	31
right	26	34	35	31	37	33
uh	56	55	58	51	57	56
uh-huh	44	31	39	21	37	35
um	48	45	57	33	45	46
well	44	46	54	48	50	48
yeah	36	36	38	37	40	37
yes	29	35	35	32	36	34
Average	39	39	43	36	50	42

A separate ANN was trained for each dialog marker and each predicted feature with the optimal predictor set (Table 5.17). However, please note that unlike with other models, the features were not z-normalized but used as is (since the ANN produced unintelligible results with the z-normalized values).

Table 6.11 shows that the overall average prediction error was reduced to 42%, thus, outperforming other models reported so far. Also, the predictions were the best for each predicted feature (Table 6.10). This model also produced the best predictions for seven out of 12 dialog markers' prosody (Table 6.11, c.f. Tables 6.5 and 6.9).

In conclusion, not only did the local context prove to be useful in predicting appropriate dialog marker prosody, but also we can improve the quality of these predictions when learning using more suitable algorithms that are trained with the best representatives of the context.

Chapter 7

Predictability for Task-Oriented Dialogs

This chapter tests the applicability of the main hypothesis of this research—the prosody of a dialog marker token can be predicted appropriately from its local dialog context prosody—for task-oriented dialogs, expanding my previous work on open-domain conversations.

7.1 Data Set

The dataset —the Harper Valley Bank Corpus (Wu et al., 2020)— consists of recorded audio dialogs that are primarily simulated task-oriented conversations between a bank's call center agent and a customer, collected using the Gridspace Mixer platform. Conversations are goal-oriented, for example, a customer ordering a new checkbook or checking the balance of an account, etc.

A sample conversation from this corpus is given below:

Agent: hello this is harper valley national bank my name is jay how can I help you today

Caller: hi my name is mary davis

Caller: [noise]

Caller: i would like to schedule an appointment

Agent: yeah sure what day what time

Caller: thursday one thirty pm

Agent: that's done anything else

Caller: that's it

Agent: have a good one.

This dataset—developed to support education and experimentation across a wide range of conversational speech and language machine-learning tasks—contains about 23 hours of audio from 1,446 human-human conversations between 59 unique speakers, encoded as 8kHz per the original telephony data. Each conversation is labeled with human transcripts, timing information, emotion and dialog act model outputs, subjective audio quality, task descriptions, and speaker identity. Table 7.1 shows the total count of tokens for each dialog marker type from this corpus that was used for the experiments in this chapter. These counts take into account only those dialog markers present in the *human transcript* field and not those in the machine-generated *transcript* field of a conversation which, though greater in number, were mostly inaccurate as was revealed by listening to the first 10 of them. The dialog marker, *uh-huh* had zero count in this corpus when the human-generated transcripts were considered.

Table 7.1: Number of instances of each dialog marker in the Harper Valley Bank Corpus

	token count
huh	38
now	77
oh	120
okay	1565
really	9
right	30
uh	672
uh-huh	0
um	464
well	126
yeah	96
yes	253
Total	3450

Table 7.2: Summary of results for predicting mean features of a dialog marker token in a task-oriented dialog corpus using 10-fold cross-validation training.

	predicting mean features			
	le	cf	p	hr
Baseline Average RMSE	0.86	0.56	0.78	0.49
Model Average RMSE	0.83	0.54	0.76	0.47
Reduction, %	3.51	2.92	2.44	3.23

7.2 Predictive Model and its Performance

Local context (3.2 sec each, from the past and future context of the target token and from both speakers) is represented by the 72-dimensional prosody vector (Table 4.3). Unlike in Chapter 4, the pitch tracker used here is the *REAPER*—Robust Epoch And Pitch EstimatoR (Talkin)—developed by David Talkin at Google. This is much faster than MATLAB's *fxrapt* and is fairly robust to recording quality. The main reason for using this pitch tracker is that *fxrapt* failed to compute pitch for most of this corpus's audios. Since there is only a limited number of instances available for each dialog marker type in the corpus, it was necessary to ensure that each one of them has some computed prosody value that could be used for k-fold cross-training.

Table 7.3: Percent reduction of prediction errors per dialog marker type for predicting dialog marker prosody in the task-oriented dialogs from the Harper Valley Bank Corpus.

	predicting mean features				
	le	cf	p	hr	Avg.
huh	4.1	2.6	2.1	0.3	2.3
now	2.4	2.6	4.3	0.9	2.6
oh	1.5	0.6	0.5	2.1	1.2
okay	3.6	4.3	0.5	-0.1	2.1
right	9.8	4.6	5.1	12.0	7.9
uh	1.4	1.1	0.9	0.8	1.1
um	1.6	1.4	0.0	1.7	1.2
well	2.6	4.7	2.1	1.3	2.7
yeah	1.2	3.0	6.8	10	5.3
yes	4.9	5.2	0.5	0.6	2.8
Average	3.5	2.9	2.4	3.2	2.9

An intra-corpus evaluation approach was followed but, unlike in Chapter 4, a k-fold, k=10 cross-training method was adopted to mitigate the effect of fewer training samples. The dialog marker *really* had too few instances (9) to be of use to any supervised machine learning algorithm and understandably, had large prediction errors, so these results are omitted from the tables below. Each predicted mean feature: log energy, cepstral flux, pitch, and harmonic ratio, had a positive error reduction rate, Table 7.2, indicating that the proposed prediction model also works for task-oriented dialogs, although the prediction results were much worse than what was achieved for the open-domain dialogs (c.f. Table 4.4). On average, *right*, and *yeah* were predicted well above the overall prediction average. *uh*, *oh* and *um* had predictions quite below the overall average.

One possible reason for such a modest prosody prediction could easily be the lack of a sufficient number of training samples for most dialog markers. This is not only because of fewer audio hours available in the corpus but also because task-oriented conversations do not feature as much variation in contextual uses of dialog markers as open-domain dialogs do (which aligns with my intuition in Section 4.1). This is probably because they are more structured and are spoken in a more formal environment.

To conclude, the original research hypothesis—context prosody is useful for predicting the target dialog marker's prosody— also holds true, albeit marginally, for task-oriented dialogs. This also goes on to show the significance of my research problem which has been shown to be solvable to a certain extent. A more general solution, though much harder to achieve, is worth pursuing in future research.

Chapter 8

Predictability of Autistic Prosody

Inspired by the cross-domain performance of the proposed prosody prediction approach in task-oriented dialogs, this chapter investigates whether this predictability can be further extended for atypical prosody in autistic children dialogs. This task is also motivated by the possible benefits that automatic comparisons of autistic with neurotypical prosody based on typical context-prosody mappings may provide in the automatic and early detection of atypical prosody (reviewed in Section 2.2).

8.1 Data Set

The experiments in this chapter used the NMSU children corpus (Lehnert-LeHouillier et al., 2020). Specifically, I used two groups of matched dialogs from the corpus: one recorded with neurotypical children and the other recorded with autistic children. Each group had 14 dialogs: each ranging from 4 to 10 minutes. An adult counselor or interviewer played a *spot the difference* game with a child in each dialog. The child was asked to help identify these differences by describing the picture given to them in detail without seeing its counterpart. The audios were post-processed to separate each speaker to a separate track. Each child's dialog segment was annotated to mark the start and end timestamps for each token of the dialog markers. The number of instances for each dialog marker in this corpus is listed in Table 8.1. It is to be noted that the neurotypical children sometimes also used *okay* and *umm* in the corpus, which were completely absent in the autistic children's dialogs and hence, not considered.

Table 8.1: Number of instances of each dialog marker in NMSU's children corpus.
CASD: Children with Autistic Spectrum Disorder and CNT: Children NeuroTypical

Dialog Marker	CASD	CNT
oh	19	16
uh	24	8
um	37	55
yeah	72	89
yes	20	52
Total	172	220

8.2 Autistic Prosody vs. NeuroTypical Prosody

Given the atypical nature of autistic children's prosody (Section 2.2), it seems likely that:

- predicting the dialog markers' prosody for autistic children would be much harder than for neuro-typical children, that is,
- autistic prosody is much less influenced by local context prosody.

In more formal terms, the following are the hypotheses tested in this section:

1. The prosody of local context in dialogs between typical adults is sufficiently informative to predict the prosody of dialog markers occurring in children's dialogs, and
2. This contextual information is less predictive in dialogs involving autistic children than those with neuro-typical ones.
3. Prosody of autistic children is harder to predict than their neuro-typical counterparts, even when trained on intra-corpus data.

Linear regression, trained with typical adult prosodic data from the Switchboard corpus (Chapter 4), is used to predict the mean for the prosodic features: log energy, cepstral flux, pitch, and harmonic ratio, for each of the above dialog marker type found in both the autistic

and neuro-typical children data, given their corresponding local context information. The set of context prosodic features used was the same as described in Chapter 4.

Contextual prosody prediction using the model derived from the adult dialogs in Switchboard corpus was proved to be not very useful in predicting children's behavior since almost all dialog marker prosody predictions were much worse than baseline for each response feature (Table 8.2 and Figure 8.1). Thus, the first hypothesis was not supported.

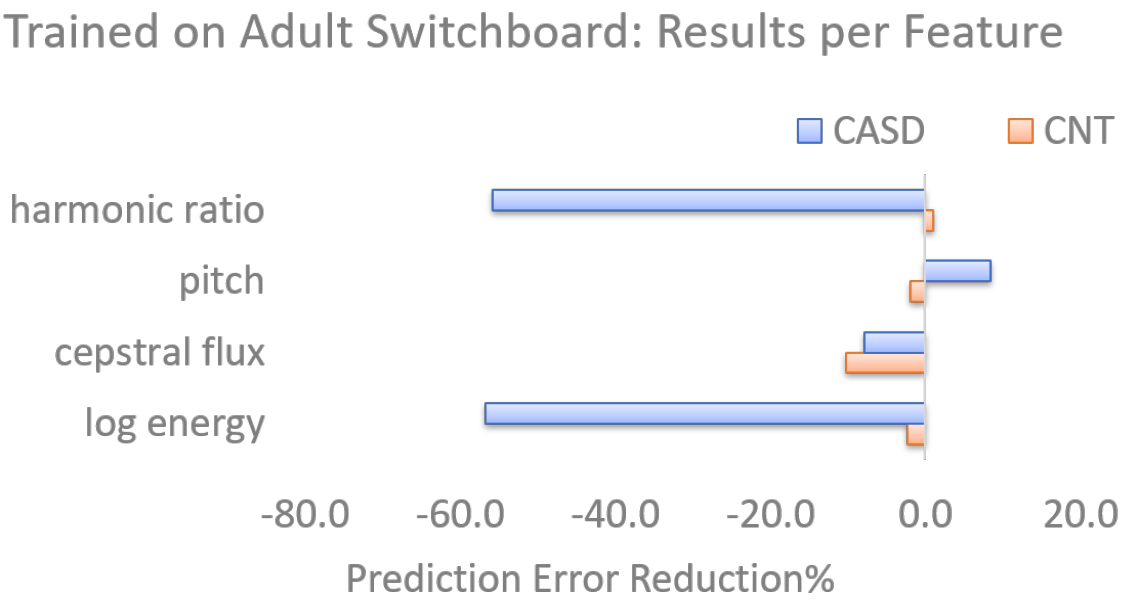

Figure 8.1: Prediction results for each predicted feature: Comparing average prediction error rate for predicting dialog marker prosody in autistic (CASD) vs. neurotypical (CNT) children for each predicted feature when trained on Switchboard adult dialog data.

For each dialog marker, the average prediction error rate of all the response features is seen to be much worse for the autistic data than for the neuro-typical group (Table: 8.3 and Figure 8.2). When the prediction error rate is considered with respect to each predicted feature (Table 8.2), mean log energy and mean harmonic ratio was found especially hard to predict for autistic children while mean cepstral flux and mean pitch predictions were somewhat better than for the neurotypical children. Additionally, the overall average prediction error rate for autistic children (-29%) is much worse than that for neurotypical children (-3%).

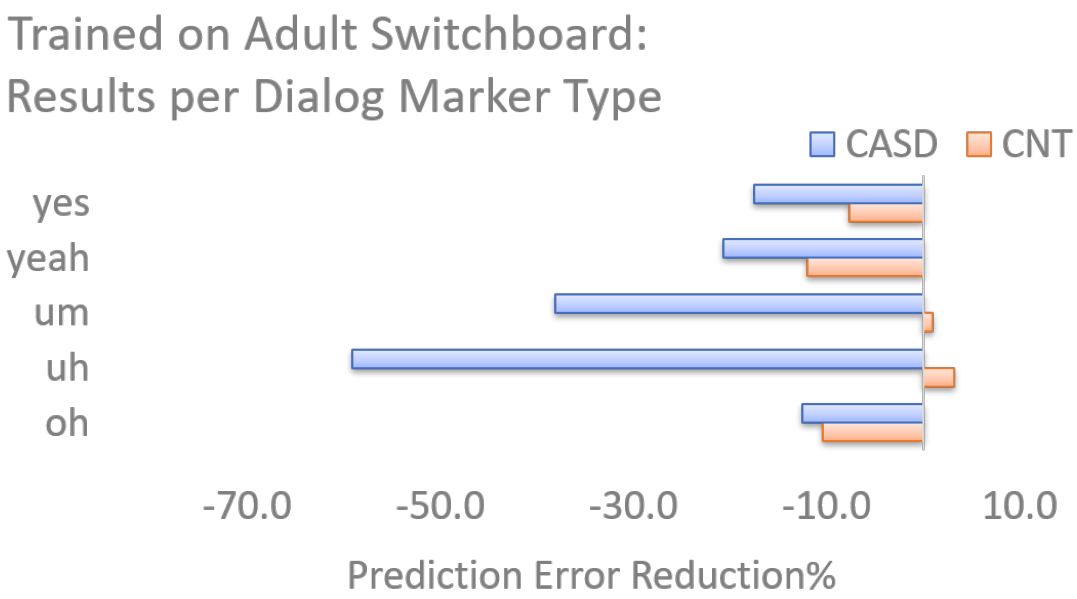

Figure 8.2: Prediction performance comparison per dialog marker type: Comparing average prediction error rate for predicting dialog marker prosody in autistic (CASD) vs. neurotypical (CNT) children for each dialog marker type when trained on Switchboard adult dialog data.

Hence, it can be concluded that the experimental evidence weakly supports that autistic children's prosody is less like that of adults than those of neuro-typical children, partially supporting the second hypothesis. Interestingly, the predicted average log energy and pitch in autistic dialogs varied the most from their neurotypical counterparts. This aligns with prior research (Section 2.2) that observed prominent and consistent atypical pitch and intensity patterns in the autism spectrum.

Table 8.2: Comparing average prediction error reduction rate for predicting mean of each of the features for autistic (CASD) vs. neurotypical (CNT) children using linear regression trained on Switchboard adult dialog data.

predicting mean features	CASD	CNT
log energy	-56.8	-2.3
cepstral flux	-7.8	-10.2
pitch	8.4	-1.9
harmonic ratio	-55.7	1.1

Table 8.3: Comparing average prediction error reduction rate for predicting dialog marker prosody in autistic (CASD) vs. neurotypical (CNT) children, for each dialog marker type using linear regression trained on adult Switchboard adult dialog data.

	CASD	CNT
oh	-12.5	-10.4
uh	-59.1	3.3
um	-38.0	1.0
yeah	-20.7	-11.9
yes	-17.4	-7.6

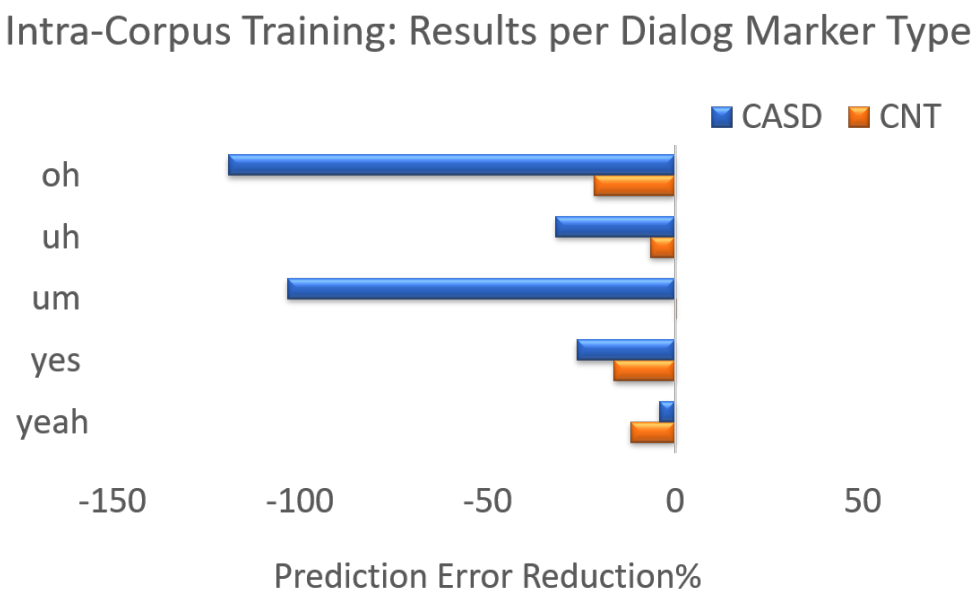

Figure 8.3: Prediction results for intra-corpus training: Comparing average prediction error rate for predicting dialog marker prosody in autistic (CASD) vs. neurotypical (CNT) children for each dialog marker type, when trained on intra-corpus children data via 5-fold cross-validation approach.

In another experiment, the local prosodic context information of a dialog marker token was derived from the corresponding children's dialogs instead of being trained on adult dialog corpus. Specifically, a linear regression model that predicted the prosody for each dialog marker of autistic children was trained on the corresponding local context via a 5-fold

cross-validation approach. A similar approach was followed for predicting the neuro-typical

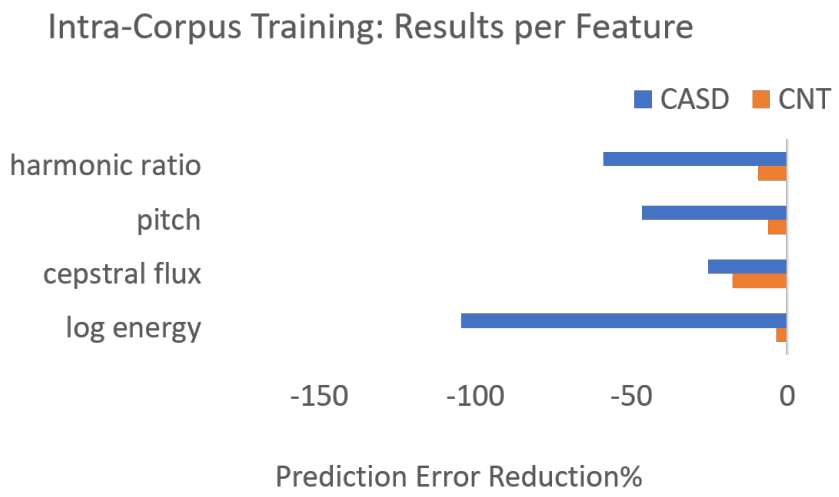

Figure 8.4: Prediction results with intra-corpus training: Comparing average prediction error reduction rate for predicting dialog marker prosody in autistic (CASD) vs. neurotypical (CNT) children for each predicted feature, when trained on intra-corpus children data via 5-fold cross-validation approach.

children's prosody. However, the prediction quality (Figure 8.3 and Figure 8.4) was bad for both the sub-corpora. This was not very surprising given the small amount of training data available in the children's corpus. Nevertheless, even in this scenario of intra-corpus training, the dialog prosody of autistic children was predicted even worse than those of their neuro-typical counterparts on an average for each dialog marker type (Figure 8.3) and for each predicted feature (Figure 8.4), indirectly supporting the third hypothesis.

To conclude, though the predictability of the children's prosody from the adult dialogs' context could not be demonstrated, it was confirmed that appropriate context-prosody modeling was even harder to achieve for autistic children's dialogs than for their neurotypical peers. Hence, typical contextual dependency of prosody should be considered an essential factor for further research on the automatic comparison between atypical and neurotypical prosodic behavior.

Chapter 9

Discussion and Future Research

The following sections summarize the significant findings of this research and provide suggestions on how to improve the prediction results, leading to future research.

9.1 Summary of Findings

The experiments that support my claim—that dialog markers' prosody can indeed be predicted directly from the prosody of the context to a fair extent, even with very limited feature sets and very simple models— are enumerated below.

- Linear regression with only a minimal set of context features (Table 4.3) reduced the error by an overall average of 26% (c.f. Table 4.5) for predicting each dialog marker's mean feature values and by 33% for predicting their maximum.
- A simple feedforward artificial neural network model—that used 10 sec of context features that included CPPS as a predictor (Table 5.17)—has the best prediction performance with an overall average prediction error of 42% (c.f. Table 6.10).
- An appropriate and optimal set of engineered predictor prosody features (Table 5.17) can ensure much improved prediction even with the simplest of machine learning algorithms such as linear regression (c.f. Table 5.18), random forests (c.f. Table 6.2) and knn (c.f. Table 6.6) than what was achieved with the original set of predictors (c.f. Tables 4.4, 5.12 and 5.15)

Following are the conclusions I have drawn from the other experiments.

- Linear regression when trained on context segmented into more windows (Table 5.6) predicted less accurate prosody (c.f. Table 5.7) than that trained on uniformly windowed context (c.f. Table 4.4).
- A generic model, trained globally on the context prosody of all dialog marker types, predicted worse (c.f Table 5.4) than those trained on the type-specific data (c.f. Table 4.4).
- Past context alone proved less informative than when used with future information for prediction (c.f. Table 5.1).
- Both speaker context is more informative than either individual speaker's context (c.f. Table 5.2).
- Use of wider context (10 sec) improves the prediction performance (Section 5.3.2).
- Adding lexical context information improves the prediction performance (Section 5.5).
- Proposed prosody prediction approach has some value also in task-oriented dialog domain (Chapter 7).
- Prosody is harder to predict for autistic children than for the neurotypical population belonging to the same age group (c.f. Tables 8.2 and 8.3).

It is also to be noted that the proposed predictive approach was able to overcome the key limitations of prior research (c.f. Sections 2.1.3 and 2.3.3) since it:

- primarily learns from prosodic context
- predicts appropriate prosody for more than a few dialog markers, twelve to be exact.
- uses both past and future local dialog context (instead of only the immediate past) and context from both the speakers.
- is tested across multiple dialog domains.

Additionally, this approach neither does rely on human annotations for context or intents nor on hand-crafted rules but follows an automated prediction technique using machine learning algorithms.

9.2 Implications

The performance of the proposed prosody prediction approach implies that, indeed, local prosodic context is directly informative for such prediction.

In the near future, this predictive modeling may be extended to generate appropriate prosodic adjustments in the responses of dialog systems to create, for example:

- highly responsive spoken language chatbots or
- more natural sounding voice assistants.

Further, this proposed approach of context-appropriate prosody prediction could be used to build intervention bots that would provide feedback to people exhibiting atypical prosody. For example, a) language learners often fail to grasp the technique of changing the prosody in their spoken words appropriately to the dialog context, or b) autistic people, who are usually less context-sensitive than their neuro-typical counterparts, often exhibit awkward prosody while communicating. They could practice their communication skills with these proposed feedback bots. Such bots could detect atypical prosody—inappropriate to the dialog context—in their user's responses and intervene as it seems fit. One possible approach could be to pause the interaction when such atypicality is detected, provide feedback to the user on what went wrong, and then repeat a revision of the user's target utterance modified to have appropriate prosody in terms of pitch, volume, and duration. This way, the users could be trained to master the typical responsive dialog patterns that would, eventually, help them improve their communication skills and be more socially effective.

However, to develop such responsive systems, further exploitation of context-based prosody prediction is required. This may involve generating tokens of utterances that exhibit fully

appropriate prosody. For example, prosodic style or pattern templates that include the prosodic configurations for pitch, volume, duration, etc., for any utterance appropriate to a dialog context need to be formed so that the prosody of synthesized speech can be modified as needed. Very recently, Fernandez et al. (2022) developed a high-fidelity scheme of translating a dedicated single-speaker corpus conversation style to a multi-speaker setting with no quality degradation using a prosody preserving voice-conversion-based data augmentation technique. Such a scheme could be followed to develop the context-appropriate prosodic styles that also need to include the context embeddings as part of the speech representation in their proposed *S2S TTS* architecture.

The extent to which the prosody adjustments recommended by a context-sensitive model have actual value in dialog still remains to be seen. Previous research suggests that improved responsiveness can increase perceived naturalness and responsiveness, and ultimately rapport, engagement, and user satisfaction (Acosta and Ward, 2011; Lubold and Pon-Barry, 2014; Li et al., 2019; Gálvez et al., 2020; Choi and Agichtein, 2020; Sadoughi et al., 2017). However, experiments with human subjects are needed to establish whether such manipulations also have value for dialog markers and, more specifically, whether using such prosodically appropriate dialog markers can improve rapport and engagement with the users.

9.3 Future Research

Future work should attempt more detailed predictions such as predicting prosody :

- not just of a dialog marker token's averages but also of contour parameters or even frame-by-frame values, and
- of full utterances.

Prediction of appropriate prosody can, perhaps, further be improved by using:

- an exhaustive list of commonly occurring dialog markers that would include more than the twelve ones considered here. More importantly, dialog markers could be segregated

based on their general role/function in dialog, and samples belonging to each role should be separately modeled. Since each specific function of a dialog marker should ideally be characterized by its associated prosodic forms, such segregation could enable better prosody predictions. Alternatively, this segregation process could be automated by employing semi-supervised approaches such as the UMAP clustering method, successfully applied by Liesenfeld and Dingemanse (2022) directly on the speech signals to represent their structure and variation across 16 languages in a bottom-up study of the behavior of the response tokens. These approaches would automatically cluster the dialog markers based on their common prosodic structure.

- acoustic word embeddings, instead of engineered prosody features, to represent the local context, following the approach that has recently been used to improve prosody of synthesized speech in neural text-to-speech (Chen et al., 2021), and

- methods to combine these context-based predictions with other factors that might affect the prosody, such as the current dialog state and the communicative intent of the system (Ward and DeVault, 2016).

- more sophisticated deep learning models that could automatically learn the prosodic feature representations, specifically, self-supervised learning approaches. Recent advances in self-supervised learning in speech processing led to the development of a benchmark framework, *SUPERB* (Yang et al., 2021), released as an open-source that aimed for a simple and more generalized solution for any speech-related task. This framework consists of several pre-trained self-supervised models and a common toolkit successfully used for various downstream and upstream tasks (Mohamed et al., 2022). It also enables fine-tuning the models to suit any unknown task. The extracted fixed representations from the pre-trained models can be fed to any prediction head for a downstream task. Specifically, in the context of this research, these representations could be extracted for each utterance in the dialog corpus and then combined to represent the entire local prosodic context. Such self-supervised contextual representation

could hugely improve the quality of prosody predictions.

Future research could also involve experiments with languages other than English to test whether the same context-dependent approach could be successfully applied to predict appropriate prosody for other languages.

To conclude, although there is much room for improvement, this research is able to build the foundation of a novel context-dependent prosody prediction model that can be expected to be used to improve the natural prosodic behavior in future spoken dialog systems.

.

References

Jaime C. Acosta and Nigel G. Ward. Achieving rapport with turn–by–turn, user–responsive emotional coloring. *Speech Communication*, 53(9):1137–1148, November 2011. doi: 10.1016/j.specom.2010.11.006.

Hans Asperger and Uta Frith. *Autistic Psychopathy in Childhood*. Cambridge University Press, 1991. doi: https://doi.org/10.1017/CBO9780511526770.002.

Christiane Baltaxe, James Q. Simmons, and Evi van der Zee. Intonation patterns in normal, autistic and aphasic children. In *Proceedings of the 10th International Congress of Phonetic Sciences*, pages 713–718, 1984. doi: https://doi.org/10.1515/9783110884685-117.

Wayne Beach. Using prosodically marked "Okays" to display epistemic stances and incongruous actions. *Journal of Pragmatics*, 169:151–164, 2020. doi: https://doi.org/10.1016/j.pragma.2020.08.019.

Timothy Bickmore and Toni Giorgino. Health dialog systems for patients and consumers. *Journal of Biomedical Informatics*, 39(5):556–571, 2006. doi: 10.1016/j.jbi.2005.12.004.

L.A. Black, M.McTear, N. Black, R.Harper, and M.Lemon. Appraisal of a conversational artefact and its utility in remote patient monitoring. In *18th IEEE Symposium on Computer–Based Medical Systems*, pages 506–508, 2005. doi: 10.1109/CBMS.2005.33.

Mike Brookes. *Voicebox: Speech Processing Toolbox for MATLAB*. Department of Electrical Electronic Engineering, Imperial College, London, UK., 2019. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

Judee Burgoon, L A. Stern, and L Dillman. Interpersonal Adaptation: Dyadic Interaction Patterns. *Cambridge University Press*, 1995. doi: https://doi.org/10.1017/CBO9780511720314.

Donna K. Byron and Peter A. Heeman. Discourse Marker Use in Task-Oriented Spoken Dialog. In *In Proceedings of 5th European Conference on Speech Communication and Technology*, pages 2223–2226, 1997. doi: 10.21437/Eurospeech.1997-586.

Vera Cabarrão, Helena Moniz, Jaime Ferreira, Fernando Batista, Isabel Trancoso, Ana Mata, and Sérgio Curto. Prosodic Classification of Discourse Markers. In *International Congress of Phonetic Sciences*, page 2105, 2015. https://www.inesc-id.pt/publications/11438/pdf.

Sasha Calhoun, Jean Carletta, Jason M Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation*, 44(4):387–419, 2010. doi: 10.1007/s10579-010-9120-1.

Liping Chen, Yan Deng, Xi Wang, Frank K. Soong, and Lei He. Speech Bert Embedding for Improving Prosody in Neural TTS. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6563–6567, 2021. doi: https://doi.org/10.48550/arXiv.2106.04312.

Nathan A. Chi, Peter Washington, Aaron Kline, Arman Husic, Cathy Hou, Chloe He, Kaitlyn Dunlap, and Dennis Wall. Classifying Autism From Crowdsourced Semistructured Speech Recordings: Machine Learning Model Comparison Study. *Journal of Medical Internet Research Pediatrics and Parenting*, 5(2):e35406, 2022. doi: 10.2196/35406.

Jason Ingyu Choi and Eugene Agichtein. Quantifying the Effects of Prosody Modulation on User Engagement and Satisfaction in Conversational Systems. In *Proceedings of the Conference on Human Information Interaction and Retrieval*, pages 417–421, 2020. doi: 10.1145/3343413.3378009.

Anne Cutler, Delphine Dahan, and Wilma van Donselaar. Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40(2):141–201, 1997. URL https://doi.org/10.1177/002383099704000203.

SvenOlof Dahlgren, Annika D Sandberg, Sofia Strömbergsson, Lena Wenhov, Maria Råstam, and Ulrika Nettelbladt. Prosodic traits in speech produced by children with autism spectrum disorders – Perceptual and acoustic measurements. *Autism & Developmental Language Impairments*, 3(3):239694151876452, 2018. doi: https://doi.org/10.1177/2396941518764527.

Helen K. Delichatsios, Robert H. Friedman, Karen Glanz, Sharon Tennstedt, Charles Smigelski, Bernardine M. Pinto, Heather Kelley, and Matthew W. Gillman. Randomized Trial of a "Talking Computer" to Improve Adults' Eating Habits. *American Journal of Health Promotion*, 15(4):215–224, 2001. doi: https://doi.org/10.4278/0890-1171-15.4.215.

Mehmet Demirel, Anabela Venancio, and Ercan Kahya. Flow forecast by SWAT model and ANN in Pracana basin, Portugal. *Advances in Engineering Software*, 40(2):467–473, 2009. doi: 10.1016/j.advengsoft.2008.08.002.

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. Resegmentation of Switchboard. In *5th International Conference on Spoken Language Processing*, pages 1543–1546, 1998. doi: 10.21437/ICSLP.1998-588.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8–9):630–645, 2008. doi: 10.1016/j.specom.2008.04.002.

Ramesh Farzanfar, Sophie Frishkopf, Jeffrey Migneault, and Robert Friedman. Telephone-linked care for physical activity: A qualitative evaluation of the use patterns of an information technology program for patients. *Journal of Biomedical Informatics*, 38(3):220–228, 2005. doi: 10.1016/j.jbi.2004.11.011.

Raul Fernandez, David Haws, Guy Lorberbom, Slava Shechtman, and Alexander Sorin. Transplantation of Conversational Speaking Style with Interjections in Sequence-to-Sequence Speech Synthesis. In *23rd Conference of the International Speech Communication Association*, 2022. doi: 10.21437/Interspeech.2022-388.

Fernanda Ferreira. Prosody . *Encyclopedia of Cognitive Science*, 2006. doi: https://doi.org/10.1002/0470018860.s00258.

Carol Figueroa, Adaeze Adigwe, Magalie Ochs, and Gabriel Skantze. Annotation of Communicative Functions of Short Feedback Tokens in Switchboard. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 1849–1859, 2022. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.197.pdf.

Marisa G. Filipe, Sónia Frota, and Selene G. Vicente. Executive Functions and Prosodic Abilities in Children With High-Functioning Autism. *Frontier Psychology*, 9(Cognition), 2018. doi: https://doi.org/10.3389/fpsyg.2018.00359.

Katherine Forbes-Riley and Diane J. Litman. Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 201–208, 2004. https://aclanthology.org/N04-1026.

Katherine Forbes-Riley and Diane J. Litman. Adapting to Multiple Affective States in Spoken Dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, 2012. doi: 10.5555/2392800.2392839.

Mary Ellen Foster, Manuel Giuliani, and Amy Isard. Task-Based Evaluation of Context-Sensitive Referring Expressions in Human-Robot Dialogue. *Language and Cognitive Processes*, 29(8):1018–1034, 2013. doi: 10.1080/01690965.2013.855802.

Rubén Fraile and Juan Ignacio Godino-Llorente. Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, 14:42–54, 2014. doi: https://doi.org/10.1016/j.bspc.2014.07.001.

Bruce Fraser. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952, 1999. doi: 10.1016/S0378-2166(98)00101-5.

Bruce Fraser. An Account of Discourse Markers. *International Review of Pragmatics*, 1(2): 293–320, 2009. doi: 10.1163/187730909X12538045489818.

Valerie Freeman, Gina-Anne Levow, Richard Wright, and Mari Ostendorf. Investigating the role of 'yeah' in stance-dense conversation. In *16th Annual Conference of the International Speech Communication Association*, pages 3076–3080, 2015. doi: 10.21437/Interspeech.2015-625.

Riccardo Fusaroli, Anna Lambrechts, Dan Bang, Dermot M. Bowler, and Sebastian B. Gaigg. Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis. *Autism Research*, 10(3):384–407, 2017. doi: 10.1002/aur.1678.

Ramiro H Gálvez, Agustín Gravano, Štefan Beňuš, Rivka Levitan, Marian Trnka, and Julia Hirschberg. An empirical study of the effect of acoustic-prosodic entrainment on the perceived trustworthiness of conversational avatars. *Speech Communication*, 124:46–67, 2020. doi: https://doi.org/10.1016/j.specom.2020.07.007.

Petra Geutner, Frank Steffens, and Dietrich Manstetten. Design of the VICO spoken dialogue system: Evaluation of user expectations by Wizard-of-Oz experiments: A speech driven in-car assistance system. In *Proceedings of International Conference on Language Resources and Evaluation*, 2002. http://www.lrec-conf.org/proceedings/lrec2002/pdf/74.pdf.

K.G. Ghanem, H.E. Hutton, J.M. Zenilman, R. Zimba, and E.J. Erbelding. Audio computer assisted self interview and face to face interview modes in assessing response bias among STD clinic patients. *Sexually Transmitted Infections*, 81(5):421–425, 2005. doi: 10.1136/sti.2004.013193.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 517–520, 1992. doi: 10.5555/1895550.1895693.

Agustín Gravano, Stefan Benus, Héctor Chávez, Julia Hirschberg, and Lauren Wilcox. On the role of context and prosody in the interpretation of 'okay'. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 800–807, 2007a. https://aclanthology.org/P07-1101.

Agustín Gravano, Stefan Benus, Julia Hirschberg, Shira Mitchell, and Ilia Vovsha. Classification of discourse functions of affirmative words in spoken dialogue. In *8th Annual Conference of the International Speech Communication Association*, pages 1613–1616, 2007b. doi: 10.7916/D81C25BF.

Haohan Guo, Shaofei Zhang, Frank K. Soong, Lei He, and Lei Xie. Conversational End-to-End TTS for Voice Agent. In *2021 IEEE Spoken Language Technology Workshop*, pages 403–409, 2021. doi: 10.1109/SLT48900.2021.9383460.

Petra Hansson. Prosodic Correlates of Discourse markers in Dialogue. In *Proceedings of the ESCA workshop on dialogue and prosody*, pages 99–104, 1999. https://www.isca-speech.org/archive_open/archive_papers/dia_pros/diap_099.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *In Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

Yolanda D Heman-Ackah, Deirdre D Michael, and George S Goding Jr. The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice*, 16(1):20–27, 2002. doi: 10.1016/s0892-1997(02)00067-x.

Thomas Hempel. *Usability of speech dialogue systems: Listening to the target audience*. Springer Publishing Company, 2010. doi: 10.1007/978-3-540-78343-5.

James Hillenbrand, Ronald A. Cleveland, and Robert L. Erickson. Acoustic correlates of breathy vocal quality. *Journal of Speech and Hearing Research*, 37(4):769–778, 1994. doi: 10.1044/jshr.3902.311.

Julia Hirschberg and Diane Litman. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–530, 1993. https://aclanthology.org/J93-3003.

Nobukatsu Hojo, Yusuke Ijima, Hiroaki Sugiyama, Noboru Miyazaki, Takahito Kawanishi, and Kunio Kashino. DNN-based speech synthesis considering dialogue act information and its evaluation with respect to illocutionary act naturalness. In *10th International Conference on Speech Prosody*, pages 975–979, 2020. doi: 10.21437/SpeechProsody.2020-199.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems*, 38(3):1–32, 2020. doi: 10.1145/3383123.

Robert C. Hubal and Ruth S. Day. Informed consent procedures: An experimental test using a virtual character in a dialog systems training application. *Journal of Biomedical Informatics*, 39(5):532–540, 2006. doi: 10.1016/j.jbi.2005.12.006.

Christina A. Irvine, Inge-Marie Eigsti, and Deborah A. Fein. Uh, Um, and Autism: Filler Disfluencies as Pragmatic Markers in Adolescents with Optimal Outcomes from Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 46(3):1061–1070, 2016. doi: 10.1007/s10803-015-2651-y.

Srinivasan Janarthanam, Oliver Lemon, Xingkun Liu, Phil Bartie, William Mackaness, and Tiphaine Dalmas. A Multithreaded Conversational Interface for Pedestrian Navigation and Question Answering. In *Proceedings of the Special Interest Group on Discourse and Dialogue Conference*, pages 151–153, 2013. https://aclanthology.org/W13-4025.

Constantijn Kaland, Marc Swerts, and Emiel Krahmer. Accounting for the listener: Comparing the production of contrastive intonation in typically-developing speakers and speakers with autism. *The Journal of the Acoustical Society of America*, 134(3):2182–2196, 2013. doi: https://doi.org/10.1121/1.4816544.

Leo Kanner. Autistic disturbances of affective contact. *Nervous Child*, 2(3):217–250, 1943. https://psycnet.apa.org/record/1943-03624-001.

Masahito Kawamori, Takeshi Kawabata, and Akira Shimazu. Discourse Markers in Spontaneous Dialogue: A Corpus based study of Japanese and English. In *Discourse Relations and Discourse Markers*, pages 93–99, 1998. https://aclanthology.org/W98-0316.

Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora. *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, 2005. doi: 10.1002/0470093366.

Janine Kleinhans, Mireia Farrús, Agustín Gravano, Juan Pérez, Catherine Lai, and Leo Wanner. Using Prosody to Classify Discourse Relations. In *The 18th Annual Conference of the International Speech Communication Association*, pages 3201–3205, 2017. doi: 10.21437/Interspeech.2017-710.

Jan Krebber, Sebastian Möller, Rosa Pegam, Ute Jekosch, Miroslav Melichar, and Martin Rajman. Wizard-of-Oz tests for a dialog system in smart homes. In *Proceedings of the joint congress CFA and Meeting of the Deutsche Arbeits Gommunity for AKUSTIK (German annual conference for acoustics)*, pages 1149–1150, 2004. https://pub.dega-akustik.de/DAGA_1999-2008/data/articles/001763.pdf.

Catherine Lai. Perceiving Surprise on Cue Words: Prosody and Semantics Interact on Right and Really. In *10th Annual Conference of the International Speech Communication Association*, pages 1963–1966, 2009. https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2009/papers/i09_1963.pdf.

Ilse Lehiste. *Suprasegmentals*. The MIT Press, 1970. doi: https://doi.org/10.1163/26660393-bja10049.

Heike Lehnert-LeHouillier, Susana Terrazas, and Steven Sandoval. Prosodic Entrainment in Conversations of Verbal Children and Teens on the Autism Spectrum. *Frontiers in Psychology*, 11:582221, 2020. doi: 10.3389/fpsyg.2020.582221.

Yuanchao Li, Carlos Toshinori Ishi, Koji Inoue, Shizuka Nakamura, and Tatsuya Kawahara. Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. *Advanced Robotics*, 33(20):1030–1041, 2019. doi: 10.1080/01691864.2019.1667872.

Andreas Liesenfeld and Mark Dingemanse. Bottom-up discovery of structure and variation in response tokens ('backchannels') across diverse languages. In *23rd Conference of the International Speech Communication Association*, pages 1126–1130, 2022. doi: 10.21437/Interspeech.2022-11288.

Diane Litman and Katherine Forbes-Riley. Evaluating a Spoken Dialogue System that Detects and Adapts to User Affective States. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 181–185, 2014. doi: 10.3115/v1/W14-4324.

Diane J. Litman. Cue Phrase Classification Using Machine Learning. *Journal of Artificial Intelligence Research*, 5:53–94, 1996. doi: 10.1613/jair.327.

Diane J. Litman. Enhancing the Effectiveness of Spoken Dialogue for STEM Education. *Proceedings Workshop on Speech and Language Technology in Education*, pages 13–14, 2013. https://www.isca-speech.org/archive_v0/slate_2013/papers/sl13_013.pdf.

Max Louwerse and Heather Mitchell. Toward a Taxonomy of a Set of Discourse Markers in Dialog: A Theoretical and Computational Linguistic Account. *Discourse Processes*, 35:199–239, 2003. doi: 10.1207/S15326950DP3503_1.

Nichola Lubold and Heather 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*, pages 5–12, 2014. doi: 10.1145/2666633.2666635.

Nichola Lubold, Heather Pon-Barry, and Erin Walker. Naturalness and rapport in a pitch adaptive learning companion. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 103–110, 2015. doi: 10.1109/ASRU.2015.7404781.

Fatik Baran Mandal. Nonverbal Communication in Humans. *Journal of Human Behavior in the Social Environment*, 24(4):417–421, 2014. doi: 10.1080/10911359.2013.831288.

Lucia Mareková and Stefan Benus. Slovak 'no' and its pragmatic meanings and functions in relation to prosody. *Topics in Linguistics*, 21:1 – 14, 2020. doi: 10.2478/topling-2020-0001.

Youri Maryn, Nelson Roy, Marc De Bodt, Paul Van Cauwenberge, and Paul Corthals. Acoustic measurement of overall voice quality: a meta-analysis. *Journal of the Acoustical Society of America*, 126(5):2619–2634, 2009. doi: 10.1121/1.3224706.

Maël Mauchand and Marc D. Pell. Emotivity in the voice: Prosodic, lexical, and cultural appraisal of complaining speech. *Frontiers in Psychology*, 11, 2021. doi: 10.3389/fpsyg.2020.619222.

Maël Mauchand, Nikos Vergis, and Marc D. Pell. Irony, prosody, and social impressions of affective stance. *Discourse Processes*, 57(2):141–157, 2020. doi: https://doi.org/10.1080/0163853X.2019.1581588.

Jeffrey P. Migneault, Ramesh Farzanfar, Julie A. Wright, and Robert H. Friedman. How to write health dialog for a talking computer. *Journal of Biomedical Informatics*, 39(5):468–481, 2006. doi: 10.1016/j.jbi.2006.02.009.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. Self-

Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16, 2022. doi: 10.1109/JSTSP.2022.3207050.

Olivia Murton, Robert Hillman, and Daryush Mehta. Cepstral Peak Prominence Values for Clinical Voice Evaluation. *American Journal of Speech-Language Pathology*, 29(3): 1596–1607, 2020. doi: 10.1044/2020_AJSLP-20-00001.

Aparna Nadig and H. Shaw. Acoustic and perceptual measurement of expressive prosody in high-functioning autism: Increased pitch range and what it means to listeners. *Journal of Autism and Developmental Disorders*, 42(4):499–511, 2011. doi: 10.1007/s10803-011-1264-3.

Yasushi Nakai, Ryoichi Takashima, Tetsuya Takiguchi, and Satoshi Takada. Speech intonation in children with autism spectrum disorder. *Brain and Development*, 36(6):516–522, 2014. doi: https://doi.org/10.1016/j.braindev.2013.07.006.

Anindita Nath. Towards Naturally Responsive Spoken Dialog Systems by Modelling Pragmatic-Prosody Correlations of Discourse Markers. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion*, pages 128—-129, 2020. doi: 10.1145/3379336.3381490.

Anindita Nath and Nigel G. Ward. On the Predictability of the Prosody of Dialog Markers from the Prosody of the Local Context. In *Speech Prosody*, pages 664–668, 2022. doi: 10.21437/SpeechProsody.2022-135.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Vishnumurthy Adiga, and E. Cambria. Recent advances in deep learning based dialogue systems: a systematic survey. *Artificial Intelligence Review*, 56(2):3055–3155, 2022. doi: 10.1007/s10462-022-10248-8.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2006.

David G. Novick and Stephen Sutton. An Empirical Model of Acknowledgment for Spoken-Language Systems. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 96–101, 1994. doi: https://doi.org/10.3115/981732.981746.

Lynne C. Nygaard, Debora S. Herold, and Laura L. Namy. The Semantics of Prosody: Acoustic and Perceptual Evidence of Prosodic Correlates to Word Meaning . *Cognitive Science*, 33(1):127–146, 2009. doi: 10.1111/j.1551-6709.2008.01007.x.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. http://www.aclweb.org/anthology/D14-1162.

Sue Peppé. Prosodic development in atypical populations. *The Development of Prosody in First Language Acquisition*, pages 343–362, 2018. doi: 10.1075/tilar.23.17pep.

Sujitha P.S. and Gopi Kishore Pebbili. Cepstral Analysis of Voice in Young Adults. *Journal of Voice*, 36(1):43–49, 2020. doi: 10.1016/j.jvoice.2020.03.010.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. Conversational AI: The Science Behind the Alexa Prize. *ArXiv*, abs/1801.03604, 2018. doi: https://doi.org/10.48550/arXiv.1801.03604.

Harley Z Ramelson, Robert H Friedman, and Judith K Ockene. An automated telephone-based smoking cessation education and counseling system. *Patient Education and Counseling*, 36(2):131–144, 1999. doi: 10.1016/S0738-3991(98)00130-X.

Tommaso Raso, Albert Rilliard, and Saulo Mendes Santos. Modeling the prosodic forms of Discourse Markers. *Domínios de Lingu@gem*, 16(4):1436–1488, 2022. doi: 10.14393/DL52-v16n4a2022-8. URL https://hal.science/hal-03775659.

Timo B Roettger and Kim Rimland. Listeners' adaptation to unreliable intonation is speaker-sensitive. *Cognitive Science*, 204:104372, 2020. doi: 10.1016/j.cognition.2020.104372.

Gabriela Rosenblau, Dorit Kliemann, Isabel Dziobek, and Hauke R Heekeren. Emotional prosody processing in autism spectrum disorder. *Social Cognitive and Affective Neuroscience*, 12(2):224–239, 2017. doi: 10.1093/scan/nsw118.

Najmeh Sadoughi, André Pereira, Rishub Jain, Iolanda Leite, and Jill Fain Lehman. Creating Prosodic Synchrony for a Robot Co-Player in a Speech-Controlled Game for Children. In *ACM/IEEE International Conference on Human-Robot Interaction*, page 91–99, 2017. doi: 10.1145/2909824.3020244.

Oscar Saz, Shou-Chun Yin, Eduardo Lleida, Richard Rose, Carlos Vaquero, and William R. Rodríguez. Tools and Technologies for Computer-Aided Speech and Language Therapy. *Speech Communication*, 51(10):948–967, 2009. doi: 10.1016/j.specom.2009.04.006.

Deborah Schiffrin. *Discourse markers*. Cambridge University Press, 1987. doi: https://doi.org/10.1017/CBO9780511611841. ISBN:9780511611841.

Iris Scholten, Eerin Engelen, and Petra Hendriks. Understanding Irony in Autism: The Role of Context and Prosody. *The Facts Matter. Essays on Logic and Cognition in Honour of Rineke Verbrugge*, pages 121–132, 2015. https://www.let.rug.nl/hendriks/papers/irony2015.pdf.

Stephen J. Sheinkopf, Peter Mundy, D. Kimbrough Oller, and Michele Steffens. Vocal atypicalities of preverbal autistic children. *Journal of Autism and Developmental Disorders*, 30(4):345–354, 2000. doi: 10.1023/a:1005531501155.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning Wavenet on

MEL Spectrogram Predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4779–4783, 2018. doi: 10.1109/ICASSP.2018.8461368.

Hadas Shintel, Howard C. Nusbaum, and Arika Okrent. Analog acoustic expression in speech communication. *Journal of Memory and Language*, 55:167–177, 2006. doi: 10.1016/j.jml.2006.03.002.

Elizabeth Shriberg and Robin Lickley. Intonation of Clause-Internal Filled Pauses. volume 50, pages 172–179, 1993. doi: 10.1159/000261937.

Jennifer Smith, Aaron Spaulding, Harry Bratt, Dimitra Vergyri, Girish Acharya, Kristin Precoda, Andreas Kathol, and Colleen Richey. Towards Conversationally Intelligent Dialog Systems. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2022. doi: 10.1145/3491101.3519842.

Vikrant Soman and Anmol Madan. Social Signalling: Predicting the Outcome of Job Interviews from Vocal Tone and Prosody. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009. https://hd.media.mit.edu/tech-reports/TR-637.pdf.

David Talkin. REAPER: Robust Epoch And Pitch EstimatoR. https://github.com/google/REAPER.

David Talkin and W Bastiaan Kleijn. A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, 495:518, 1995. https://www.ee.columbia.edu/ dpwe/papers/Talkin95-rapt.pdf.

Jesse Thomason, Huy V. Nguyen, and Diane J. Litman. Prosodic Entrainment and Tutoring Dialogue Success. In *16th International Conference on Artificial Intelligence in Education*, pages 750–753, 2013. doi: 10.1007/978-3-642-39112-5_104.

Fatemeh Torabi Asr and Vera Demberg. On the Information Conveyed by Discourse Markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics*, pages 84–93, 2013. https://aclanthology.org/W13-2610.

Pirros Tsiakoulis, Catherine Breslin, Milica Gašić, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, and Steve Young. Dialogue context sensitive HMM-based speech synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2554–2558, 2014. doi: 10.1109/ICASSP.2014.6854061.

Christina Y. Tzeng, Laura L. Namy, and Lynne C. Nygaard. Communicative Context Affects Use of Referential Prosody. *Cognitive Science*, 43(11):12799, 2019. doi: https://doi.org/10.1111/cogs.12799.

Ilenia Tonetti Tübben and Daniela Landert. Uh and Um as Pragmatic Markers in Dialogues: A Contrastive Perspective on the Functions of Planners in Fiction and Conversation. *Contrastive Pragmatics*, pages 1–32, 2022. doi: https://doi.org/10.1163/26660393-bja10049.

Darinka Verdonik, Andrej Zgank, and Agnes Pisanski Peterlin. The impact of context on discourse marker use in two conversational genres. *Discourse Studies*, 10(6):759–775, 2008. doi: https://www.jstor.org/stable/24049381.

Sarenne Wallbridge, Peter Bell, and Catherine Lai. It's Not What You Said, it's How You Said it: Discriminative Perception of Speech as a Multichannel Communication System. In *22nd Annual Conference of the International Speech Communication Association*, pages 2386—2390, 2021. doi: 10.21437/Interspeech.2021-1658.

Karen Ward and David Novick. Prosodic cues to word usage. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 620–623, 1995. doi: 10.1109/ICASSP.1995.479674.

Nigel G. Ward. Midlevel Prosodic Features Toolkit. https://github.com/nigelgward/midlevel, 2015—2022.

Nigel G. Ward. A corpus-based exploration of the functions of disaligned pitch peaks in American English dialog. In *Speech Prosody*, pages 349–353, 2018. doi: 10.21437/SpeechProsody.2018-71.

Nigel G. Ward. *Prosodic Pattterns in English Conversation.* Cambridge University Press, 2019. doi: 10.1017/9781316848265.019.

Nigel G. Ward and David DeVault. Challenges in Building Highly Interactive Dialogue Systems. *AI Magazine*, 37:7–18, 12 2016. doi: 10.1609/aimag.v37i4.2687.

Nigel G. Ward, Anais G. Rivera, Karen Ward, and David G. Novick. Root causes of lost time and user stress in a simple dialog system. In *International Speech Communication Association*, pages 1565–1568, 2005. doi: 10.21437/Interspeech.2005-458.

Nigel G. Ward, Ambika Kirkland, Marcin Włodarczak, and Eva Székely. Two Pragmatic Functions of Breathy Voice in American English Conversation. In *Speech Prosody*, pages 82–86, 2022. doi: 10.21437/SpeechProsody.2022-17.

Virginia Wolfe and David Martin. Acoustic correlates of dysphonia: type and severity. *Journal of Communication Disorders*, 30(5):403–416, 1997. doi: 10.1016/s0021-9924(96)00112-8.

Mike Wu, Jonathan Nafziger, Anthony Scodary, and Andrew L. Maas. HarperValleyBank: A Domain-Specific Spoken Dialog Corpus. *ArXiv*, abs/2010.13929, 2020. https://deepai.org/publication/harpervalleybank-a-domain-specific-spoken-dialog-corpus.

Xin Xie, Andrés Buxó-Lugo, and Chigusa Kurumada. Encoding and Decoding of Meaning through Structured Variability in Intonational Speech Prosody. *Cognition*, 211(5):104619, June 2021. doi: 10.1016/j.cognition.2021.104619.

Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito. Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems. *22nd Annual Conference of the International Speech Communication Association*, pages 246–250, 2021. doi: 10.21437/Interspeech.2021-381.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-

Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing Universal PERformance Benchmark. In *22nd Annual Conference of the International SpeecCommunication Association*, pages 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

Atef Ben Youssef, Mathieu Chollet, Hazaël Jones, Nicolas Sabouret, Catherine Pelachaud, and Magalie Ochs. Towards a Socially Adaptive Virtual Agent. In *Intelligent Virtual Agents*, pages 3–16, 2015. doi: 10.1007/978-3-319-21996-7.

Casey J Zampella, Kelsey D Csumitta, Emily Simon, and Loisa Bennetto. Interactional Synchrony and Its Association with Social and Communication Ability in Children With and Without Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 50:3195–3206, 2020. doi: 10.1007/s10803-020-04412-8.

Fatemeh Zarei. Discourse Markers in English. *International Research Journal of Applied and Basic Sciences*, 4(1):101–117, 2013.

Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. Spontaneous Conversational Speech Synthesis from Found Data. In *20th Annual Conference of the International Speech Communication Association*, pages 4435–4439, 2019. doi: 10.21437/Interspeech.2019-2836.

Curriculum Vitae

Anindita Nath started her Ph.D. in Computer Science at the University of Texas at El Paso in Spring 2017. Her research interests include predictive prosody modeling, affect computing, and improving responsiveness in human-machine interactions. Specifically, her dissertation proposes to predict context-appropriate prosody in dialog markers by modeling their relationship with local context prosody. Her works have been published in peer-reviewed conferences: ACM ICMI 2018, ACM International Conference on Intelligent User Interfaces Companion 2020, and Speech Prosody 2022.

Anindita's projects as a Research Assistant included optimizing prosody-based machine learning solutions to predict: i) appropriate turn-taking in senior citizen conversations, both in American English and Japanese (Toyota Motor Corporation, Japan) and ii) relevance, resolution, and location of disasters in low-resource language news broadcasts (DARPA).

She has also been actively involved in Biomedical Informatics projects, such as applying–on NIH's data–process mining techniques to develop a maternal care pathway and data analytics to predict maternal morbidity in American pregnant women.

As an intern, she improved the accuracy of medical transcriptions via prosody-based LSTM RNN solutions (3M, 2019), employed Bert multi-task learning in health data mining projects (UPenn, 2020), and developed voice assistants that supported the caregivers of Alzheimer's patients by customizing the state-of-art models (BentenTech, 2022).

In India, Anindita earned her Masters in Computer Applications (2010) and MBA in Systems and IT Management (2015). She also worked as a software developer (2010) and an Information Technology executive (2011-2016).

At UTEP, Anindita was President of Upsilon Pi Epsilon (2019-2021) and ACM-W's Secretary (2021-2023). She was a UPE scholar in 2019 and the recipient of the prestigious Generation Google Scholarship in 2020.

Email: nath.anindita2110@gmail.com