Prosody in Human Communication and Machine Understanding

Sara B. Ng

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Richard A. Wright, Chair
Mari Ostendorf, Chair
Gina-Anne Levow
Pamela E. Souza

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

Prosody in Human Communication and Machine Understanding

Sara B. Ng

Co-Chairs of the Supervisory Committee:
Professor Richard A. Wright
Department of Linguistics

Professor Mari Ostendorf
Department of Electrical and Computer Engineering

Speech technology is a ubiquitous part of the modern world, from the voice-enabled assistants in smartphones to bespoke tools used by language researchers. Technological advances and the curation of large speech datasets have enabled these systems to identify words with remarkable quality. However, the black-box nature of large commercial speech understanding systems brings into question the extent to which they can take advantage of cues from prosody.

Prosody has great potential as an untapped source of linguistic information for speech understanding that is not surfaced in other aspects of language. Previous work has shown that prosodic information can be exploited computationally to resolve ambiguity for linguistic structures in computational models, and to perform tasks which are considered prosodically significant, such as sarcasm detection. However, computational systems do not benefit from the same social and conversational context that humans have in processing this kind of communication, making such tasks more challenging and further motivating the careful study of prosodic input.

This work investigates the hypothesis that explicit encoding of acoustic-prosodic features is a benefit to speech understanding technology. From the domain of punctuation prediction

in automatic speech recognition, I show that adding acoustic-prosodic measures can improve the performance of punctuation prediction models for speech transcripts compared to a system that uses only the word sequence. I provide a potential use case for prosodic modeling in the domain of speech entrainment. Finally, I show how computational methods can be used to understand human behavior in prosodically marked speech within the domains of speech timing and regions of presumed hyper-articulation.

This work bridges the gap between linguistic questions about prosody, and computational questions about the use of or need for linguistically-motivated acoustic features. Understanding how prosody influences the quality of speech understanding systems is vital in enhancing their utility across various domains and for diverse speakers. The synthesis of the these research strands provides a bird's eye view of the methodologies and challenges that can be involved in computational processing of prosody.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

ASR: Automatic Speech Recognition

ATAROS: Automatic Tagging and Recognition of Stance

CNN: Convolutional Neural Network

EDF: Effective Degrees of Freedom

F0: Fundamental frequency

F(1-3): Formant (1-3)

FTO: Floor-Transfer Offset

GAM: Generalized Additive Model

GAMM: Generalized Additive Mixed Model

GRU: Gated Recurrent Unit

IP: Interruption Point

IPU: Inter-pausal Unit

LSTM: Long Short-Term Memory Network

MFCC: Mel-frequency Cepstral Coefficient

NHR: Noise-to-Harmonics Ratio

NLP: Natural Language Processing

RMSE: Root Mean Square Error

RNN: Recurrent Neural Network

TOBI: Tone and Break Indices

WER: Word Error Rate

ACKNOWLEDGMENTS

Thank you to my Ph.D. advisors Richard Wright and Mari Ostendorf for their mentorship, guidance, support and generosity throughout my time at the University of Washington.

Thank you to my committee member Gina-Anne Levow for sharing her vast knowledge of speech technology, and for being a frequent and always pleasant mentor and collaborator. I am thankful for Pam Souza's open perspective on collaboration, and her encouragement when I wanted to explore projects that tested our collective understanding in new ways. I am also grateful to Gregory Ellis and Kendra Marks from Northwestern, who fielded my data questions with incredible patience.

The administrative staff of the linguistics department have been dutifully responsible for my financial and logistical well-being these past seven years, for which I am extremely grateful. Thank you Joyce Parvi, Karoliina Kuisma, Brandon Graves, Zach Phelps, Mike Furr, and others.

Thank you to Katie Vadella and Andrew Hard, my intern hosts at Google in 2019 and 2021. I am so grateful for the mentorship and cameraderie I experienced in my time at Meta Reality Labs. Christi Miller was the best intern manager one could ask for, and I am truly glad for the opportunity to learn from her. My peer mentors Calvin Murdock, Hao Lu, Vamsi Krishna Ithapu, and Khia Johnson were all so generous with their time and expertise. I also learned so much thanks to Women's Working Lunch at Reality Labs, who gave their unique support and career advice.

I have so many people to be thankful for from the UW student body. Naomi Tachikawa Shapiro and Amandalynne Paullada-Won were persistent mentors and friends in the early years of my Ph.D., and gave generously with their time and expertise to help me learn

how to be an academic. The members of the TIAL lab have been a source of support and knowledge transfer throughout my time at UW. Vicky Zayats and Trang Tran spent many hours patiently walking through code when I first joined the lab. Kevin Lybarger, Ellen Wu, Roy Lu, Sitong Zhou, Chia-Hsuan Lee, Kevin Everson, Jenny Cho, Yushi Hu, Junkai Wu, and Yuling Gu have been fantastic lab mates and collaborators. Thank you to the Phonetics Lab members for allowing me to learn from your scholarship. Alicia Beckford Wassink and the Sociolinx working group welcomed me into their research group this past year, where I have learned so much more about the human impacts of speech technology. Thank you all.

My cup runneth over with friendship, gossip, memes, put-downs, and machine learning know-how thanks to Agatha. I am eternally indebted Ray Gagné to for his many hours of last-minute proofreading. I am astounded that after so much paper editing he does "on the clock," he has still made time to help my in my usual state of procrastinated writing. Emily Proch Ahn has been a stalwart collaborator, sounding board, and friend to me. Thank you to everyone in "407" who helped me explore Seattle: Katie Lindekugel, Cassie Maz, Dr. Downey, Yadi Peng, Ty Gill-Saucier, Sunny Ananthanarayan, and many others. I am also very fortunate to have Courtney Mansfield and Leanne Rolston in my corner; they both freely and promptly lent their expertise on interacting with data and understanding surprisal and stance, respectively. Rik Koncel-Kedziorski and Dhanush BK have been study buddies, idea bouncers, and irreplaceable friends. Thank you all.

Thank you to the hundreds of students I have had the pleasure of working with as a TA, instructor, and classmate. I've learned so much from you. Thank you to Gita Dhungana and Ted Kye; it has been a pleasure to teach with both of you.

I have been overwhelmed with the support and community that the Seattle Sacred Harp Singers has been for me. And importantly, thank you Leland Ross for reminding me that Esperanto is more than a conlang, it's a way of life.

I am grateful to my brother Alex for this thoughtful gifts that have kept me fed in

a variety of novelty forms (especially the Super Mario mushroom waffles for their special energy boost). Thank you to my parents for everything they've done to support me through this process, and to my extended family for sending their support from afar.

DEDICATION

For my cat, without whom I would not have survived the Ph.D.

So science spreads her lucid ray
O'er lands which long in darkness lay:
She visits fair Columbia,
And sets her sons among the stars.

Ode on Science, Jezaniah Sumner
Denson Scared Harp 242

Chapter 1

INTRODUCTION

Language exists in many forms, from face-to-face communication to interfacing with automated customer service agents through text; and humans are capable of adapting their language to suit various modalities and language environments. However, at the core of all produced language is a need to transmit information between entities. Prosody, the "study of the tune and rhythm of speech and how these features contribute to meaning" [1], is often used to transmit information that cannot be derived from words alone. For example, when a speaker is expressing feelings of anger they may explicitly express their emotions via the words they choose to say, and amplify and reinforce their ire through a harsh tone of voice. Prosodic variation also allows speakers to create meaning that alters the semantic interpretation of the words themselves, for instance in expressions of sarcasm.

Despite its many utilities in conveying information beyond the basic word choice and grammar of speech, prosody is less-studied than other sub-fields of linguistics. There are many practical reasons for this. Collecting empirical data is much more difficult in the audio modality than in text, but prosody primarily manifests in spoken language. What's more, prosodic strategies can vary wildly from person to person. For example, sociolinguistic factors such as gender [2], ethnic identity [3], other social identities [4], and speech environment or register [5, 6] are all factors that influence which prosodic strategies are used to create a given effect. Prosodic features are also not consistent between languages [7] or even between dialects [e.g., 2, 8, 9]. And while speech data is large and available, methods for quantizing prosodic behaviors into interpretable units do not have consensus in the field.

However, there is pioneering experimental work showing the many uses of prosody in natural language. Within pragmatics, the discourse behavior of stance-taking (investigated

in Chapters 5 and 6) has been shown to manifest in the acoustic signal as prosodic variation [10]. Systems for prosodic annotation, based on the Tone and Break Indices framework [11] as well as other cross-linguistic systems of annotation [e.g., 12, 13] are continuously being developed for understudied language varieties.

Fluency in prosodic behavior is required by both humans and machines. In human-human interactions, we are able to process the subtle prosodic cues produced by our interlocuters, and use these cues to process and understand what is being conveyed. In turn, as speakers we naturally make use of prosodic changes to convey meaning to our interlocutors, with the subconscious expectation that they will correctly interpret meaning from our modifications to the acoustic signal.

For machines, naturalistic prosody is a necessary early goal post in producing normal-sounding speech. For example, text-to-speech systems such as Amazon's Alexa or Google Assistant can make use of Speech Synthesis Markup Language, a programmatic method for encoding prosodic templates onto text [14]. Other computer interfaces may process the prosody of input audio in order to determine when to break transcriptions into smaller units [15], diarize text [16], or disambiguate between sentences types (e.g., questions versus declaratives for languages such as French where the surface word order can be ambiguous) [17].

While prosody is well-studied in applications such as speech synthesis, it is less explored for applications in speech understanding. A challenge specific to the study of prosody for speech understanding systems is that there is not a consensus on how to classify prosodic phenomena. Compared to a field such as syntax, where units and labels tend to be discrete, clearly defined, and typologically consistent, prosody is noisy and difficult to discretize. Annotation standards can vary based on the language or dialect, and are often focused on specific phenomena. Another challenge in researching prosody within the computational understanding space is the ideological divide between linguistic perspectives on prosody and its application in technology. Linguists tend to focus on specific speech phenomena and the local acoustic changes that convey meaning and emotion. However, when audio is modeled

computationally without hand-selecting of features, there is no way for practitioners to embed linguistic knowledge of prosody into the learning process. Bridging this divide is crucial for achieving the dual goals of understanding prosody's role in natural speech and developing computational methods to accurately represent these mechanisms.

The work I present in this thesis comprises two complementary goals. First, I seek to use computational methods to understand prosody from the human perspective. I empirically analyze large corpus speech data in order to understand human prosodic behavior at scale. Second, I investigate how knowledge of prosody can improve the performance of modern speech understanding technology, and show ways in which the use of prosody for speech technology is not being exploited to its fullest potential.

Specifically for improving speech understanding technology, I show that acoustic-prosodic features can improve the prediction of punctuation for transcribed text. I also explore a method for modeling the entrainment of prosodic features between interlocutors, and discuss the limitations of such an approach. From the perspective of understanding human behavior, I show that speech timing is influenced by stance-taking behaviors in discourse, and further that its effect on hyper- and hypo-articulation interacts with the informational load of speech.

This document is organized as follows: In Chapter 2, I provide a theoretical overview of prosody and related speech concepts. In Chapters 3 and 4, I show two experiments using explicit encoding of prosody to improve performance on natural language processing tasks. Finally, in Chapters 5 and 6, I empirically demonstrate how communicative needs can alter the flow of speech via changes in time and articulation, respectively.

Chapter 2
BACKGROUND

Before discussing the role of speech prosody in language technology, a theoretical understanding of prosody must be established. Further, it is beneficial to review what is known about the function and manifestation of prosody in natural speech. In the following sections, I present an overview of the linguistic theory of prosody. Section 2.1 overviews the theoretical conceptions of prosody in linguistic research, including systems for measuring and categorizing prosodic behavior, functions of prosody in discourse, and cross-linguistic variation. In Section 2.2, I describe some of the current common uses of prosody in speech technology. In Section 2.3, I will introduce the notions of spontaneous and conversational speech, and discuss why this type of language is ideal for the study of prosody. Section 2.4 outlines the notion of stance-taking, a prosodically-encoded discourse function which I explore in Chapters 5 and 6. Finally, Section 2.5 describes the corpora which are used in various experiments in Chapters 3 to 6.

2.1 Foundations of Speech Prosody

In the web series of introductory linguistics topics from Macquarie University, prosody is broadly defined as "the study of the tune and rhythm of speech and how these features contribute to meaning" [1]. In their textbook for Natural Language Processing (NLP) practitioners, Jurafsky & Martin define prosody as "the study of the intonational and rhythmic aspects of language, and in particular the use of F0, energy, and duration to convey pragmatic, affective, or conversation-interactional meanings" [18, ch. H p. 7]. For the purposes of this dissertation, I consider English prosody to be any suprasegmental changes to the speech signal, especially those produced via acoustic perturbations from the norm, that contribute

new meaning or reinforce the intended meaning of language. This definition is purposefully broad: prosody must not be only acoustically defined, lest it exclude written or signed languages (though the latter is not explored in this work). It must also not rely on specific acoustic cues; the acoustic-prosodic strategies I've observed are not uniform across tasks or languages. And finally, there must be a purpose behind changes to the signal [19].

2.1.1 Prosody as Suprasegmental Phonology

Consensus views of prosody can be separated into three generalizations: 1) that prosody is a part of the linguistic signal separate from other linguistic levels such as morphemes and syntax, 2) that prosody contributes to the meaning or purpose of speech, and 3) that prosody manifests as observable patterns in the acoustic speech signal.

The Autosegmental framework of phonology organizes speech sounds into what are seen as the smallest discrete units of language, called *segments*, and describes features that can be associated to the segments [20]. For research agendas that follow the autosegmental view, prosody must be described relative to the segment. In their discussion of the wide range of definitions for prosody, Cutler and Ladd assert that the Autosegmentalist views prosody as "any phenomena that involve phonological organization at levels above the segment" ([19, p. 2]). In this view, prosodic phenomena are a unique metrical system above the segment, which may fortuitously appear to apply to single segments but more often apply to larger prosody-bearing units [21].

From a practitioner perspective, the components of prosody may be more related to the labeling scheme or acoustic measures employed. The most common labelling scheme for prosody across languages is the Tone and Break Indices System (ToBI) [11], which enables annotators to describe melodic contours and breaks at multiple levels of granularity using discrete labels. Annotation standards vary from language to language and even between varieties of language [7], however the notation remains more-or-less consistent across languages.

2.1.2 Acoustic Correlates of Prosody

The most commonly attributed acoustic correlates of prosody are fundamental frequency, duration, and intensity [19, p. 1]. These measurements are often described as local changes to general contours, for example a peak or trough in an intensity contour, unusually long syllable duration, or a rising pitch contour in the final words of a sentence. Measurements can be associated to any linguistic level: phones, syllables, words, intonational phrases, etc. Other acoustic correlates of prosody have been observed as well. For example, changes in voice quality are observed in regions of prosodic interest. Davidson [22] notes that creaky voicing can be used to establish sentential prosodic boundaries in English, Mandarin, and Finnish.

2.1.3 The Function of Prosody in Discourse

One well-known discourse function of prosody in English is in the creation of focus constructions. A *focus* in discourse is a region of semantic importance that introduces new or contrasting information, in other words "the nonpresupposed part of the sentence" [23, p. 1]. Example (1) shows an example where on the topic of what a cat eats for breakfast, the word *duck* is emphasized to identify a variety of broth. One possible context for this word to be focused is when the interlocutor fails to hear a part of a previous utterance, and so it must be re-introduced to the discourse. Another common type of focus construction is *contrast focus*, where a constituent is emphasized in particular to contrast it against some other conflicting constituent [24]. Example (2) shows the word 'salmon' is contrasted as a correction for the work 'duck,' which was previously introduced to the discourse.

(1) The cat eats kibble and DUCK broth for breakfast.

(2) No, the cat wants SALMON broth, not duck broth!

Notice that in (1) in particular, the fact that the word *duck* is being focused cannot be deduced from the words or word order alone; something about the pronunciation of the phrase must be changed.

2.1.4 Cross-linguistic Variation

In general, prosody is used for managing turn-taking, signaling discourse structure, and facilitating mutual understanding. Through prosodic cues, speakers can indicate the boundaries of their utterances [25, 26], as well as highlight key points and express emphasis or surprise [8]. Moreover, prosody enables listeners to resolve ambiguous lower-level speech cues [27], detect sarcasm or irony [28], and infer the speaker's emotional state or intentions. The ability to interpret and respond to these prosodic cues is a fundamental aspect of natural human conversation.

The English language has a particular poverty of distinguishing prosodic phenomena when compared to other languages. Some prosodic phenomena are fossilized to specific lexemes, for instance the case of negative 'yeah' [29]. Like other languages, though, English prosody is still a valuable too for speakers and listeners in spontaneous conversation. For example, [30] found differing distributions of speech timing, loudness and measures of "vocal intensity" between speakers who were judged by interlocutors to be bad and good conversationalists. However other languages more robustly use prosodic information as a primary information stream of both semantic and pragmatic information.

One obvious typological distinction influencing prosody is the dichotomy of tonal and non-tonal languages. In tonal languages, pitch patterns are associated with lexemes. Variations in these pitch patterns distinguish word forms from one another. For example, in Toisan Chinese, tone is used to distinguish word forms such that words may have identical phone sequences but vary in tone and have very different meanings. This is shown in Examples (3–7). The numerals at the end of each pronunciation indicate the tonal pattern.

(3)	凉 lianŋ22 'cool'	(4)	领 lianŋ33 'to receive'	(5)	两 lianŋ55 'tael' (50g)	(6)	领 lianŋ21 'collar'	(7)	亮 lianŋ32 'bright' [31]

The distinction between tonal and non-tonal systems of language is important for assessing other uses of pitch for suprasegmental prosody. Within tonal languages of the same variety,

different prosodic strategies may be used to convey the same pragmatic information [32, 33]. In addition, languages within the same tonal category may include or exclude suprasegmental prosodic distinctions; for example, Mandarin Chinese encodes stress, while Cantonese does not [7, p. 431].

Languages also vary in what structure of language forms the metrical unit; this is known as language *isochrony* [34]. Languages can be divided into *syllable-*, *stress-*, and *mora-timed* groups. Syllable-timed languages include Italian and Spanish, while English and many of Germanic languages are stress-timed [7, p. 432]. The mora-timed category was created expressly for the analysis of Japanese [35, p. 252]. The distinction between metrical timing structures can be subtle, and the existence of such discrete categories is controversial [36]. However, they influence prosodic features such as the degree of vowel lengthening that can occur [7, p. 432]. Importantly, the different timing structures of languages makes it dubious that observed prosodic norms will hold between languages with very different metrical structures.

2.2 Prosody as Input to Speech Technology

Prosodic information has been widely used in speech synthesis applications. In Wavenet, a popular generative speech model, speech synthesis functions are learned from log pitch and phone duration [37]. Another domain where prosodic information is actively used is for creating speaker personas. Researchers used prosodic representations in [38] to create synthesized speech with learned patterns of simple emotion categories, and found that these emotional personas were preferred over systems that did not learn the affective prosody. An example of explicit prosodic encoding improving speech understanding comes from [39], where researchers used transformers with a prosody encoder to improve classification performance on tasks where affective prosody is particularly useful.

Explicit prosodic features have been used in a variety of natural language processing tasks, such as constituency parsing [e.g., 40], identification of stance-taking [e.g., 41], punctuation prediction [e.g., 42], and dialogue act prediction [e.g., 43]. A variety of approaches have

been used previously to create models of prosody in conversation. For example, [44] used an LSTM model (as do we; see Section 4.3.3) trained on text-based measures to jointly predict pitch and phone duration. Within the scope of creating realistic prosody for speech synthesis, [45] used contextual word embeddings to create a prosody predictor for novel speech. Based on subjective judgments of synthesized speech using their predictor, they outperformed a comparable system which had access to oracle prosodic encoding, however there is still a gap between their model's rating and the raters' ranking of natural prosody samples.

End-to-end models of speech are thought to implicitly learn prosodic representations. Some research has attempted to quantify the relationship between prosodic cues and end-to-end system performance. For example, [46] probed the self-attention layers of an emotion recognition system. They found that the impact of fine-tuning on attention to voice quality and pitch features was greater for higher layers. Other work has focused on hybrid approaches, where systems can be front-loaded with learned prosodic representations [e.g., 47].

Prosodic cues convey meaning, manage turn-taking, and express emotions. Recognizing the significance of prosody, computational models have incorporated prosodic features to improve performance in various NLP tasks, including speech recognition, sentiment analysis, speaker identification, and dialogue systems. These cues are an underutilized asset to speech processing and understanding tasks.

2.3 Spontaneous and Conversational Speech

One important distinction for studies of prosody is the register and environment in which speech occurs. Spontaneous, conversational speech is a context where prosodic information carries more information than in other forms of speech. Spontaneous speech is unprompted speech where speech planning is entirely the responsibility of the speaker and no prompts such as word lists or prose are provided. Conversational speech occurs between more than one speaker. Compared to more formal registers of speech such as read speech or word list elicitation, spontaneous, conversational speech contains richer linguistic information. This

includes the social context and dynamics of an interlocutor or interlocutors, casual registers which may have more acoustic reduction [48], and non-standard forms of language. In the context of prosodic research, conversational speech exposes more pragmatic communication needs than other forms of speech, making it ideal for study. There can be challenges to performing empirical research on spontaneous, conversational data: data collection at scale is difficult and may induce artificial biases versus truly naturalistic data. Data can also be much noisier, and target phenomenon may not always be present. Finally, the nature of conversational data introduces many possible confounds which can be difficult to account for, such as visual cues, speaker and interlocutor demographics, and other interpersonal context which may not be available to the researcher.

2.4 Stance

Stance-taking is a pragmatic information-sharing strategy which speakers use to disclose "attitudes and opinions about the topic of discussion" [49, p. 1]. In conversation, stance-taking occurs when speakers convey information about their internal state to their interlocutors, and these internal states can then adapt within the conversation [50]. For example, stance-taking may be evaluative, i.e., assigning an attribute such as *pretty* or *ugly*, or aligning, i.e., showing that speakers agree with one another [51]. Thus in order to correctly process stance-taking, information about the speaker state, interlocutor state, and the linguistic object of stance must be interpreted.

Stance-taking acts can be categorized into meaningful groups based on their goal and intensity. In this work, I follow the coarse dichotomy defined in [41, 49, 52]. In this system, the object or goal or stance-taking is not considered. However, stance acts are grouped according to their strength and polarity. Strength refers to the fervor of the attitude or opinion being expressed. A purely informative declaration such as "The cat is sitting on her pillow" has no strength, as there is nothing in particular about the speaker's state evident in the text. In contrast, "I adore listening to the cat purr" has much stronger stance, making obvious the evaluative opinion of the speaker towards the object (cat purring). Stance

polarity exists on a categorical continuum with Negative, Neutral, and Positive variants, and refers to the sentiment of the stance being expressed. Table 2.4 provides examples of the combinations of stance strength and polarity.

Table 2.1: Examples of stance-taking in ATAROS by strength and polarity.

Stance → Polarity ↓	None	Weak	Moderate	Strong
Negative	n/a	I don't like cutting things in general	I would say no no no keep it	I would definitely not wanna get rid of the poetry
Neutral	I'm wondering if it's like the thing	That could bring in big bucks	People can still get taxis in the most important areas anyway	They're getting run out by like all these ride share programs
Positive	n/a	Let's just leave it there for now	You're right I didn't think about it that way	You could just drink water that's what I would cut too

2.5 Datasets

2.5.1 ATAROS

Many of the experiments in this work rely on the Automatic Tagging and Recognition of Stance (ATAROS) dataset,[1] a collection of task-based dyadic conversations created to elicit stance-taking speech behaviors [49]. Adult native English speakers from the Pacific Northwest Region of the United States (Washington, Oregon, and Idaho) were paired in age-matched dyads [49, p. 22] and tasked with completing collaborative activities.

Previous work using this dataset has quantitatively shown that dynamic phonetic and prosodic changes occur with stance-taking behaviors [49, pp. 57-66]. This data has also been

[1] https://depts.washington.edu/phonlab/projects/ataros.php

used to create computational predictors of stance from phonetic and lexical information [41, 53].

There are six tasks completed by each dyad, with the third and sixth tasks being targets of stance elicitation. The third task (the *inventory* task) was designed to be a low motivator of speaker engagement, and thus elicit examples of weak stance-taking. In this task, speakers are given a list of items and must collaboratively arrange them into plausible aisles for a grocery store. The sixth task (the *budget* task) is assumed to be a high motivator for speaker engagement, and thus elicit more instances of strong stance-taking. Participants are provided with a hypothetical county budget, and asked to agree on which line items to cut in order to balance the budget. Close talking microphones captured mono audio for each speaker [52]. Speech is subdivided into *spurts* of continuous speech where no internal pause is greater than 500 milliseconds, following [54]. Human annotators transcribed the speech at the word level, and marked speech spurts as having one of four stance strength levels (none, weak, moderate, or strong), and one of three stance polarities (neutral, negative, positive) [49, p. 31].

2.5.2 Telephone Conversations

Two other conversational datasets appear in the experiments in Chapters 3 to 5, the Switchboard and Fisher datasets [55, 56]. These are two collections of telephone calls, randomly initiated between strangers located in the United States on a small set of assigned topics. These datasets both have widespread use in many domains including speech recognition. While these datasets both comprise approximately 10 minute telephone conversations on a provided topic, there are some distinctions. In Fisher, subjects were explicitly allowed to take part in up to 3 recording sessions [56]; in Switchboard there was no explicit permission for participants to take part in multiple calls, thus some speakers appear in multiple conversations [55]. In contrast to ATAROS, Switchboard and Fisher contain an order of magnitude more paired audio and transcription data, on a wider dialectical pool. This makes these datasets better suited for training neural language models, which benefit from large training

data. For Switchboard in particular, many useful annotations have been contributed by various research groups, including annotations for dialogue act (with labels for backchanneling) [57], high fidelity speech timing [58], and disfluencies [55]. The dataset also offers relative gender balance, and other demographic information such as dialect region and age of all participants. Table 2.2 gives an overview of the size of Switchboard, with gender information about the participants.

Table 2.2: Count of conversations by gender in Switchboard. Note that these counts do not reflect the true number of unique participants in the dataset, but the number of conversations matching the demographic criteria. This is because some participants took part in more than one conversation.

Subset	Speaker Gender	Total Count	with Female interlocutor	with Male interlocutor
Train	Female	2053	1202	851
	Male	1669	851	818
Dev	Female	134	48	86
	Male	222	86	134
Test	Female	122	36	86
	Male	276	86	190

In telephone audio collected at the time these data were collected, band pass filtering removed information from the signal at high and low frequencies. Variables such as environmental noise, signal quality, distance from speaker to phone receiver, and other uncontrolled factors such as phone make and model are also an inherent challenges of working from telephone data.

Chapter 3

AUTOMATIC PUNCTUATION PREDICTION FOR SPONTANEOUS SPEECH

3.1 Prosody and Punctuation

Thus far, I have described prosody as a process occurring primarily in *spoken* language, one that is tied to the high-level acoustic properties of speech. However, artifacts of prosody are also present in written language as well. Chiefly, punctuation can be viewed as a proxy for many prosodic functions. For example, commas can be used to indicate clause structure. Exclamation points, periods, and question marks are a shorthand for the prosodic contours that define exclamations, declarations, and questions, respectively. Thus punctuation is a crucial part of many written languages, because it has the power to influence the same types of meaning changes as prosody does to spoken language.

3.2 Automatic Prediction of Punctuation for Speech Recognition

In Automatic Speech Recognition (ASR), machines are tasked with transforming language from a spoken modality to a written modality. In order for readers to recover the same meaning from the written language as in the original speech, the transcription must include information from prosody, such as sentence structure. By effectively using punctuation within the transcribed text, systems are able to preserve some of the prosodic information

---

The work presented in this chapter was completed in equal collaboration with Yeonjin Cho, as published in "Leveraging Prosody for Punctuation Prediction of Spontaneous Speech" at INTERSPEECH 2022 [42]. My specific contributions were the research question, experimental design, and equal contributions to analysis.

that would otherwise be lost.

At the time of this work, many ASR systems did not have mechanisms for including punctuation in transcribed text. This results in transcriptions that can be difficult to correctly interpret, or even to read fluently [59]. These difficulties are magnified when other challenges such as word recognition errors co-occur. Accurate punctuation prediction for automatically transcribed text is thus vital in providing users with readable, interpretable text.

As discussed in 2.3, the expected register of speech has a massive impact on its acoustic and grammatical properties. For more formal and regularized registers of language, e.g., read speech, the "correct" punctuation can be predicted with high fidelity from the word sequence alone. This is especially true for modern large language models. However, prosodic features may still be useful for less formal, spontaneous styles of speech. In these registers, speech structure is less likely to match the text that the language models were trained on. For instance, spontaneous speech contains many partial or incomplete sentence fragments, which are uncommon in text and do not have standardized terminal punctuation and may require a novel punctuation inventory. In these cases, prosody can be more helpful. However, it is also possible that prosodic features are susceptible to the conditions that induce ASR errors.

In this chapter, I propose a framework for incorporation of prosodic features into a model for predicting punctuation for transcribed speech. I include punctuation marking for interruption points and incomplete utterances, which are necessary for the domain of transcribed speech. I will show how using prosody improves performance over text alone, and the extent to which punctuation quality is influenced by transcription quality. My research questions are as follows:

1. How much does use of prosody improve punctuation prediction?

2. Are some types of punctuation easier or more difficult to predict?

3. How does transcription quality effect punctuation prediction?

3.3 *Methods*

I train models that learn associations between text transcriptions and acoustic features, and word-level punctuation tags, and evaluate on unseen test data. I compare models that learn these associations from various types of acoustic features.

3.3.1 Modeling Procedure

I base a punctuation prediction model on the dialog act recognition model from [40, 43], which is an RNN encoder-decoder with attention [60] with a CNN for the acoustic features [61]. In contrast to [43], previous context labels are not used since punctuation is less dependent on the labels of previous turns. A graph of the model structure is given in Figure 3.1.

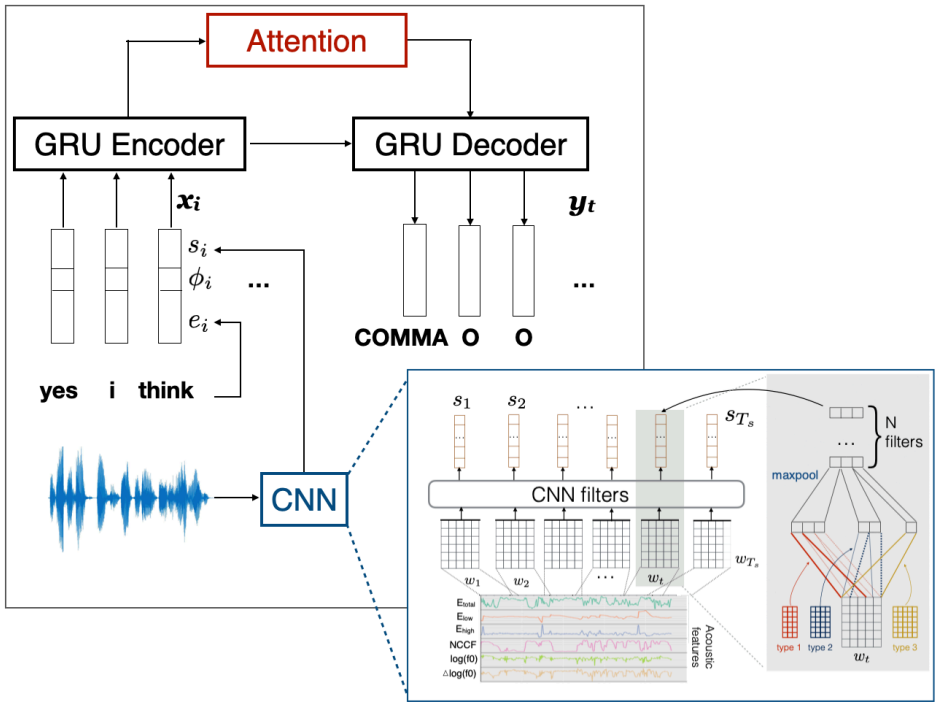

Figure 3.1: Schematic of the punctuation model. Each turn u is encoded via embeddings of the BERT-tokenized text, (optional) pause and duration embeddings, and (optional) convolved acoustic features. Image from [43].

3.3.2 Input features

The input to the model varies in two dimensions: the type of transcriptions (either human-created or ASR), and the acoustic features used. A model with a complete set of prosody features includes word embeddings, pause and duration features, and learned energy and pitch features. In sum, these comprise a prosodically-contextualized word vector. Word embeddings are derived from pre-trained BERT embeddings [62], specifically the base uncased version of BERT.

The pause and duration features include a raw and categorical (6 categories as in [40]) measure of the pause duration after each word, and word duration normalized by the mean duration of the word type in the training set.

The acoustic features are learned via a CNN from energy and pitch contours as described in [40]. In contrast to [40], the center of the convolution window is placed at the end of words, since changes in prosody that are associated with syntactic structure are more likely to occur at the word boundary.

Pitch and energy measures were extracted at the frame level using Kaldi [63], normalized by speaker, and aligned to the word-level. Aligned frames are convolved with N filters of m sizes (a total of mN filters). This allows features to capture information at different time scales. Each filter has a 1-D convolution over the pitch and energy features with a stride of 1, and the result is max-pooled to output mN-dimensional speech vectors.

3.3.3 Data

I use the Switchboard dataset [55] as the setting of experimentation. The Switchboard dataset has been transcribed and corrected by expert annotators, including annotations for dialog acts and disfluencies. As a practical consideration, I limit analysis to the subset of sessions which have been annotated for dialog acts [64], since the system I build from [40] was originally designed for dialog act segmentation and recognition. I also follow the train-test

split described in [64]. I use the text from the transcription set with disfluency annotations[1] when available. The Mississippi State transcriptions [58] correct time alignments from previous transcriptions and are considered the most temporally-faithful available transcriptions. I use the Mississippi State timings when alignments are available. The test set defined in [64] includes some sessions which have not been annotated for disfluency. Where experiments predict interruption points, the reported performance is based on the subset of the test set that has been annotated for disfluency. I label this the "IP test" set.

Table 3.1 reports the relative size of these subsets of data.

Table 3.1: Dataset statistics of Switchboard, from [42].

Split	# Dialogues	# Turns	# Sentences	# Tokens
train	1.1K	107K	194K	1.4M
dev	21	1.6K	3.2K	25K
full test	19	2.4K	4.1K	29K
IP test	14	1.7K	2.9K	21K

3.3.4 Preprocessing

The speech unit processed by the model is a speaker turn, i.e. the concatenation of uninterrupted utterances from a speaker, ignoring backchannels. Backchannels are treated as separate turns. Non-linguistic vocalizations (e.g. laughter or yawning) are removed from the transcriptions. Text is tokenized at the word level, with no separation of contractions. While many tokenizers for this kind of data and task split contractions (`you're` → `you 're`), on the basis of semantic and syntactic distinctiveness, I choose not to split on the basis that non-final constituents of a contraction are never punctuated and further that the contraction as a whole comprises a prosodic unit.

Punctuation labels are taken from the Switchboard transcriptions with some exceptions

[1] `https://doi.org/10.35111/gq1x-j780`

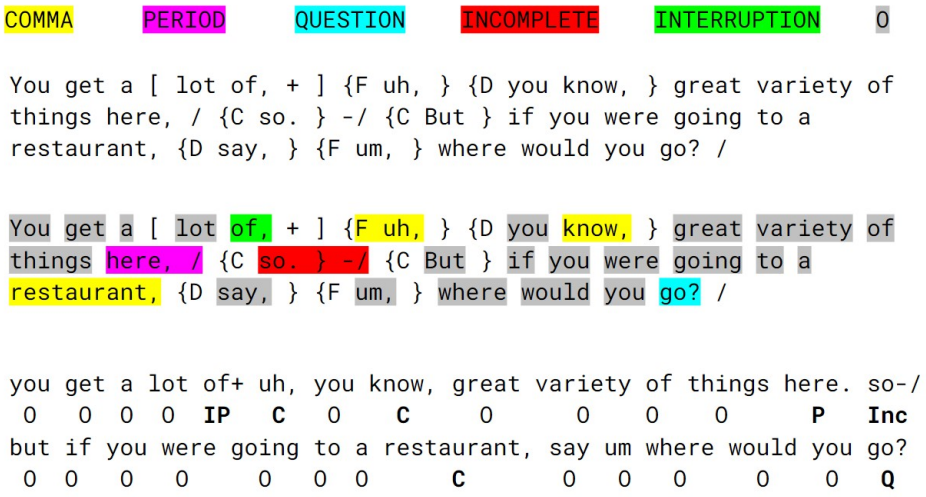

Figure 3.2: Example of Switchboard preprocessing. The top block shows the raw data, including original disfluency annotations; the middle block shows the mapping of annotations to punctuation tags; and the bottom block shows the resulting labels. Image by Yeonjin Cho [42].

as shown in Figure 3.2. The general tag set is:

- Period - full stop, terminal punctuation for sentence-like units
- Question - terminal punctuation for interrogative utterances
- Comma - non-terminal punctuation
- Incomplete - premature termination of an utterance ("-" in Switchboard notation)
- Interruption Point (IP) - the demarcation between reparandum and repair in a disfluency ("+" in Switchboard notation)

Disfluency span markers, annotations for coordination, and discourse markers are ignored. In the standard Switchboard transcriptions, filled pauses and interruption points are always enclosed by commas. This practice artificially biases the tag set towards non-standard comma

prediction, and so commas in those regions are removed. Periods are inserted at the end of slash units (sentence-like units of speech from one speaker). In addition, commas attached to *uh* and *um* are removed, unless they precede the phrase *you know*. Non-terminal punctuation such as exclamation marks and ellipses were infrequent in the dataset, and are rarely considered in punctuation prediction systems [c.f., 65, 66, 67, 68]. Thus, all terminal punctuation except question mark is replaced with a period. Punctuation in general is associated to the word directly preceding it.

The two sets of experiments differ in how interruption points are labeled. In the first set, interruption points are treated as having no punctuation. I will also refer to these as the 4-class experiments, since they will predict only comma, period, question, and incomplete. Counts of tokens labeled using the 4-class set are given in Table 3.2. In the second set, the model can explicitly predict interruption points. These are the 5-class experiments. In both settings, a dummy "other" (O) label is predicted for all words without punctuation.

Table 3.2: Counts of 4-class punctuation types: 'C,'=comma; 'P.'=period; 'Inc-'=incomplete; 'Q?'=question; 'O'=no punctuation. Roughly 4% of the "O" tokens from the 4-class correspond to IPs in the 5-class system. From [42].

Split	C,	P.	Inc-	Q?	O	Total
train	128K	127K	9.0K	7.8K	1.1M	1.4M
dev	3.2K	2.2K	144	92	19K	25K
full test	2.8K	2.7K	175	197	23K	29K
IP test	1.8K	1.9K	134	125	16K	21K

3.3.5 Automatic Speech Recognizer

I use ASPiRE [69], a standard standard benchmark for ASR systems at the time of this work, available in Kaldi's [63] model suite and trained on Fisher [56], to create automatic transcriptions for Switchboard. This system has a word error rate of 21% on the development and 24% on the test set of Switchboard.

3.3.6 Performance Evaluation

Prediction quality is evaluated using macro F1 scores. Some care must be taken when comparing the performance of predictions based on human-created transcriptions and those based on ASR transcriptions, since the ASR may insert or delete tokens from the reference transcript. Following [67], if a deleted word is assigned punctuation *and* the previous word was also assigned the same punctuation, then it is considered correct. Equations 3.1 and 3.2 provide the formulae for computing precision and recall of question mark ("?") as an example; the metrics for other punctuation types are computed in the same fashion.

$$P = \frac{|TP(?)|}{|? \text{ in ASR}|} \tag{3.1}$$

$$R = \frac{|TP(?)|}{|? \text{ in reference}|} \tag{3.2}$$

I report performance of models run on subsets of the input features, including only the word embeddings and the word embeddings plus the categorical pause feature. The RNN I base all models on is a uni-directional GRU [70], and parameters were learned using an Adam optimizer [71] with initial learning rate 0.0001, halving when the performance on the development set does not improve every 3 epochs. The best performance occurred with 12-dimensional pause embeddings; the CNN has $N = 32$ sets of filters of widths $[5, 10, 25, 50]$, i.e. $m = 4$, totaling 128 filters.

3.3.7 4-class prediction

First, I compare the performance of 4-class models given various sets of input features. Table 3.3 provides the test F1 scores for 4-class prediction. F0 and energy features give a boost over the pause in the macro scores for hand transcripts, due to improved prediction of "incomplete." Commas are predicted least accurately. Performance degrades for the ASR transcripts, with a large regression for predictions of "incomplete." Inclusion of prosodic features impacted performance more with ASR transcripts, and was most useful for recognizing

instance of “incomplete.”

Table 3.3: F1 scores for prediction of 4-class punctuation types on the full test set, using different features with hand vs. ASR transcripts. From [42].

Hand	C,	P.	Inc-	Q?	Macro
text only	.60	.81	.80	.73	.736
pause only	.60	.82	.80	.74	.739
all features	.60	.82	.82	.73	.744
ASR	C,	P.	Inc-	Q?	Macro
text only	.53	.77	.49	.60	.597
pause only	.53	.77	.52	.62	.612
all features	.53	.78	.53	.62	.615

3.3.8 Interruption Point Prediction

Table 3.4: F1 scores for prediction of 5-class punctuation types on the IP test set, using different features with hand vs. ASR transcripts. From [42].

Hand	C,	P.	Inc-	Q?	IP+	Macro
text only	.63	.81	.80	.79	.77	.761
pause only	.63	.82	.81	.78	.76	.759
all features	.65	.82	.82	.80	.78	.773
ASR	C,	P.	Inc-	Q?	IP+	Macro
text only	.56	.76	.49	.65	.54	.600
pause only	.56	.77	.47	.63	.54	.595
all features	.57	.77	.52	.65	.55	.611

Table 3.4 gives test results for the 5-class punctuation task. The prediction of interruption points is sensitive to transcription type. For human transcripts, the interruption point prediction has relatively high precision and thus does not effect the prediction quality of other punctuation types. For the ASR transcripts, commas and question marks are predicted more faithfully when the interruption point is included. In contrast, when interruption points are

included in systems with acoustic features, F1 was lower compared to the 4-class model except for comma prediction. The categorical pause feature is not as helpful for predicting from the 5-class set compared to the gains for the 4-class prediction.

3.4 Conclusion

With respect to the utility of prosody in predicting punctuation, I found that using any prosodic information either modestly improved or had no effect for the majority of predictions. Using all prosodic features tended to be more successful than using pause features alone. Both simple and learned acoustic cues increase the computational load of the models, however they do not degrade performance. Concerning the ease with which different punctuation can be predicted, I found that periods and question marks were easier to predict than commas. The relative accuracy of interruption point and incomplete predictions were sensitive to transcription type. The learned acoustic-prosodic features were more helpful in predicting incomplete boundaries, especially when applied to ASR transcripts. In experiments where interruption points were explicitly modeled, prediction quality of the other punctuation types improved. As to whether transcription quality influenced prediction performance, I found automatically-generated transcripts induce noise that enhances the difficulty of detecting punctuation. These results show that explicit acoustic-prosodic features have potential for modelling prosodically-salient NLP tasks, although more work is needed to understand the magnitude of potential benefits.

Chapter 4

MODELING ACOUSTIC-PROSODIC ENTRAINMENT

Computational representations of prosodic information are important for a variety of speech technology applications. However, the acoustic manifestations of prosodic information are influenced by many interacting factors that make it difficult to predict. This is especially true in spontaneous conversation, where speakers will adapt their speaking styles based on conversational context and the behavior of their interlocutors. In this work, I present a method for predicting salient prosodic features of spontaneous speech, using the prosodic information from a speaker's and interlocutor's speech history. Motivated by studies of entrainment, I use the models' ability to accurately predict speakers pitch in the Switchboard corpus as a method for exploring interdependence of speaker prosody. The models showed no significant effects for the specific prosody features I explored.

4.1 Experimental Goals

One well-studied phenomena in dyadic conversation is entrainment. Also called accommodation, this is the process of speakers converging on a common speech style, for example using the same words when synonyms are available, or adopting similar phrasal structures [72]. However, the complexity of human communication, including complex interactions between multiple sociolinguistic, environmental, and other contextual factors makes the degree of entrainment speakers will engage in difficult to predict. Of interest to this work is the degree to which the prosodic entrainment between speakers can be quantified and predicted using one traditional acoustic measure, speaker pitch.

In this chapter, I propose a method for predicting salient pitch features of spontaneous speech by leveraging the prosodic information extracted from both a speaker's own speech

history and that of their interlocutor. My hypothesis is that by explicitly encoding aspects of conversational dynamics, such as turn taking behaviors, and speakers' differences in pitch, I will be able to leverage natural prosodic entrainment in the data to more accurately predict prosodic features of future speech.

I use an LSTM architecture to learn prosodic properties extracted from speech turns of a single speaker, augmented with salient information about their interlocutor's previous turns. This system is used to predict pitch features about the next turn in the time series. I present the performance of this model on the Switchboard corpus [58]. I explore how providing different combinations of prosodic features to the models impacts their performance. I use fine-tuning of models on data subsets, specifically considering high versus low entrainment, speaker gender, and inter-speaker gender concord to assess predictability based on those factors.

Of principal interest to this study is the success with which aggregate acoustic measures of prosody across an utterance can be modeled using a neural architecture. This comprises two research questions:

1. To what extent does the pitch prosody an interlocutor's speech improve the prediction of the pitch prosody of future speech?

2. Does speaker entrainment, disentrainment, or interlocutor gender make a speaker's pitch more predictable?

4.2 Entrainment

Conversational participants adjust their communicative behaviors based on the behavior of their interlocutors. This behavior mentioned is often assumed to be unconscious [73]. This happens in many modalities. For example, [74] found that nonverbal synchrony between speakers in conversational dyads, such as mirroring facial muscle activation or posture, was predictive of conversational rapport. It has also been suggested that nonverbal synchrony

is a way for listeners to activate similar neural pathways as their interlocutors to facilitate fostering of empathetic conversation [75].

Speakers alter their speech to sound more like the speech of their interlocutor(s). This type of behavior has been described as a communicative device to establish *conceptual pacts*, the norms that inform meaning given the context of the conversation [76]. It is also associated with speakers' sociolinguistic indexing as a member of an in-group [72]. The opposite behavior, disentrainment, where speakers alter their speech to be less like their interlocutors, is also observed in spontaneous conversation. Disentrainment or lower levels of entrainment on some features has been associated with lower conversational success [73], although some traditional heuristics for disentrainment may be associated with high speaker engagement [77]. A lack of appropriate baseline speech prosody and prosodic entrainment can be associated with conversation difficulty or communication disorder and can be perceived as unnatural by interlocutors [78].

The work of Sarah and Rivka Levitan and colleagues [e.g., 73, 79, 80], has sought to computationally model acoustic-prosodic entrainment. In [73, 79], the Columbia Games Corpus [81], which consists of dyads collaboratively completing computer games, is used to compare the similarity of speech between and within speakers. Specifically, the researchers measured combinations of pitch, intensity, jitter, shimmer, noise-to-harmonics ratio, and speaking rate of speakers across sessions. ([79] also made comparisons at the turn-level.) The experiment that is most applicable to my use case is found in [82], where entrainment is computed at the session level between interlocutors. The negative sum of absolute differences of the means of each acoustic measure between speakers is computed, and conversations are labeled as entraining if the difference between speakers in the session is greater than the difference compared to speakers in different sessions. This general method of computing entrainment by comparing acoustic features between different pairs of speakers has been shown to be correlated with social behaviors and task success [73], however the non-goal-oriented nature of Switchboard Experiments means that there is less consistency of style and content across conversations than in the Columbia Games corpus, making between-

conversation comparisons fraught. Using a measure of within-speaker proximity between recordings of the same speaker with different interlocutors, as in [79], removes the need to compare across very different kinds of conversations, provided that speakers can be compared between different key points in the same conversation. I describe such a method in 4.3.5.

4.3 Experimental Design

4.3.1 Dataset

For training the prosody model, I use the Mississippi State version of transcriptions for the Switchboard Corpus [58] which contains recordings of spontaneous speech between unfamiliar interlocutors, where speakers were given a prompt and asked to hold a conversation using that prompt for about ten minutes. The gender balance of the dataset is given in Table 2.2.

The data were each divided into train, validation/development, and test sets following the conventions of [83]: conversations with ids matching 'sw0[23].+' were used for training, 'sw04[5-9].+' for validation/development, and 'sw04[01].+' for testing. Other conversations were unused.

4.3.2 Feature Extraction

For each conversation in the Switchboard dataset, I separate the turns by speaker. Thus, the pitch features for each speaker are predicted separately. However, that prediction is based on some information about the interlocutor's previous speech as well.

Signal features were extracted from the dialogues using the Kaldi Speech Toolkit, using standard configuration for this dataset (sampling frequency at 8000 Hz, energy not used in MFCC calculation) [63]. The log of the non-normalized pitch (in Hz, base e) is extracted from single-channel audio for each speaker. Each feature is computed at the frame level, where overlapping frames are 10 ms apart with 25 ms duration. A number of experimental environments were considered. Preliminary experiments conducted using the methodology described in this section, speech from the Fisher dataset [56], and a large set of acoustic

features showed negligible benefits to an expansive acoustic feature space. Thus, the presented methodology includes a minimal set of acoustic features which showed most promise for model training. From the raw signal features and transcriptions, I extract 3 measures of speaker pitch, 4 measures of interlocutor pitch, and 4 timing measures. Note that all mentions of pitch refer to the log pitch provided by Kaldi.

Three features are used to represent the overall behavior of pitch throughout an utterance: the mean pitch across all relevant frames, the mean pitch across the first 200ms of the utterance, and the range of unpadded median-filtered permutation of all pitch measures for the utterance, with a window length of 7 (see Equation 4.1).

$$\begin{aligned} \text{range} &= \max(W_i) - \min(W_i), \text{where} \\ W_i &= [\text{median}(\text{frames}[i : i + 7]) \text{for } i \text{ from 0 to (number frames - 7)}] \end{aligned} \tag{4.1}$$

A primary hypothesis of this work is that in spontaneous conversation, prosodic accommodation, i.e., entrainment, will influence prosody production. To allow the model to learn the extent of entrainment in the data, I include four features about the interlocutor: the mean pitch of their most recent complete utterance, the mean pitch of the final 200 ms of their most recent complete utterance, and the mean pitch of the first and final 200 ms of their most recent utterance (complete or incomplete). It is suspected that the features computed from end 200 ms of previous interlocutions will be especially useful in predicting what occurs at the beginning of the next utterance of the speaker.

To characterize local turn-taking dynamics, I include four timing features summarized in Table 4.1. The timestamps used to compute these features were taken from the word-level timing information available in the Mississippi State version of transcriptions.

4.3.3 Model Architecture

An LSTM was implemented for each configuration under investigation using the `nn` module of `pytorch` [84], having hidden dimension 256, a learning rate 0.001, a dropout rate of 0.1,

Table 4.1: Timing features for modelling utterance pitch

Timing Feature	Description
duration	Total speaking time, including inter-word pauses, of speaker's most recent turn.
duration since last of speaker	The difference in time between the start time of speaker's most recent turn and end time of their interlocutor's most recent turn.
interruption	A binary value denoting whether the speaker's most recent turn was an interruption, i.e. whether it began in the middle of the interlocutor's previous turn.
speaking rate	Speaking rate of the speaker's most recent turn, computed as number of words per second. No post-processing is done to change the definition of "word" as it is in the Mississippi transcriptions. Contractions, e.g., "wasn't," are considered a single word.

batch size of 64, and a maximum of 70 epochs with early stopping on a patience of 3 epochs and threshold 0.0. The models each have two LSTM layers and a single linear output layer. The development set was used to create a metric for early stopping.

Conversations were truncated to a maximum sequence length of 206 utterances (within 2 standard deviations of mean sequence length for the Fisher dataset), and then shorter conversations padded to this length. Before training, the data is normalized using a MinMax scaler transformation (as implemented in MinMaxScaler module of `sklearn` [85]) to values between -1 and 1, fit on the set of training data.

The base models include as input: the 3 pitch features and 4 timing features associated with the target speaker, plus the features associated with the interlocutor. Separate models

are trained for predicting the mean pitch of the first 200ms and the pitch range as defined in Equation 4.1. Contrasting models are trained with no interlocutor features.

Model performance is recorded as root mean square error between the ground truth pitch measure and predicted value.

4.3.4 Fine-tuning conditions

To investigate whether characteristics of certain types of conversations are more informative of pitch trajectories, I follow a split fine-tuning procedure based on speaker characteristics. Along the dimensions of entrainment, speaker gender, and inter-speaker gender concord (defined below), I split the training data into two equally-sized subsets. I fine-tune each of the base models on each of the subsets separately, and report performance on the complete test set.

Fine-tuning dimensions:

- **Entrainment**: for each speaker, I compute an entrainment score using the method described in Section 4.3.5. Speakers are sorted and evenly split by their entrainment value, creating *low* and *high* entrainment subsets.

- **Speaker gender**: One subset contains all *female* speakers, the other subset contains all *male* speakers.

- **Inter-speaker gender concord**: Data is subset into four categories based on both the gender of the speaker and the gender of their interlocutor. I abbreviate these subsets as female speaker, female interlocutor (*F-F*); female speaker, male interlocutor (*F-M*); male speaker, female interlocutor (*M-F*); and male speaker, male interlocutor (*M-M*).

To separate the effect of the criteria of the data subsets from the reduced training size, I also fine-tune on subsets of the data of the same size as those in the three fine-tuning conditions, but with speakers randomly assigned to the subsets.

Models that are fine-tuned on splits according to entrainment are initially trained on a subset of the training data (referred to as *null entrainment excluded*) that does not include any speakers for whom an entrainment score was not available (e.g., due to pitch tracking errors).

4.3.5 Measuring Entrainment

I base the calculation of prosodic entrainment on the measurement used in [73]. For each speaker in the dataset, I divide the number of their turns in three, and compare the voice quality features of the first third and the last third of turns in the conversation. The motivation for this comparison is that I expect that in the initial phase of conversation, strangers are learning the speaking style of their interlocutor and are less likely to be influenced by their speech. At the end of the ≈10 minute conversation, the speakers are most likely to be familiar with each others' speaking styles, and adaptation may make their speech during this period shift away from their style at the beginning of the conversation. I assume that changes in voice quality between the beginning and end of conversation are due to prosodic entrainment or disentrainment.

For each turn in either the first or third sections of conversation, I extract the following features using Praat [86] with default settings:

- Maximum f0
- Average intensity
- Maximum intensity
- Jitter, as Relative Average Perturbation (RAP)
- Shimmer, as "the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude" [86]

- Speaking rate, as syllables per second. Pronunciations are taken from the CMU Pronouncing Dictionary[1]. Out-of-vocabulary words do not contribute to speaking rate.

Unlike [73], I do not include Noise-to-harmonics ratio (NHR) as a voice quality measure. This feature is especially sensitive to errors in pitch tracking, which happens often with Switchboard audio due to signal quality. Since many conversations would have been excluded from analysis even though NHR was the only voice quality measure not able to be computed, I choose instead to deviate from [73]'s definition further, with the hope that similar information will still be encoded in the other voice quality features which can be extracted from the majority of conversations.

A single entrainment score is calculated for each speaker by taking the absolute difference of the average of each feature between the first and final third partitions of speech. These average differences are normalized by the averages of the first partition, and summed to yield a single measure. The sum is negated to make greater differences (i.e. non-entrainment or disentrainment) have the lower value. This procedure is formalized in Equation 4.2, where P_n is a matrix of voice quality measure computed on each turn in partition n, I is the set of voice quality features, J is the number of turns in the final third partition, and K is the number of turns in the initial third partition. While J and K should be approximately equal, they may differ between 0 and 2 depending on the total number of turns in the conversation.

$$-\sum_{i=1}^{I}\left|\frac{\frac{\sum_{j=1}^{J}\left(P_{3_{i,j}}\right)}{J}-\frac{\sum_{k=1}^{K}P_{1_{i,k}}}{K}}{\frac{\sum_{k=1}^{K}P_{1_{i,k}}}{K}}\right| \tag{4.2}$$

Because this score encodes information about speaker change over time, I hypothesize it could be a proxy for predictability of pitch prosody features.

[1] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

Table 4.2: Model evaluations using minimal prosody features, timing information, and interlocutor features to predict pitch at turn onset and median-filtered pitch range. Cells with multiple values have been fine-tuned on more than one training subset, and then evaluated on all of the data from the test set.

Expt.	Model	Predicting mean f0, first 200ms	Predicting range
1	base	0.24	0.46
2	base without interlocutor features	0.24	0.46
3	base, null entrainment excluded	0.24	0.45
4	fine-tune on arbitrary 2-way split	1: 0.24; 2: 0.24	1: 0.45; 2: 0.45
5	fine-tune on split by entrainment	low: 0.24; high: 0.25	low: 0.45; high: 0.45
6	fine-tune on split by speaker gender	F: 0.28; M: 0.26	F: 0.45; M 0.46
7	fine-tune on arbitrary 4-way split	1: 0.24; 2: 0.24; 3: 0.24; 4: 0.24	1: 0.45; 2: 0.45; 3: 0.45; 4: 0.46
8	fine-tune on split by gender concord	F-F: 0.27; F-M: 0.28; M-F 0.26; M-M: 0.27	F-F: 0.46; F-M: 0.46; M-F: 0.45; M-M: 0.45

4.4 Prediction Performance of Prosodic Features

Table 4.2 provides the RMSE for all configurations on the held-out test data. When all of the speakers are pooled for training, we find no difference between the base model (Expt. 1) and the contrasting version without interlocutor features. Thus, these interlocutor features are not useful for pitch prediction. This is not expected for the subset of speakers with high entrainment, so I choose to compare that subset of speakers to the low entrainment speakers.

Since training on a subset of speakers could give degraded performance, I fine-tune specialized models on an initial model trained on all the data. For the entrainment case, there are some speakers with entrainment measurement errors, so those were excluded from the initial model. Experiment (3) shows that this had no effect on performance relative to the initial model. As another control for training set size, I also fine-tuned models on random splits of the data. There was no significant difference in performance for any model (Expts. 4, 5), leading to the conclusion that there is no observable effect of entrainment, either because of the specific input features chosen or the target pitch prediction features.

I analogously test the predictability of pitch features relative to gender and gender concordance between interlocutors. Again, I use the same random controls (Expts. 4, 7). When predicting pitch range, there is no difference among the different conditions. For predicting the mean pitch onset, I observe an increase in error for both speaker gender and gender concord compared to the random control. This suggests that models fine-tuned on gendered data may be over-training.

4.5 Discussion

4.5.1 Distribution of Entrainment

Figure 4.1 illustrates the distribution of entrainment values in this dataset. The x-axis represents the entrainment values, while the y-axis corresponds to the frequency of occurrences. Values ranged from -112.62 to -0.22. From the figure, I observe a skewed distribution where speakers generally trend toward increased consistency in voice quality features between par-

titions of the conversation. Outliers are extreme and negative; more investigation is needed to determine what is the primary cause of this phenomenon. They may be be due to pitch tracking errors, use of out-of-vocabulary language, or more complicated patterns of speaker accommodation than can be modeled with aggregate voice quality measures.

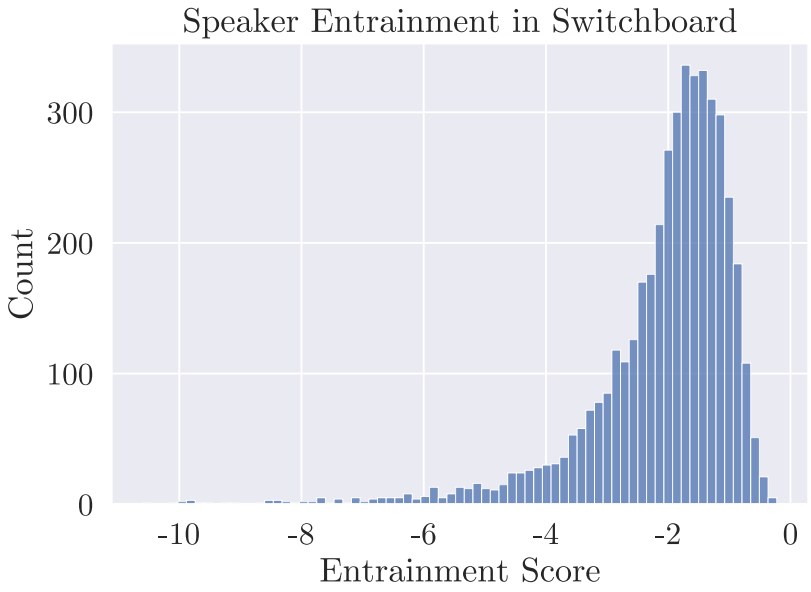

Figure 4.1: Distribution of entrainment scores in the dataset.

4.5.2 Subset Analysis: Speaker Gender

Model performance showed no differences when trained and evaluated on gendered subsets of data. The proxy entrainment measure also does not appear to show any difference in distribution between Male and Female speakers. This is shown in Figure 4.2. I performed a 1-way Welch's ANOVA on the distributions of entrainment scores relative to speaker gender, and found no significant effect.

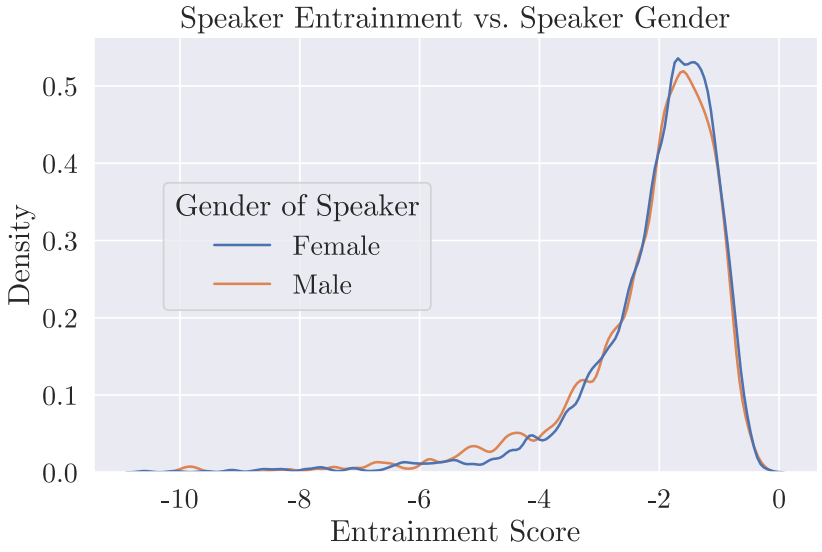

Figure 4.2: Densities of entrainment scores relative to speaker gender estimated using kernel density. Densities are normed individually to remove any differences contributed by sample size. Outliers with z-scores > 3 are not pictured.

4.5.3 Subset Analysis: Gender Concordance

While inter-speaker gender concord did not improve pitch predictability, it has been shown previously to impact the extent to which speakers entrain to one another [73]. Figure 4.3 shows the distribution of entrainment scores for speakers in the Switchboard. Densities are normalized individually to remove any differences contributed by sample size. Outliers with z-scores > 3 are not pictured. All groups showed similar unimodal distributions with long left tails. However, there does seem to be a slightly different shape for the Female-Female cohort, which has a more peaked distribution than the others. This intuitively makes sense, as previous work on entrainment has found that women, especially when speaking to other women, can be an edge case for prosodic entrainment [73]. A 1-way Welch's ANOVA was performed on the distributions of entrainment scores relative to inter-speaker gender concord, which returned a strong significant effect ($F = 10.94757$, $p < 0.0001$).

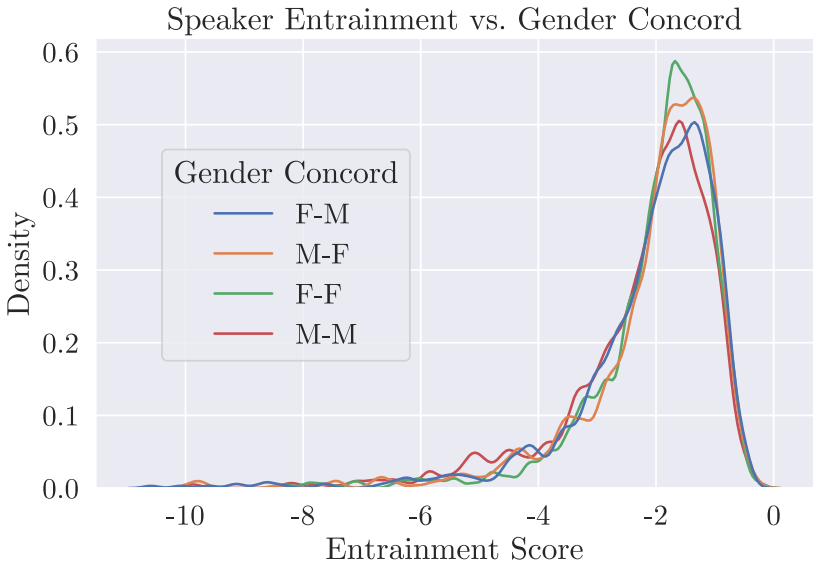

Figure 4.3: Densities of entrainment scores relative inter-speaker gender concord estimated using kernel density. Densities are normed individually to remove any differences contributed by sample size. Outliers with z-scores > 3 are not pictured.

4.6 Conclusion

This study has presented a method of representing speaker prosody using information about both the speaker and their interlocutor, for predicting salient prosodic features in spontaneous speech. Leveraging a modified quantification of acoustic-prosodic entrainment computed on conversational data, and incorporating both a speaker's own speech context and that of their interlocutor, the proposed approach explored multiple plausible manifestations of entraining behavior to improve the ability to predict the range and onset pitch behaviors a speaker will surface given their conversational context.

The results obtained using the LSTM architecture on the Switchboard corpus yielded no findings. However, other prosodic features might lead to different conclusions.

The incorporation of entrainment scores based on quantitative measures of voice quality as a criteria in the fine-tuning pipeline is novel, and further investigation is need to see whether

the measure's relationship with inter-speaker gender will make it useful for the prediction task in some other form.

The outcomes of this study provide a foundation for future research on prosodic modeling. Further investigations should revisit the model choice and architecture, especially if the entrainment score can be used in a more direct way as part of the pipeline. In addition to looking for speech sources with more modern language, the voice quality extraction process would benefit from a more controlled recording environment to reduce fatal pitch tracking errors. I hypothesize that with practical adjustments to the machine learning modules and with careful dataset selection, the features described in this work can meaningfully inform understanding of prosody and its trajectory in spontaneous dyadic conversation.

Chapter 5

MODELING INFORMATION TRANSFER THROUGH SPEECH TIMING PATTERNS

5.1 Conversation Analysis

In language activities with more than one participant, speakers must navigate the shared speech timing between all interlocutors, deciding who should speak next and when, when it is acceptable to interrupt the current speaker, and how to interpret deviations from the expected timing of conversation. Thus conversational timing includes the duration of speech, any pauses a speaker makes, and the pauses or overlaps in speech between speakers. Quantitative analysis of timing interplay between speakers often follows the Jeffersonian Conversation Analysis style described by [88]. Central to this description system is the notion of turn-taking, where conversations are divided temporally by minimally overlapping turns [88]. A single speaker *should* hold the floor at any given point in time. This *floor* is sometimes somewhat circularly defined as the ability to have a turn in the conversation [89]; there is also a body of researchers who choose to eschew strict definitions of floor in favor of functional descriptions of speaker and listener behaviors in turns [90]. Nonetheless, the timing of conversation in the Conversation Analysis framework is relative to the units that comprise and exclude speech from speaker turns. In the canonical methodology of this style of analysis, the timing of turn-taking is described by expert listening and stopwatch-based counting of timing in tenths of a second [88, 91, p. 81], and in fact the speech timing of any

---

This chapter is revised from "Investigating the Influence of Stance-Taking on Conversational Timing of Task-Oriented Speech," which appeared in INTERSPEECH 2024 [87]. My specific contributions were the research question, experiment design and implementation, and equal contributions to analysis.

analysis in this tradition is not advised to be compared between practitioners for this reason [92]. However some more recent efforts have included finer-grained or automated extraction of timing features [e.g., 93, 94]. These methods allow conversation analysis to be applied to modern, large corpora without the massive expense of annotating turns by hand. Thanks to this advantage, I will apply conversation analysis to a large corpora of dyadic speech, and show one use case where these methods can aid in understanding factors that change the course of speech timing.

5.2 Stance

As discussed in Chapter 2, stance-taking may be motivated by a desire for "the expression of internal psychological states of an individual speaker" [95], to agree with their interlocutor [96], built rapport [49], or other subjective and inter-subjective goals [97]. Previous work has shown that the phonetic and prosodic properties of speech are affected by stance [49, 98]. What's more, stance-taking has been shown to significantly affect speech spurt duration in dyadic speech [49]. However, the timing of conversation is much more than the timing of a single speaker. The Jeffersonian analysis method for discourse can be utilized to make sense of the dual timing between speakers and interlocutors.

5.3 Research Questions

In this chapter, I investigate the influence of stance-taking behaviors on the timing of turn-taking in dyadic task-oriented conversation. I specifically ask whether the strength and polarity of stance-taking affects duration measures related to the taking of turns in collaborative task-oriented conversation, and answer the following questions:

1. What distributional variation exists in the conversational timing of speech when stance-taking is varied in strength and polarity?

2. Is there an impact of speaker gender on the timing of stance-marked speech?

5.4 Discourse and Stance Behaviors

Hyper-articulation has been shown to be motivated by stance-taking [99], and other prosodic qualities of speech may also be affected. For instance, Valerie Freeman[1] found significant differences in both utterance duration and speaking rate in different stance conditions in the ATAROS dataset, also finding that these differences were confounded by speaker gender. What's more, it has been shown that there is a relationship between the coordination of linguistic style and speaker stance [100].

5.5 Experimental Design: Stance as modifier of turn-taking

5.5.1 Stance Annotation

For this study, I consider the third task and sixth tasks from the ATAROS dataset (see 2.5.1 for review). I exclude from the original dataset instances where annotators marked the stance of a spurt as ambiguous, as well as any conversations from the inventory and budget tasks which did not have stance annotations. Given these omissions, this analysis is based on 116 conversations (57 dyads with both inventory and budget recordings, 2 dyads with only one task type).

5.5.2 Speech Timing Measures

I consider the following common measures for turn-taking:

- **Floor-transfer Offset (FTO)** - the duration of time during a floor transfer between the end of the previous speaker's turn and the beginning of the next speaker's turn. FTO can be negative, representing interrupted speech.

- **Inter-pausal Unit (IPU) Duration** - the inter-pausal unit is a unit of speech from a single speaker with no internal pauses greater than a threshold. I use a threshold of 180 ms, consistent with previous work on conversation dynamics [e.g., 93, 101].

[1] V. Freeman 2014, Presentation at the University of Illinois Linguistics Seminar, Urbana-Champaign, IL

- **Pause Duration** - the duration of pauses between speech units, where no floor transfer occurs.

- **Turn Duration** - the time elapsed from the start of the first, to the end of the last of consecutive IPUs from a single speaker. When turns consist of a single IPU, the turn duration is equal to the IPU duration.

For each conversation, I create an interleaved transcript by using the word-level timings to compose a set of IPUs for each speaker and concatenating all consecutive words that are separated by no more than 180 ms. This is consistent with other efforts to automatically annotate turn-taking behaviors [e.g., 93]. IPUs are composed into turns by comparing both their timing and content. Consecutive IPUs are combined into a turn, provided that no speech from the interlocutor occurs in the same time. If there is overlapping speech, and neither IPU is determined to be a backchannel (see 5.5.2), then the floor is transferred to the speaker whose IPU ends latest. If there is no intervening speech from the interlocutor, the speaker who has the floor maintains it even if they are not actively speaking. A toy conversation provided as reference in Figure 5.1.

To map the spurt-level stance annotations to the definitions of IPU and turn from the Jeffersonian framework, I first consider all speech spurts whose timings overlap with IPU- or turn-level timings. For all such spurts, I label the IPU or turn with the strongest stance annotation of any matched spurt, and the non-neutral polarity where available.[2] For instance, an IPU composed of spurts with annotations "Moderate Neutral" and "Weak Positive" will be given the label "Moderate Positive." In this way, speech timing and stance features can be compared at the same level of granularity. To explore a possible effect of gender on the timing of turn taking [102], I associate the gender of each speaker to the timing measures as well.

[2]There were no instances of a turn or IPU having both positive and negative polarity.

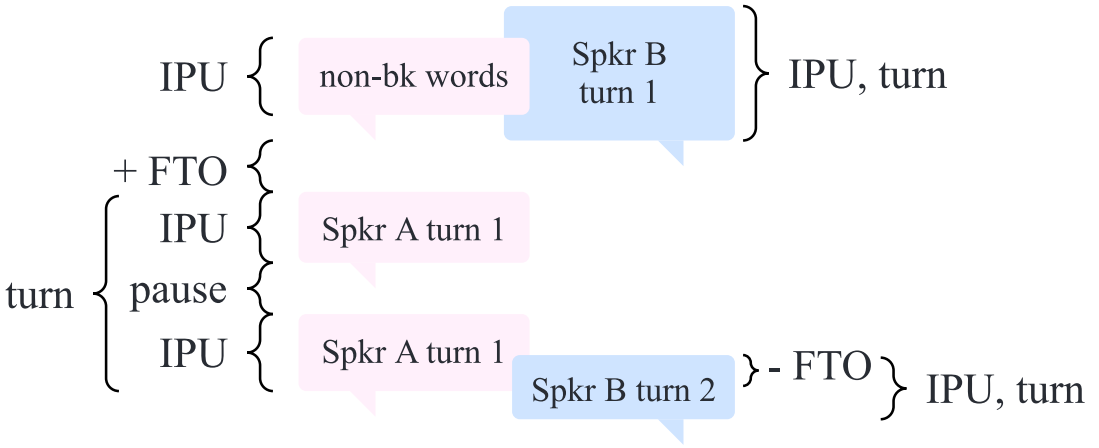

Figure 5.1: A toy conversation, annotated with FTO, IPU, pause and turn annotations. In this conversation, there are 5 IPUs, 3 turns, and one pause. This illustrates the novel condition where speech matches the timing of a backchannel but not the text criteria, and thus acts as a turn-less IPU.

Backchannel Filtering

Backchannels are spurts of speech which occur while a different speaker holds the floor, and are used to signal to the floor-holder that the backchanneler is listening and supporting their continued speech [103]. Backchannels do not represent a desire to take the floor [104]; therefore, they are excluded from turn-based measures of speech timing. While often backchannels are described solely by their occurrence entirely within the duration of the interlocutor's utterance [e.g., 93], this simple definition will erroneously capture speech acts such as failed attempts at taking the floor. To more conservatively filter backchannels from the corpus, I first compose a set of backchannel n-grams[3] by taking the top 15 most common phrases from all dialog acts labeled as backchannels in the Switchboard Dialog Act Corpus [105]. I consider speech a backchannel if 1) the text is composed of a combination of the 15 backchannel n-grams, and 2) the timing is either entirely within or interrupts the interlocutor's speech. When backchannels are observed in the ATAROS data, they are ignored for the purposes of

[3]In order of frequency: ['okay', 'oh', 'I see', 'uh-huh', 'yeah', 'all right', 'uh', 'right', 'um', 'huh', 'yes', 'oh okay', 'no', 'ooh', 'well']

computing speech timing measures. Table 5.1 shows the counts of speech timing measures, grouped by stance type.

Table 5.1: Counts of timing measures in the dataset, grouped by stance type. The correspondence for the timing measure labels is: Neu. : Neutral; +: Positive; -: Negative; 0: None; 1: Weak; 2: Moderate; 3: Strong.

	Strength				Polarity			Total
Meas.	0	1	2	3	Neu.	-	+	
FTO	3,262	6,752	3,256	84	8,342	734	4,278	13,354
IPU	5,401	11,117	6,016	170	15,847	1,366	5,491	22,704
Pause	2,116	4,340	2,756	86	7,453	632	1,213	9,298
Turn	2,824	6,017	2,933	74	7,417	651	3780	11,848

5.5.3 Statistical Comparison

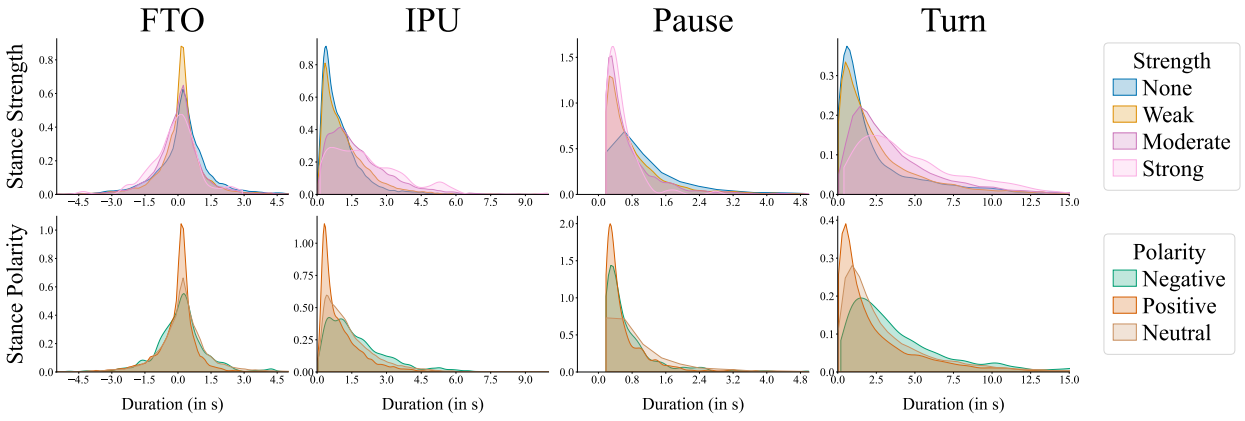

Figure 5.2: Distribution of Duration of Speech Timing Phenomena. From left to right: Floor Transfer Offset (FTO), Inter-Pausal Unit (IPU) duration, pause duration, and turn duration. The top row shows stance strength and the bottom shows stance polarity. For clarity of presentation, the duration range of these plots is clipped to exclude long right tails with very low density. All measures except FTO comprise only positive values.

To evaluate the importance of stance strength and polarity on speech timing, I conduct Linear Mixed Effects Regressions using `lme4` in R [106] where stance strength, polarity, and speaker gender are taken as fixed effects of the dependent timing variables. Speaker is treated as a random effect.

5.6 Results

Figure 5.2 presents the distribution of speech timing measures, grouped by stance strength and polarity. Across strength and polarity conditions the shapes of distributions for each timing measure follow generally the same shape, with IPU, turn and pause having right-tailed unimodal distributions with slightly positive modes. The distributions for FTO are also unimodal and fairly symmetric, with modes near zero.

There is a general trend where the distribution of measures for positive polarity phenomena have lower variance compared to the other polarity conditions. Similarly for FTO, IPU and turn, the strong strength conditions have higher variance compared to the other strength conditions (with the exception of the no stance condition for FTO). The opposite effect is seen for pause duration: The stronger stance conditions show lower variance than the weaker conditions. There also seems to be a general paired relationship for stance strength, where the distributions of Moderate and Strong stance look much more like one another than either the None or Weak conditions, and vice versa.

Table 5.2 shows the output of the statistical analysis. For each condition, I run a Linear Mixed Effects Regression with the formula `{FTO|IPU|pause|turn} ~ stance strength + stance polarity + gender + (1|speaker)`, where stance strength, polarity, and speaker gender are treated as fixed effects, and speaker is considered a random effect. The intercept conditions were None (for strength), Neutral (for polarity), and male (for speaker gender). Data points greater than 2 standard deviations from the mean are discarded.

The majority of the stance conditions tested showed significant effects across the speech timing measures. In contrast, there were no significant effects for speaker gender. Negative stance polarity was significant ($p < 0.005$) with respect to pause duration, but none of the

Table 5.2: Statistical results of comparison of stance conditions and duration measures. Each test was based on the normalized duration measures, and controlled for speaker gender. Significance is labeled with codes *** for $p < 0.001$ and ** for $p < 0.01$, and values abbreviated to "< 0.001" when the value was less than 0.0001.

Meas.	Effect	t-value	Pr(>t)	Sig.
FTO	Intercept	7.1E+01	6.61	<0.001 ***
	+	1.3E+04	-8.43	<0.001 ***
	-	1.3E+04	0.38	0.705826
	1	1.3E+04	-1.51	0.132191
	2	1.3E+04	-6.08	<0.001 ***
	3	1.3E+04	-3.60	<0.001 ***
	Female	5.6E+01	0.005	0.995830
IPU	Intercept	7.2E+01	19.38	<0.001 ***
	+	1.1E+04	-12.39	<0.001 ***
	-	1.1E+04	0.00	0.999
	1	1.1E+04	8.41	<0.001 ***
	2	1.1E+04	18.40	<0.001 ***
	3	1.1E+04	7.77	<0.001 ***
	Female	5.8E+01	1.396	0.168
Pause	Intercept	9.0E+01	34.39	<0.001 ***
	+	9.0E+03	-8.77	<0.001 ***
	-	9.0E+03	-2.78	0.00553 **
	1	9.1E+03	-9.57	<0.001 ***
	2	9.0E+03	-12.45	<0.001 ***
	3	9.1E+03	-6.27	<0.001 ***
	Female	6.0E+01	0.154	0.87837
Turn	Intercept	7.2E+01	19.38	<0.001 ***
	+	1.1E+04	-12.39	<0.001 ***
	-	1.1E+04	0.00	0.999
	1	1.1E+04	8.41	<0.001 ***
	2	1.1E+04	18.40	<0.001 ***
	3	1.1E+04	7.77	<0.001 ***
	Female	5.7E+01	0.324	0.747

other measures. Weak stance showed a significant effect for all speech timing measures except FTO. Significant effects for all stance types on FTO and pause (excluding intercept condition) resulted in shorter durations. For turn and IPU duration, significant stance strength conditions were associated with longer durations, while significant polarity conditions were associated with shorter durations.

5.7 Discussion

These results suggests statistically significant relationships between the type of categorical stance described in the ATAROS dataset, and the timing of turn-taking associated with this type of dyadic, task-oriented speech.

A pattern emerged between the no stance and weak stance conditions, with speech timing measures between these two conditions having similar distributions. Likewise, moderate and strong stances demonstrated coordinated timing patterns. This observation suggests that a certain level of strong stance may be essential before substantial alterations occur in speakers' production models, although further study is needed to know whether this theory has bearing on the cognitive processes of speech production and perception.

5.7.1 Negative Polarity

The regressions yielded compelling evidence supporting the hypothesis that speaker stance may significantly influence the coordination of turns in conversation. With most stance conditions showing significant effects for all speech timing measures, the lack of significance for negative polarity with respect to pause and turn may be due to a lack of power (only 5.5 - 6.8% of the instances of each timing measure had negative polarity). More research is needed to determine whether this difference suggests a similarity between neutral and negative polarity, or whether the lack of significance is due to other factors (e.g. annotator bias, etc.).

5.7.2 Speaker Gender

The lack of significance for the interaction of speaker gender and speech timing measures is unexpected given previous findings. In a gender-balanced 20-dyad sample of the ATAROS dataset, [49, p. 26] found that while there were not overall differences in the spurt duration and speaking rates between male and female participants, women were found to have higher relative speaking rates in the inventory (low stance eliciting) task. From a smaller 12-dyad sample, they found that spurt length differed more for men between the inventory and budget (high stance eliciting) tasks [49, pp. 24-25]. While it is possible there are interactional sociolinguistic factors that may affect gendered manifestations of stance, in follow-up experiments where both speaker and interlocutor gender were accounted for I similarly found no significant effects.

5.7.3 Long Negative Floor Transfer Offsets

In this paper I present a method for automatically labeling and interleaving turns in conversation, based on the standard annotation method for this type of turn-taking analysis and augmented with a lexical check for backchannels.

The expectation from the field of Conversation Analysis is that the majority of floor transfer offset times should be positive and near zero [88]. It should be noted that there is dissent among some researchers whether this norm is an adequate or relevant motivation for the patterns found in conversational speech [c.f. 107]). However, in other corpus evaluations of the timing of turn-taking [e.g., 93, 101, 108]), they indeed find patterns of floor-transfer offset that match this assumption.

While the modes of the densities of FTO under all stance conditions were near zero, I observed many instances of long negative FTO in all conditions as well. To investigate this behavior I randomly sample 25 instances of IPUs from the budget task preceded by an FTO less than -2 seconds and I impressionistically observe four categories of speech. These are summarized in Table 5.3. The most frequent speech type with long negative FTO were failed

attempts to take the floor. There were also five instances of speech spurts which behaved most like backchannels, but didn't match the phrases extracted from the Switchboard Dialog Act Corpus. A common cause of mismatch was differences in the spelling conventions (e.g. *alright* vs. *all right*; *Mmhmm* vs. *Mm hmm*).

Table 5.3: Speech types of sample turns preceded by an interruption (negative FTO) of more than 2 seconds. Categories were determined by a single author listening to the samples of 25 speech samples from the budget task of the ATAROS corpus.

Speech Type	Count
Failure to take the floor	14
Backchannels with non-Switchboard tokens	5
Expression of agreement/paraphrasing	4
Floor ownership unclear	2

5.8 Limitations and Future Work

While Linear Mixed Effects modeling provides convenient and interpretable comparison of features, appropriate use of these models requires that the residuals of features be normally-distributed. From the distributions provided in 5.2, it is clear that for most of the duration features this assumption does not hold. In the future, I would like to replicate this experiment with a more appropriate non-parametric modeling paradigm to confirm the significance of these findings.

While I have shown a strong relationship between stance and the timing of conversation (accepting questions of statistical validity), I do not suggest that the timing of conversation can be anticipated via stance alone. In fact, previous work has shown that there is variety of linguistic and non-linguistic behaviors related to turn-taking, and further that the acoustic

changes to speech because of turn-taking can also be induced by other phenomena. For example, the degree to which conversation follows the norms of speech timing (such as the majority of floor transfers having near-zero FTO [88]) is confounded by the familiarity of interlocutors [72, 109], and further by physiological factors such as hearing impairment [110].

There are some known influences of turn-taking and dialogue structure that cannot be assessed using the ATAROS dataset. Jokinen et al. [95], for example, used a corpus-based analysis on an audiovisual dataset to map gesture and eye gaze to the occurrence of turn-taking. However visual cues are not available in ATAROS.

Surprisal, the predictability of speech input given context, is another confound meriting further investigation. Regions of high surprisal in speech have been show to have a pattern of hyper-articulation and local changes in speech timing [111], in the same way that hyper-articulation is observed with strong-taking [99]. It is therefore natural to question whether a compounding effect emerges when both high surprisal and strong stance are present in speech, or whether some other mitigating patterns emerge.

5.9 Conclusion

This work uses corpus-based, automatic methods to analyze how the timing of turn-taking in conversation is influenced by expressions of stance in dyadic, task-oriented conversation in English. I find that there are significant differences in speech timing measures when compared with stance strength and polarity, with moderate and strong, positive stance being consistently associated with significant duration differences in speech timing measures. I also find that speaker gender does not significantly effect the distribution of speech timing measures.

This work supports previous findings on the timing and acoustics of stance, and highlights the importance of inter-subjective timing information in understanding the impact of stance on speech.

Chapter 6
STANCE-TAKING AND VOWEL EXPANSION

6.1 Information in the Speech Signal

Many types of linguistic information are simultaneously encoded in speech. Speech can include information about the sociolinguistic background of speakers and interlocutors, the discourse function of an utterance, attitudes and opinions, and of course lexical meaning. The challenge in studying how these meanings are created and perceived is that they are happening simultaneously and often using the same kinds of acoustic permutations. However, one consensus opinion is that when speech is more information-laden, it will induce hyper-articulation [113]. Conversely, speech with a small information load will be more liable to reduction.

Strong stance-taking, a part of the pragmatic information carried in speech, has been shown to motivate hyper-articulation [99]. What's more, research from Jurafsky et al. [57], among others, has shown that when words are less predictable in a given speech context, then they are more likely to be hyper-articulated. Both of these findings are compatible with one another, and compatible with the Lindblom's theory of hypo- and hyper-articulation [113], which he asserts is driven by communicative load of all types. However, the effects of strong stance and predictability on hyper-articulation have not been studied in tandem. In this chapter, I consider the interaction between stance-taking and lexical predictability (operationalized as surprisal) and their effect on two measures of hyper-articulation in vowels:

This chapter is revised from "Effects of Information Load and Pragmatic Load on the Hypo-Hyper Continuum," which appeared in Fonetik 2024 [112]. My specific contributions were the experiment design and implementation, and equal contributions to analysis.

vowel space and duration. Using Generalized Additive Modeling, I compare the effect of stance and surprisal on the repulsive force within the vowel space and on the duration of vowels. I show that there are significant interactions between some types of stance-taking and surprisal, and discuss implications for future study in hyper-articulation.

6.2 Information Motivates Hyper-articulation

In 1990, Lindblom [113] posited an explanation for systematic variation, where redundancy in the information signal (chiefly, sufficient context in the surrounding speech to recover missing information) induces hypo-articulation. Conversely, when communicative needs are higher or when there isn't sufficient context, hyper-articulation is more likely to occur. This theory has been tested and supported in a variety of experimental paradigms [e.g., 114, 115, 116, 117, 118, 119]. These works expand the study of articulation to more the channels of information within speech.

An important consideration for these new channels is the extent to which they overlap acoustically and influence one another. The pragmatic information-bearing of stance-taking is a less-studied motivator of hyper-articulation, although previous work has shown a positive relationship between some types of stance-taking and hyper-articulation [99]. Following this work, I investigate how surprisal (bearing lexical information) and stance-taking (bearing pragmatic information) contribute to vowel space articulation. My work answers two research questions to this point:

1. What individual effects do surprisal and stance have on the vowel space? On vowel duration?

2. Are there any mitigating/reinforcing interactions between surprisal and stance that change the degree of hyper-articulation?

6.3 Methods

6.3.1 Dataset and Stance Annotation

The ATAROS dataset [49, 52] is ideal for investigating these questions, due to its coarse- and fine-grained stance annotations, and phone level alignments from P2FA [120]. I consider the recordings from 34 pairs of participants, completing the inventory (participants arranging inventory in a department store) and budget (participants balancing a hypothetical municipal budget) tasks.

As in Chapter 5 and [49], I define coarse stance in two categorical dimensions: relative strength (None, Weak, Moderate, Strong) and polarity (Negative, Neutral, or Positive). I exclude all data except stressed vowels in content words (nouns, verbs, adverbs, and adjectives).[1] Stress labels are taken from dictionary pronunciations from the CMU pronouncing dictionary.[2] Words are annotated for part-of-speech using the pre-trained Greedy Averaged Perceptron tagger from NLTK [121].[3]

The counts of stressed vowels in content words are grouped by stance type in Table 6.1.

Table 6.1: Number of vowel instances in the ATAROS budget and inventory tasks, grouped by stance strength (rows) and polarity (columns).

	-	Neutral	+
None	0	15374	0
Weak	403	19751	12064
Moderate	3749	15003	3344
Strong	282	341	51

[1]Specifically, I include the following Penn Treebank categories: 'CD', 'FW', 'JJ', 'JJR', 'JJS', 'NN', 'NNS', 'NNPS', 'PDT', 'RB', 'RBR', 'RBS' , 'UH', 'VB', 'VBD', 'VBG', 'VBN', 'VBP', 'VBZ.'

[2]http://www.speech.cs.cmu.edu/cgi-bin/cmudict/

[3]https://www.nltk.org/_modules/nltk/tag.html#pos_tag

6.3.2 Lexical Surprisal

I use lexical surprisal to approximate the lexical information of speech, by training a language model on a source of comparable data. In order to derive word-level probabilities for the data, I train an LSTM on a next-word prediction task, using the Fisher training dataset [56].

This neural network's objective is to learn to predict the next word in a sequence of words and its architecture is known for its ability to capture sentential context and ability to train with a moderate amount of data. I use all transcripts from the Fisher dataset as training, which contain spontaneous dyadic conversations like ATAROS. I train a model with hidden dimension 1024 and embedding dimension 50, using gradient clipping with a maximum norm of 1 for 50 epochs with batch size 128. I use cross-entropy loss with a weighted mean reduction as the training objective. A learning rate scheduler decreases the learning rate every 10 epochs, with multiplicative factor $\gamma = 0.1$. Once trained, utterance transcripts can be passed to the model, and it will output associated probabilities for all words, given the prior words in the utterance. To evaluate word probabilities, I split transcripts into utterances, and use the basic English tokenizer in pytorch (breaking words on white space and punctuation) [84].

For the purposes of this work, the surprisal of a sentence with word sequence W is defined as the negative average log probability of the *content* words in the sequence, as produced by passing the *entire* sentence through the language model. Let W_c be the set of content words in W. Equation 6.1 computes the surprisal of W.

$$s\left(W\right) = -\frac{\sum_{w \in W_c} \log_2\left(p\left(w\right)\right)}{|W_c|} \tag{6.1}$$

6.3.3 Repulsive Force Ratio

Following the description of gravitational force in the vowel space presented in [119, 122, 123], I describe the expansion of the vowel space with respect to the repulsive force of vowels originating from each speaker individually. In order to compare the expansion of individual

vowels rather than the vowel space in general, I modify the definition of repulsive force from [123], normalizing the repulsive force by the average repulsive force of all instances of the same vowel.

Let l be the vowel index $l \in \{\text{ʌ, ʊ, aɪ, u, oʊ, æ, ɛ, i, ɑ, ɪ, ɝ, ɔ, eɪ, aʊ, ɔɪ}\}$ and for each vowel l (inscribed as a superscript), i is a single instance of the vowel $i = 1, \ldots, n_l$ (inscribed as subscript), where n_l is the number of vowels with label l. A vowel instance v_i^l is represented by the mean F1 and F2 across its duration. Let $d\left(v_i^l, v_j^k\right)$ be the Euclidean distance between vowels v_i^l and v_j^k in the F1xF2 plane.

The repulsive force $r(v_i^l)$ for the vowel instance v_i^l is the sum of the inverse square distances between v_i^l and all other vowel instances with a different label:

$$r\left(v_i^l\right) = \sum_{j:v_j^k,\ k \neq l} \frac{1}{d\left(v_i^l, v_j^k\right)^2} \tag{6.2}$$

The normalizing factor of repulsive force for vowel v_i^l is the average repulsive force for all vowels with the same label:

$$\overline{r}\left(v_i^l\right) = \frac{r\left(v_i^l\right)}{\frac{1}{n_l}\sum_{j:v_j^k,\ k=l} r\left(v_j^l\right)} \tag{6.3}$$

Note that the repulsive force of a vowel becomes greater when it is nearer in the vowel space to other vowels of the same type. Thus, a smaller average repulsive force across vowels indicates an expanded vowel space.

6.3.4 Duration

I use timings from phone-level transcriptions force-aligned using the Penn Phonetics Lab Forced Aligner [P2FA, 120] to compute vowel duration. Where lexical items in ATAROS were out-of-vocabulary for P2FA, they were manually added to the dictionary. For comparability, I analyze all vowels that were included according to the criteria of the repulsive force ratio.

6.3.5 Statistical Comparison

I use Wood's Generalized Additive Modeling (GAM) [124] as implemented in the `mgcv` package [125] in R (using Gamma model family) to model the relationships between surprisal and stance and the two measures of vowel hyper-articulation. GAMs are a non-parametric modeling method that can model non-linear relationships between data types. They have been used in similar studies of acoustic distinctiveness such as Matt Kelley's dissertation [126] (the author used the variant Generalized Additive Mixed Model (GAMM) with mixed effects to study the impact of various acoustic measures on listener response times).

In addition to fixed effects of stance and surprisal (smoothed via thin plate regression splines [127]) variables, I include smoothed interaction terms between stance and surprisal. I fit separate models for strength and polarity (both including surprisal and the respective interaction term).[4] Neutral polarity, zero stance strength are treated as the intercept conditions.

6.4 Results

Figure 6.1 shows the distributions of the repulsive force ratio and vowel duration for the vowels of interest. Both distributions are right-skewed with long right tails, which is expected for these features.

I report significance in tables using the following notation: $*** = p < 0.001$, $** = p < 0.01$, $* = p < 0.05$. Significance of parametric coefficients is derived from the Bayesian estimated covariance matrix of the parameter estimators. For smoothed parameters, the effective degrees of freedom (EDF) estimates the degree polynomial needed to fit each term, and the associated significance is derived from ANOVA tests.

Tables 6.2 and 6.3 summarize the estimated linear coefficients of stance strength and surprisal EDF when modeling repulsive force ratio and duration. The coefficients indicate

[4] In the syntax of `mgcv`, the formula that defines the model is `gam(repulsive force|distance) ~ s(surprisal) + stance {polarity|strength} + s(surprisal, by = stance {polarity|strength})`

the impact of the stance conditions on the predicted value (e.g., vowel duration). Interaction terms are abbreviated with the syntax `surprisal.x`, where `x` is the stance condition. For example, `surprisal.moderate` is the interaction between surprisal and moderate stance strength. I compare models which include (right) and exclude (left) the interaction term.[5]

Tables 6.4 and 6.5 similarly report the coefficients and EDF of stance polarity-based models of repulsive force and duration, respectively.

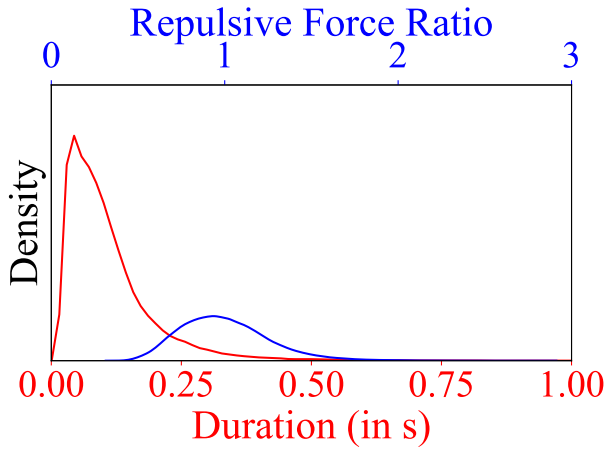

Figure 6.1: Kernel density estimates of repulsive force ratio (blue) and duration (red).

6.5 Discussion

The statistical evidence from the GAMs supports previous findings that both stance-taking and lexical surprisal are sources of hyper-articulation, measured as repulsive force and duration of vowels.

Models without interaction terms showed highly significant effects for stronger stance strength and surprisal, for both repulsive force ratio and duration. Consistent with hyper-articulation, increasing stance reduces the repulsive force ratio and increases duration. Non-zero levels of strength, as well as non-neutral polarity, reduced the repulsive force and ex-

[5]In notation, `s(surprisal, by=stance polarity|strength)`

Table 6.2: Summary of GAM linear stance strength coefficients and surprisal EDF for predicting the repulsive force ratio, with and without interacting effects. Stance strength 0 is intercept. Only significant interactions included: surprisal.x, where x is stance strength.

Prediction Variables		− Interactions		+ Interactions	
		Est.	Signif.	Est.	Signif.
Coeff	0	1.011	***	1.011	***
	1	-0.004	ns	-0.004	ns
	2	-0.028	***	-0.028	***
	3	-0.058	***	-0.56	**
EDF	surprisal	6.6	***	8.3	***
	surprisal.moderate	n/a		7.1	*
	surprisal.strong	n/a		3.4	**

Table 6.3: Summary of GAM linear stance strength coefficients and surprisal EDF for predicting duration, with and without interacting effects. Stance strength 0 is intercept. No interaction terms were significant.

Prediction Variables		- Interactions		+ Interactions	
		Est.	Signif.	Est.	Signif.
Coeff	0	8.43	***	8.42	***
	1	0.52	***	0.54	***
	2	1.55	***	1.54	***
	3	1.20	***	1.16	ns
EDF	surprisal	8.8	***	7.8	ns

Table 6.4: Summary of GAM linear stance polarity coefficients and surprisal EDF for predicting the repulsive force ratio, with and without interacting effects. Stance polarity neutral (0) is intercept. No interaction terms were significant.

Prediction Variables		- Interactions		+ Interactions	
		Est.	Signif.	Est.	Signif.
Coeff	0	1.000	***	1.011	***
	-	-0.023	***	-0.022	***
	+	-0.005	*	-0.005	*
EDF	surprisal	6.4	***	1.7	ns

Table 6.5: Summary of GAM linear stance polarity coefficients and surprisal EDF for predicting duration, with and without interacting effects. Stance polarity neutral (0) is intercept. No interaction terms were significant.

Prediction Variables		- Interactions		+ Interactions	
		Est.	Signif.	Est.	Signif.
Coeff	0	9.33	***	9.34	***
	-	0.16	ns	0.10	ns
	+	-0.88	***	-0.84	***
EDF	surprisal	8.8	***	8.6	**

panded the vowel space. When interaction terms are included in models with stance strength, they are only significant between surprisal and moderate and strong stance (see Table 6.2). Use of interaction terms may have reduced the significance of other terms where there are interactions, i.e. between surprisal and strong stance. For duration, no interaction terms are significant, their inclusion forces overfitting and reduces significance of strong stance which is associated with fewer instances.

Stance polarity was also a significant predictor of repulsive force and duration, with some caveats. Negative polarity was not significant for duration (see Table 6.5). Surprisal was only significant when interaction terms were excluded in predicting repulsive force, and significance was weaker with interactions in the duration model. This is also likely due to overfitting.

An unexpected finding is the significant interactions between surprisal and stronger levels of stance, but a lack of significant interaction for any kind of stance polarity. One issue in understanding this discrepancy is the interdependence between stance and polarity: a spurt cannot have polarity without positive stance strength, and the distribution of negative and positive polarity is not equal within strength level. Further study is needed to understand how this two-part conception of stance effects the significance of the results.

6.5.1 Limitations and Future Work

The effects presented in this chapter represent a specific speech style at one point in time (Pacific Northwest English speakers in the early 21st century). It is unknown whether the interactions observed in this data would hold in other linguistic contexts, such as in languages other than English, conversations between familiar interlocutors, etc.

The differing findings for the two hyper-articulation measures motivates further work on appropriate acoustic measures of hyper-articulation in this context. A limitation of using only the ATAROS dataset is that the stance-taking tends to be on the weaker side (see 6.1). The artificial nature of this dataset is unlikely to be representative of speech behavior in the wild, especially because of the unusual nature of the lexicon that the tasks require. It may be the case that the kinds of stance elicited in a laboratory environment like the ATAROS collection design are too weak to see large hyper-articulation effects.

One concern with respect to the vowel duration is that vowel identity is not controlled for in this description of duration. Separate pilot experiments with identity-normalized vowel duration did not show any noticeable differences compared to the un-normalized duration. A more fruitful venue for re-testing the impact of surprisal and stance on vowel duration would be to check the alignments output by P2FA; while including only averaged values for vowels mitigates some of the more fine-grained alignment issues that can occur with forced alignment, it's still possible that alignment errors contributed to the fewer significant findings with respect to duration.

6.6 Conclusion

I have shown that stance-taking and lexical surprisal are significant mediators of changes in articulation, particularly in the gravitational force of the vowel space. Lexical surprisal and stance-taking were also found to interactively influence repulsive force in the vowel space. Both information streams impacted vowel duration, however I did not find definitive evidence towards an interaction with respect to duration. The competition induced onto the acoustic

signal from the various simultaneous sources of linguistic information is a manifestation of the complexity of information expression through language.

Chapter 7

CONCLUSION

I have demonstrated through a series of human- and machine-facing experiments the wide range of behaviors and applications of speech prosody in the landscape of linguistic technology. Surprisal and stance, both instigators of prosodic variation, have been shown to significantly interact with one another. The timing of discourse, intimately related to prosodic timing, was found to vary significantly with respect to speaker stance as well.

While the results of experiments using prosodic features for existing speech processing tasks do not suggest across-the-board benefits to modeling prosody, I believe their strength is in exposing the types of data where including prosodic information is most likely to show a benefit. Incorporating prosodic cues into computational models, whether implicitly or explicitly, enables speech understanding that would not be possible with words alone. In looking toward the future of artificial intelligence, the complex social and interpersonal information transfer facilitated by language can only be understood by machines when the full context of speech, including all linguistic modalities, are taken into account. The acoustics I've used to represent prosody (pitch, duration, and intensity) are a small part of that.

A challenge of researching prosody in the computational space is the ideological divide between how linguists conceive of prosody and its utility, and how prosodic information is utilized in technological applications. The linguistic conception of prosody takes many forms and interpretations, based in understanding specific instances of speech and local changes to the acoustics therein. In contrast, the black-box approach of the state-of-the-art in computational speech understandings prevents practitioners from understanding where prosody is being used in speech tasks, and to what extent. I believe bridging this divide will enable the two-fold goal of understanding how prosody functions in natural speech, and how

to computationally represent those mechanisms.

BIBLIOGRAPHY

[1] Robert Mannel, Felicity Cox, and Johnathon Harrington. Introduction to prosody theories and models. Macquarie University, 2014. https://www.mq.edu.au/about/about-the-university/our-faculties/medicine-and-health-sciences/departments-and-centres/department-of-linguistics/our-research/phonetics-and-phonology/speech/phonetics-and-phonology/intonation-prosody.

[2] Cynthia G. Clopper and Rajka Smiljanic. Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, 39(2):237–245, 2011.

[3] Nicole Holliday. Prosody and sociolinguistic variation in American Englishes. *Annual Review of Linguistics*, 7(1):55–68, 2021.

[4] Paul E. Reed. *Sounding Appalachian:/ai/ monophthongization, rising pitch accents, and rootedness*. PhD Thesis, University of South Carolina, Columbia, SC, 2016.

[5] Malcah Yaeger-Dror. Register and prosodic variation, a cross language comparison. *Journal of Pragmatics*, 34(10):1495–1536, 2002.

[6] Douglas Biber. *Variation across Speech and Writing*. Cambridge University Press, 1988.

[7] Sun-Ah Jun. *Prosodic typology: The phonology of intonation and phrasing*, volume 1. Oxford University Press, 2006.

[8] Yi Xu, Szu-wei Chen, and Bei Wang. Prosodic focus with and without post-focus compression: A typological divide within the same language family? *The Linguistic Review*, 29(1):131–147, 2012.

[9] Phillip M. Carter, Lydda López Valdez, and Nandi Sims. New dialect formation through language contact: Vocalic and prosodic developments in Miami English. *American Speech*, 95(2):119–148, 2020.

[10] Valerie D. Freeman. Using acoustic measures of hyperarticulation to quantify novelty and evaluation in a corpus of political talk shows. Master's Thesis, University of Washington, Seattle, WA, 2010.

[11] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. ToBI: A standard for labeling English prosody. In *Proc. 2nd International Conference on Spoken Language Processing*, 1992.

[12] Sun-Ah Jun. Prosodic typology: By prominence type, word prosody, and macrorhythm. *Prosodic typology II*, pages 520–539, 2014.

[13] Jennifer Cole and Stefanie Shattuck-Hufnagel. New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, 7(1), 2016.

[14] Paul Taylor and Amy Isard. SSML: A speech synthesis markup language. *Speech Communication*, 21(1-2):123–133, 1997.

[15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. 40th International Conference on Machine Learning*, pages 28492–28518, 2023.

[16] Abraham Woubie, Jordi Luque, and Javier Hernando. Using voice-quality measurements with prosodic and spectral features for speaker diarization. In *Proc. INTERSPEECH 2015*, pages 3100–3104, 2015.

[17] Caroline Smith. Marking the boundary: utterance-final prosody in French questions and statements. In *Proc. 14th International Congress of Phonetic Sciences*, volume 5, pages 1181–1184, 1999.

[18] Daniel Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Pearson Prentice Hall, 3rd edition, 2024.

[19] Anne Cutler and D. Robert Ladd. *Prosody: Models and measurements*, volume 14. Springer Science & Business Media, 2013.

[20] John Anton Goldsmith. *Autosegmental phonology.* PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1976.

[21] John A. Goldsmith. *Autosegmental and metrical phonology*, volume 1. Basil Blackwell Cambridge, 1990.

[22] Lisa Davidson. The versatility of creaky phonation: Segmental, prosodic, and sociolinguistic uses in the world's languages. *WIREs Cognitive Science*, 12(3):e1547, 2021.

[23] Maria Luisa Zubizarreta. *Prosody, focus, and word order*. Linguistic Inquiry Monographs. MIT Press, 1998.

[24] Elisabeth Selkirk. Contrastive FOCUS vs. presentational focus: Prosodic evidence from right node raising in English. In *Proc. Speech Prosody 2002*, 2002.

[25] Sara Bögels and Francisco Torreira. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57, 2015.

[26] Jan-Peter De Ruiter, Holger Mitterer, and Nick J. Enfield. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535, 2006.

[27] Lynne C. Nygaard, Neelam Patel, and Jennifer S. Queen. The link between prosody and meaning in the production of emotional homophones. *The Journal of the Acoustical Society of America*, 112(5_Supplement):2444, 2002.

[28] Henry S. Cheang and Marc D. Pell. The sound of sarcasm. *Speech Communication*, 50(5):366–381, 2008.

[29] Valerie Freeman, Richard Wright, and Gina-Anne Levow. The prosody of negative yeah. In *LSA Annual Meeting Extended Abstracts*, volume 6, 2015.

[30] Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13), 2023.

[31] Stephen Li. Taishanese Language Home 台山话资源网, 2008.

[32] Wing Li Wu. *Cantonese prosody: Sentence-final particles and prosodic focus*. PhD Thesis, University College London, London, England, 2013.

[33] Wing Li Wu and Yi Xu. Prosodic focus in Hong Kong Cantonese without post-focus compression. In *Proc. Speech Prosody 2010*, 2010.

[34] Marina Nespor, Mohinish Shukla, and Jacques Mehler. Stress-timed vs. Syllable- timed Languages. In *The Blackwell Companion to Phonology*, volume 2, pages 1147–1159. John Wiley & Sons, Ltd, 2011.

[35] Peter Ladefoged, Keith Johnson, and Peter Ladefoged. *A course in phonetics*. Thomson Wadsworth Boston, 2006.

[36] Pier Marco Bertinetto. Reflections on the dichotomy 'stress' vs. 'syllable-timing' . *Revue de phonétique appliquée*, 91(93):99–130, 1989.

[37] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *Proc. 9th ISCA Workshop on Speech Synthesis*, 2016.

[38] Chunhui Lu, Xue Wen, Ruolan Liu, and Xiao Chen. Multi-Speaker Emotional Speech Synthesis with Fine-Grained Prosody Modeling. In *Proc. ICASSP 2021*, pages 5729–5733, June 2021.

[39] Jack Weston, Raphael Lenain, Udeepa Meepegama, and Emil Fristed. Learning de-identified representations of prosody from raw audio. In *Proc. 38th International Conference on Machine Learning*, pages 11134–11145, 2021. ISSN: 2640-3498.

[40] Trang Tran, Shubham Toshniwal, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Mari Ostendorf. Parsing speech: A neural approach to integrating lexical and acoustic-prosodic information. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long Papers), pages 69–81, New Orleans, Louisiana, 2018.

[41] Gina-Anne Levow, Valerie Freeman, Alena Hrynkevich, Mari Ostendorf, Richard Wright, Julian Chan, Yi Luan, and Trang Tran. Recognition of stance strength and polarity in spontaneous speech. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 236–241, 2014.

[42] Yeonjin Cho, Sara Ng, Trang Tran, and Mari Ostendorf. Leveraging prosody for punctuation prediction of spontaneous speech. In *Proc. INTERSPEECH 2022*, pages 555–559, 2022.

[43] Trang Tran. *Neural Models for Integrating Prosody in Spoken Language Understanding*. PhD Thesis, University of Washington, Seattle, WA, 2020.

[44] Raul Fernandez, Asaf Rendel, Bhuvana Ramabhadran, and Ron Hoory. Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In *Proc. INTERSPEECH 2014*, 2014.

[45] Zack Hodari, Alexis Moinet, Sri Karlapati, Jaime Lorenzo-Trueba, Thomas Merritt, Arnaud Joly, Ammar Abbas, Penny Karanasou, and Thomas Drugman. Camp: A two-stage approach to modelling prosody in context. In *Proc. ICASSP 2021*, pages 6578–6582, 2021.

[46] Andreas Triantafyllopoulos, Johannes Wagner, Hagen Wierstorf, Maximilian Schmitt, Uwe Reichel, Florian Eyben, Felix Burkhardt, and Björn W. Schuller. Probing speech emotion recognition transformers for linguistic knowledge. In *Proc. INTERSPEECH 2022*, pages 146–150, 2022.

[47] Rose Sloan, Adaeze Adigwe, Sahana Mohandoss, and Julia Hirschberg. Incorporating prosodic events in text-to-speech synthesis. In *Proc. Speech Prosody 2022*, pages 287–291, 2022.

[48] Susanne Brouwer. *Processing strongly reduced forms in casual speech.* PhD Thesis, Radboud University Nijmegen, Nijmegen, Netherlands, 2010.

[49] Valerie Freeman. *The phonetics of stance-taking.* PhD Thesis, University of Washington, Seattle, WA, 2015.

[50] Elise Kärkkäinen. Stance taking in conversation: From subjectivity to intersubjectivity. *Text & Talk*, 26(6):699–731, 2006.

[51] John W. Du Bois. The stance triangle. In Robert Englebretson, editor, *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, volume 164 of *Pragmatics & Beyond New Series*, pages 139–182. John Benjamins Publishing Company, 2007.

[52] Valerie Freeman, Julian Chan, Gina-Anne Levow, Richard Wright, Mari Ostendorf, and Victoria Zayats. Manipulating stance and involvement using collaborative tasks: An exploratory comparison. In *Proc. INTERSPEECH 2014*, pages 303–307, 2014.

[53] Gina-Anne Levow and Richard A. Wright. Exploring dynamic measures of stance in spoken interaction. In *Proc. INTERSPEECH 2017*, pages 1452–1456, 2017.

[54] Elizabeth Shriberg, Andreas Stolcke, and Don Baron. Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation, disfluencies, and overlapping speech. In *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[55] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. 1992 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520, 1992.

[56] Christopher Cieri, David Miller, and Kevin Walker. Fisher English training speech parts 1 and 2. *Linguistic Data Consortium*, 2004.

[57] D. Jurafsky, A. Bell, M. Gregory, and W.D. Raymond. The effect of language model probability on pronunciation reduction. In *Proc. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 801–804, 2001.

[58] Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. Resegmentation of SWITCHBOARD. In *Proc. 5th International Conference on Spoken Language Processing*. Syndey, 1998.

[59] Douglas A. Jones, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas A. Reynolds, and Marc A. Zissman. Measuring the readability of automatic speech-to-text transcripts. In *Proc. INTERSPEECH 2003*, pages 1585–1588, 2003.

[60] Tianyu Zhao and Tatsuya Kawahara. Joint dialog act segmentation and recognition in human conversations using attention to dialog context. *Computer Speech & Language*, 57:108–127, 2019.

[61] Trang Tran, Jiahong Yuan, Yang Liu, and Mari Ostendorf. On the role of style in parsing speech with neural models. In *Proc. INTERSPEECH 2019*, pages 4190–4194, 2019.

[62] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[63] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. The Kaldi Speech Recognition Toolkit. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[64] Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, 1997.

[65] Heidi Christensen, Yoshihiko Gotoh, and Steve Renals. Punctuation annotation using statistical prosody models. In *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001.

[66] Tal Levy, Vered Silber-Varod, and Ami Moyal. The effect of pitch, intensity and pause duration in punctuation detection. In *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, pages 1–4, 2012.

[67] Monica Sunkara, Srikanth Ronanki, Kalpit Dixit, Sravan Bodapati, and Katrin Kirchhoff. Robust Prediction of Punctuation and Truecasing for Medical ASR. In *Proc. Workshop on NLP for Medical Conversations*, 2020.

[68] P. Zelasko, P. Szymanski, J. Mizgajski A. Szymczak, Y. Carmiel, and N. Dehak. Punctuation Prediction Model for Conversational Speech. In *Proc. INTERSPEECH 2018*, pages 2633–2637, 2018.

[69] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Proc. INTERSPEECH 2016*, pages 2751–2755, 2016.

[70] Kyunghyun Cho, Bart Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proc. Empirical Methods in Natural Language Processing*, 2014.

[71] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. 3rd International Conference for Learning Representations*, 2015.

[72] Allan Bell. Language style as audience design. *Language in Society*, 13(2):145–204, 1984.

[73] Rivka Levitan, Agustín Gravano, Laura Willson, Ştefan Beňuš, Julia Hirschberg, and Ani Nenkova. Acoustic-prosodic entrainment and social behavior. In *Proc. 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11–19, Montréal, Canada, June 2012.

[74] Frank J. Bernieri, John S. Gillis, Janet M. Davis, and Jon E. Grahe. Dyad rapport and the accuracy of its judgment across situations: A lens model analysis. *Journal of Personality and Social Psychology*, 71(1):110, 1996.

[75] Adrienne Wood, Jennie Lipson, Olivia Zhao, and Paula Niedenthal. Forms and functions of affective synchrony. In Michael D. Robinson and Laura E. Thomas, editors, *Handbook of Embodied Psychology: Thinking, Feeling, and Acting*, pages 381–402. Springer International Publishing, 2021.

[76] Susan E. Brennan. Lexical entrainment in spontaneous dialog. *Proc. ISSD*, 96:41–44, 1996.

[77] Juan Pérez, Ramiro Gálvez, and Agustín Gravano. Disentrainment may be a Positive Thing: A Novel Measure of Unsigned Acoustic-Prosodic Synchrony, and its Relation to Speaker Engagement. In *Proc. INTERSPEECH 2016*, 2016.

[78] 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, 2020.

[79] Rivka Levitan and Julia Bell Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proc. INTERSPEECH 2011*, 2011.

[80] Sarah Ita Levitan, Jessica Xiang, and Julia Hirschberg. Acoustic-Prosodic and Lexical Entrainment in Deceptive Dialogue. In *Proc. Speech Prosody 2018*, pages 532–536, 2018.

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

[82] Renzo Mora, Barbara Crippa, Edoardo Cervoni, Valentina Santomauro, and Luca Guastini. Acoustic features of voice in patients with severe hearing loss. *Journal of Otolaryngology-Head & Neck Surgery*, 41(1):8–13, 2012.

[83] Eugene Charniak and Mark Johnson. Edit detection and parsing for transcribed speech. In *Proc. Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.

[84] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. Pytorch: An imperative style, high-performance deep learning library. In *Proc. 33rd International Conference on Neural Information Processing Systems*, 2019.

[85] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[86] Paul Boersma. Praat: Doing phonetics by computer [Computer program]. *http://www.praat. org/*, 2011.

[87] Sara Ng, Gina-Anne Levow, Mari Ostendorf, and Richard Wright. Investigating the influence of stance-taking on conversational timing of task-oriented Speech. In *Proc. INTERSPEECH 2024*, 2024.

[88] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

[89] Jeffrey J. Shultz, Susan Florio, and Frederick Erickson. Where's the floor? Aspects of the cultural organization of social relationships in communication at home and in school. *Children in and out of school: Ethnography and education*, pages 88–123, 1982.

[90] Carole Edelsky. Who's got the floor? *Language in Society*, 10(3):383–421, 1981.

[91] Ian Hutchby and Robin Wooffitt. *Conversation analysis*. Polity Press, 2008.

[92] George Psathas and Timothy Anderson. The 'practices' of transcription in conversation analysis. *Semiotica*, 78(1-2):75–100, 1990.

[93] A. Josefine Munch Sørensen. *The effects of noise and hearing loss on conversational dynamics*. PhD Thesis, DTU Health Tech, Kongens Lyngby, Denmark, 2021.

[94] Louis Ten Bosch, Nelleke Oostdijk, and Jan Peter De Ruiter. Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 3206 of *Lecture Notes in Computer Science*, pages 563–570. Springer, Berlin, Heidelberg, 2004.

[95] Kristiina Jokinen. Non-verbal signals for turn-taking and feedback. In *Proc. 7th International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

[96] John Heritage. Oh-prefaced responses to assessments: A method of modifying agreement/disagreement. In Cecilia E. Ford, Barbara A. Fox, and Sandra A. Thompson, editors, *The Language of Turn and Sequence*, Oxford Studies in Sociolinguistics, pages 196–224. Oxford University Press, 2002.

[97] Joanne Scheibman. Subjective and intersubjective uses of generalizations in English conversations. In Robert Englebretson, editor, *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, volume 164 of *Pragmatics & Beyond New Series*, pages 111–137. John Benjamins, 2007.

[98] Pentti Haddington. Positioning and alignment as activities of stancetaking in news interviews. In Robert Englebretson, editor, *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, volume 164 of *Pragmatics & Beyond New Series*, pages 283–317. John Benjamins, 2007.

[99] Valerie Freeman. Hyperarticulation as a signal of stance. *Journal of Phonetics*, 45:1–11, 2014.

[100] Leanne Elizabeth Rolston. *Dialogical signals of stance taking in spontaneous conversation.* PhD Thesis, University of Washington, Seattle, WA, 2020.

[101] Lauren V. Hadley, W. Owen Brimijoin, and William M. Whitmer. Speech, movement, and gaze behaviours during dyadic conversation in noise. *Scientific Reports*, 9(1):10451, 2019.

[102] Jennifer Coates. One-at-a-time: The organization of men's talk. In Sally Johnson and Ulrike Hanna Meinhof, editors, *Language and Masculinity*, pages 107–129. Blackwell, 1997.

[103] Shigeo Uematsu. The use of back channels between native and non-native speakers in English and Japanese. *Intercultural Communication Studies*, 10(2):85–98, 2001.

[104] Vojtěch Pipek. *On backchannels in English conversation.* PhD Thesis, Masarykova univerzita, Pedagogická fakulta, Brno-střed, Czech Republic, 2007.

[105] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[106] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1 – 48, 2015.

[107] Daniel C. O' Connell and Sabine Kowal. Social responsibility in spoken dialogue. In *Dialogical Genres*, pages 189–196. Springer, New York, NY, 2012.

[108] A. Josefine Munch Sørensen, Ewen N. MacDonald, and Thomas Lunner. Timing of turn taking between normal-hearing and hearing-impaired interlocutors. In *International Symposium on Auditory and Audiological Research: Auditory Learning in Biological and Artificial Systems*, pages 37–44. ISAAR, 2020.

[109] Susan R. Fussell and Robert M. Krauss. Understanding friends and strangers: The effects of audience design on message comprehension. *European Journal of Social Psychology*, 19(6):509–525, 1989.

[110] Pamela Souza, Namita Gehani, Richard Wright, and Daniel McCloy. The advantage of knowing the talker. *Journal of the American Academy of Audiology*, 24(08):689–700, 2013.

[111] Matthew Aylett and Alice Turk. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(Pt 1):31–56, 2004.

[112] Sara Ng, Valerie Freeman, Gina-Anne Levow, Mari Ostendorf, and Richard Wright. Effects of information load and pragmatic load on the hypo-hyper continuum. In *Phonetik 2024*, 2024.

[113] Björn Lindblom. Explaining phonetic variation: a sketch of the H and H theory. In *Speech Production and Speech Modelling.*, number 55 in NATO ASI Series, pages 403–439. Kluwer Academic Publishers Dordrecht/London, 1990. doi:10.1007/978-94-009-2037-8_16.

[114] Anne Anderson, E. Bard, C. Sotillo, A. Newlands, and Gwyneth Doherty-Sneddon. Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception & Psychophysics*, 39:580–592, 1997.

[115] Matthew Aylett and Alice Turk. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the Acoustical Society of America*, 119(5):3048–3058, 2006.

[116] Rachel Baker and Ann Bradlow. Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, 52(4):391–413, December 2009.

[117] Jonah Katz and Elisabeth Selkirk. Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language*, pages 771–816, 2011.

[118] Kaoru Tomita. Effects of word familiarity in contexts on speaker's vowel articulation. *Bulletin of Yamagata University: Humanities*, 16(3):55–64, 2008.

[119] Richard Wright. Factors of lexical competition in vowel articulation. In *Phonetic Interpretation*, number VI in Papers in Laboratory Phonology, pages 75–87. Cambridge University Press, Cambridge, UK, 2004.

[120] Jiahong Yuan and Mark Liberman. Speaker identification on the SCOTUS corpus. *Acoustical Society of America Journal*, 123(5):3878, 2008.

[121] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[122] Daniel McCloy, Richard Wright, and Pamela Souza. Talker versus dialect effects on speech intelligibility: A symmetrical study. *Language and Speech*, 58(3):371–386, 2015.

[123] Daniel McCloy. *Prosody, intelligibility and familiarity in speech perception*. Doctoral Dissertation, University of Washington, 2013.

[124] Simon Wood. *Generalized additive models: an introduction with R*. Texts in Statistical Science. Chapman and Hall/CRC, New York, 2nd edition, 2017.

[125] Simon Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

[126] Matthew C Kelley. *Acoustic Distance, Acoustic Absement, and the Lexicon*. PhD Thesis, University of Alberta, Edmonton, Canada, 2021.

[127] Simon Wood. Thin plate regression splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(1):95–114, 2003.