# TOWARDS A MODEL OF THE MAPPING BETWEEN ENGLISH AND SPANISH PROSODY

JONATHAN AVILA

Doctoral Program in Computer Science

APPROVED:

Nigel Ward, Chair, Ph.D.

David Novick, Ph.D.

Olac Fuentes, Ph.D.

Natalia Mazzaro, Ph.D.

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

# TOWARDS A MODEL OF THE MAPPING BETWEEN ENGLISH AND SPANISH PROSODY

by

JONATHAN AVILA, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

December 2023

# Acknowledgments

I would like to express my gratitude to the following individuals and organizations who have contributed to the completion of this research project:

- Professor Nigel Ward for his guidance and support. His mentorship and expertise have been invaluable in my development as a researcher and the execution of this project.
- Emilia Rivas and Divette Marco for their support in the data collection process. Their dedicated efforts in recruiting participants, leading data collection sessions, and assisting with listening analyses have enriched the quality of this research.
- Ann Lee, Benjamin Peloquin, and Justine Kao for discussions. Their perspectives and thoughtful input have enhanced the depth and breadth of this research.
- Meta, the National Science Foundation, and UTEP's University Research Institute for their funding. Their financial support have enabled the execution of this project.

This material is based in part on work supported under the AI Research Institutes program by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award #2229873 – National AI Institute for Exceptional Education. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

# Abstract

Current speech-to-speech translation systems face challenges in effectively translating the nuances of prosody, which plays a pivotal role in conveying speaker intent and stance in dialog. This limitation restricts cross-lingual communication, especially in situations demanding deeper interpersonal understanding. To address this, this research delves into the relationships between prosody and its pragmatic functions, in English and Spanish. First, I discuss a data collection protocol in which bilingual speakers re-enact utterances from an earlier conversation in their other language, then describe an English-Spanish corpus, consisting of 3816 matched utterance pairs. Second, I describe a prosodic dissimilarity metric based on Euclidean distance over a broad set of prosodic features. I then used these to investigate cross-language prosodic differences, and create three simple models for mapping prosody from one language to another to identify phenomena which will require more powerful modeling. These findings should inform future research on cross-language prosody and the design of speech-to-speech translation systems capable of effective prosody translation.

# Table of Contents

**Chapter**

**Appendix**

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Motivation

Challenges in communication frequently emerge when people do not speak the same language. Overcoming these language barriers often requires the assistance of a translator. Besides requiring knowledge of the language in question, translators also require familiarity with specific contexts, which in practice limits their availability. To address this issue, there are ongoing efforts to automate the translation process, leading to the development of speech-to-speech translation systems.

Speech-to-speech translation systems are already valuable tools used in cross-language communication. Today, these systems are available as standalone software or are integrated into other technologies, such as communication applications like Google Translate. These systems allow people to communicate information quickly when a human translator is not available, such as in business settings. Despite the impressive strides in making speech-to-speech translation more accessible, there remains room for improvement.

In particular, while speech-to-speech translation systems are already useful for facilitating short, transactional interactions, they are less useful when applied to long-form conversation [41]. A key factor contributing to these limitations is an inadequate translation of prosody, encompassing elements such as rhythm, intonation, stress, and pitch.

Current speech-to-speech translation systems generally aim to produce prosody that appears natural, but this is not always sufficient. A translation might exhibit prosody that appears natural but may not suit the conversational context. For instance, the prosody of the utterance *yeah* spans a spectrum of interpretations, ranging from a request for clarification to an empathetic acknowledgment, or even a subtle disagreement[1]. Such distinctions can be lost in translation, potentially leading to misunderstandings and miscommunications.

Thus, without adequate prosody translation, speech-to-speech translation systems are unable to reliably convey many intents and stances. Consequently, users of these systems are unable to participate in natural conversation with interlocutors who speak a different language. This constraint extends to an array of scenarios — including, for instance, conversations with a neighbor, a hairdresser, or a fellow attendee at a social gathering — and is a barrier to deepening interpersonal relationships and achieving social inclusion.

The relationship between prosody and its pragmatic functions in dialog is an area that remains relatively unexplored. For instance, many common prosodic patterns in English conversation have only recently been identified [77]. A thorough study of the prosodic differences across languages has yet to be conducted.

While the linguistic literature provides some insights into prosody differences across languages, the majority of work on this topic primarily focuses on prosody at the syllable, lexical, and syntactic levels. In particular, there is relatively little work on differences in how prosody conveys pragmatic functions. Even for languages as globally prevalent as English and Spanish, our knowledge is sparse beyond a few topics such as turn-taking [5], questions and declaratives [22, 85], the expression of certainty [56] and sarcasm and sincerity [57]. However, these certainly do not exhaust the prosodic meanings important for dialog.

Taking an alternative avenue of exploration, other investigations specific to the English-Spanish pair have focused on their cross-linguistic influences. This includes studies on the production of uptalk [37], the perception of intelligibility and accentedness [20], and the production of final boundary tones in declarative utterances [85]. While these studies provide valuable insights, they diverge from the focus of pragmatic prosody differences in dialog.

Further, the scope of these studies have generally been restricted to examining differ-

[1] Audio examples of these interpretations are available at `https://jonavila.dev/dissertation`.

ences in intonation and duration, leaving out most prosodic features.

Thus, an important goal of this dissertation is to improve the fidelity of speech-to-speech translation for dialog. This research explores the prosody mappings across languages, using a parallel corpus of English and Spanish utterances.

The contributions of this research are:

1. A protocol for collecting parallel utterances from spontaneous and re-enacted conversations, resulting in a corpus of English and Spanish utterances

2. A novel representation of utterance prosody

3. The first metric of prosodic similarity between utterances, with use for identifying prominent failures of models of prosody translation

4. A reduced dimensionality representation of utterance prosody and interpretation of the first five dimensions of this representation

5. An assessment of simple models for English-to-Spanish and Spanish-to-English prosody translation, including descriptions of aspects of prosody and associated pragmatic function that pose challenges in translation

6. A comparative analysis of the similarities and differences between English and Spanish prosody in conversation, including a description of some key aspects of prosody as it conveys pragmatic functions in the two languages

The remainder of this chapter provides an overview of related work on speech-to-speech translation relevant to this research, specifically in the areas of multilingual speech corpora, representations of speech, speech-to-speech translation evaluation metrics, and speech-to-speech translation modeling strategies.

## 1.2 Multilingual Speech Corpora

For the purpose of improving the pragmatic fidelity of speech-to-speech translation for dialog, researchers require speech data. This section presents an overview of existing multilingual speech corpora, with a focus on how broadly they represent prosody as used in dialog.

Many multilingual speech corpora currently serve as resources for speech research. These corpora have been developed using various data collection methods. A common data collection method is recording individuals as they read aloud from text. For instance, the *Multilingual LibriSpeech* corpus was derived from recordings of read audiobooks [55]. Notably, this corpus is not parallel, meaning that speech in one language is not matched with speech in another. For speech-to-speech translation research, parallel speech data has the advantage of being easier to compare varied languages and identify their similarities and differences. Examples of non-parallel speech generally do not convey the same intent, making this comparison more difficult.

Recordings of read speech are a common source of parallel speech data. For instance, the *MaSS* corpus was created from readings of a religious text [84]. A much larger corpus, *Common Voice*, is crowdsourced from volunteers who read prompts in their native languages [1]. Taking advantage of its considerable size, the *CoVoST 2* corpus was derived from *Common Voice* by aligning recordings of identical prompts in various languages [72].

Another corpus, *CVSS*, synthesizes speech from translation texts in *Common Voice* and *CoVoST 2* [31]. (Also see [52]; synthesizing translations is a common approach.) However, read speech inherently lacks spontaneity, as it is planned and rehearsed, therefore missing the nuances typical of spontaneous speech; some, but not all, pragmatic uses of prosody can be reliably reproduced in read speech [71]. Moreover, synthesized speech is not entirely natural and lacks prosodic fidelity, rendering it unfit for learning about human speech.

In contrast to reading-based corpora, there exist corpora comprising speech that does not rigidly adhere to a script. For instance, the *VoxPopuli* corpus was derived from recordings of legislative speeches [73]. Publicly available conference talks, particularly those from TED, are a source for multiple corpora. The *Multilingual TEDx (mTEDx)* [63] and *MuST-C* [15] corpora were both created from TED talks and their respective translations. Another corpus adds speech from recorded academic lectures and press conferences [18]. However, the speech in these corpora is primarily intended to inform or instruct vast audiences, and thus, while semi-improvised, it still lacks the nuances of one-on-one interactions.

More reflective of real dialog is acted dialog. The *Heroes* corpus was derived from original and dubbed dialog segments from a television series [47]. Similarly, another corpus [7] was created from multiple television shows. Despite being conversational, these interactions are theatrical and dramatized for entertainment purposes.

At least one corpus was designed specifically for English and Spanish translation. The *Fisher and Callhome Spanish-English* corpus consists of recorded Spanish telephone conversations, and their transcriptions with English translations [53]. However, this corpus does not provide English speech translations and is non-public.

To my knowledge, no previous corpora consist of speech from dialog that is simultaneously spontaneous, parallel, and public. To provide a corpus that meets this need, we created the *Dialogs Re-enacted Across Languages* corpus, which I describe in Chapter 2.

## 1.3 Speech Representations

A speech representation encodes the underlying factors of speech most relevant to its intended application. The selection of such a representation is largely governed by the research objectives and the manner in which the speech is recorded (for instance, an audio recording or a transcription). A straightforward representation of speech is a digital recording, which involves sampling an analog speech signal at regular intervals and encoding these samples into digital data. However, using a recording as the representation directly is impractical due to its size. For instance, an 8 kHz low-quality recording involves 8 thousand samples per second.

A more manageable representation selectively encodes information, but producing such a representation is a non-trivial task. Speech contains an abundance of information, encompassing acoustic, lexical, and semantic information. This information is intertwined in the same signal and only some of it is relevant to a given application.

The features in a representation can be either engineered features or learned features. Engineered features are obtained through carefully chosen algorithms based on prior knowledge and understanding of speech. In contrast, learned features are obtained through machine learning algorithms based on patterns discovered automatically from speech data.

An alternative to feature-based representations are embeddings, which are representations that transform inputs into a lower-dimensional vector space such that similar inputs are closer together in the vector space.

A common type of representation are those based on the words in the speech. Early methods include bag-of-words and TF-IDF, while more recent methods use word embeddings such as word2vec [45] and GloVe [50]. Large language models such as ELMo [51] and BERT [14] advance this by generating contextualized word embeddings.

However, word-based representations, in generalizing speech, overlook prosody. An alternative to word-based representations are representations based on perceptual qualities such as pitch or energy. For example, a pitch contour encodes the fundamental frequency variations over time. Still, this only covers one aspect of speech. Other representations encode a broader range of speech characteristics. Two common representations are spectrograms, which encode the frequency spectrum over time, and mel-frequency cepstral coefficients (MFCCs), which encode the distribution of energy across different frequencies.

Since spectrograms and MFCCs shallowly encode all aspects of speech, models using these representations as input must be able to determine which are relevant. In contrast, other representations are designed for specific tasks, such as speech recognition [3, 28], style control and transfer [86], cross-lingual text and speech translation retrieval [36], speaker de-identification [82], and non-semantic tasks [64].

Machine learning approaches are often limited by the availability of labeled speech data.

While large-scale, labeled speech datasets are ideal, they are relatively scarce, limiting the effectiveness of purely supervised learning methods. This scarcity of labeled data has prompted the development of self-supervised learning methods. These methods leverage unlabeled data to pretrain a model on a related task, which can then be fine-tuned on specific tasks with limited labeled data [46].

Although self-supervised models are effective for many tasks, the extent to which they encode prosodic information has not been well studied. For instance, the extent to which they encode pitch and energy has only recently been investigated [42]. The development of self-supervised methods highlights a possibly significant gap in the field: the lack of a representation specifically designed to capture the prosody as used in dialog. As dialog-driven interfaces gain prominence in technology and applications, representations of utterance prosody could be pivotal in contexts beyond speech-to-speech translation, such as accessibility tools, education and training, and healthcare assistants.

The challenges in interpreting representations may stem from factors not unique to speech, such as their high dimensionality, their ability to discover emergent properties, and the scarcity of labeled data to establish a ground truth. Prosody, in particular, presents its own set of challenges due to its ability to serve multiple functions. These functions can be broadly categorized as paralinguistic, phonological, or pragmatic [42].

Phonological functions relate to the sounds and patterns of a language, including marking syllables and words. Paralinguistic functions relate to speaker identity and self-expression, including identity cues such as a speaker's vocal tract anatomy, pitch range, and accent [16]. Lastly, pragmatic functions relate to expressing an individual's feelings, thoughts, intentions, and attitudes.

The prosodic configurations associated with pragmatic functions can vary within an utterance [42]. Individuals may use these functions to influence the direction and outcome of an interaction, for example, in managing turn-taking, indicating topic and information structure, and conveying stance. These functions are especially important in dialog and are the focus of this research. While existing representations are effective in several respects, they lack focus on representing prosody in a dialog context, thereby potentially overlooking the aspects critical in these interactions.

This research explores a representation of utterance prosody constructed from prosodic features that are easily computed and interpretable. This prosody representation is the subject of Chapter 3. A reduced dimensionality prosody representation is the subject of Chapter 7.

## 1.4 Speech-to-Speech Translation Evaluation Metrics

Assessing the performance of speech-to-speech translation models is essential to their improvement. Throughout the training phase of a model, loss functions guide a learning algorithm to optimize the model's parameters. Following the training phase, evaluation metrics measure the model's performance on unseen data. Evaluation metrics may serve as loss functions, but this is not always feasible or appropriate. Providing a more comprehensive view of performance, evaluation metrics play a critical role in gauging a model's utility for real-world scenarios, identifying its limitations and potential areas for refinement, and benchmarking its performance against other models. Such comparative evaluations are deployed to gauge the implications of modifications to a model, or to underscore advances in the field by juxtaposing it with others.

The evaluation of speech-to-speech translations often relies on human judgments. In Mean Opinion Scores (MOS) evaluations, human evaluators rate translations on specific qualities, such as naturalness, accentedness, and fidelity. Translations are assessed independently or in comparison to a reference.

Traditional MOS scales do not delve into the prosodic aspects that evaluators might consider when judging translations. A finer-grained MOS evaluation might consider aspects such as emotion, overall manner, and meaning [29]. However, there is a limit to how detailed these evaluations can become before becoming too tiresome for evaluators.

The process of human evaluation is resource-intensive. Moreover, human evaluation lacks consistency diminishing its reproducibility [44]. Automatic metrics aim to supplement or be a proxy for human evaluators while being cheaper and more consistent.

Transcription-based automatic metrics involve comparing a transcription of an output speech-to-speech translation with a reference text. These metrics were originally designed for the evaluation of machine translation, but have been adopted for speech-to-speech translation. The most prevalent of these metrics is BLEU, which estimates the similarity of two sentences based on their overlap in contiguous sequences of words [49]. Subsequent metrics offer enhancements over BLEU. For instance, METEOR uses language-specific resources to match more than just exact words, such as synonyms and paraphrases [4].

Semantic similarity metrics go beyond surface-level measures of overlap and instead aim to assess the quality of translations based on how well they preserve the original meaning. Transformer-based metrics, such as BERTScore [87] and MoverScore [88], estimate the semantic similarity of two sentences based on the distance between their contextual embeddings. The most recent of these metrics, COMET, is instead trained on human translations and quality scores, resulting in higher correlation with human judgments of quality [59].

Evaluations using text-based metrics often rely on automatic transcriptions from ASR systems, which may have lower quality for some languages. The text-free metric BLASER avoids this issue by estimating semantic similarity from embeddings directly from speech [11].

These metrics are limited in not properly capturing the pragmatic meaning of speech from dialog. Text-based metrics, like text-based representations, ignore prosody, and are therefore unable to capture the differences that may come from prosody. Semantic similarity metrics are more robust to lexical differences, but are not substitutes for a metric for pragmatic similarity. Semantic similarity measures the extent to which a translation maintains the literal meaning of the original speech, whereas pragmatic similarity delves into the realm of functional meaning. For instance, a translation may be semantically similar to the original speech, but may not be appropriate for the conversational context.

Lastly, there are metrics for estimating similarity from prosodic representations, such as those adapting dynamic time warping [60, 43] and those adapting acoustic correlates of prosody used in general audio processing [66]. These primarily focus on fundamental frequency or otherwise consider a limited range of prosodic features.

At present, no automatic metric exists for estimating the pragmatic similarity in prosody between two utterances. This research presents and evaluates a prosody similarity metric based on the Euclidean distance of the prosody representation mentioned above. The prosody similarity metric is the subject of Chapter 5.

## 1.5 Speech-to-Speech Translation Modeling Strategies

Cascaded speech-to-speech translation systems perform multiple sequential stages of processing, where the output of one stage becomes the input for the next. An example cascaded architecture, shown in Figure 1.1, includes an automatic speech recognition module that recognizes and translates speech into text, a machine translation module that translates the text into the target language, and a text-to-speech module that synthesizes that text into speech.

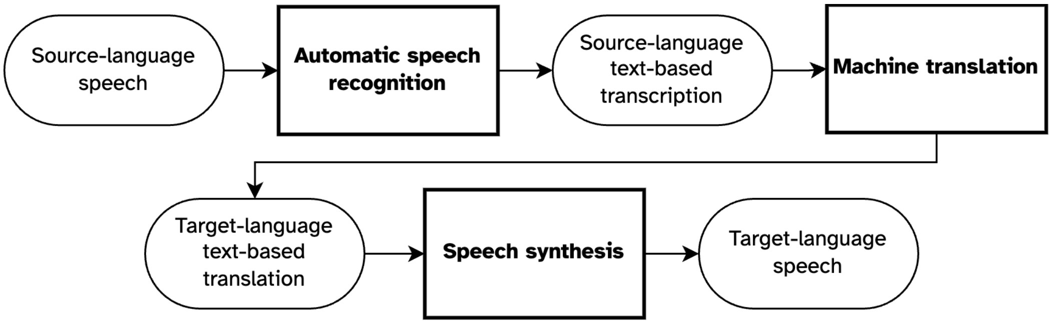

Source-language speech
Automatic speech recognition
Source-language text-based transcription
Machine translation
Target-language text-based translation
Speech synthesis
Target-language speech

Figure 1.1: Example cascaded speech-to-speech translation architecture.

The earliest cascaded systems were designed for a single domain or scenario, such as negotiation [9], travel planning [38], and hotel reservations [68]. This specialization improved performance at the cost of usability. Users were required to speak clearly, limit their vocabulary, and simplify their speech, all of which deviated from their usual speech.

As speech-to-speech translation was still in its infancy, the focus was on accurate word translation rather than prosody. However, prosody was used by these systems to some degree, for instance, in determining phrase boundaries [9], discriminating statements from questions [68], and disambiguation [9, 38].

The cascaded structure of these systems had the benefit of allowing the reuse and adaptation of existing software. However, each processing stage would pass only a few candidates to the next one, and much information from the source speech was lost during intermediate steps. At best, these systems have proved useful for strict, transactional tasks with clearly defined goals.

The field has since advanced, leading to the development of end-to-end models. These models enable joint training of the entire system and the simultaneous learning of speech recognition, translation, and synthesis. The potential to convey additional information provided by source-language prosody was one motivation for the development of direct end-to-end models [33]. By bypassing intermediate steps, these models directly translate speech from one language to another, allowing for an extensive flow of information throughout the system.

Despite rapid recent advances [52, 39, 40, 19, 32, 7], the ability of such models to translate prosody inherent in dialogs remains relatively unexplored. Current approaches predominantly address prosody translation via specific modules [17, 34, 29]. These methods target only specific functions of prosody, notably its roles in conveying paralinguistic and emotional state, emphasis, and syntactic structure. Moreover, these methods target only a few prosodic features, notably $F_0$, pauses, and word duration. Very recent work has shown that this translation of prosody can significantly improve perceived translation quality [29, 67], but also that these methods so far only close less than half of the perceived gap between default prosody and the human reference.

Understanding when and how speech-to-speech translation systems fail is crucial for their improvement. In this research, I explore this concept with an aim to contribute towards the development of more accurate, nuanced speech-to-speech translation systems.

## 1.6 Overview

In this dissertation, I begin by introducing the Dialogs Re-enacted Across Dialogs corpus, a corpus of parallel English and Spanish utterances from spontaneous and re-enacted dialogs. I then propose a representation of utterance prosody to quantify the prosody of these utterances, followed by a metric for evaluating the prosodic similarity between two utterances. Equipped with a corpus, a representation of prosody, and an evaluation metric, I then create simple models for translating the prosody of utterances between the two languages, to investigate the pragmatic functions of prosody of utterances where the models' prediction of prosody representation is least similar to the reference. To conclude, I present the potential implications of these observations and for improving the pragmatic fidelity of speech-to-speech translation models for dialog translation.

In the next chapter, I introduce the Dialogs Re-enacted Across Dialogs corpus.

# Chapter 2

# A Corpus for Speech-to-Speech Translation Research

As discussed in Chapter 1, many multilingual speech corpora are available as resources for speech research. However, these corpora consist of monologs, scripted dialogs, and speech synthesis, which fail to accurately represent the speech used in real dialog. Consequently, their capacity to serve as resources for gaining insights into cross-lingual prosody is limited.

In particular, the research community has lacked a corpus of parallel utterances from real dialogs. Accordingly, we developed the Dialogs Re-enacted Across Languages (DRAL) protocol. Following this protocol, we created the DRAL corpus, a corpus of parallel English and Spanish utterances from recorded conversations and re-enacted utterances. The DRAL corpus is the first of its kind, representing a variety of prosody used in English and Spanish dialog. The DRAL corpus enables a more comprehensive comparison of English and Spanish prosody and serves as the primary resource for this research.

In this chapter, I motivate the corpus-based approach in this research and describe the DRAL protocol and corpus.

## 2.1 Paradigm Shift to Data-Driven Modeling

Today, most methods of speech-to-speech translation adopt a data-driven approach, using machine learning models trained on large corpora. Prior to the advent of such corpora and advances in machine learning, researchers turned to alternative strategies. Early speech-to-speech translation systems integrated distinct modules, such as modules for automatic speech recognition or machine translation. These encompass, for instance: rule-based models which harness linguistic knowledge to formulate handcrafted rules for the translation of phonemes, word sequences, or sentences from one language to another [69, 62]; statistical and finite-state models which leverage a probability distribution over a space of possible translations [10]; and pivot-based models which rely on an intermediary language [24].

In recent years, deep learning models have proven effective in learning the complexities of speech, moving them into mainstream adoption for speech-to-speech translation. However, the accuracy of deep learning models is contingent on the availability of high quality training data. This reliance poses significant challenges for modeling speech with insufficient exemplary data. While techniques used to increase the size and diversity of a dataset, such as data augmentation and synthetic data generation, do exist, these are not always effective and cannot fully replace real data.

Speech data has become an invaluable resource, leading to progress in speech-to-speech translation research. The creation of the DRAL corpus aligns with the evolving needs of the research community, offering an innovative resource for speech-to-speech translation with an emphasis on prosody and its role in dialog.

## 2.2 The Dialogs Re-enacted Across Languages Protocol

The DRAL protocol was designed to collect matched utterances from dialogs. In summary, the DRAL protocol involves pairs of nonprofessional, bilingual participants who have a short conversation in one language and subsequently reenact parts of it in the other. A comprehensive description of the protocol can be found in our technical report [80].

Participants begin by completing a language background form, rating their proficiency in both languages on a 5-point scale and specifying their dialects. Only those deemed sufficiently bilingual are allowed to proceed. The participants then have a ten-minute conversation, recorded by an operator. These conversations are mostly unscripted, with the operator occasionally suggesting topics for conversation, encouraging pragmatic diversity and spontaneous interactions. The participants mostly get to know each other, catch up on recent happenings, and/or share personal experiences, although the nature of the conversation can depend on the relationship between participants.

Subsequently, under the direction of an operator, they listen to utterances or exchanges from the recorded conversation and closely re-enact them in their other language. The operator guides the participants to use the equivalent prosody, allowing different word choice.

The objective is not to mirror the original utterance's prosody but to use the equivalent prosody, which may or may not match with the original utterance's prosody. The operator instructs participants to *keep the same feeling*. When possible, the participants recreate overlaps, pauses and other disfluencies. Re-enacting an utterance or exchange may take several attempts to get right. The one-hour session typically yields a few dozen matched pairs with overall good pragmatic diversity.

To further encourage high quality data collection, the operator implements additional measures, such as alternating the starting language to avoid order bias and instructing participants in the same language as their conversation to prevent cross-lingual influence.

Post-session, the operator scrubs through the audio recordings and matches the reenacted utterances back to the original utterances. A post-processing script subsequently collects the matched utterances, their source conversation recordings, and associated metadata into a corpus. The post-processing code, along with other project-related code, is available at the DRAL GitHub repository[1].

[1] `https://github.com/joneavila/DRAL`

## 2.3 The Dialogs Re-enacted Across Languages Corpus

The DRAL corpus consists of 3816 pairs of English and Spanish utterances collected from 128 conversation pairs.

Each pair of utterances was produced by the same speaker to avoid complications due to speaker differences. In addition to maintaining quality control during the data collection phase, the two operators (both research colleagues) participated in a quality control check we conducted as a group, by which point only a few additional poor-quality utterances needed to be excluded from the corpus. Thus, while the utterance pairs do not represent all possible translations, I believe their prosody is translated faithfully and use them as the ground truth for prosody translation in this research.

The matched utterances exhibit overall good pragmatic diversity, as suggested by the examples in Figure 5.3 and Appendix B.

We have made the DRAL corpus public as a resource for speech research, including for evaluating speech-to-speech translation models. The most recent version of the DRAL corpus and our technical report are available at its home page[2]. Additional statistics are included in Table 2.1.

Creating DRAL was a joint effort, in which my specific contributions include writing the post-processing code, assisting with the design of the protocol, the training of research assistants for data collection, checking for quality, and writing the technical report.

The DRAL corpus is a novel resource for speech-to-speech translation research, useful for analyzing prosody in English and Spanish dialog. It serves as the primary resource for the research presented herein.

[2] `https://cs.utep.edu/nigel/dral/`

Table 2.1: DRAL corpus statistics, subset of English-Spanish utterances pairs.

| | |
|---|---|
| Conversation pairs | 128 |
| Unique participants | 69 |
| Utterance pairs | 3816 |
| Mean duration of utterance | 2.7s |
| Total duration of corpus | 344.5 m |

# Chapter 3

# A Representation of Utterance Prosody

A representation of speech encodes the aspects relevant to the task at hand. Like the representations mentioned in Chapter 1, representations have trade-offs, balancing between interpretability, abstraction, utility for learning, and precision. For examining the prosody of utterances from dialog, a representation should encode the human-relevant aspects of speech and be relevant to dialog.

In this chapter, I first describe different categories of prosodic features and software tools for extracting these features. I then present a new representation of utterance prosody constructed from various prosodic features and describe how I verified these features.

## 3.1 Prosodic Feature Computation Software

Prosodic features are generally classified based on their computation method and degree of abstraction.

Features can be categorized into acoustic or perceptual features. Acoustic features are objective measures. These features are directly measured from the audio signal and include fundamental frequency ($F_0$) and intensity. In contrast, perceptual features are subjective measures. These features are human percepts, i.e., may be perceived differently from listener to listener, and include pitch and volume. In practice, perceptual features are derived from acoustic features.

Features are also categorized by their level of abstraction into low-level, mid-level, and high-level features. This categorization is not as strictly defined as acoustic and perceptual features.

Low-level features are computed at the frame-level, typically every 10 ms. These include pitch and energy. Mid-level features are derived from low-level features and span longer time windows. High-level features represent an even greater level of abstraction. For instance, TOBI (Tones and Break Indices) labels represent English prosody used in (things like) indicating the important words or in distinguishing a statement from a question [65].

High-level features may be language-dependent and may not be easily computed automatically. Such is the case with TOBI labels, which are specific to English intonation, and can be computed automatically with high accuracy for some labels and poor accuracy for others [61].

The proposed prosody representation is based on mid-level features because they can be easily computed from low-level features, are language-independent, and can be computed over spans suitable for a relatively compact fixed-length representation.

Many software tools exist for extracting features from audio, most being part of a toolkit, or a collection of software tools. Noteworthy among these are: *openSMILE*, an audio analysis and processing toolkit targeted at speech and music applications [21]; *Praat*, a speech analysis toolkit with additional support for speech manipulation [6]; *SpeechBrain*, a toolkit for developing a variety of speech systems, including speech recognition, speaker recognition, and text-to-speech [58]; *COVAREP* a speech processing toolkit, for speech analysis, synthesis, conversion, and others [13]; *Kaldi*, a speech recognition toolkit [54]; and *Proscript*, for linguistically aligned prosodic features ($F_0$, intensity, speech rate) [47].

For this research, I chose to use the Mid-level Prosodic Features Toolkit [74] (henceforth referred to as the Mid-level Toolkit). Its features were designed to be robust for dialog data, generally perceptually relevant, and normalized per speaker. I expand on its feature computation and the importance of normalization in the next section.

## 3.2 Prosodic Feature Set

From the available features in the Mid-level Toolkit, I selected ten base features based on previous utility for many tasks across several languages [77]. Specifically, the ten features are: intensity, lengthening, creakiness, speaking rate, pitch highness, pitch lowness, pitch wideness, pitch narrowness, peak disalignment (primarily late peak), and cepstral peak prominence smoothed (CPPS). These features are detailed in Figure 3.2. Such features have been previously shown to be useful for investigating the extent to which prosody can be used to infer stances as they occur in radio news stories [75], as well as the role of prosody in coordinating action [81].

As the aim of this research is exploratory, I use a simple, fixed-length representation, meaning all utterances are represented by the same number of features. A fixed-length representation will make analyses easier, for example, when comparing the prosody of two utterances.

The goal is to represent any utterance with enough features to encode its prosody while being manageable enough for interpretation. To characterize the prosody of an utterance, each base feature is computed over ten non-overlapping windows, together spanning the whole utterance. Thus, each utterance is represented by 100 features. The window sizes are proportional to an utterance's duration and span fixed percentages of its duration: 0–5%, 5–10%, 10–20%, 20–30%, 30–50%, 50–70%, 70–80%, 80–90%, 90–95%, 95–100%. This prosody representation is illustrated in Figure 3.1.

My decision to compute prosodic features towards the boundaries of an utterance over smaller spans was rooted in the observation that prosodic configurations towards these boundaries often exist in shorter durations. In contrast, the prosodic configurations towards the center of an utterance tend to vary more slowly. Thus, using spans of varied sizes encodes these changes more accurately. The resulting representation is not aligned to words or syllables, instead aiming to represent the sorts of overall levels and contours that are most often associated with pragmatic functions.

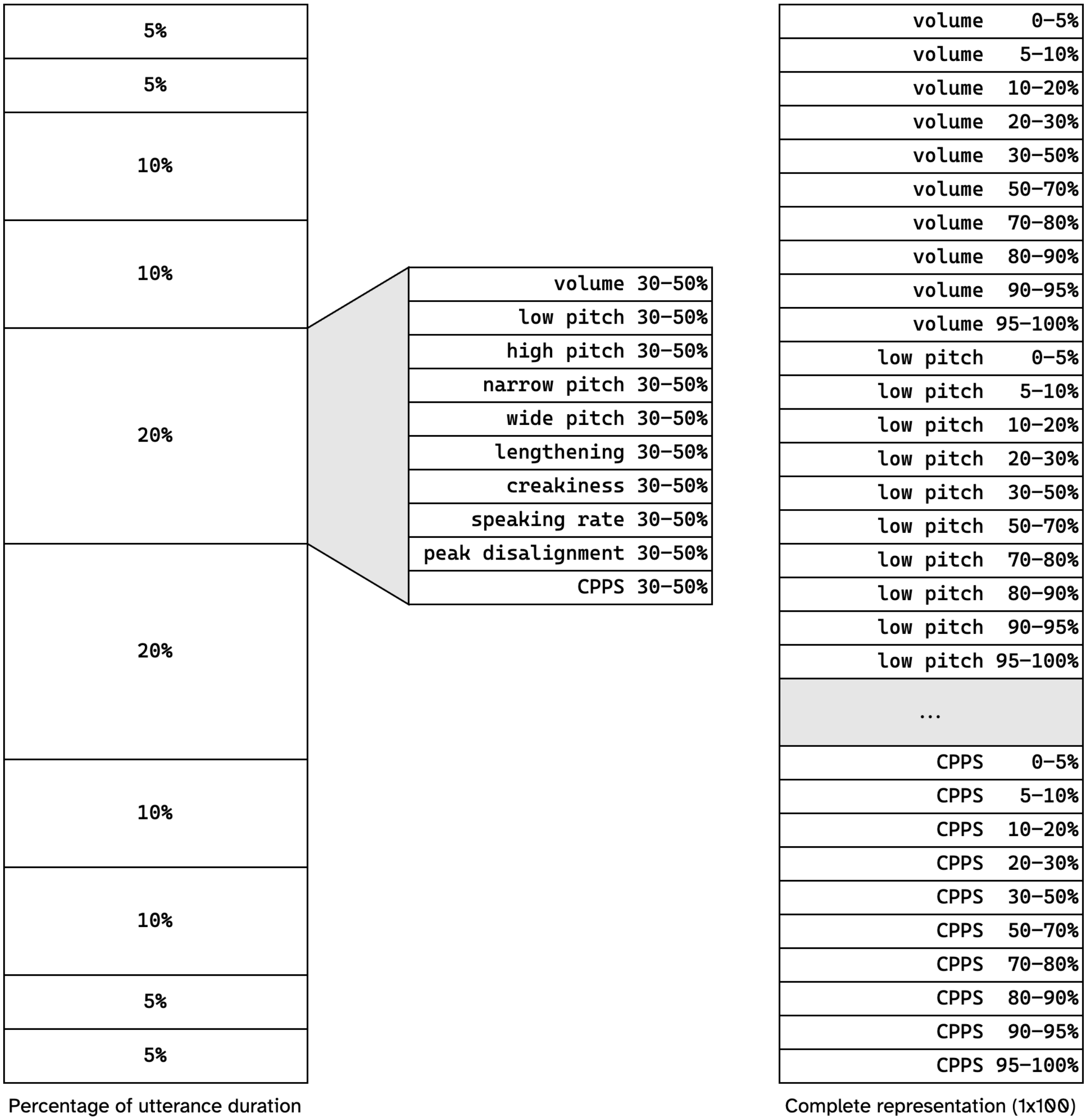


Figure 3.1: Prosody representation. Ten base prosodic features are computed at ten non-overlapping windows spanning the duration of the utterance. Base features: intensity, pitch lowness, pitch highness, narrow pitch range, wide pitch range, lengthening, creakiness, speaking rate, peak disalignment, cepstral peak prominence smoothed.

We can learn from studies investigating prosody across different segments of utterances, and not only limited to English and Spanish. These studies yield observations that are relevant to the modeling of cross-lingual prosody mappings. These observations include the tendency to lengthen the final vowel preceding a pause [8], the tendency to pause during an utterance when planning upcoming utterances [23], and the use of pitch contour of the final syllable to differentiate between declarative and interrogative statements [70].

I use the Mid-level Prosodic Features Toolkit for most of the prosodic feature computation. My modifications to the Mid-level Toolkit include those for computing fixed-length representations from variable-length utterances. I describe all modifications I implemented in more detail in Appendix A.

Normalization occurs at two stages during the feature computation: the first to address individual speaker differences, and the second to align the features to a similar scale. The first normalization is applied to the low-level (frame-level) features, where each low-level feature is normalized on a per-track basis. As a result of this normalization, for example, features measuring wide pitch range are adjusted to accommodate a speaker with a typically dynamic pitch compared to a speaker with a typically flat pitch. The second normalization is applied to the mid-level features. After computing each of the 100 features for all utterances within a track, these features are z-normalized, such that each has a mean of zero and a standard deviation of one.

## 3.3 Verifying Utterance-Level Prosodic Features

After implementing the modifications to the Mid-level Prosodic Features Toolkit mentioned above, I wanted to verify the computed features. To do this, I recorded a contrived English conversation in which I assumed the role of one of the two interlocutors. Given the objective was to create exemplars of utterances with low or high feature values, these utterances do not mirror the naturalness of utterances drawn from the DRAL corpus.

For the purpose of visual analysis, I plotted the features of these exemplar utterances.

1. **Creakiness** is a measure of jittery variations in pitch that indicate weak periodicity. It is also known as "vocal fry."

2. **Cepstral peak prominence smoothened (CPPS)** is an effective measure of breathy voice in clinical applications [27] and is relevant to production of pragmatic functions[26]. Low CPPS correlates with breathy voice, while high CPPS correlates with more harmonic voice.

3. **Narrow pitch range** measures how strongly the speakers' pitch is in a specified narrow range.

4. **Wide pitch range** measures how strongly the speakers' pitch is in a specified wide range.

5. **Speaking rate** is a crude estimate of how slowly or quickly the speaker is talking. It is based on spectral flux, which measures the rate of change of the power distribution of different frequency components in the speech signal. These changes correspond to the transitions between speech sounds, such as consonants and vowels.

6. **Pitch lowness** measures how strongly the speakers' pitch is in a specified low band.

7. **Pitch highness** measures how strongly the speakers' pitch is in a specified high band.

8. **Intensity** measures how quietly or loudly the speaker is talking.

9. **Lengthening** crudely measures lengthening of vowels.

10. **Peak disalignment** measures the degree of disalignment between pitch and intensity peaks. Low values for peak disalignment (or high peak alignment) are typical of English vowels. High peak disalignment, specifically late pitch peak, serves many functions in English, including making suggestions, making offers, and grounding [77, Chapter 6].

Figure 3.2: Prosody representation base features.

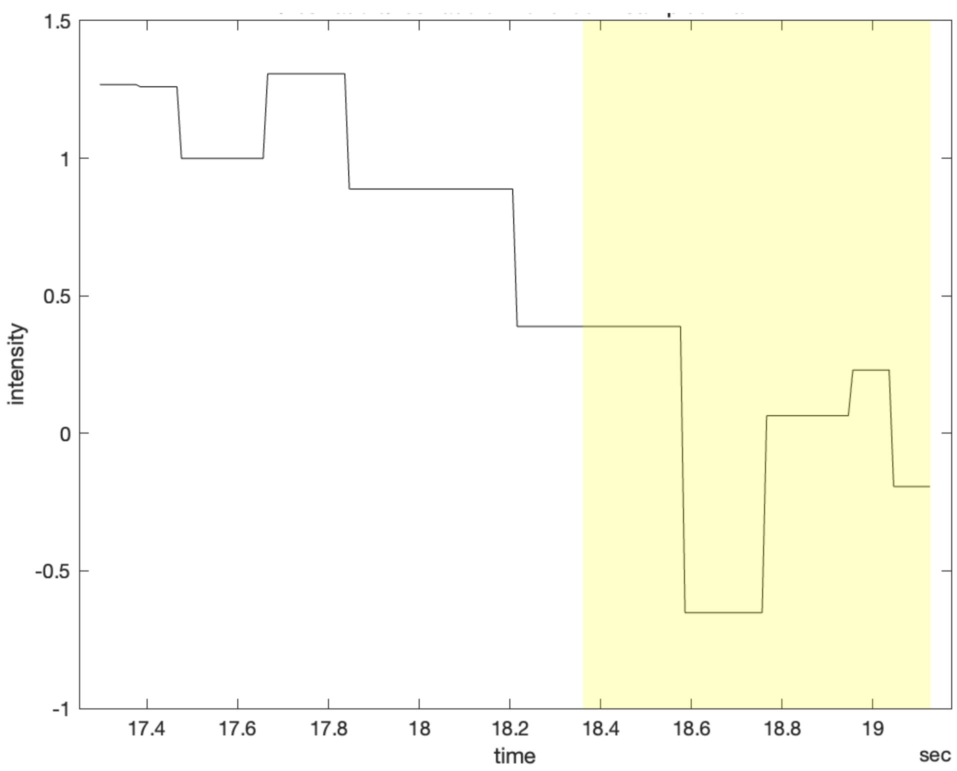


Figure 3.3: Example utterance intensity plot. The utterance *avoid spoilers* begins with high intensity and ends with low intensity, relative to the individual speaker's baseline. The highlighted segment was previously annotated for low intensity, confirming that the feature computation is consistent with the plotted data. Audio of the utterance is available at `https://jonavila.dev/dissertation`.

More specifically, for each exemplar, I plotted the base feature for which the utterance was intended to exhibit both low and high values, relative to the average value of that feature for the source conversation. As an example, Figure 3.3 shows the plot of an utterance with both low and high intensity. From this analysis, I concluded that the feature computation for utterances was working as expected. The conversation audio and remaining plots are included in the repository from Appendix A.

# Chapter 4

# Cross-Language Prosodic Feature Correlations

In this chapter, I present an analysis of the correlations between the prosodic features of equivalent English and Spanish utterances. To conduct this analysis, I use the DRAL corpus of English and Spanish utterances from dialog, as presented in Chapter 2, and the representation of utterance prosody, as presented in Chapter 3.

The pragmatic diversity of the corpus and wide range of prosodic features of the representation provide a rich set of data for this analysis. This data allows for a comparative analysis of English and Spanish prosody as used in dialog, the first of its kind to examine a wide range of prosodic features.

An examination of English and Spanish prosodic features can yield insights into how pragmatic meaning is conveyed with prosody across these languages. These insights may characterize prosodic equivalences, by which I mean the manner in which the prosody of one language corresponds with that of the other. Additionally, these insights might characterize language-specific phenomena that do not possess direct equivalents in the other.

Specifically, I examine the correlations between English and Spanish prosodic features. Because these are computed from utterances with examples of many pragmatic functions, the observed correlations are likely to imply patterns that are broadly prevalent throughout the utterances, rather than patterns that are specific to a particular pragmatic function. These observations provide a sense of the overarching relationship between English and Spanish prosody and serve as a first glimpse at the mappings of their prosodic patterns.

## 4.1 Measures of Dependence

As with any correlation computation, it is important to consider the characteristics of the data to select an appropriate measure of dependence. Factors to consider include the distribution of prosodic feature values and the degree of outliers. To inform my selection of the measure of dependence, I created histograms for each corresponding English and Spanish feature. The feature values do not generally adhere to a normal distribution. Given the non-normality of the feature values, I chose to use Spearman's rank correlation coefficient as the dependence measure. This measure does not predicate a specific distribution and serves as a non-parametric alternative to Pearson's correlation coefficient, another common measure dependence measure.

The value of Spearman's rank correlation ranges between $-1$ and $+1$, with $+1$ indicating a perfect monotonically increasing relationship and $-1$ indicating perfect monotonically decreasing one. Rather than using raw values, the variables are converted into rank variables, with each value being categorized as higher than, lower than, or equivalent to all other values.

The Spearman's rank correlation coefficient, $r_s$ is computed using the formula:

$$r_s = \frac{cov(R(X), R(Y))}{\sigma R(X) \sigma R(Y)}$$

where $R$ is a rank variable, $cov$ is covariance, and $\sigma$ is standard deviation.

## 4.2 Observations

In this section, I present observations from the analysis of the correlations between English and Spanish prosodic features[1].

Specifically, I computed the Spearman correlations between the 100 prosodic features across all matched English and Spanish pairs in the DRAL corpus (3816 pairs, sourced

[1] These observations were presented as part of our Interspeech paper [2].

from Conversation 001–029, 032–040, 043–045, 050–136). I also computed the correlations for feature values within each language for comparison.

Figure 4.1 shows the correlations between English and Spanish feature values. Figure 4.2 and Figure 4.3 show the correlations among English feature values and among Spanish feature values, respectively. Lastly, Figure 4.4 shows the difference between Figure 4.1 and Figure 4.2.

Were English and Spanish prosodically identical, we would expect each English prosodic feature to correlate perfectly with its Spanish counterpart. In fact the correlations were far more modest, but always positive and often substantial: more than half the features with the same base feature and span correlate positively $\rho \geq 0.3$. Thus, overall, English and Spanish prosody is quite similar. Pitch highness is generally the most similar, especially towards the middle of utterances (e.g., 30–50%, $\rho = 0.56$).

While some features, such as pitch highness, have stronger span-for-span correlations, other features, notably speaking rate, lengthening, and CPPS, have correlations that are strong throughout the utterances. For example, speaking rate at every span in an English utterance correlates with speaking rate at every span in the corresponding Spanish utterance (Figure 4.1). These findings are compatible with the idea that English and Spanish prosody is overall roughly similar, but that the locations of local prosodic events can vary, likely due to differences in word order and lexical accents.

However, some correlations were much weaker. The lowest cross-language correlations for the same features were for creakiness and peak disalignment, suggesting that these are likely to have different functions in the two languages. There were also many off-diagonal correlations. Most of these were unsurprising, such as the negative correlation between the speaking rate and lengthening features, since as speech rate decreases, the duration of individual sounds, including vowels, increases.

However, not all off-diagonal correlations were expected. As an example, I expand on the correlations between English intensity and Spanish CPPS in the next section.

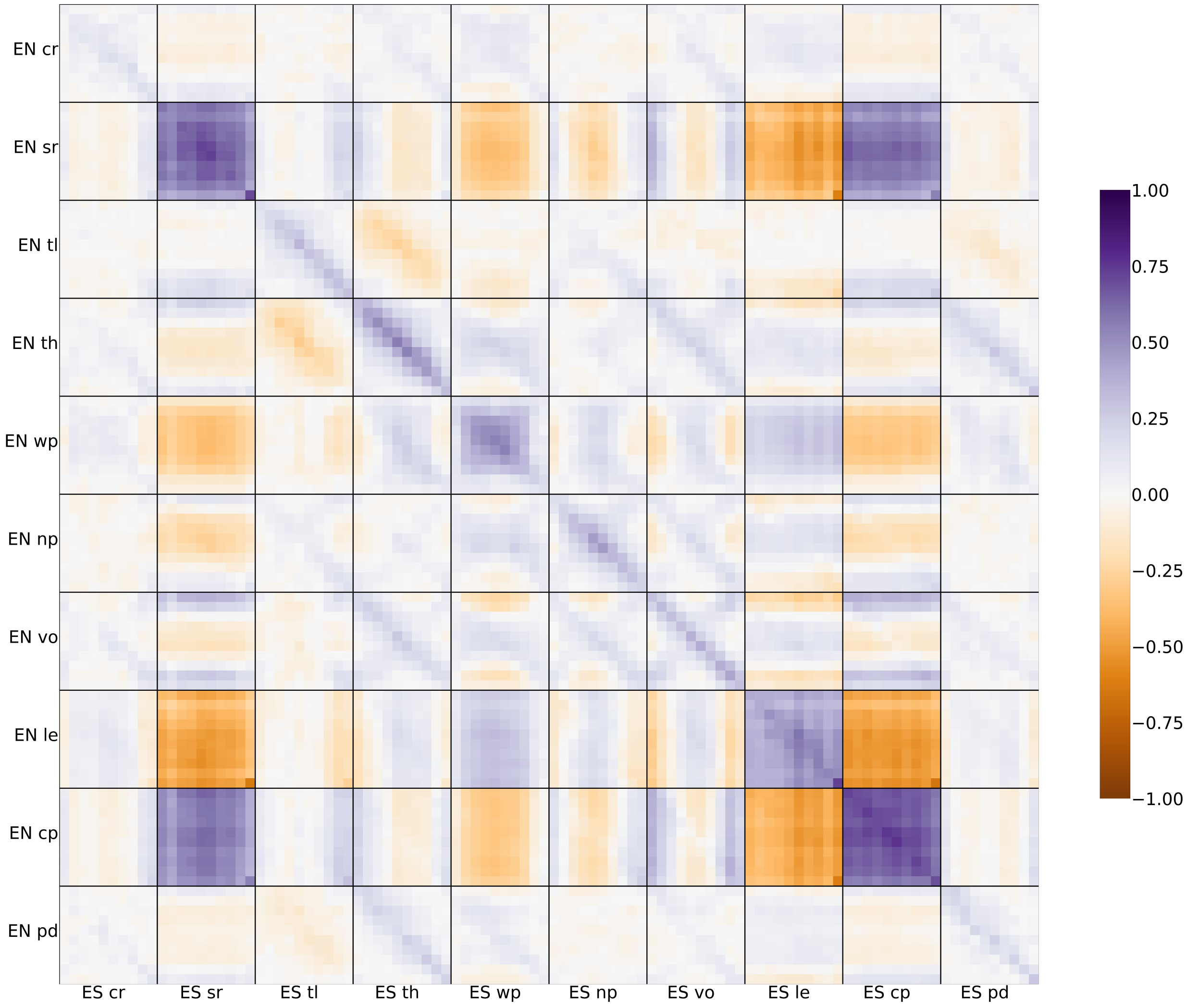


Figure 4.1: Spearman correlations between English and Spanish prosodic feature values.

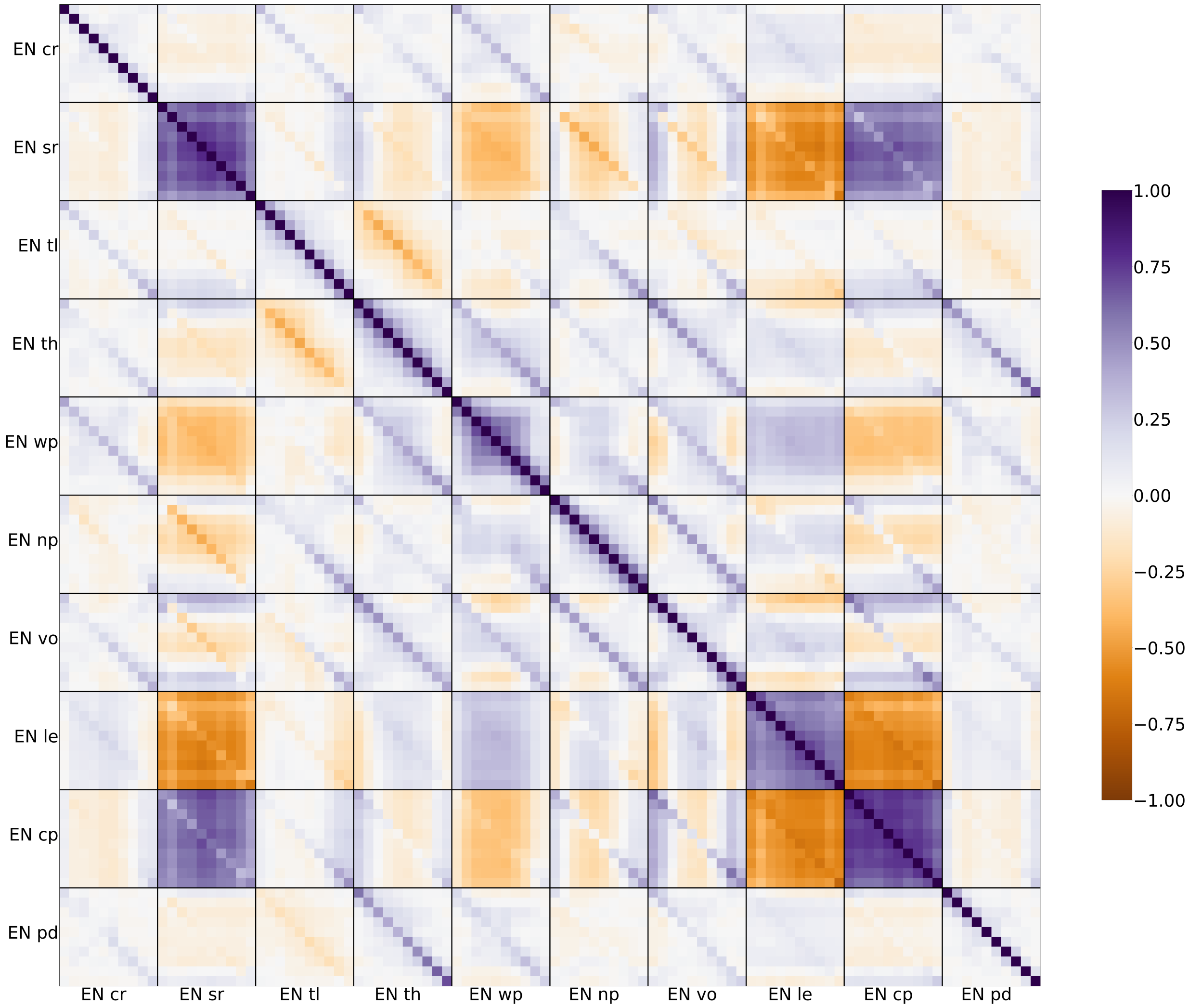


Figure 4.2: Spearman correlations among English prosodic feature values.

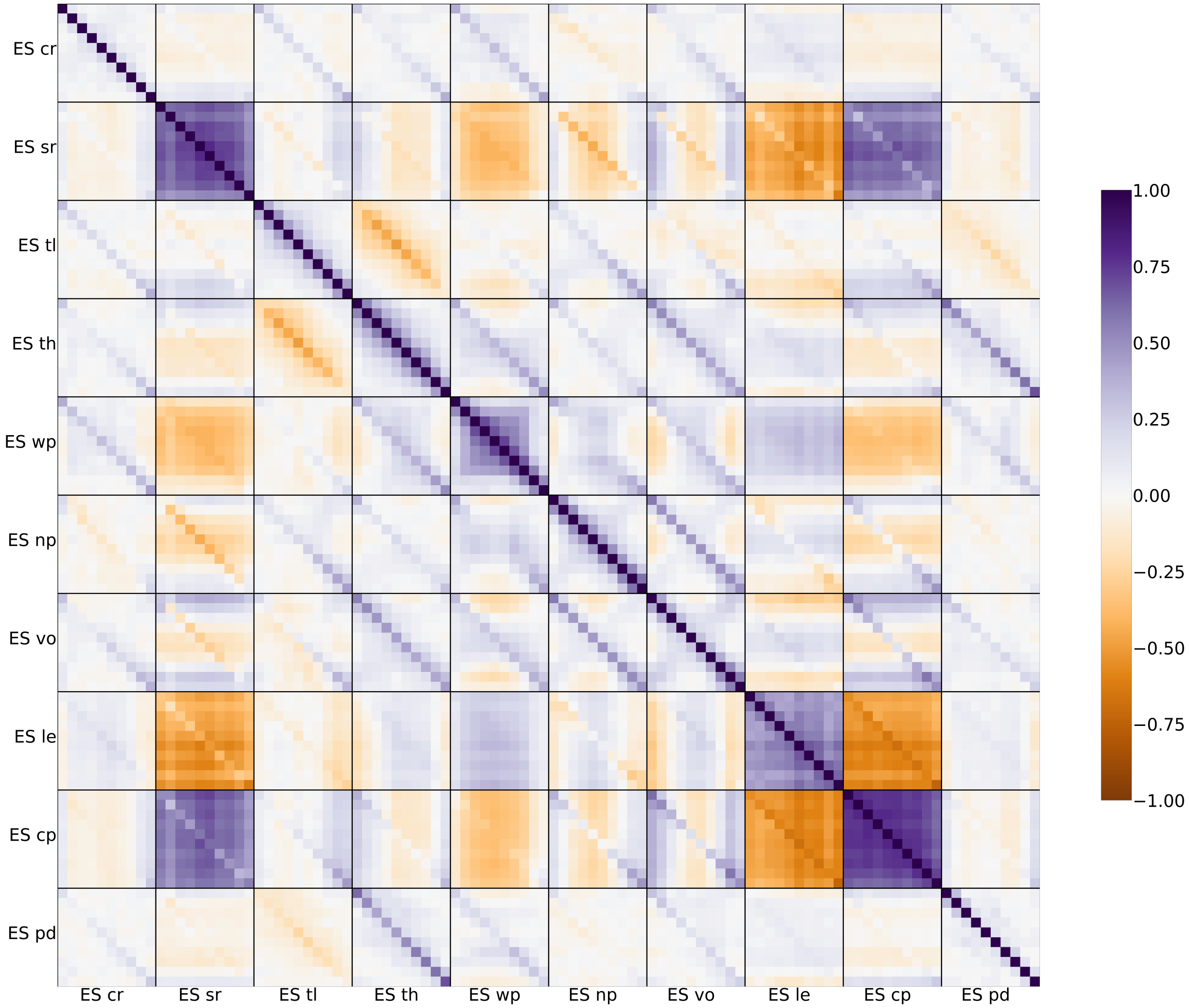


Figure 4.3: Spearman correlations among Spanish prosodic feature values.

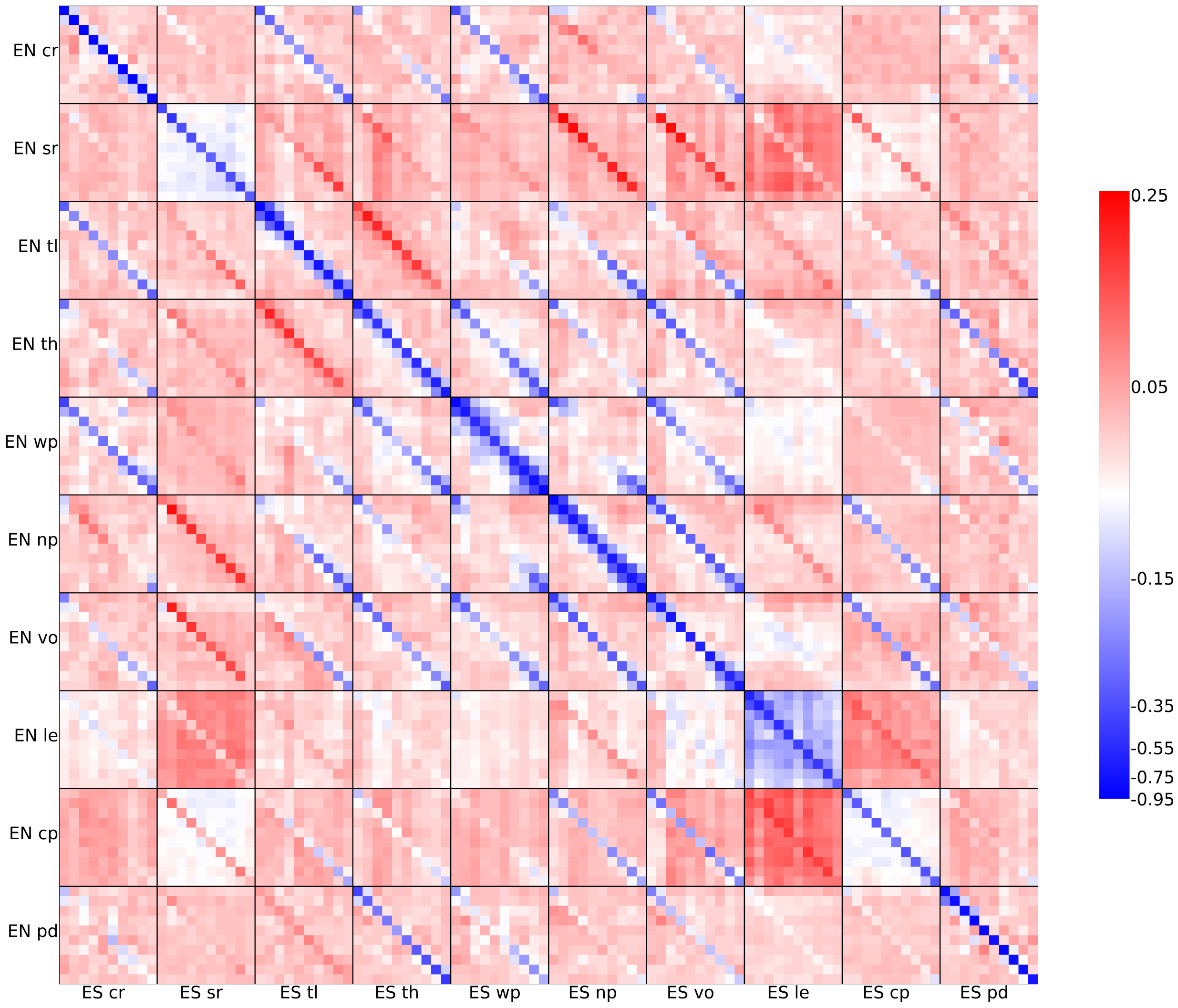


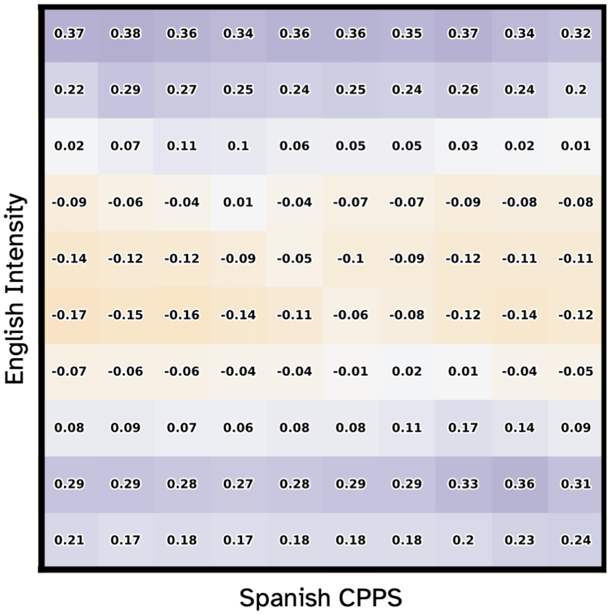
Figure 4.4: Difference between: (a) Spearman correlations between English and Spanish prosodic feature values, and (b) Spearman correlations among English prosodic feature values. This subtraction of correlations is purely for descriptive analysis.

## 4.3 Correlations Between English Intensity and Spanish CPPS

Thus far, I have discussed correlations between English and Spanish prosodic features that are found within each language. Additionally, I have discussed correlations between prosodic features differing in base feature and span. These correlations have been "symmetrical" in the sense that they hold true irrespective of whether the first set of features correspond to English and the second to Spanish, or the reverse. In this section, I discuss a set of correlations that are "asymmetrical", valid solely in one direction: the correlations between English intensity and Spanish CPPS.

Intensity at the start and end of an English utterance correlates with CPPS throughout a Spanish utterance (EN 90–95% vs. ES 80–100%, $\rho \geq 0.3$) (Figure 4.5), while no such relationship was found within either language. Examination of the ten pairs that most closely reflect this pattern (English high near-final intensity and Spanish high CPPS), showed that in half the speaker is preparing a follow-up explanation. Indeed, all of these utterances could be sensibly followed by the word *because*. Thus, we have identified a pragmatic function that seems to be prosodically marked differently in English and Spanish. Figure 4.6 shows the values for these two features for one such pair.

The cross-language prosodic feature correlations also showed that many of the features had weak correlations. Reducing the dimensionality of the representation might improve the performance of the models. A reduced representation is the topic of Chapter 7.



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.37 | 0.38 | 0.36 | 0.34 | 0.36 | 0.36 | 0.35 | 0.37 | 0.34 | 0.32 |
| 0.22 | 0.29 | 0.27 | 0.25 | 0.24 | 0.25 | 0.24 | 0.26 | 0.24 | 0.2 |
| 0.02 | 0.07 | 0.11 | 0.1 | 0.06 | 0.05 | 0.05 | 0.03 | 0.02 | 0.01 |
| -0.09 | -0.06 | -0.04 | 0.01 | -0.04 | -0.07 | -0.07 | -0.09 | -0.08 | -0.08 |
| -0.14 | -0.12 | -0.12 | -0.09 | -0.05 | -0.1 | -0.09 | -0.12 | -0.11 | -0.11 |
| -0.17 | -0.15 | -0.16 | -0.14 | -0.11 | -0.06 | -0.08 | -0.12 | -0.14 | -0.12 |
| -0.07 | -0.06 | -0.06 | -0.04 | -0.04 | -0.01 | 0.02 | 0.01 | -0.04 | -0.05 |
| 0.08 | 0.09 | 0.07 | 0.06 | 0.08 | 0.08 | 0.11 | 0.17 | 0.14 | 0.09 |
| 0.29 | 0.29 | 0.28 | 0.27 | 0.28 | 0.29 | 0.29 | 0.33 | 0.36 | 0.31 |
| 0.21 | 0.17 | 0.18 | 0.17 | 0.18 | 0.18 | 0.18 | 0.2 | 0.23 | 0.24 |

Figure 4.5: Spearman correlations between English intensity and Spanish CPPS feature values.

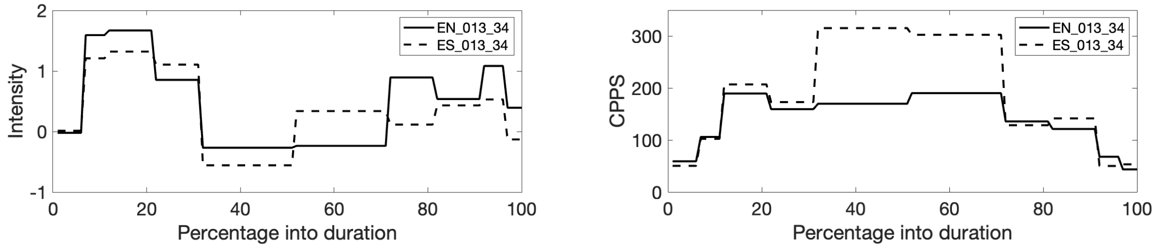


Figure 4.6: Example utterance pair with English high near-final intensity and Spanish high CPPS. EN_013_34: *If you have an undergrad in anything, you can just, skip to a Master's in anything else.* ES_013_34: *Si tienes carrera en cualquier cosa, puedes brincar a la maestría en lo que sea.* Audios of utterances are available at `https://jonavila.dev/dissertation`.

# Chapter 5

# A Metric for Prosodic Similarity Between Utterances

To identify utterances that pose challenges for modeling the mapping of English and Spanish prosody, I will need a metric for quantifying the prosodic similarity between a pair of utterances. Specifically, I will use this metric to compare the predicted target-language prosody, as inferred from the source-language prosody, against the target-language prosody of the reference human-produced utterance. In this chapter, I propose a simple metric for gauging the prosodic similarity of two utterances, and assess its reliability as a proxy for human judgment.

A metric for similarity should consider the complex relationship between prosody and its pragmatic functions. For instance, two utterances may convey similar meanings despite having different content. Conversely, two utterances with similar content may convey different meanings, such as the two *yeah* utterances from Chapter 1, where the prosody can convey a request for clarification, an empathetic acknowledgment, or subtle disagreement[1].

As discussed in Section 1.4, existing metrics commonly used in evaluating speech-to-speech translations compare utterances based solely on their content, ignoring the prosody that contributes to the utterance's meaning. Thus, a metric of similarity should consider prosody beyond that inherently tied to specific words, whether these belong to the content or its surrounding context. While metrics based on prosody exist, these are primarily limited to pitch and overlook other aspects of prosody contributing to perceptions of similarity. To date, no metric is designed to estimate the prosodic similarity of utterances, particularly

[1] Audios of these utterances are available at `https://joneavila.github.io/dissertation/`.

those from dialog. Such a metric would pivot towards improving pragmatic fidelity, where the aim is not merely to produce prosody that is perceived as natural, but also perceived as conveying the appropriate meaning. I therefore propose a metric for estimating the similarity of two utterances, potentially with a different word sequence, incorporating aspects of prosody beyond pitch.

## 5.1 Metric Definition

The proposed metric estimates the similarity of two utterances as the inverse Euclidean distance between their respective prosody representations, as computed in Chapter 3. Thus, all features contribute equally in estimating similarity. The metric is defined as follows:

$$s(p, q) = \frac{1}{d(p, q)}$$

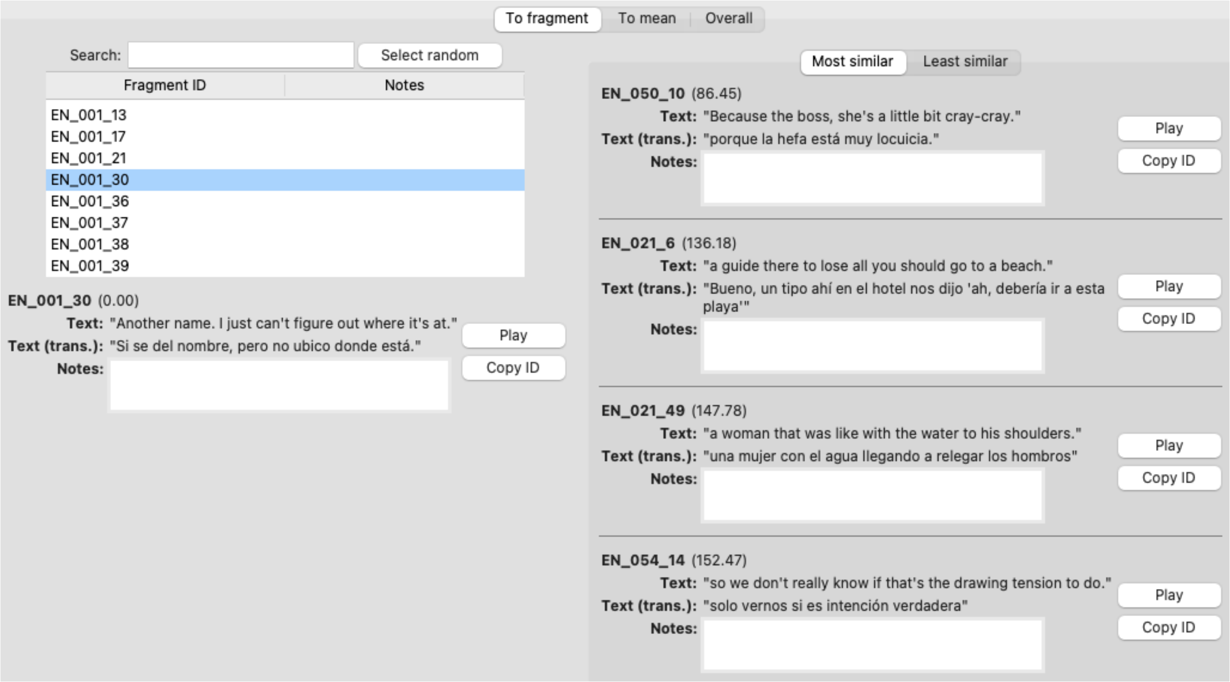
$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_{100} - q_{100})^2}$$

The similarity $s$ is computed from the prosody representations $p$ and $q$. The similarity is a non-negative floating-point value, where values closer to zero indicate greater similarity.

While this metric considers many prosodic features relevant to dialog, I do not expect it to accurately match human perceptions of similarity. Instead, I expect for the metric to be useful in identifying prosody that is challenging in translation. To identify the utterances where the predicted prosody diverges most from the reference prosody, this metric should at least be able to distinguish highly similar utterances from highly dissimilar utterances.

## 5.2 Performance as a Proxy for Human Judgments

To assess the reliability of the proposed prosody similarity metric, I compared the predicted highly similar ("close") and highly dissimilar ("far") within-language utterances to my



| Fragment ID | Notes |
|---|---|
| EN_001_13 | |
| EN_001_17 | |
| EN_001_21 | |
| EN_001_30 | |
| EN_001_36 | |
| EN_001_37 | |
| EN_001_38 | |
| EN_001_39 | |

**EN_001_30** (0.00)
**Text:** "Another name. I just can't figure out where it's at."
**Text (trans.):** "Si se del nombre, pero no ubico donde está."
**Notes:**

**EN_050_10** (86.45)
**Text:** "Because the boss, she's a little bit cray-cray."
**Text (trans.):** "porque la hefa está muy locuicia."
**Notes:**

**EN_021_6** (136.18)
**Text:** "a guide there to lose all you should go to a beach."
**Text (trans.):** "Bueno, un tipo ahí en el hotel nos dijo 'ah, debería ir a esta playa'"
**Notes:**

**EN_021_49** (147.78)
**Text:** "a woman that was like with the water to his shoulders."
**Text (trans.):** "una mujer con el agua llegando a relegar los hombros"
**Notes:**

**EN_054_14** (152.47)
**Text:** "so we don't really know if that's the drawing tension to do."
**Text (trans.):** "solo vernos si es intención verdadera"
**Notes:**

Figure 5.1: Screenshot of custom application used in evaluation of prosody similarity metric. The left panel displays the selected utterance. The right panel displays the utterances most similar or most dissimilar to the selected utterance as estimated by the similarity metric.

judgments, which I formed from discussing observations with one other native speaker of English.

To structure this process, I wrote a custom application to randomly select an utterance from the data and retrieve the four utterances most similar to, and four utterances most dissimilar to, the selected utterance as estimated by the metric. This application, pictured in Figure 5.1, and its additional functionalities is documented in the repository from Appendix A.

In the observation stage, we listened to the selected focus utterance and its close and far utterances. We repeatedly listened and identified any similarities and dissimilarities we could note, taking 2 or 3 minutes per pair to do so. Most of these observations were at the level of pragmatic function, rather than prosodic features.

After these discussions, I judged whether each estimate of close or far utterance was indeed significantly more similar or dissimilar compared to the opposite class. I labeled each of the estimates as “truly close,” “truly far,” “falsely close,” or “falsely far.” While perceptions were almost always shared, sometimes weakly similar pairs made it difficult to judge one way or the other and some final judgments were not in agreement.

This procedure is depicted in Figure 5.2. We completed the process above for seven focus utterances and eight comparisons utterances each, all from the English half of the data. Observation notes on all sets of utterances are included as Appendix C.

The observations indicated that the metric captures many aspects of pragmatic similarity, including speaker confidence, revisiting unpleasant experiences, discussing future plans, describing sequences of events, and describing personal feelings. All of these aspects of pragmatic similarly were also generally prosodically similar.

Figure 5.3 shows one set of utterances to illustrate. The prosody of this focus utterance suggested that the topic is personal feelings: an inverted U-shape speaking rate, a pause, and occasional use of creaky voice. Each utterance estimated as similar by the metric shared these qualities to varying degrees.

The similarities found were not generally lexically governed. While some words and syntactic structures have characteristic prosody, and some pairs considered similar by the metric shared lexical content, generally prosodic similarity seemed to be not be subsumed by lexical similarity. This observation aligns with the aim of the metric, as the pragmatic similarity of utterances does not always correspond to their lexical similarity.

Consider, for instance, two utterances with prosody suggesting different functions despite their lexical similarities: utterances EN_025_1 and EN_025_7, as illustrated in Figure 5.3. The prosody of EN_025_1 suggests that the speaker is interested in learning about the interlocutor and is careful in asking about their preferences. In comparison, the prosody of EN_025_7 more saliently suggests that the speaker is holding the floor and leading up to something more interesting: a faster changing speaking rate, a higher number of pauses, and a use of creaky voice throughout the utterance.

For 50 of the 56 utterances examined, the judgments aligned with the metric’s estimates,

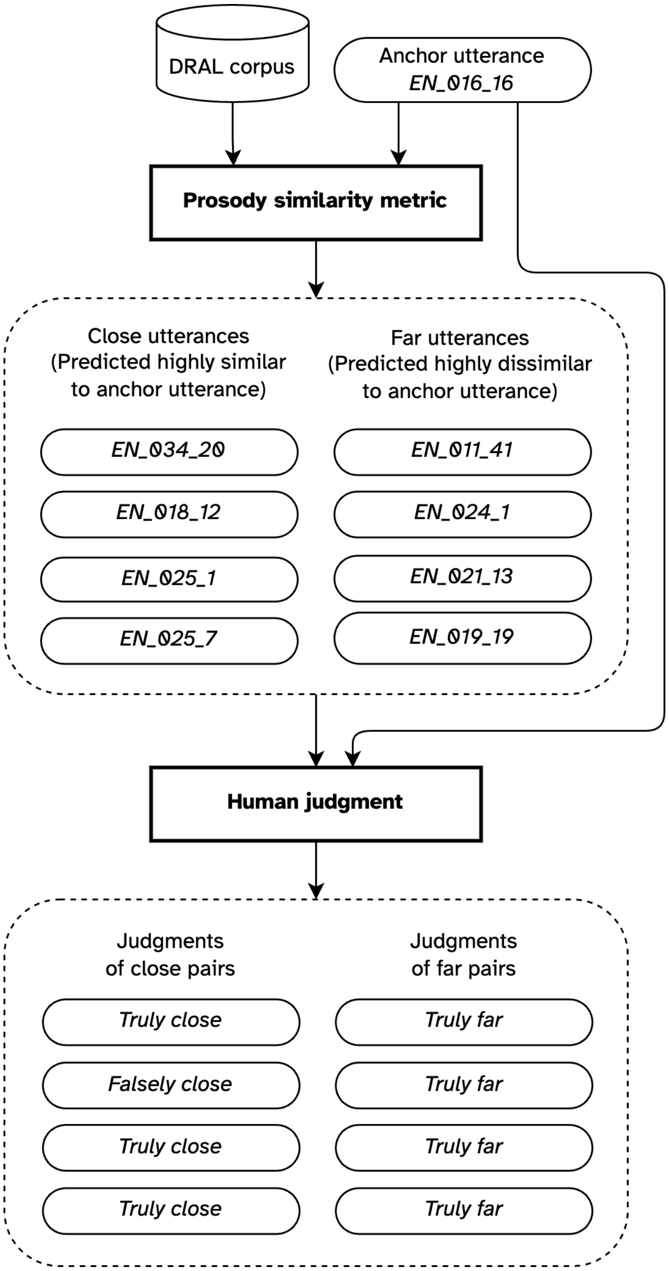


Figure 5.2: Similarity metric analysis procedure.

EN_016_16 *I would be kind of scared to ask questions to the professor or…*

(a) Focus utterance.

EN_034_20 *It's like, I would do meds, but in a lotion form.*

EN_018_12 *What have been like, some challenges for you in your career?*

EN_025_1 *So overall, what music do you prefer to listen to?*

EN_025_7 *So I have to pick music that I like, but also that people…*

(b) Close utterances, or utterances estimated as highly similar to the focus utterance.

EN_011_41 *And I really like Mejia because he is the one always like telling me "Hey, you should apply to this, you should apply to this"*

EN_024_1 *So uh yesterday you were telling me about, like, a weird, like, experience you had with the cops in Mexico, right?*

EN_021_13 *And the beach is really strange because it's like a, you see, like the beach is not like a straight line. It was like a doughnut.*

EN_019_19 *But do you think that someone who hasn't seen a Marvel move can just watch any movie? Or is there any specific movies they have to watch?*

(c) Far utterances, or utterances estimated as highly dissimilar to the focus utterance.

Figure 5.3: An utterance (a) and utterances estimated as highly similar (b) and highly dissimilar (c) by the prosody similarity metric

Table 5.1: Prosody similarity metric performance matrix ($n = 56$, $\chi^2 = 34.7$, $p < .01$).

| | Predicted similar | Predicted dissimilar |
|---|---|---|
| Judged similar | 24 | 2 |
| Judged dissimilar | 4 | 26 |

as summarized in Table 5.1. The similarity metric clearly performs better than chance in its estimates of the most similar and most dissimilar utterances. From this analysis I deemed the similarity metric reliable enough to use it for evaluating the match between target language prosody as inferred from the source-language prosody and the human-produced target language reference prosody.

However, the metric does not appear to always match perceptions. To try to understand the limitations of the metric and identify potential areas for improvement, we examined the six utterances where our judgments diverged most from the metric's estimates of their similarity to the respective focus utterance. Four of these utterances, including EN_025_1 in Figure 5.3 were estimated as highly similar to the focus utterance yet sounded rather different to us. The remaining two utterances, including EN_024_1 in Figure 5.3, were estimated as highly dissimilar to the focus utterance despite me feeling they had significant similarities. Among these six utterances, two had very salient differences in nasality — not directly represented by the features of the prosody representation — and sounded very different in terms of pragmatic function, specifically with respect to presumption of common ground.

For three of the utterances, the disparity between the utterance and focus utterance seemed to be attributable to differences in syllable-aligned pitch and energy contours, i.e., the variation or pattern of pitch and energy associated with the syllables in the speech.

These aspects are also not directly represented by the features of the prosody representation.

## 5.3 Correlations Between Prosodic Feature Values and Utterance Duration

During the analysis of the similarity metric, I noticed that utterances estimated as highly similar often shared similar durations. Ideally, the duration of an utterance should not obscure the other prosodic features of similarity; otherwise, the metric would be less reliable when similarity in duration does not reflect similarity in prosody.

Of the pairs of utterances we examined, I judged most pairs estimated as highly similar that shared similar durations to be truly similar. Additionally, I did not find examples of pairs with perceptually similar prosody yet significant different duration. However, the subset of utterances I examined was too small to be representative of all of English.

To investigate the relationship between utterance duration and the various prosodic features, I computed their correlations. Correlations between utterance duration and prosodic features would suggest the possibility of a misleading relationship between utterance duration and estimated similarity.

Figure 5.4 shows the correlations between utterance duration and the prosodic features, as well as the average correlations for features sharing the same base feature and features sharing the same span.

The first observation is that utterance duration correlates, on average, with some prosodic features sharing the same base feature:

- The correlation with speaking rate features (mean($\rho$) $= 0.62$) may be due to speakers' tendency for lowering their speaking rate in shorter utterances. For instance, [83] found that speaking rate in English conversational speech, which most of these are, rises rapidly for turns of one to seven words (although it remains level or falls gradually with duration for turns of eight to about 30 words).

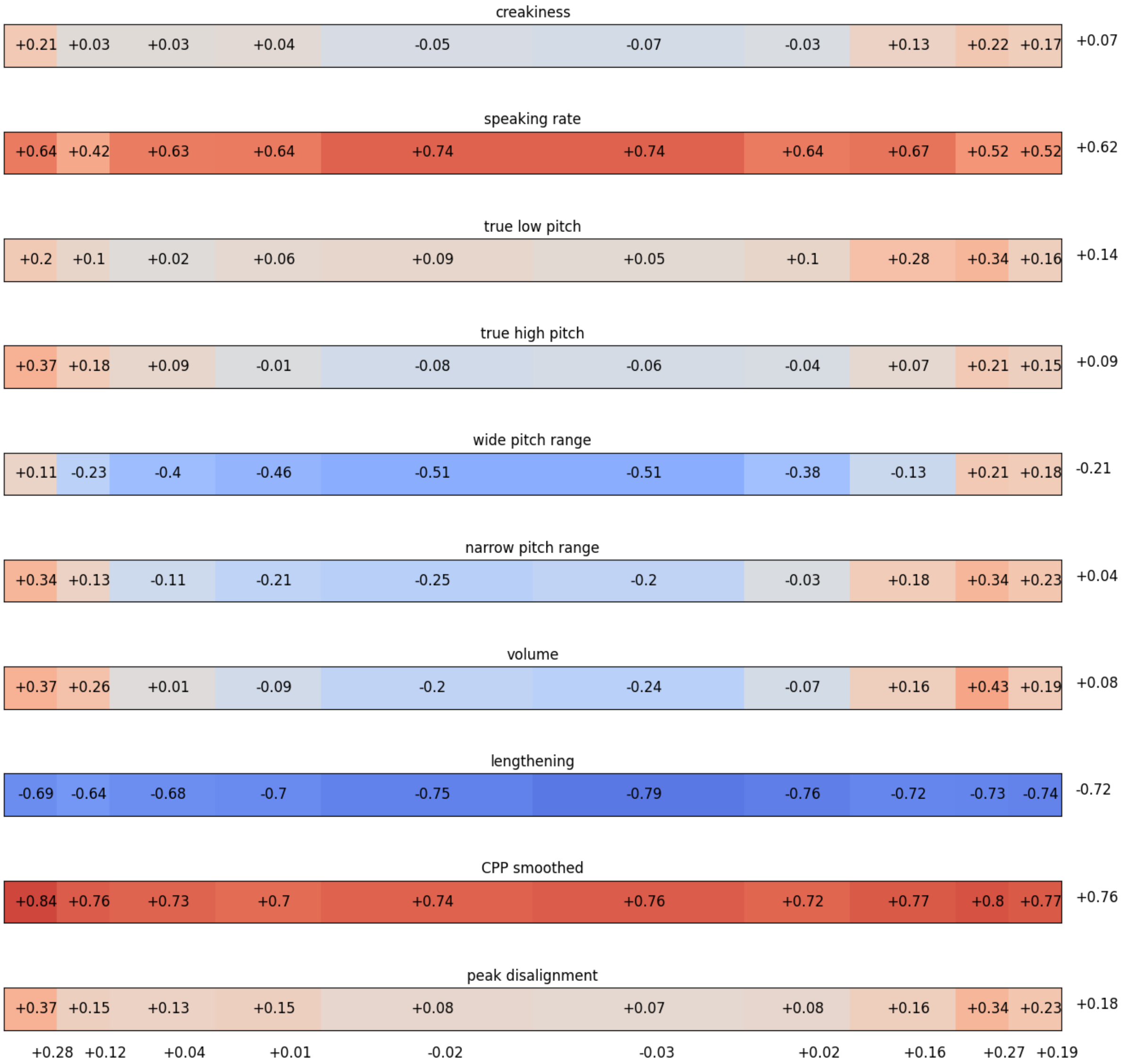


Figure 5.4: Spearman correlations between utterance duration and prosodic feature values. Mean correlations of spans are along the bottom. Mean correlations of base features across all spans are along the right.

- Similarly, the anticorrelation with lengthening features (mean($\rho$) $= -0.72$) may be due to speakers' tendency to lengthen their vowels in shorter utterances. For instance, [35] found that in English read speech, vowel duration increases as speaking rate decreases.

The second observation is that utterance duration on average correlates with some prosodic features sharing the same span. The strongest correlations are among features with spans towards the beginning or end of utterances (`0--5` $\rho = .28$, `90--95` $\rho = .27$, `95--100` $\rho = .19$).

Given that the spans towards the beginning and end of utterances are also the smallest, I speculated whether these correlations were more a result of their size rather than their location. Specifically, I speculated whether these spans were not large enough to reliably include measurable variation in prosody, resulting in the stronger correlations.

To determine if the size of the spans was the primary factor affecting their correlation with utterance duration, rather than their location, I experimented with an alternative set of spans: I recalculated the correlations after repositioning the smallest spans towards the center of an utterance, resulting in a modified feature set. The correlations for this modified set are depicted in Figure 5.5.

My underlying rationale was as follows: If the size of the span is the main factor in its correlation with utterance duration, then we would expect to see significant differences in correlations between the original and the modified feature sets. However, the minor differences observed between these two sets suggest that the size of the span is not the primary determinant in its correlation with utterance duration.

To form a more comprehensive understanding of the relationship between utterance duration and estimated similarity, I examined instances where certain utterances were estimated as highly dissimilar despite not being among the longest utterances in the data, for instance, utterances EN_021_13 and EN_019_19, as seen in the observation notes in Appendix C. I suspected these utterances were outliers, and distant to all other utterances in the data. To test this suspicion, I estimated the similarity of all utterances to the

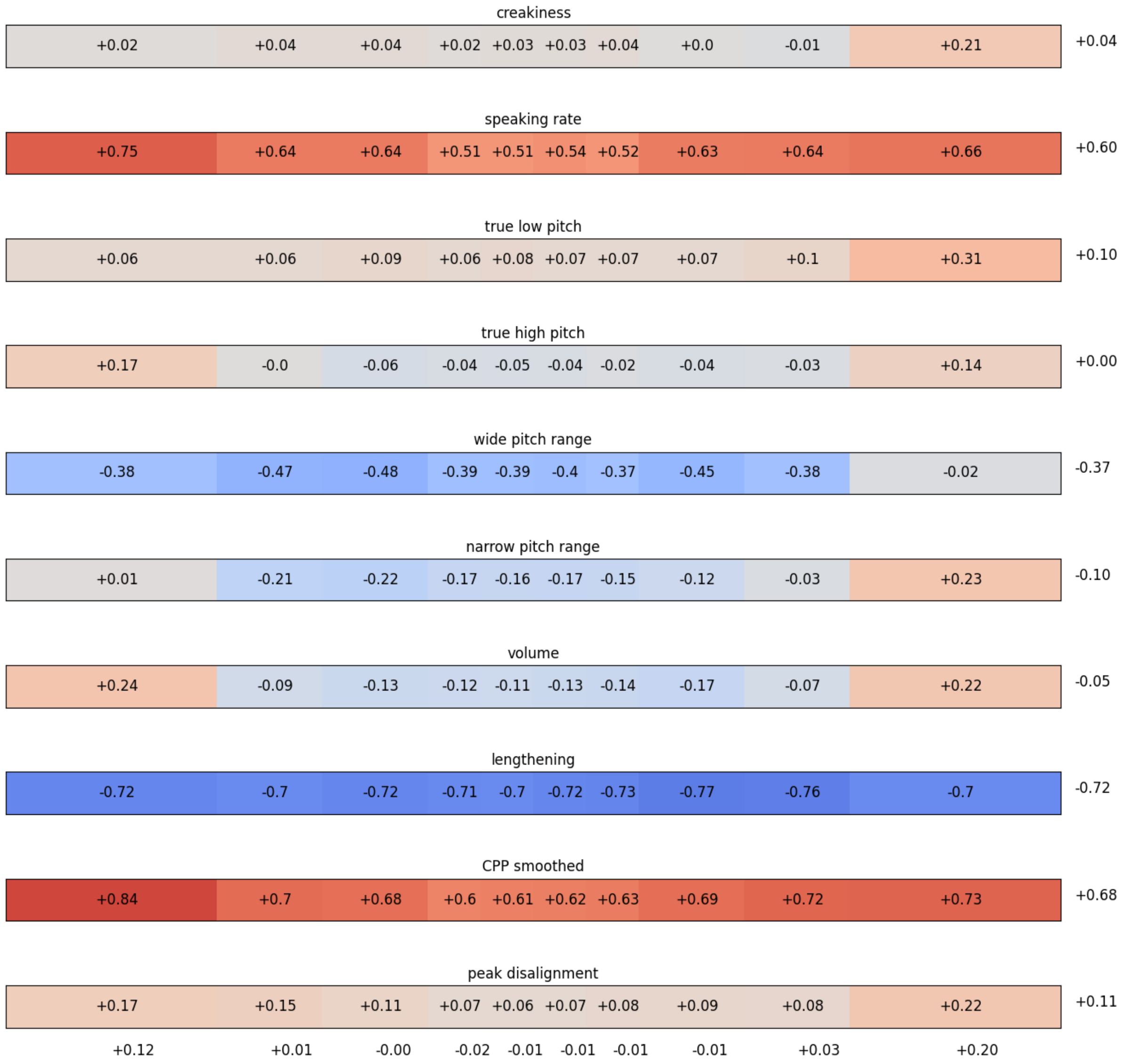


Figure 5.5: Spearman correlations between utterance duration and feature values for a modified feature set. The original spans (Figure 5.4) are: 5% 5% 10% 10% 20% 20% 10% 10% 5% 5%. The modified spans are: 20% 10% 10% 5% 5% 5% 5% 10% 10% 20%. Mean correlations of spans are along the bottom. Mean correlations of base features across all spans are along the right.

centroid of the data, or the "average" utterance, using the same metric. I used the same custom application previously used to gauge the reliability of the metric (Section 5.2), which includes an "average" view to display the utterances most similar to the centroid. Indeed, these utterances were among the utterances estimated as highly dissimilar to the centroid.

These findings suggest that utterances estimated as highly similar tend to share similar duration, but their similar duration is not the only reason for their similarity.

# Chapter 6

# Analysis of Cross-Language Prosody Mapping Modeling Approaches

In this chapter, I compare various approaches for modeling the mappings of English and Spanish prosody with an aim to identify the aspects of prosody and its pragmatic functions that pose challenges for modeling.

For this analysis, I use English and Spanish utterances from the DRAL corpus, as described in Chapter 2, in conjunction with their corresponding prosody representations, as computed in Chapter 3, and the similarity metric, as defined in Chapter 5.

## 6.1 Prosody Translation Task Definition

The task of a prosody translation model is to predict the target-language prosody representation given the prosody representation of a source-language utterance. Thus, additional data commonly integrated into complete speech-to-speech translation systems, such as lexical content and surrounding context, are ignored. While incorporating this additional data may be useful in improving the pragmatic fidelity of speech-to-speech translation models, insights gained by this simplified task may guide the development methods that make better use of this additional data for this purpose.

I partition a subset of the DRAL corpus as shown in Table 6.1. The training set is used for training the models, while the test set is used for their evaluation. The English-to-Spanish task and Spanish-to-English task use the same partitions, and only the source language and target language are exchanged. These partitions are nearly speaker-independent,

Table 6.1: Data partitions used in evaluation of prosody translation models, with an 80/20 split and sharing at most one unique speaker.

| Partition | Number of utterance pairs | Number of unique speakers |
|---|---|---|
| Training | 912 (Conversation 001–013, 015–029, 032–036, 038–040, 043–045, 050–054) | 20 |
| Testing | 227 (Conversation 008, 010, 012, 014, 017–020, 032–034, 036–038, 044, 045, 051, 055, 056) | 7 |

with at most one unique speaker shared across the partitions. This partitioning scheme is intended to simulate real-world applications of speech-to-speech translation systems, where systems translate speech of speakers not included within their training data.

The prediction error of a model is quantified by the similarity between its predicted prosody representation and the prosody representation of the human-produced reference utterance, as estimated by the prosody similarity metric defined in Chapter 5.

The prosody translation task is visualized in Figure 6.1.

## 6.2 Hypotheses

I compare the following models, each grounded in unique modeling approaches: direct-transfer baseline model, a source-ignoring baseline model, a linear regression model, a $k$-nearest neighbors model, and a shallow neural network model. Each of these models is described in further detail in the next section.

The two baseline models serve as a reference point for the other three models. Notably, these baseline models disregard potential patterns within the training data when inferring

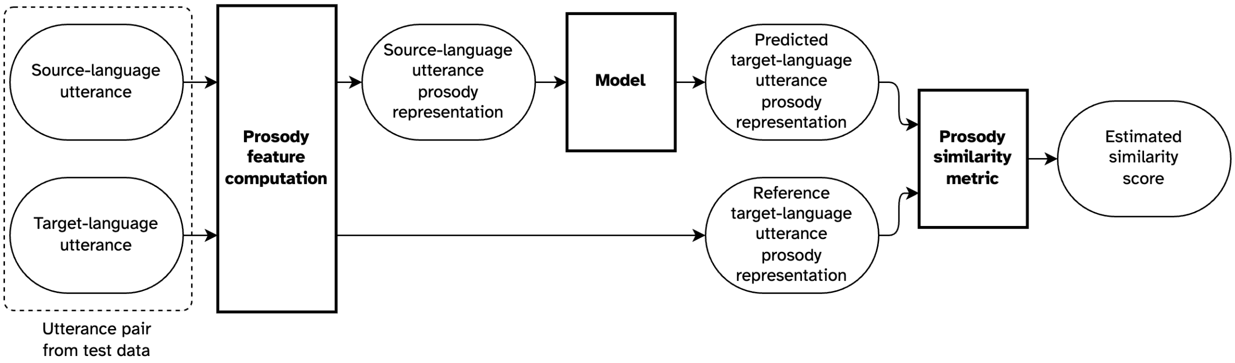

Source-language utterance
Target-language utterance
Utterance pair from test data
Prosody feature computation
Source-language utterance prosody representation
Model
Predicted target-language utterance prosody representation
Reference target-language utterance prosody representation
Prosody similarity metric
Estimated similarity score

Figure 6.1: Prosody translation task.

the target-language prosody representation.

Recognizing and integrating potential patterns from the training data can offer the following advantages. First, while English and Spanish prosody share many similarities, they are not identical, as already shown by the prosodic feature correlations examined in Chapter 4. Second, the prosody of a target-language utterance is not tied to its lexical content, as exemplified by instances of utterances where the perceived meaning may change based on prosody.

**Hypothesis 1** Predicting the prosody representation of a target-language utterance from cross-language patterns will yield, on average, a higher similarity compared to predicting the representation as identical to that of the source-language utterance.

**Hypothesis 2** Predicting the prosody representation of a target-language utterance from cross-language patterns will yield, on average, a higher similarity compared to predicting it based solely on the lexical content of the source-language utterance.

These hypotheses highlight the potential advantages of pattern-informed prosody mapping models moving beyond the baseline approaches.

## 6.3 Description of Models

I evaluate models for prosody translation which are based on existing algorithms and are simple compared to those commonly used in speech-to-speech translation research. This was a deliberate choice, as simpler models tend to be easier to interpret. This interpretability will be helpful in understanding the decision-making of the models, and in turn, the possible reasons for their failures. I describe each of the models in the following subsections.

### 6.3.1 Direct-Transfer Baseline

The direct-transfer baseline model represents an approach of directly transferring the prosody of the source-language utterance to the target-language utterance. Unlike the next models, this model does not rely on patterns learned from speech data.

Given the prosody representation of a source-language utterance, this model trivially outputs the same prosody representation as its prediction. It operates on the assumption that the prosody of a target-language utterance should be identical to that of the source-language utterance.

### 6.3.2 Source-Ignoring Baseline Model

The source-ignoring baseline model is intended to represent the optimal achievable performance of a typical cascaded speech-to-speech model with a speech synthesizer that ignores the source-language prosody.

To avoid the impact of source-language automatic speech recognition or machine translation errors, the implementation of the source-ignoring baseline model is based on a lookup of the human-produced translation in the target language. The model is thus provided the lexical content of the target-language utterance directly, rather than having to infer it from the input source-language utterance.

To translate prosody, the model first transcribes the reference target-language utterance into a sequence of words and punctuation. Following this, the transcription is synthesized

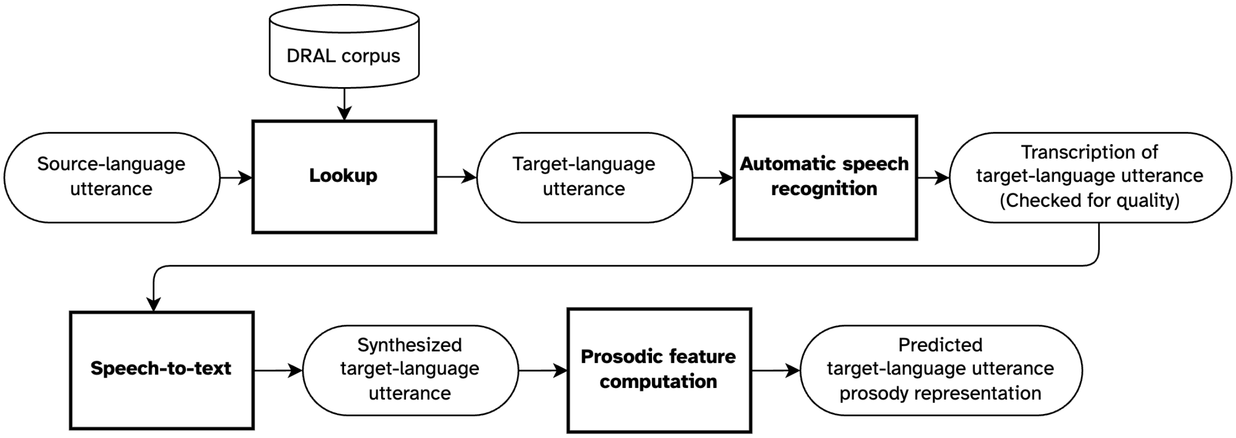

DRAL corpus
Source-language utterance
Lookup
Target-language utterance
Automatic speech recognition
Transcription of target-language utterance (Checked for quality)
Speech-to-text
Synthesized target-language utterance
Prosodic feature computation
Predicted target-language utterance prosody representation

Figure 6.2: Source-ignoring baseline model during inference.

into speech. Lastly, the prosody representation of the synthesized speech is computed using the same method used to compute the reference prosody representation. To ensure a fair comparison with the other models, incorrectly transcribed utterances, as determined by a human check, were excluded from the data. Table 6.1 accounts for the 252 utterances excluded. Figure 6.2 illustrates the source-ignoring baseline model during inference.

The source-ignoring baseline model is the only model that does not receive as input a prosody representation; the task defined in Section 6.1 is thus modified to predict the prosody representation of the target-language utterance given only the transcription of a source-language utterance.

This model is constructed from readily available resources, using pre-trained, open-source models for transcription and speech synthesis that are well-supported actively maintained. Transcription is performed with Whisper [48] pre-trained speech recognition models, using an English-specific model for English utterances and a multilingual model for Spanish utterances. Synthesis is performed with Coqui TTS [12] pre-trained text-to-speech models, using same-language models built on the Tacotron 2 architecture and trained on corpora of read books. The English model was trained using the LJ Speech Dataset [30], while the Spanish model was trained on the M-AILABS Speech Dataset [25].

### 6.3.3 Linear Regression Model

The multiple linear regression model embodies a parametric approach. This model takes into account the distribution of prosodic feature values and uses a fixed set of parameters to map the relationship between English and Spanish prosodic features as a linear function.

Each target-language feature is expressed as a linear combination of the source-language features and a set of estimated parameters, or coefficients. Thus, each feature of the target-language prosody representation is predicted as a linear function of the 100 features of the source-language prosody representation:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{100} x_{100}$$

The predicted target-language prosodic feature value is represented by $\hat{y}$. The source-language prosodic feature values are represented by $x_1, \ldots, x_{100}$. The regression coefficients, or contributions of the source-language features, are represented by $\beta_1, \ldots, \beta_{100}$. Lastly, the intercept, or constant value of the target-language feature not explained by the source-language features, is represented by $\beta_0$.

### 6.3.4 *k*-Nearest Neighbor Regression Model

The $k$-nearest neighbor regression model embodies a local approach. This model predicts the target-language prosody representation of an utterance based on the proximity of its source-language prosody representation in a feature space representation of the training data. Thus, this approach does not rely on overall patterns in the training data and instead relies on local patterns to predict the target prosody.

This model first constructs a feature space representation of the source-language training data. To predict the target-language representation of a query source-language representation, it first finds its $k$-nearest neighbors, or utterances closest to the query utterance in the feature space, using a distance metric. It then predicts the target-language representation of the query utterance as the average of the target-language representations of the $k$-nearest

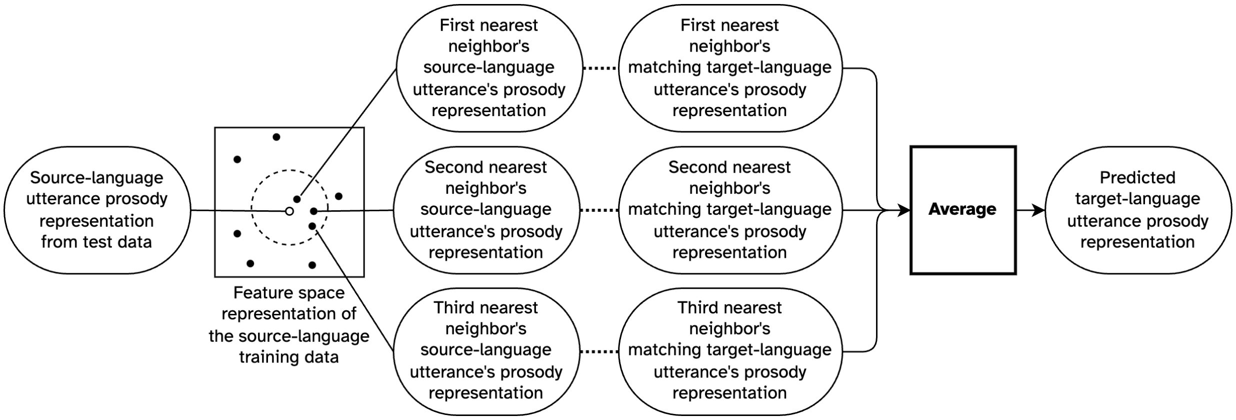

Source-language utterance prosody representation from test data
Feature space representation of the source-language training data
First nearest neighbor's source-language utterance's prosody representation
Second nearest neighbor's source-language utterance's prosody representation
Third nearest neighbor's source-language utterance's prosody representation
First nearest neighbor's matching target-language utterance's prosody representation
Second nearest neighbor's matching target-language utterance's prosody representation
Third nearest neighbor's matching target-language utterance's prosody representation
Average
Predicted target-language utterance prosody representation

Figure 6.3: $k$-nearest neighbors model during inference.

neighbors. Here, I use $k = 3$ and the proposed prosody similarity metric as the distance metric. Figure 6.3 illustrates the $k$-nearest neighbor regression model during inference.

## 6.4 Comparison of Model Performance: Validation of Hypotheses

Table 6.2 presents the overall average error for each model. The direct-transfer model outperformed the source-ignoring model in both English-to-Spanish and Spanish-to-English tasks, confirming Hypothesis 1. This outcome indicates that predicting the prosody representation of a target-language utterance from cross-language patterns yields, on average, a higher similarity compared to predicting the representation as identical to that of the source-language utterance.

Furthermore, the linear regression model outperformed the direct-transfer model in both English-to-Spanish and Spanish-to-English tasks, confirming Hypothesis 2. This outcome indicates that predicting the prosody representation of a target-language utterance from cross-language patterns yields, on average, a higher similarity compared to predicting the representation based solely on the lexical content of the source-language utterance.

The success of the linear regression model is evidence that even a relatively simple

Table 6.2: Average error of prosody translation models.

| Model | English-to-Spanish | Spanish-to-English |
|---|---:|---:|
| Source-Ignoring | 12.65 | 12.32 |
| Direct-Transfer | 11.35 | 11.35 |
| Linear regression | 9.23 | 9.37 |

predictive model is capable of learning aspects of the prosodic mapping between English and Spanish.

## 6.5 Failure Analysis

The results above pertain to the average error across all utterances in the training data, but do not account for the quality of prosody translation for individual utterances. To obtain a deeper understanding of the challenges associated with cross-language prosody modeling, this section examines the performance of the various models.

### 6.5.1 Source-Ignoring Baseline Model

I examined the 16 utterances in each translation direction whose synthesized prosody was least similar to the human-produced target. The most common and salient differences were due to failure to: lengthen vowels and varying the speaking rate for utterances where speakers are thinking or expressing uncertainty or hesitation, failure to change pitch at turn ends, and generally sounding read or rehearsed and thus unnatural for conversational speech.

### 6.5.2 Direct-Transfer Baseline Model

I examined the 16 utterance pairs for which the direct-transfer model's predicted prosody representation diverged most from the reference representation. Often there were salient differences in a few common patterns, such as Spanish utterances being creakier than the English (three pairs), English but not Spanish utterances ending with rising pitch (three pairs), and English utterances being breathier in some regions (in five pairs).

The latter two differences may reflect the common use of uptalk in English to establish common ground regarding a referent, which is characterized by the use of breathy voice and rising pitch [76]. This pattern is uncommon in the Spanish dialects of the DRAL corpus. In other instances, there were no highly salient differences; presumably these involved smaller differences which added up to a larger difference according to the metric.

### 6.5.3 Linear Regression Model

I examined the examples where linear regression model provided the most improvement relative to the direct-transfer baseline. Unsurprisingly, these were often examples where the direct-transfer baseline diverged most from the reference, mentioned in Section 6.5.2. Of the 16 source-language utterances where the linear regression model most improved over the direct-transfer baseline, 7 of these inputs were part of the same 16 inputs where the direct-transfer baseline worst performed.

Lastly, I examined the highest-magnitude coefficients of both the English-to-Spanish and Spanish-to-English variants of the model. Given that most coefficients were close to zero, I focused on the 16 coefficients with the greatest magnitude for the linear regression models. Most were unsurprising and reflected correlations noted in Chapter 4. However, there was a –.32 coefficient relating English lengthening over 5%–10% to Spanish CPPS over 0%–5%. This may reflect the tendency for English speakers to start turns with fast speech (low lengthening) but not Spanish speakers [79], who perhaps tend instead to start turns with more harmonic, or higher CPPS, speech.

Table 6.3: Cross-language linear regressor coefficients with the largest magnitude for features with different base feature and span.

| Source-language prosodic feature | Target-language prosodic feature | Coefficient |
|---|---|---:|
| English speaking rate 95–100% | Spanish lengthening 95–100% | -0.33 |
| Spanish lengthening 95–100% | English speaking rate 95–100% | -0.32 |
| Spanish CPPS 0–5% | English lengthening 5–10% | -0.32 |
| English speaking rate 95–100% | Spanish lengthening 95–100% | -0.33 |
| Spanish CPPS 50–70% | English creakiness 10–20% | 0.31 |

The relationship between English and Spanish prosody tends to be local. Of the features examined, most were dependent on a corresponding other-language feature sharing the same base feature and span or a neighboring span (12 of the 16 English features, and 15 of the 16 Spanish features). This was not too surprising, as the feature correlation analysis in Chapter 4 showed that most features of one language correlated with the same feature of the other.

While the prosody at one region of a translation is always best predicted by the same region of the original utterance, the strongest predictors depend on the source and target languages. The remaining coefficients are shown in Table 6.3. English final speaking rate is dependent on Spanish final lengthening. Likewise, Spanish final lengthening is dependent on English final speaking rate. Less commonly, the prosody at one region of a translation is best predicted by a different region of the original utterance. English near beginning creakiness is dependent on Spanish mid CPPS.

The linear model shows advantages for modeling the relationship between English and Spanish prosody. However, its prediction error is still high. This may be due to the mappings being too complex for a linear model. Several other factors may contribute to the prediction error, including: insufficient training data, dependencies of target-language prosody on the source-language utterance context that have not been captured including its lexical content, speaker-specific tendencies in prosody behavior, and the existence of free variation, which implies a permissible margin of error for the similarity metric.

In the next chapter, I explore the benefits of using a representation of utterance prosody with reduced dimensionality.

# Chapter 7

# A Reduced Dimensionality Representation of Utterance Prosody

In this chapter, I present a corpus-driven approach to creating a representation of utterance prosody with reduced dimensionality.

The representation of utterance prosody from Chapter 3 encodes many aspects of prosody relevant to its pragmatic functions in dialog. In this representation, the prosody of an utterance is represented by 100 prosodic features, consisting of ten base prosodic features measured over ten non-overlapping windows spanning the duration of the utterance. Such representation can be used to model cross-language mappings of prosody, as explored by the modeling of English and Spanish prosody in Chapter 6. However, this representation is not without limitations and presents opportunities for improvement.

First, the current representation lacks interpretability. While the individual features of the representation are generally interpretable, its large number of features are difficult to interpret simultaneously. A representation that is more interpretable would enable analysts to use their knowledge of prosody to make informed decisions on how to improve a speech-to-speech translation model.

Second, the representation may be susceptible to noise. This noise can arise from speakers' limited fluency, use of rare dialect, or recording environment. A representation that is more robust may improve a model's learning ability by reducing the amount of irrelevant information.

Last, the representation may encode redundant information if aspects of prosody are superfluously encoded by multiple features. Additionally, redundancy can arise from encoding aspects of prosody that are ubiquitous across all utterances in a language, i.e., those that do not contribute to differentiating one utterance from another. A representation that is less redundant may improve a model's ability to learn the relationship between source and target language prosody.

Thus, researchers can benefit from a representation that is more interpretable, more robust, and less redundant. Achieving this goal involves representing utterance prosody with fewer features while retaining essential information.

To create a representation of utterance prosody with reduced dimensionality, I use the prosody representation (Chapter 3) of utterances from the DRAL corpus (Chapter 2). I apply a dimensionality reduction technique on this prosody representation to derive dimensional models for both English and Spanish utterance prosody. I then interpret the first five dimensions within these models to identify their functions in dialog. This methodology parallels the one used to develop a dimensional model of interaction styles in dialog [78]. Here, I adopt a simpler version of this methodology for prosody at the utterance level.

## 7.1 Method

Many techniques for dimensionality reduction exist, each with advantages that motivate their use in different contexts. These include Independent Component Analysis (ICA), for creating mutually independent features; t-Distributed Stochastic Neighboring Embedding (t-SNE), for representing non-linear relationships within data; and Linear Discriminant Analysis (LDA), for categorizing data into distinct classifications.

To create a representation of utterance prosody, I selected a different technique, Principal Component Analysis (PCA), due to its interpretability. PCA functions by transforming the data into a lower-dimensional space, with the aim of preserving the majority of the variance from the original data. The output of this process is a set of principal components, each of which represents a linear combination of the original features, and ordered by the amount of variance they account for in the data.

For the analyses conducted in this chapter, I partitioned a subset of the DRAL corpus, consisting of same-language utterances from Conversation 1 through 56. I partitioned this subset into a training set and a test set, following an 80/20 split. I computed the principal components using the training set, then applied the learned transformation to both the training set and test sets.

To obtain a reduced dimensionality representation of an utterance, I first compute the principal components from training data corresponding to the same language. This creates a PCA model, which I then apply to the input representation of the utterance to transform it into the space of principal components. I then truncate the transformed representation to retain only the principal components that explain a large amount of variance in the training data. The amounts of variance explained by the first five principal components in the training data of each language are shown in Figure 7.1.

The first five principal components explain 39% of the variance in both the English data and Spanish data. Interpreting what each of these principal components may represent could provide insights into the variance of prosody in each language, its pragmatic functions, and the differences between the two languages.

To arrive at my interpretations of these five principal components, or dimensions, I examined their loadings and extremes. The loadings of a dimension represent the contribution from the original features, thereby revealing features that highly correlate with that dimension.

For each dimension, I examined the utterances with the highest and lowest values for that dimension. I refer to the low extremes and high extremes of a dimension as its negative pole and its positive pole, respectively. For each dimension, I examined its 16 positive and 16 negative extremes. The lexical content of these extremes is provided in Appendix B.

For Spanish utterances, I also give the lexical content of their matching English utterances. Each utterance pair was produced by the same speaker, who was instructed to translate utterances with a focus on faithfully translating their feeling (including prosodic aspects), rather than their words. Therefore, the lexical content of English and Spanish

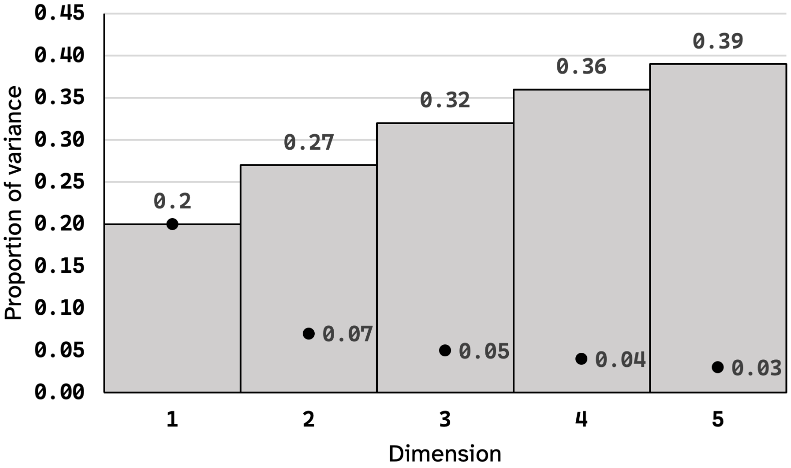


Figure 7.1: Variance explained by the first five dimensions of English and Spanish reduced dimensionality prosody representation. Individual contributions are indicated by the points, while cumulative variance is indicated by the bars. As the values for both English and Spanish representations are the same to two significant figures, they are consolidated into a single figure for clarity.

pairs may not be one-to-one translations.

In my descriptions, I mention aspects of prosody that distinguish the utterances of a particular pole from all others, being selective in which I discuss. Specifically, I discuss the aspects that I found easy to hear, easy to interpret, part of a larger pattern across other utterances, and/or related to findings of previous research.

After completing my interpretations, I compared these to the interpretations of one other person. Comparing our interpretations, we found that we had named the first couple dimensions of each language differently, but our interpretations were compatible. Progressing to subsequent dimensions, our interpretations exhibited a greater degree of divergence. This outcome was anticipated, as the salience of prosodic features in an utterance is subjective to individuals, and the interpretation of their functions can also vary.

## 7.2 Interpretation of English Dimensions

### 7.2.1 English Dimension 1: Focus on Speaker

The **positive** pole of English Dimension 1 represents "focus is directed towards the speaker."

The speaker may hold the floor by using fillers such as *um*, *uh*, and *like* (Appx. B.1: 1, 2, 3, 4, 6, 9, 10, 14, 16), for instance, *Ah no that's-that's actually, like, that's really sad because like, the gang violence down there is terrible* (Appx. B.1: 6). Alternatively, the speaker may use pauses instead of fillers (Appx. B.1: 5, 7, 11, 12). Fillers and pauses are used to signal that the speaker is not finished speaking, and that the focus should remain on them.

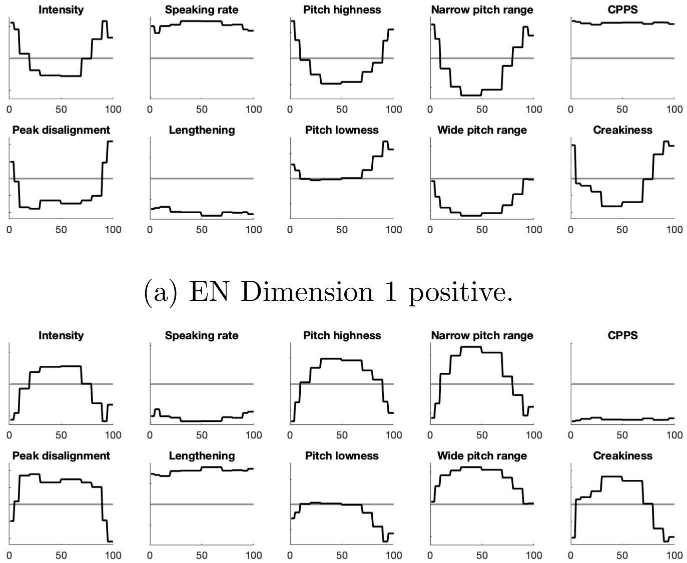
Utterances near the positive pole are characterized by high speaking rate and U-shaped pitch, lowest at the middle of the utterance (Figure 7.2a).

The **negative** pole of English Dimension 1 represents "focus is directed towards the interlocutor."

When the focus is already on the interlocutor, the speaker maintains the focus on the interlocutor by not drawing focus to themselves. The speaker may maintain the focus on the interlocutor by responding with a short expression of surprise (Appx. B.2: 1, 5, 11, 12) or a backchannel (Appx. B.2: 8, 9, 10). The speaker may also complete the interlocutor's thought (Appx. B.2: 16). Alternatively, the speaker may respond in a way that does not warrant further speaking (Appx. B.2: 4, 7, 14).

When the focus has been directed towards the speaker, they may redirect focus towards the interlocutor by asking them a short question (Appx. B.2: 2, 6). Utterances near the negative pole are significantly shorter compared to utterances near the positive pole, with at most five words.

Utterances near the negative pole are characterized by low speaking rate, use of breathy voice, disaligned pitch and intensity peaks, and inverted U-shaped pitch, highest at the middle of the utterance (Figure 7.2b).



(a) EN Dimension 1 positive.

(b) EN Dimension 1 negative.

Figure 7.2: EN Dimension 1 loadings.

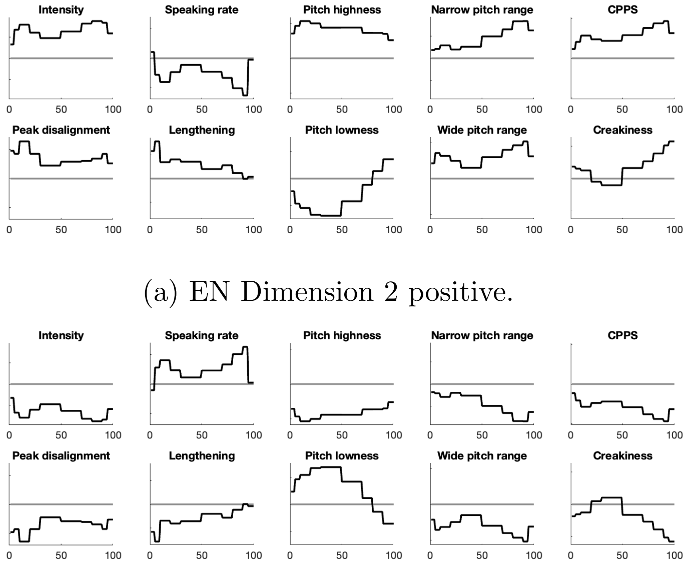
### 7.2.2 English Dimension 2: Engaged/Animated

The **positive** pole of English Dimension 2 represents "high engagement, highly animated."

The speaker may exhibit high engagement when expressing agreement (Appx. B.3: 1, 9, 11, 12) or surprise (Appx. B.3: 2, 5, 16), for instance, *Yeah, yeah* (Appx. B.3: 1) and *That's weird!* (Appx. B.3: 2). The speaker may also exhibit high engagement when recounting events (Appx. B.3: 4, 6, 7, 8, 10). For instance, *My shoulders are, like, killing [me]* (Appx. B.3: 6). The speaker's high engagement may be evident by their continuous, uninterrupted speech (Appx. B.3: 3, 11, 14).

Utterances near the positive pole are characterized by high intensity and wide pitch range throughout the utterance (Figure 7.3a). Unlike utterances near the positive pole of English Dimension 1 (Figure 7.2a) where the pitch drops at the middle of the utterance, utterances near the positive pole of English Dimension 2 use high pitch throughout the utterance.

The **negative** pole of English Dimension 2 represents "low engagement, not animated."



(a) EN Dimension 2 positive.

(b) EN Dimension 2 negative.

Figure 7.3: EN Dimension 2 loadings.

The speaker may exhibit low engagement when sympathizing with the interlocutor (Appx. B.4: 4, 5, 7, 9, 16), for instance, *It's better to ask questions and learn than to stay confused* (Appx. B.4: 5). The speaker may also exhibit low engagement when fatalistically accepting a past or future outcome (Appx. B.4: 11, 13, 14), for instance, *And aspects that I couldn't, like, see back then* (Appx. B.4: 11). A low engagement utterance may be as short as one or two words, especially when completing the interlocutor's thought (Appx. B.4: 2, 3, 6, 15), for instance, *Interrupt?* (Appx. B.4: 15).

Like utterances near the positive pole, utterances near the negative pole also commonly use only the word *yeah*. However, utterances near the negative pole (Appx. B.4: 4, 7) are much less lively than utterances near the positive pole (Appx. B.3: 1, 9).

Utterances near the negative pole are characterized by low intensity, narrow pitch range, and breathy voice (Figure 7.3b).

### 7.2.3 English Dimension 3: Existence of Shared Understanding

The **positive** pole of English Dimension 3 represents “lack of shared understanding.”

The speaker may address a misunderstanding, or a presumed misunderstanding, from the interlocutor (Appx. B.5: 1, 4, 7, 8, 9, 10). When addressing a misunderstanding, the speaker may use the words *but*, *no*, or *actually*, for instance, *No, it was actually, there's-there's some in, um, El Paso* (Appx. B.5: 10). The speaker may indicate that they just received new information from the interlocutor, for instance, *Okay, okay* (Appx. B.5: 2) and *Oh my god, okay* (Appx. B.5: 3). The speaker may ask for an explanation, for instance, *Oh, you get dizzy? Or what?* (Appx. B.5: 5). Lastly, the speaker may provide an explanation or offer a possible explanation (Appx. B.5: 12, 14), for instance, *They probably need more space* (Appx. B.5: 14).

Utterances near the positive pole are characterized by initial high pitch and creaky voice, followed by an increasing speaking rate and a decreasing intensity as the utterance progresses (Figure 7.4a).

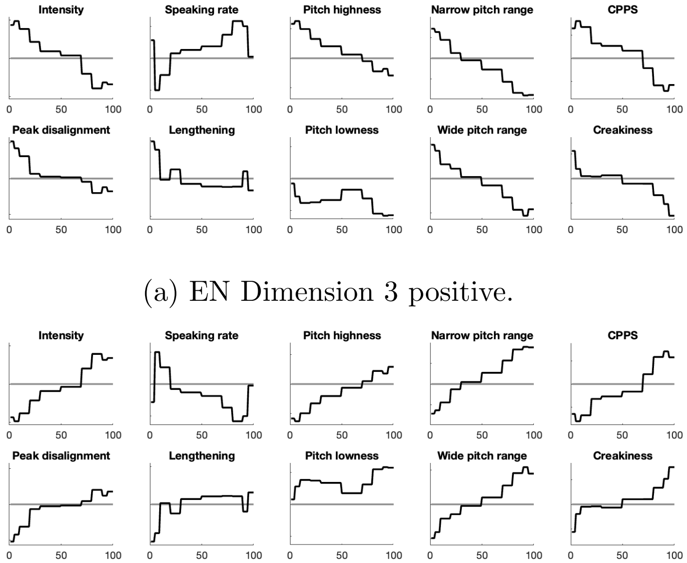
The **negative** pole of English Dimension 3 represents “shared understanding.”

The speaker may omit details because they assume that the interlocutor already has the required information, likely because it was discussed earlier in the conversation (Appx. B.6: 4, 6, 11, 13), for instance, *Psychology undergrad and then I'm* (Appx. B.6: 4). The speaker may imply elaboration from the interlocutor is unnecessary (Appx. B.6: 2, 5, 10, 14). For instance, *Yeah, yeah* (Appx. B.6: 2). Lastly, the speaker may not elaborate on a topic because they do not know more information, thus the level of limited understanding is shared, for instance, *So, I don't really remember* (Appx. B.6: 12).

Utterances near the negative pole are characterized by low pitch, increasing intensity throughout the utterance, and final creaky voice (Figure 7.4b).

### 7.2.4 English Dimension 4: Intent to Continue Topic

The **positive** pole of English Dimension 4 represents “intent to close the current topic.”



(a) EN Dimension 3 positive.

(b) EN Dimension 3 negative.

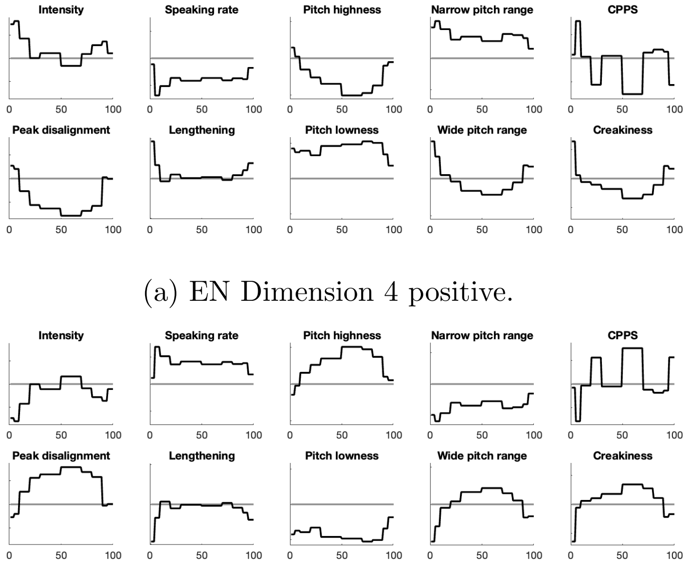
Figure 7.4: EN Dimension 3 loadings.

The speaker may downplay some aspect of the current topic of discussion, minimizing its important or significance (Appx. B.7: 3, 4, 7, 9, 10, 13, 15, 16). Downplaying may be evident by the speaker's choice of words, for instance, *Lab or something* (Appx. B.7: 2), *I was sixteen, we were partying, you know* (Appx. B.7: 8), *Like in the warehouses and all that* (Appx. B.7: 11), and *Yeah, just my dad* (Appx. B.7: 12).

Utterances near the positive pole are characterized by low pitch and low speaking rate, and aligned pitch peaks towards the middle of the utterance (Figure 7.5a).

The **negative** pole of English Dimension 4 represents "intent to continue the topic."

The speaker may ask a short, one or two word question to allow the interlocutor to continue the topic (Appx. B.8: 3, 5, 8, 9, 10), for instance, *Why?* (Appx. B.8: 3) and (Appx. B.8: 8) *And you?* The speaker may respond with agreement (Appx. B.8: 1, 2, 11), for instance, *Oh, yeah yeah* (Appx. B.8: 1) and *Well, yeah* (Appx. B.8: 11). Agreement may be evident by the speaker's choice of words, for instance, *I kind of agree with you* (Appx. B.8: 12, 16). Lastly, the speaker may respond with a simple backchannel, for



(a) EN Dimension 4 positive.

(b) EN Dimension 4 negative.

Figure 7.5: EN Dimension 4 loadings.

instance, *Yeah* (Appx. B.8: 13).

Utterances near the negative pole are characterized by low pitch and high speaking rate throughout the utterance (Figure 7.5b).

### 7.2.5 English Dimension 5: Checking Existence of Shared Knowledge

The **positive** pole of English Dimension 5 represents "checking whether the speaker and interlocutor share knowledge."

The speaker may prompt the interlocutor to confirm the interlocutor's understanding of information (Appx. B.9: 2, 3, 4, 7, 10, 14, 16), for instance, *So yeah, yeah, like, I go out* (Appx. B.9: 4), *Well the thing is that it was really alone* (Appx. B.9: 14), and *I went with my dad, my mom, and my sister* (Appx. B.9: 16). In these instances, the interlocutor can fittingly respond with a backchannel such as a brief *uh huh.* Alternatively, the speaker may prompt the interlocutor to confirm the speaker's understanding of information (Appx. B.9:

8, 9, 11, 15). For instance, *Turn it in?* (Appx. B.9: 9), *And you want to work with, kids?* (Appx. B.9: 11), and *You shower here?* (Appx. B.9: 15).

Utterances near the positive pole are characterized by ending with creaky voice, lengthening, and pitch rise (Figure 7.6a).

The **negative** pole of English Dimension 5 represents "lack of checking whether the speaker and interlocutor share knowledge."

The speaker may hold their turn to follow up with information only they know, thus checking whether the interlocutor shares knowledge is unnecessary (Appx. B.10: 1, 2, 3, 8, 15), for instance, *Psychology undergrad and then I'm* (Appx. B.10: 1) and *Professors are not on top of you and nothing like that, and* (Appx. B.10: 2). The speaker may assume the interlocutor already understands the speaker's view (Appx. B.10:10, 13), for instance, *I think it would completely, like, mess up, like, my perception of him, you know* (Appx. B.10: 13). The speaker may answer the interlocutor's question without elaboration, for instance, *I don't know, I've always been pretty bad at placing blame* (Appx. B.10: 7) and *I'm nineteen, I'm turning twenty in January* (Appx. B.10: 14). Lastly, the speaker may change the topic, for instance, *Uh, no but actually what I was gonna tell you was* (Appx. B.10: 4).

Compared to utterances near the positive pole, in utterances near the negative pole the speaker is less likely to cue for backchannel responses from the interlocutor and is more likely to present information without pausing.

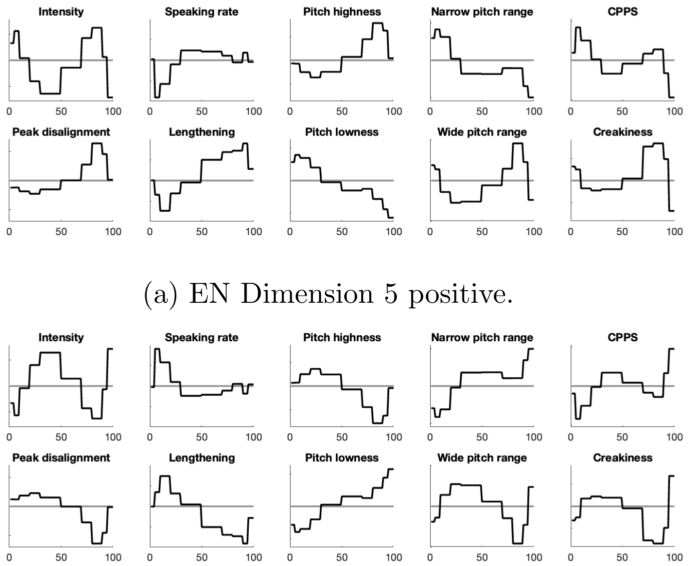
Utterances near the negative pole are characterized by ending with non-creaky voice and pitch drop (Figure 7.6b).

## 7.3 Interpretation of Spanish Dimensions

### 7.3.1 Spanish Dimension 1: Focus on Speaker

The **positive** pole of Spanish Dimension 1 represents "focus is directed towards the speaker."

The speaker may use fillers such as *pues* or *bueno* (similar to English *well*), *o sea* (similar



(a) EN Dimension 5 positive.

(b) EN Dimension 5 negative.

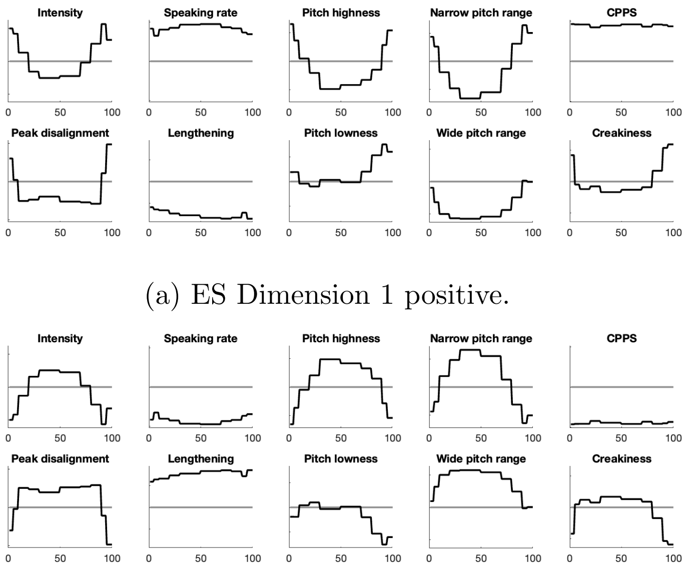
Figure 7.6: EN Dimension 5 loadings.

to English *I mean*), or *como que* (similar to English *like*) to hold the floor (Appx. B.11: 6, 7, 9, 11, 12, 14, 15, 16), for instance, *¿En qué época está, pues si, en que- en que época sucede o sea, después el imperio? ¿Durante el imperio?* (English *So in what time is it? Like, yeah, what time period is it in? Uh, is it, after the empire? During the empire?*) (Appx. B.11: 7). Alternatively, the speaker may use pauses instead of fillers (Appx. B.11: 1, 2, 4, 5, 10, 13). Lastly, the speaker's continuous speech may not allow the interlocutor to interject (Appx. B.11: 3, 8).

Utterances near the positive pole are fast, with U-shaped pitch (Figure 7.7a).

The **negative** pole of Spanish Dimension 1 represents "focus is directed towards the interlocutor."

The speaker may redirect focus towards the interlocutor by asking them a question (Appx. B.12: 8, 12, 13), for instance, *¿Y no te gusto?* (English *And you didn't like it?*) (Appx. B.12: 13). Alternatively, the speaker may respond briefly without intending to continue speaking, specifically when expressing surprise (Appx. B.12: 3, 10, 11), agreement



(a) ES Dimension 1 positive.

(b) ES Dimension 1 negative.

Figure 7.7: ES Dimension 1 loadings.

or positive assessment (Appx. B.12: 2, 9, 14), or when completing the interlocutor's thought (Appx. B.12: 16), for instance, *Oh wow* (Appx. B.12: 3) and *Que interesante* (English *That's interesting*) (Appx. B.12: 9). Lastly, the speaker may respond in a way not intending to redirect the attention towards themselves (Appx. B.12: 1, 5, 6, 7, 15), for instance, *Tiene que ser* (English *It's gotta be*) (Appx. B.12: 5). Utterances towards the negative pole are significantly shorter compared to utterances towards the positive pole, with at most three words.

Utterances near the negative pole are slow, breathy, with disaligned pitch and intensity peaks, and with inverted U-shaped pitch (Figure 7.7b).

### 7.3.2 Spanish Dimension 2: Engaged/Animated

The **positive** pole of Spanish Dimension 2 represents "high engagement, highly animated."

A speaker's high engagement (or low engagement) is not specific to any particular discourse function. The speaker may exhibit high engagement when recounting events

(Appx. B.13: 1, 3, 4, 10, 14, 15), for instance, *Casi me dijo, "Hazlo otra vez"* (English *She almost told me like, "Do it again"*) (Appx. B.13: 1). The speaker may also exhibit high engagement when correcting the interlocutor (Appx. B.13: 11, 16), disagreeing with the interlocutor (Appx. B.13: 8, 12), or agreeing with the interlocutor (Appx. B.13: 6), for instance, *No, no fue por eso* (English *Mm, mm, no it wasn't because of that*) (Appx. B.13: 11) and *Pues sí, sabe mejor* (English *But yeah, I like the flavor*) (Appx. B.13: 6).

Utterances near the positive pole are characterized by high intensity and wide pitch range (Figure 7.8a).

The **negative** pole of Spanish Dimension 2 represents "low engagement, not animated."

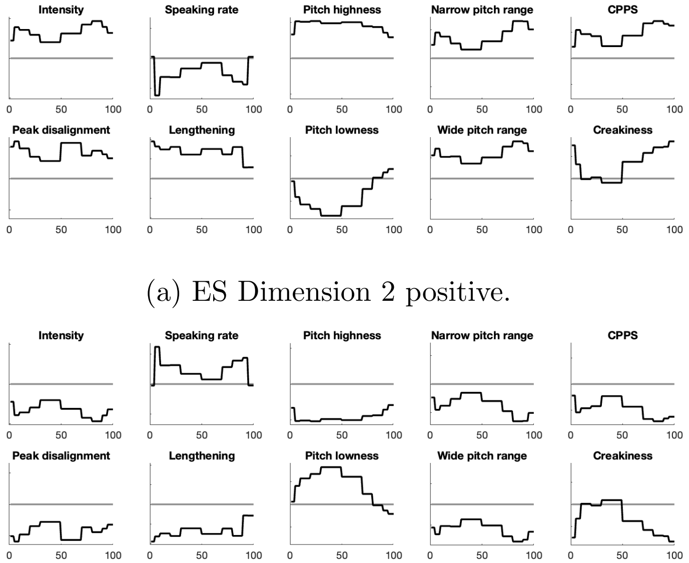
The speaker may exhibit low engagement when indicating they understand what the interlocutor just said (Appx. B.14: 1, 5), for instance, *Ah* (English *Oh, okay*) (Appx. B.14: 1) and *Ah, ok* (English *Okay*) (Appx. B.14: 5). The speaker may agree with the interlocutor (Appx. B.14: 10, 14, 15), for instance, *Padre* (English *Nice*) (Appx. B.14: 10) and *Sí* (English *Yes*) (Appx. B.14: 14). The speaker may also ask the interlocutor for clarification (Appx. B.14: 7, 9), answer the interlocutor's question (Appx. B.14: 11, 13, 16), or suggest a word for the interlocutor (Appx. B.14: 12), for instance, *¿Rápido?* (English *Upbeat?*) (Appx. B.14:7), *Pues, yo diría que Monterrey* (English *Um, I think Monterrey*) (Appx. B.14: 11), and *Mecatrónica* (English *Mechatronics*) (Appx. B.14: 12). Lastly, the speaker may close the topic or end their turn (Appx. B.14: 2, 3, 6, 8).

Utterances near the negative pole are characterized by low intensity, low pitch, and use of breathy voice (Figure 7.8b).

### 7.3.3 Spanish Dimension 3: Predictability

The **positive** pole of Spanish Dimension 3 represents "predictable information."

In general, utterances near the positive pole convey predictable or unsurprising info (Appx. B.15: 4, 5, 7, 10, 12, 13, 16). The speaker may confirm that they understand the interlocutor's motivation (Appx. B.15: 3, 8, 11) Lastly, the speaker may respond in an unsurprised way (Appx. B.15: 1).



(a) ES Dimension 2 positive.

(b) ES Dimension 2 negative.

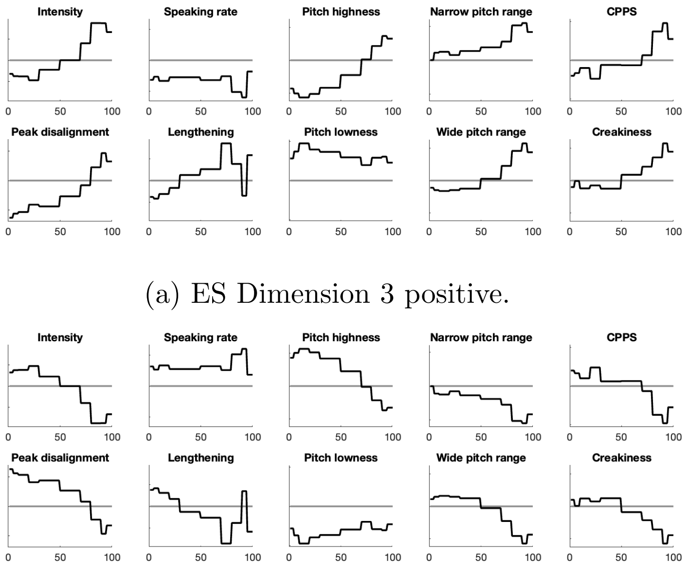
Figure 7.8: ES Dimension 2 loadings.

Utterances near the positive pole are characterized by slow speaking rate and ending with rising pitch and intensity (Figure 7.9a).

The **negative** pole of Spanish Dimension 3 represents "unpredictable information."

The speaker may respond with surprise because they find some information strange (Appx. B.16: 1, 9), or because they did not expect some new information from the interlocutor (Appx. B.16: 2, 5). The speaker may have not expected the interlocutor's question, first repeating the question before answering it (Appx. B.16: 4, 8), for instance, *¿Todo de Juárez? Riquísimo* (English *Everything from Juárez? Peak*) (Appx. B.16: 8). The speaker may address the interlocutor's confusion after saying something that mislead the interlocutor (Appx. B.16: 15). Lastly, the speaker may be surprised by the interlocutor's concern that the speaker may have been offended by something they said (Appx. B.16: 16).

Utterances near the negative pole are characterized by high speaking rate, early pitch and intensity peak disalignment, and ending with dropping pitch and intensity (Figure 7.9b).



(a) ES Dimension 3 positive.

(b) ES Dimension 3 negative.

Figure 7.9: ES Dimension 3 loadings.

### 7.3.4 Spanish Dimension 4: Experience and Knowledge

The **positive** pole of Spanish Dimension 4 represents "speaker is more knowledgeable or experienced in topic."

Consequently, the speaker may not feel obligated to provide a comprehensive answer to a question (Appx. B.17: 1, 6, 15), for instance, *Uh, como todo el día* (English *Uh, like all day*) (Appx. B.17: 15). The speaker may make a correction because they previously mislead the interlocutor, who took the speaker's word as truth, for instance, *Ah no te creas, perdón* (English *Oh no, just kidding*) (Appx. B.17: 10) and *No, no te creas, vimos como una cada día* (English *No, just kidding, we watched like one a day*) (Appx. B.17: 13). The speaker may agree with the interlocutor with little enthusiasm, because they do not fully support the interlocutor's answer or opinion but maintain politeness (Appx. B.17: 3, 5, 7), for instance, *Wow, qué loco* (English *Wow, that's crazy*) (Appx. B.17: 7). Lastly, the speaker may respond confidently because they are certain they are correct or because they believe the interlocutor will take their response as truth, regardless of what it is (Appx. B.17: 2, 4, 9).

Utterances near the positive pole are characterized by low pitch and beginning with non-breathy, creaky, high intensity voice (Figure 7.10a).

The **negative** pole of Spanish Dimension 4 represents "speaker is less knowledgeable or experienced in topic."

The speaker may turn to the interlocutor for the factual truth (Appx. B.18: 10, 12) or ask the interlocutor to confirm some information (Appx. B.18: 16), for instance, *¿Aquí a Juárez?* (English *Here in Juárez?*) (Appx. B.18: 10) and *¿Cómo las casas verdad?* (English *Like the houses right?*) (Appx. B.18: 16). The speaker may admit to not knowing some information (Appx. B.18: 2) or admit to getting some information wrong (Appx. B.18: 4). For instance, *Es que, no me acuerdo como después de eso como lo tomo* (English *I really don't remember after that how he handled it*) (Appx. B.18: 2). The speaker may disagree with the interlocutor but agree anyway as to not offend them (Appx. B.18: 9, 15). The speaker may recall when a third party had the higher authority, for instance, *Siempre nos decía "Ah, que no están haciendo esto bien" o que nos gritaba* (English *He would always tell us "Ah, you guys aren't doing this right" or he would scream at us*) (Appx. B.18: 14). Lastly, the speaker may let the interlocutor know they got the information wrong, politely, to not offend them (Appx. B.18: 1).

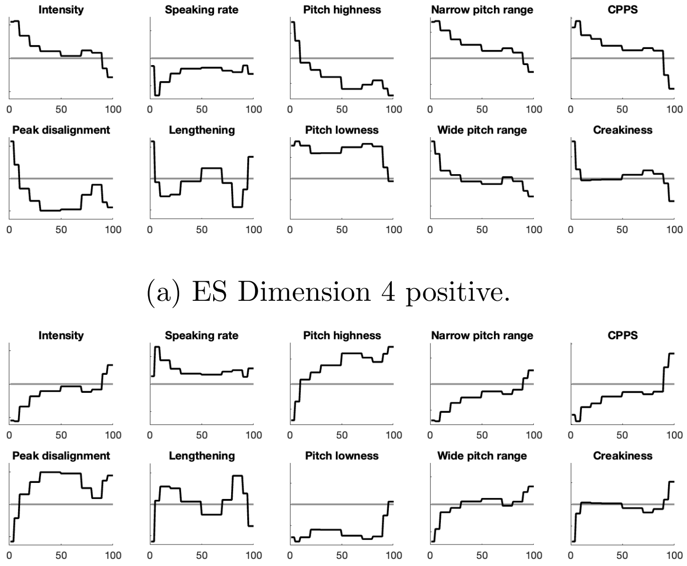
Compared to utterances near the positive pole, in utterances near the negative pole the speaker is willing to admit that they do not know something, got some information wrong, or is less confident with respect to information, thought, or place of authority.

Utterances near the negative pole are characterized by high pitch and ending with non-breathy, creaky, high intensity voice (Figure 7.10b).

### 7.3.5 Spanish Dimension 5: Certainty

The **positive** pole of Spanish Dimension 5 represents "speaker is certain about the information they are delivering."

The speaker may slow down to make sure the interlocutor understands some information



(a) ES Dimension 4 positive.

(b) ES Dimension 4 negative.

Figure 7.10: ES Dimension 4 loadings.

the first time, delivered at the end of the utterance in low pitch (Appx. B.19: 2, 6, 9, 14, 15), for instance, *He visto Bambi antes, pero no me acuerdo de la historia ni nada* (English *I've watched Bambi before, but, like, a long time ago, like, I don't remember the plot or anything*) (Appx. B.19: 15). The speaker may already know what their opinion is or what the requested information is, but slows down before providing it to sound less assertive, for instance *Mm, no* (Appx. B.19: 1), *Uh, no* (Appx. B.19: 12), and *Tengo diecinueve. Voy a cumplir veinte en enero* (English *I'm nineteen, I'm turning twenty in January*) (Appx. B.19: 13). The speaker may think they know some information but ask the interlocutor to confirm it (Appx. B.19: 3, 7), for instance, *¿Ah, rentaste la cabina? ¿O era como* (English *Oh did you rented the cabin? Or like*) (Appx. B.19: 7).

Utterances near the positive pole are characterized by beginning with high speaking rate, then slowing down at the middle of the utterance, and ending with low pitch and creaky voice (Figure 7.11a).

The **negative** pole of Spanish Dimension 5 represents "speaker is uncertain about the information they are delivering.”

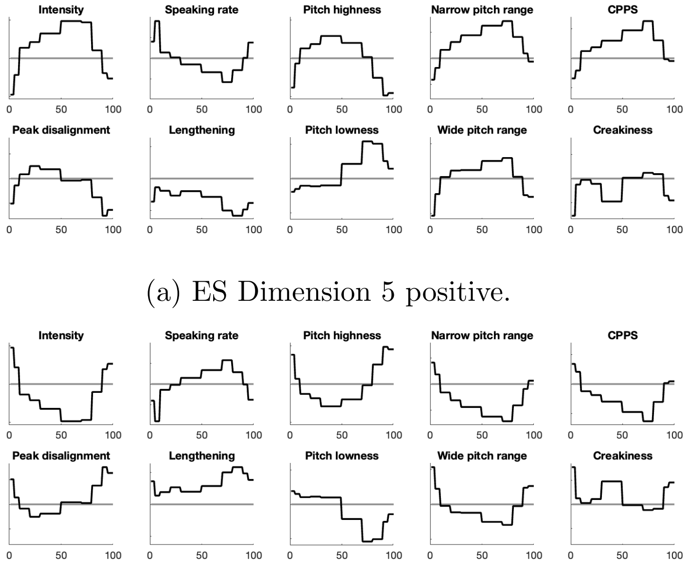
The speaker may need to think about what their opinion truly is (Appx. B.20: 1, 12), for instance, *Ah, es- es- está padre* (English *That-that’s cool*) (Appx. B.20: 1). The speaker may need to think about how to best describe something or how to best deliver some information (Appx. B.20: 4, 13, 15, 16), for instance, *Para desahogar, como si tienes un, día malo* (English *To release your feeling, like if you had like a rough day*) (Appx. B.20: 15). The speaker may need to recall some details or word them carefully to remain truthful (Appx. B.20: 5, 11). For instance, *No tocamos, por culpa del otro guitarrista* (English *We didn’t play because of the other guitarist’s fault*) (Appx. B.20: 11). Lastly, the speaker may lose certainty about previous information they felt was true after learning new information (Appx. B.20: 7, 10), for instance, *¿No e-, eres de Horizon? ¿Votaste?* (English *No, you’re from Horizon right? Did you vote?*) (Appx. B.20: 7).

Utterances near the negative pole are characterized by flattened U-shape intensity, lowest at the middle of the utterance, and low pitch and high speaking rate near the end of the utterance (Figure 7.11b).

## 7.4 Comparison of English and Spanish Dimensions

The five dimensions explain more of the variance in the data. Comparing the dimensions would tell us more about how prosody varies differently between the two languages: whether dimensions sharing function also share prosody, or the prosody of a dimension in one language is similar that of a dimension in the other.

The same pragmatic functions are equally present in both languages’ data, as the data collection protocol specifically requires participants to re-enact translations with the same feeling. Whether the associated prosody is similar for some or all functions in both languages is not well researched. This is particularly important for analysts working on speech-to-speech translation systems, who can use this knowledge to adapt the prosody of translations depending on the source and target languages.



(a) ES Dimension 5 positive.

(b) ES Dimension 5 negative.

Figure 7.11: ES Dimension 5 loadings.

The pragmatic functions of the first five dimensions of English and Spanish are summarized in Figure 7.12.

To facilitate the comparison of English and Spanish dimensions, I computed the cosine similarity of their respective loadings. By treating the loadings as vectors, the cosine of the angle between them can range from -1 to +1. A cosine similarity of +1 indicates that the loadings are identical, while a cosine similarity of -1 indicates that the loadings are in opposite directions.

If two loadings are aligned, this does not necessarily mean that their associated prosody of any two specific utterances is highly similar.

The cosine similarity between the loadings of the first five dimensions of English and Spanish are shown in Table 7.1.

Dimension 1 of English and Spanish are highly similar. Both dimensions also share function, representing the direction of focus. On the positive pole, attention is directed towards the speaker, and on the negative pole, attentions is directed towards the interlocutor.

1. **Focus on speaker.** Focus is directed towards the speaker *vs.* focus is directed towards the interlocutor.
2. **Engaged/Animated.** High engagement, highly animated *vs.* low engagement, not animated.
3. **Existence of shared understanding.** Lack *vs.* existence of shared understanding.
4. **Intent to continue topic.** Intent to close the current topic *vs.* intent to continue the topic.
5. **Checking existence of shared understanding.** Checking whether the speaker and interlocutor share knowledge *vs.* lack of checking.

(a) Pragmatic function of English dimensions.

1. **Focus on speaker.** Focus is directed towards the speaker *vs.* focus is directed towards the interlocutor.
2. Engaged/Animated. High engagement, highly animated *vs.* low engagement, not animated.
3. **Predictability.** Predictable information *vs.* unpredictable information.
4. **Authority.** The speaker holds higher authority compared to the interlocutor or a third party *vs.* the speaker holds lower authority.
5. **Certainty.** The speaker is certain about the information they are delivering *vs.* the speaker is uncertain.

(b) Pragmatic function of Spanish dimensions.

Figure 7.12: Pragmatic function of the first five dimensions of the English and Spanish

Table 7.1: Pairwise cosine similarity of the first five dimensions of English (EN) and Spanish (ES) reduced dimensionality prosody representation. Similarities of the same index dimensions are in bold.

| | ES Dim. 1 | ES Dim. 2 | ES Dim. 3 | ES Dim. 4 | ES Dim. 5 |
|---|---:|---:|---:|---:|---:|
| EN Dim. 1 | **0.997** | -0.009 | 0.002 | 0.016 | -0.012 |
| EN Dim. 2 | 0.015 | **0.978** | 0.063 | -0.021 | 0.084 |
| EN Dim. 3 | -0.007 | 0.077 | **-0.841** | 0.456 | -0.186 |
| EN Dim. 4 | -0.002 | -0.043 | 0.448 | **0.776** | -0.033 |
| EN Dim. 5 | -0.032 | 0.091 | 0.220 | 0.215 | **-0.393** |

Prosodic features have similar contribution to the dimensions (loadings cosine similarity = 0.99). However, the prosody is not identical. English utterances near the positive pole, compared to Spanish utterances, tend to avoid creakiness in the middle of long utterances.

Dimension 2 of English and Spanish are also highly similar. Both dimensions share function, representing engagement or how animated the speaker is. Prosodic features have similar contribution to the dimensions (cosine similarity = 0.97).

Dimension 3 of English and Spanish are similar. Unlike the previous two dimensions, their cosine similarity is negative (cosine similarity = –0.84). The negative, high magnitude means their loadings are largely antiparallel, or opposites. In other words, the poles of one dimension correspond to the opposite poles of the other. Considering them as opposites, the dimensions are not identical, but are similar. The existence of shared understanding (English Dimension 3) is related to the predictability of information (Spanish Dimension 3). If the speaker and interlocutor share understanding, the information they exchange becomes more predictable.

Dimension 4 of English and Spanish are moderately similar, with loadings cosine similarity = 0.77. High authority (Spanish Dimension 4, negative) might be related to the intent to close topic (English Dimension 4, positive). For example, if a speaker has high authority, they may feel more comfortable closing the topic since they have more control over the direction of the conversation. However, the authority in Spanish Dimension 4 is not always with respect to the interlocutor, so this may be true only in some cases.

Dimension 5 of English and Spanish are the least similar. Checking the existence of shared knowledge (English Dimension 5) is loosely related to certainty (Spanish Dimension 5). Their respective loadings are the least similar (cosine similarity = -0.39).

## 7.5 Utility for Modeling Cross-Language Prosody Mappings

Reducing the dimensionality of the prosody representation may improve the accuracy of a prosody mapping model. Eliminating noise may help prevent overfitting. To test the utility of the reduced dimensionality representation for modeling the mapping of prosody, I use it for same task from Chapter 6.

The output is the same, but the input to the linear regression model is different. After computing the prosody representation of an utterance, the transformation learned from the training data is applied to it, then truncated to the number of dimensions. The task becomes: Predict the full-dimensional prosody representation of an utterance in the target language from the reduced-dimensionality prosody representation of the utterance in the source language, as shown in Figure 7.13.

I recorded the average error as the number of predictors increased from 1 to 100, the maximum number of predictors, shown in Figure 7.14 and Figure 7.15. In the English-to-Spanish translation task, the average error was lowest with 34 predictors. Conversely, in the Spanish-to-English translation task, the average error was lowest with 31 predictors.

The reduced dimensionality representation outperformed the full-dimensional representation in both tasks. These observations suggest there is an optimal number of predictors for mapping prosody, and is influenced by the specific source and target language pair.

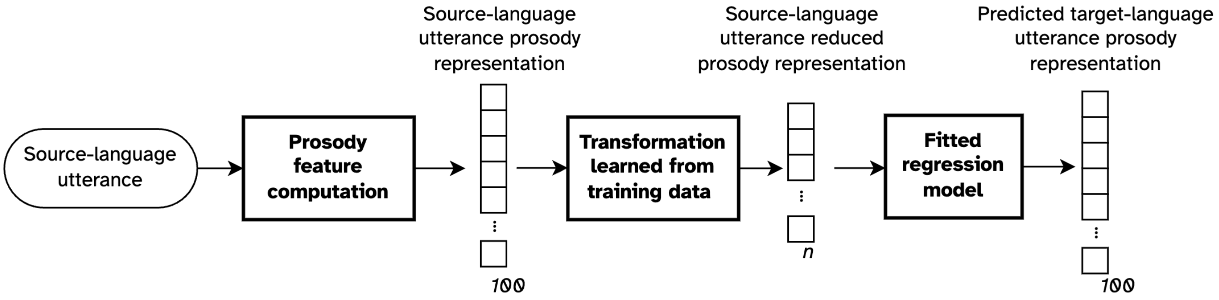

Source-language utterance
Prosody feature computation
Source-language utterance prosody representation
100
Transformation learned from training data
Source-language utterance reduced prosody representation
n
Fitted regression model
Predicted target-language utterance prosody representation
100

Figure 7.13: Linear regression model with reduced dimensionality representation as input.

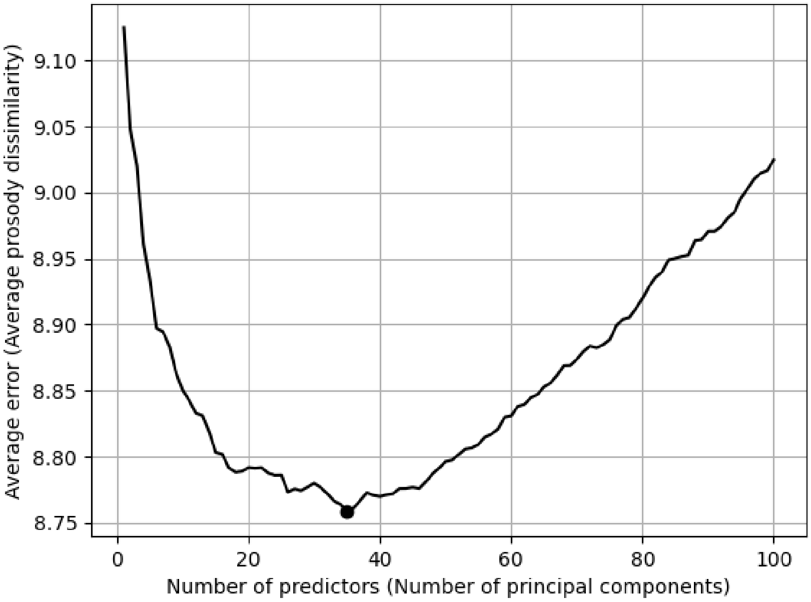


Figure 7.14: Linear regression model English-to-Spanish average error with increasing number of dimensions. Average error at 100 is 9.02, lowest at 34 is 8.76.

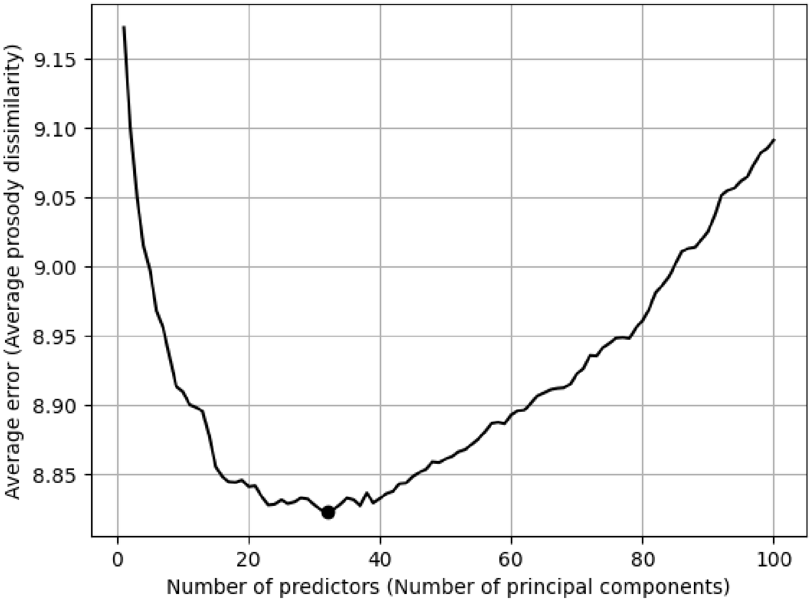


Figure 7.15: Linear regression model Spanish-to-English average error with increasing number of dimensions. Average error at 100 is 9.09, lowest at 31 is 8.82.

# Chapter 8

# Significance

## 8.1 Summary of the Problems Addressed

Speech-to-speech translation systems are valuable tools for cross-lingual communication, helping individuals overcome language barriers by providing quick, accessible translations. Today, such systems are effective for short, transactional exchanges, but are less effective for long-form conversations. One reason for this limitation is their general inability to adequately translate the nuances of prosody essential to its pragmatic functions, such as conveying intents and stances.

Without reliable context-appropriate prosody, users of these systems face challenges in engaging in natural conversation with others who do not speak the same language, thereby posing a barrier to deepening interpersonal relationships and achieving social inclusion.

The relationship between prosody and its pragmatic functions across languages has remained a relatively unexplored research area. This holds true even for globally prevalent languages such as English and Spanish, where our understanding of prosodic differences is sparse beyond a few topics such as turn-taking, questions, and declaratives, and is primarily focused on intonation and duration.

Accordingly, the aim of this research was to identify and characterize aspects of prosody and their pragmatic functions that pose challenges in speech-to-speech translation. The ultimate goal is to improve the pragmatic fidelity of speech-to-speech translation systems, thus extending their functional scope.

## 8.2 Contributions and Implications

**The Dialogs Re-enacted Across Languages corpus and protocol.** The Dialogs Re-enacted Across Languages (DRAL) corpus consists of 3816 matched English and Spanish utterances from spontaneous and re-enacted dialogs. The corpus was sourced from conversations between pairs of bilingual speakers, who subsequently re-enacted utterances to produce translations with equivalent prosody. It thus offers examples of many of the pragmatic functions of prosody in dialog and their corresponding translations across the two languages.

Multilingual corpora have predominantly relied on speech sources such as individuals reading texts, delivering presentations, or engaging in scripted conversations, or synthesized speech. Such corpora lack adequate representation of the spontaneity and prosodic nuances and fidelity of natural one-on-one interactions.

In light of the considerable advancements made possible by data-driven approaches, speech corpora have become invaluable resources for speech-to-speech translation research. As a publicly accessible resource, the DRAL corpus promotes progress in the field. The corpus serves as a resource for investigations of English and Spanish prosody, and the evaluation of speech-to-speech translation systems.

**A representation of utterance prosody.** The representation of utterance prosody developed here encodes many aspects of prosody important to its pragmatic functions in dialog. Its prosodic features are easily computed, designed to be robust to speaker differences, and interpretable.

Existing sentence and utterance representations frequently overlook prosody, which may be instrumental in modeling the pragmatic meaning of speech. These representations entangle these aspects amidst others, or instead focus on other aspects of speech, such as its lexical aspects.

The broad set of prosodic features incorporated into this representation is advantageous for comparative examinations of cross-lingual prosody. For example, the prosodic feature correlation analysis yielded observations on the general relationship between English and Spanish, such as the correlation identified between English intensity and Spanish CPPS found in utterances where the speaker is planning a follow-up utterance.

**A metric for prosodic similarity of utterances.** The baseline metric developed here estimates the prosodic similarity of two utterances. By comparing their respective representations of prosody, this metric focuses on aspects of prosody important to its pragmatic functions in dialog.

Evaluations of the performance of speech-to-speech translation models often rely on human judgments. Human evaluation is expensive, is often limited to assessing qualities related to naturalness, and lacks consistent reproducibility. This has motivated the development of automatic metrics to supplement human evaluation. These metrics primarily focus on lexical content, or otherwise gauge semantic similarity, focusing on the literal meaning of utterances rather than their contextual meaning within a conversation. A metric suitable for the evaluation of prosodic similarity of utterances from dialog has been absent.

The proposed metric addresses this absence, incorporating many prosodic aspects of speech that contribute to perceptions of similarity. Ultimately, this metric is a step towards the development of a standard metric for the evaluation of the pragmatic fidelity of speech-to-speech translations.

**A reduced dimensionality representation of utterance prosody.** The representation of utterance prosody with reduced dimensionality offers a compact alternative to the baseline representation. It condenses the number of features down to those contributing much of the variance observed in English and Spanish utterances, each a linear combination of the original features. Given the absence of a standard representation of utterance prosody, the reduced-dimensionality representation explores this possibility.

A reduced-dimensionality representation may be used to reduce the number of parameters speech-to-speech translation models. Using this reduced-dimensionality representation, I built and tested a more streamlined model for prosody translation and achieved improved performance compared to the original model. This gain in performance may apply to a full-fledged speech-to-speech translation system, wherein its modeling of cross-lingual prosody can be condensed for efficiency.

Further, this representation is interpretable. I identify meanings for the first five dimensions of the English and Spanish representations, and compare the dimensions of the two languages.

**Analysis of cross-language prosody mapping modeling approaches.** This research contributes observations from the assessment of simple models mapping English and Spanish prosody. The models are based solely on representations of utterance prosody, ignoring cues such as an utterance's lexical context or surrounding context. Thus, their performance provides an indication of the extent to which prosody can be directly translated between languages.

The analysis has two implications for future improvements to speech-to-speech translation. First, ignoring the source-language prosody proves insufficient. Observations showed that effective cross-language translation requires attention to prosodic features beyond pitch and duration. These include, at least, breathy voice, creaky voice, and intensity.

Second, direct transfer of the source-language prosody also falls short. The prosody of certain pragmatic functions as they occur in dialog differs in previously unsuspected ways across languages. These differences include, at least, prosody involved in grounding (Section 6.5.2), leading into a topic or point (Section 4.3), expressing uncertainty or hesitation (Section 6.5.1), and taking the turn (Section 6.5.1).

## 8.3 Limitations and Future Work

This research relied on a novel corpus of matched English and Spanish utterances, a novel representation of utterance prosody, and a novel metric for prosodic similarity. These enabled the analysis of cross-language prosody mapping models to obtain observations of challenges in mapping English and Spanish prosody presents. These observations include those on what prosody is conveying in the two languages, how it does it, and what the differences are. Extensions and improvements would enable future research to produce a clearer and broader picture of the pragmatic functions of prosody in dialog.

The DRAL protocol faces challenges of the costs associated with collecting high-quality matched speech data, challenges not unique to this corpus. Future work, however, could see an expansion of the corpus to include additional languages and dialects. The technical report [80] describes the protocol in detail, as well as our design experiences, as a reference for fellow researchers. Future work may expand the corpus or create similar corpora, for example, by following the DRAL protocol. Future work may overcome the obstacle of high costs by making efficient use of limited data, such as by exploiting per-language or joint self-supervised training techniques. Such an expansion would enhance the corpus's utility, such as its application in training speech-to-speech systems and testing the generalizability of translation techniques.

The baseline representation of utterance prosody is fixed-length, based on linear scaling. Future work may explore other methods for designing or learning representations. These include, for example, leveraging the surrounding context to incorporate the prosody that precedes or follows an utterance, and adapting speech embeddings specifically for dialog. Other possible improvements include incorporating features not currently present, such as vibrato, nasalization, and reduction.

The baseline metric operates under the assumption that all prosodic features equally contribute to the perception of similarity. Future work may improve the metric by collecting human judgments of similarity of utterance pairs, or judgments of the metric's estimates of similarity. These judgments could then be used to improve the correlation between the metric's estimates and human judgments, by reweighing features' contributions to the estimate. Beyond use for evaluation of speech-to-speech translations, an improved metric may potentially be incorporated into a model's loss function for use during training. Additionally, it may be used to curate a dataset with particular prosody, by collecting similar utterances from one or more existing corpora.

The analysis techniques and findings presented here could inform the design of a specific prosody-translation module, and inspire the development of synthesizers capable of following a rich prosody specification and thereby conveying a wide range of pragmatic functions.

Well-designed prosody translation techniques will be important for effective speech-to-speech translation. Developing these techniques has the potential to convey many more pragmatic functions that have been previously addressed. Improving the translation of dialog will bring us closer to realizing translations with reliable, context-appropriate prosody. This work has brought us one step closer to this goal.

# Bibliography

[1] Rosana Ardila et al. "Common Voice: A Massively-Multilingual Speech Corpus". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, 2020, pp. 4218–4222. ISBN: 979-10-95546-34-4.

[2] Jonathan E. Avila and Nigel G. Ward. "Towards Cross-Language Prosody Transfer for Dialog". In: *Interspeech 2023*. 2023, pp. 2143–2147. DOI: `10.21437/Interspeech.2023-1152`.

[3] Alexei Baevski et al. "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS'20. Curran Associates Inc., 2020, pp. 12449–12460. ISBN: 978-1-71382-954-6.

[4] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, 2005, pp. 65–72. URL: `https://aclanthology.org/W05-0909`.

[5] Anne Berry. *Spanish and American Turn-Taking Styles: A Comparative Study*. Tech. rep. ED398747. Education Resources Information Center, 1994.

[6] Paul Boersma and David Weenink. *Praat: Doing Phonetics by Computer*. Version 6.3.10. 2023. URL: `http://www.praat.org/`.

[7] William Brannon, Yogesh Virkar, and Brian Thompson. "Dubbing in Practice: A Large Scale Study of Human Localization With Insights for Automatic Dubbing". In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 419–435. ISSN: 2307-387X. DOI: `10.1162/tacl_a_00551`.

[8] Gillian Brown. "Prosodic Structure and the Given/New Distinction". In: *Prosody: Models and Measurements*. Ed. by Willem J. M. Levelt, Anne Cutler, and D. Robert Ladd. Vol. 14. Springer Berlin Heidelberg, 1983, pp. 67–77. ISBN: 978-3-642-69105-8 978-3-642-69103-4. DOI: 10.1007/978-3-642-69103-4_6.

[9] T. Bub, W. Wahlster, and A. Waibel. "Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. 1997, pp. 71–74. ISBN: 978-0-8186-7919-3. DOI: 10.1109/ICASSP.1997.607199.

[10] F. Casacuberta et al. "Some Approaches to Statistical and Finite-State Speech-to-Speech Translation". In: *Computer Speech & Language* 18.1 (2004), pp. 25–47. ISSN: 08852308. DOI: 10.1016/S0885-2308(03)00028-7.

[11] Mingda Chen et al. "BLASER: A Text-Free Speech-to-Speech Translation Evaluation Metric". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023, pp. 9064–9079. DOI: 10.18653/v1/2023.acl-long.504.

[12] Coqui. *Coqui TTS*. Coqui. 2023. DOI: 10.5281/zenodo.6334862. URL: https://github.com/coqui-ai/TTS.

[13] Gilles Degottex et al. "COVAREP — A Collaborative Voice Analysis Repository for Speech Technologies". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 960–964. ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6853739.

[14] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[15] Mattia A. Di Gangi et al. "MuST-C: A Multilingual Speech Translation Corpus". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019, pp. 2012–2017. DOI: 10.18653/v1/N19-1202.

[16] Shaojin Ding et al. "Golden Speaker Builder – An Interactive Tool for Pronunciation Training". In: *Speech Communication* 115 (2019), pp. 51–66. ISSN: 01676393. DOI: 10.1016/j.specom.2019.10.005.

[17] Quoc Truong Do et al. "Improving Translation of Emphasis with Pause Prediction in Speech-to-Speech Translation Systems". In: *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers*. 2015, pp. 204–208.

[18] Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. "Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data". In: *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*. Association for Computational Linguistics, 2021, pp. 226–235. DOI: 10.18653/v1/2021.iwslt-1.27.

[19] Qianqian Dong et al. "Leveraging Pseudo-labeled Data to Improve Direct Speech-to-Speech Translation". In: *Interspeech 2022*. ISCA, 2022, pp. 1781–1785. DOI: 10.21437/Interspeech.2022-10011.

[20] Paul Edmunds. "Relationship of Prosody by Spanish Speakers of English as a Second Language on the Perception of Intelligibility and Accentedness by Native English Listeners." In: *The Journal of the Acoustical Society of America* 125.4 (2009), pp. 2766–2766. ISSN: 0001-4966. DOI: 10.1121/1.4784704.

[21] Florian Eyben, Martin Wöllmer, and Björn Schuller. "openSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor". In: *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 2010, pp. 1459–1462. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874246.

[22] Maria Gabriela Valenzuela Farías. “A Comparative Analysis of Intonation between Spanish and English Speakers in Tag Questions, Wh-Questions, Inverted Questions, and Repetition Questions”. In: *Revista Brasileira de Linguística Aplicada* 13.4 (2013), pp. 1061–1083. ISSN: 1984-6398. DOI: 10.1590/S1984-63982013005000021.

[23] Susanne Fuchs et al. “Acoustic and Respiratory Evidence for Utterance Planning in German”. In: *Journal of Phonetics* 41.1 (2013), pp. 29–47. ISSN: 00954470. DOI: 10.1016/j.wocn.2012.08.007.

[24] Federico Gaspari and John Hutchins. “Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects”. In: *Proceedings of Machine Translation Summit XI: Papers*. 2007.

[25] Munich Artificial Intelligence Laboratories GmbH. *The M-AILABS Speech Dataset – Caito*. https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/. 2019.

[26] Mattias Heldner et al. “Voice Quality as a Turn-Taking Cue”. In: *Interspeech 2019*. The International Speech Communication Association (ISCA), 2019, pp. 4165–4169. DOI: 10.21437/Interspeech.2019-1592.

[27] Yolanda D. Heman-Ackah et al. “Cepstral Peak Prominence: A More Reliable Measure of Dysphonia”. In: *Annals of Otology, Rhinology & Laryngology* 112.4 (2003), pp. 324–333. ISSN: 0003-4894. DOI: 10.1177/000348940311200406.

[28] Wei-Ning Hsu et al. “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3451–3460. ISSN: 2329-9290, 2329-9304. DOI: 10.1109/TASLP.2021.3122291.

[29] Wen-Chin Huang et al. *A Holistic Cascade System, Benchmark, and Human Evaluation Protocol for Expressive Speech-to-Speech Translation*. 2023. DOI: 10.48550/arXiv.2301.10606.

[30] Keith Ito and Linda Johnson. *The LJ Speech Dataset*. `https://keithito.com/LJ-Speech-Dataset/`. 2017.

[31] Ye Jia et al. "CVSS Corpus and Massively Multilingual Speech-to-Speech Translation". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, 2022, pp. 6691–6703.

[32] Ye Jia et al. "Leveraging Unsupervised and Weakly-Supervised Data to Improve Direct Speech-to-Speech Translation". In: *Interspeech 2022*. ISCA, 2022, pp. 1721–1725. DOI: `10.21437/Interspeech.2022-10938`.

[33] Ye Jia et al. "Translatotron 2: High-quality Direct Speech-to-Speech Translation with Voice Preservation". In: *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022, pp. 10120–10134.

[34] Takatomo Kano et al. "A Method for Translation of Paralinguistic Information". In: *Proceedings of the 9th International Workshop on Spoken Language Translation: Papers*. 2012, pp. 158–163.

[35] Rachel H. Kessinger and Sheila E. Bumstein. "Effects of Speaking Rate on Voice-Onset Time and Vowel Production: Some Implications for Perception Studies". In: *Journal of Phonetics* 26.2 (1998), pp. 117–128. DOI: `10.1006/jpho.1997.0069`.

[36] Sameer Khurana, Antoine Laurent, and James Glass. "SAMU-XLSR: Semantically-Aligned Multimodal Utterance-Level Cross-Lingual Speech Representation". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1493–1504. ISSN: 1932-4553, 1941-0484. DOI: `10.1109/JSTSP.2022.3192714`.

[37] Ji Young Kim. "Spanish–English Cross-Linguistic Influence on Heritage Bilinguals' Production of Uptalk". In: *Languages* 8.1 (2023), p. 22. ISSN: 2226-471X. DOI: `10.3390/languages8010022`.

[38] A. Lavie et al. “Janus-III: Speech-to-Speech Translation in Multiple Languages”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE Comput. Soc. Press, 1997, pp. 99–102. ISBN: 978-0-8186-7919-3. DOI: 10.1109/ICASSP.1997.599557.

[39] Ann Lee et al. “Direct Speech-to-Speech Translation With Discrete Units”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022, pp. 3327–3339. DOI: 10.18653/v1/2022.acl-long.235.

[40] Ann Lee et al. “Textless Speech-to-Speech Translation on Real Data”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2022, pp. 860–872. DOI: 10.18653/v1/2022.naacl-main.63.

[41] Daniel J. Liebling et al. “Unmet Needs and Opportunities for Mobile Translation AI”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–13. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376261.

[42] Guan-Ting Lin et al. “On the Utility of Self-Supervised Models for Prosody-Related Tasks”. In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1104–1111. ISBN: 9798350396904. DOI: 10.1109/SLT54892.2023.10023234.

[43] Leena Mary et al. “Evaluation of Mimicked Speech Using Prosodic Features”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7189–7193. ISBN: 978-1-4799-0356-6. DOI: 10.1109/ICASSP.2013.6639058.

[44] Shikib Mehri et al. *Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges*. Tech. rep. arXiv:2203.10012. arXiv, 2022.

[45] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. DOI: 10.48550/arXiv.1301.3781.

[46] Abdelrahman Mohamed et al. "Self-Supervised Speech Representation Learning: A Review". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pp. 1179–1210. ISSN: 1932-4553, 1941-0484. DOI: 10.1109/JSTSP.2022.3207050.

[47] Alp Öktem, Mireia Farrús, and Antonio Bonafonte. "Corpora Compilation for Prosody-Informed Speech Processing". In: *Language Resources and Evaluation* 55.4 (2021), pp. 925–946. ISSN: 1574-0218. DOI: 10.1007/s10579-021-09556-2.

[48] OpenAI. *Whisper*. OpenAI. 2023. URL: https://github.com/openai/whisper.

[49] Kishore Papineni et al. "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Association for Computational Linguistics, 2001, pp. 311–318. DOI: 10.3115/1073083.1073135.

[50] Jeffrey Pennington, Richard Socher, and Christopher Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[51] Matthew Peters et al. "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202.

[52] Sravya Popuri et al. "Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation". In: *Proc. Interspeech 2022*. ISCA, 2022, pp. 5195–5199. DOI: 10.21437/Interspeech.2022-11032.

[53] Matt Post et al. "Improved Speech-to-Text Translation with the Fisher and Callhome Spanish-English Speech Translation Corpus". In: *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*. 2013.

[54] Daniel Povey et al. "The Kaldi Speech Recognition Toolkit". In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[55] Vineel Pratap et al. "MLS: A Large-Scale Multilingual Dataset for Speech Research". In: *Proc. Interspeech 2020*. ISCA, 2020, pp. 2757–2761. DOI: 10.21437/Interspeech.2020-2826.

[56] Dolores Ramírez Verdugo. "The Nature and Patterning of Native and Non-Native Intonation in the Expression of Certainty and Uncertainty: Pragmatic Effects". In: *Journal of Pragmatics* 37.12 (2005), pp. 2086–2115. ISSN: 03782166. DOI: 10.1016/j.pragma.2005.02.012.

[57] Rajiv Rao, Ting Ye, and Brianna Butera. "The Prosodic Expression of Sarcasm vs. Sincerity by Heritage Speakers of Spanish". In: *Languages* 7.1 (2022), p. 17. ISSN: 2226-471X. DOI: 10.3390/languages7010017.

[58] Mirco Ravanelli et al. *SpeechBrain: A General-Purpose Speech Toolkit*. 2021. DOI: 10.48550/arXiv.2106.04624. arXiv: 2106.04624.

[59] Ricardo Rei et al. "COMET: A Neural Framework for MT Evaluation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213.

[60] Albert Rilliard, Alexandre Allauzen, and Philippe Boula De Mareüil. "Using Dynamic Time Warping to Compute Prosodic Similarity Measures". In: *Interspeech 2011*. ISCA, 2011, pp. 2021–2024. DOI: 10.21437/Interspeech.2011-531.

[61] Andrew Rosenberg. "AuToBI — A Tool for Automatic ToBI Annotation". In: *Eleventh Annual Conference of the International Speech Communication Association*. ISCA, 2010, pp. 146–149. DOI: 10.21437/Interspeech.2010-71.

[62] Fatiha Sadat et al. “PORTAGE: A Phrase-Based Machine Translation System”. In: *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Association for Computational Linguistics, 2005, pp. 129–132.

[63] Elizabeth Salesky et al. “The Multilingual TEDx Corpus for Speech Recognition and Translation”. In: *Proc. Interspeech 2021*. ISCA, 2021, pp. 3655–3659. DOI: 10.21437/Interspeech.2021-11.

[64] Joel Shor et al. “Towards Learning a Universal Non-Semantic Representation of Speech”. In: *Proc. Interspeech 2020*. ISCA, 2020, pp. 140–144. DOI: 10.21437/Interspeech.2020-1242.

[65] Kim Silverman et al. “TOBI: A Standard for Labeling English Prosody”. In: *2nd International Conference on Spoken Language Processing (ICSLP 1992)*. ISCA, 1992, pp. 867–870. DOI: 10.21437/ICSLP.1992-260.

[66] R. J. Skerry-Ryan et al. “Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron”. In: *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 4693–4702.

[67] Jakub Swiatkowski et al. “Expressive Machine Dubbing Through Phrase-level Cross-lingual Prosody Transfer”. In: (2023). DOI: 10.48550/arXiv.2306.11662. arXiv: 2306.11662.

[68] Toshiyuki Takezawa et al. “A Japanese-to-English Speech Translation System: ATR-MATRIX”. In: *5th International Conference on Spoken Language Processing (ICSLP)*. ISCA, 1998, paper 0957. DOI: 10.21437/ICSLP.1998-581.

[69] Peter Toma. “Systran as a Multilingual Machine Translation System”. In: *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the Language Barrier*. 1977, pp. 569–581.

[70] Jacqueline Vaissière. “Language-Independent Prosodic Features”. In: *Prosody: Models and Measurements*. Ed. by Willem J. M. Levelt, Anne Cutler, and D. Robert Ladd. Vol. 14. Springer Berlin Heidelberg, 1983, pp. 53–66. ISBN: 978-3-642-69105-8 978-3-642-69103-4. DOI: `10.1007/978-3-642-69103-4\_5`.

[71] Petra Wagner and Andreas Windmann. “Re-Enacted and Spontaneous Conversational Prosody: How Different?” In: *Speech Prosody 2016*. ISCA, 2016, pp. 518–522. DOI: `10.21437/SpeechProsody.2016-106`.

[72] Changhan Wang et al. “CoVoST 2 and Massively Multilingual Speech Translation”. In: *Proc. Interspeech 2021*. ISCA, 2021, pp. 2247–2251. DOI: `10.21437/Interspeech.2021-2027`.

[73] Changhan Wang et al. “VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021, pp. 993–1003. DOI: `10.18653/v1/2021.acl-long.80`.

[74] Nigel Ward. *Midlevel Prosodic Features Toolkit*. 2022. URL: `https://github.com/nigelgward/midlevel`.

[75] Nigel Ward and Saiful Abu. “Action-Coordinating Prosody”. In: *Speech Prosody*. ISCA, 2016, pp. 629–633. DOI: `10.21437/SpeechProsody.2016-129`.

[76] Nigel Ward et al. “Two Pragmatic Functions of Breathy Voice in American English Conversation”. In: *Speech Prosody*. 2022, pp. 82–86. DOI: `10.21437/SpeechProsody.2022-17`.

[77] Nigel G. Ward. *Prosodic Patterns in English Conversation*. Cambridge University Press, 2019. ISBN: 978-1-316-86355-8.

[78] Nigel G. Ward and Jonathan E. Avila. "A Dimensional Model of Interaction Style Variation in Spoken Dialog". In: *Speech Communication* 149 (2023), pp. 47–62. ISSN: 01676393. DOI: 10.1016/j.specom.2023.03.002.

[79] Nigel G. Ward and Paola Gallardo. "Non-Native Differences in Prosodic-Construction Use". In: *Dialogue & Discourse* 8.1 (2017), pp. 1–30. ISSN: 2152-9620. DOI: 10.5087/dad.2017.101.

[80] Nigel G. Ward et al. *Dialogs Re-enacted Across Languages, Version 2*. Tech. rep. UTEP-CS-23-27. University of Texas at El Paso, 2023. URL: https://www.cs.utep.edu/nigel/abstracts/dral-techreport2.html.

[81] Nigel G. Ward et al. "Inferring Stance from Prosody". In: *Proc. Interspeech 2017*. ISCA, 2017, pp. 1447–1451. DOI: 10.21437/Interspeech.2017-159.

[82] Jack Weston et al. "Learning De-Identified Representations of Prosody from Raw Audio". In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 11134–11145.

[83] Jiahong Yuan, Mark Liberman, and Christopher Cieri. "Towards an Integrated Understanding of Speaking Rate in Conversation". In: *Proc. Interspeech 2006*. ISCA, 2006. DOI: 10.21437/Interspeech.2006-204.

[84] Marcely Zanon Boito et al. "MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 2020, pp. 6486–6493. ISBN: 979-10-95546-34-4.

[85] Germán Zárate-Sández. "Production of Final Boundary Tones in Declarative Utterances by English-speaking Learners of Spanish". In: *Speech Prosody 2018*. ISCA, 2018, pp. 927–931. DOI: 10.21437/SpeechProsody.2018-187.

[86] Ya-Jie Zhang et al. "Learning Latent Representations for Style Control and Transfer in End-to-end Speech Synthesis". In: *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949. ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8683623.

[87] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: 2019. DOI: 10.48550/arXiv.1904.09675.

[88] Wei Zhao et al. "MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019, pp. 563–578. DOI: 10.18653/v1/D19-1053.

# Appendix A

# Project Repository Description

The GitHub repository *Code for Dialogs Re-enacted Across Languages (DRAL)* is available at: `https://github.com/joneavila/DRAL`

## A.1 Overview of Repository Contents

Contained within this repository are the code and its associated documentation for various tasks, including:

- Post-processing of the DRAL corpus,
- Computation of prosodic features from utterances (modifications to the Mid-level Prosodic Features Toolkit [74])
- Analysis of prosodic feature value correlations
- Viewing utterances estimated as highly similar or dissimilar
- Execution of models, including transcription and synthesis for baseline models

## A.2 Modifications and Improvements to the Mid-level Prosodic Features Toolkit

### A.2.1 Adapting Feature Computation for Utterances

The individual utterances from the DRAL corpus are significantly shorter than the speech the Mid-level Prosodic Features Toolkit was designed and tested on. Computing the features for DRAL utterances required some modifications to the code.

The full length original and re-enactment conversation audios were impractical for this purpose due to large portions of silence, background noise, and other irrelevant speech. Including this in the normalization described in Section 3.2 would have made the feature computation less reliable.

Instead of using the full length conversation audios, I created a new pair of audios for both conversation audios by concatenating the utterances of each speaker. This transforms the initial two conversation audios into four more concise audios. Each of the new audios contain only audio from one speaker, and only the utterances that were selected and re-enacted in the other language.

I proceed with the feature computation as described in Section 3.2.

Similar modifications were made to the code for computing PCA, as used in Chapter 7.

### A.2.2 Optimizing the CPPS Feature Computation

I improved the speed of the function for computing the CPPS feature by eliminating redundant computations. Specifically, the original function for computing CPPS averages over 5 CPPS (which come from 5 spectrogram windows) values that span 10 ms. The modified version eliminates the extra computations and the need for averaging.

Comparing the original and modified versions of the function on utterances from DRAL Conversation 1–18, the modified version ran between 5 and 6 times faster. This modification would significantly improve the computation time over the entire corpus, However, the computed CPPS of the two functions had a mean correlation of $\rho = 0.913$, and was too weak for some audios to comfortably use the modified version in place of the original.

This might be due to using concatenated audios, since some windows used in the computation overlap the splices. To test this, I ran the test on the same first 18 conversations, using the original conversations. This also resulted in weak correlations for some audios.

The modified version might be more accurate, but testing this would require comparing the computed CPPS with perceptions of breathy voice, and such annotation does not exist, and conducting a data collection is beyond the scope of this project. Ultimately, I continued using the original function but include the modified version in the repository.

### A.2.3 Enforcing REAPER for Pitch Computation

The pitch computation enforces a more robust pitch tracker, transitioning from RAPT (Robust Algorithm for Pitch Tracking) to REAPER (Robust Epoch and Pitch EstimatoR).

### A.2.4 Standardizing Feature Vector Lengths

The original code allows for feature vectors of different lengths, usually different in length by one or two frames. In the modified code, feature vectors are padded to the expected length so that all features have the same length.

# Appendix B

# Lexical Content of English and Spanish Dimension Extremes

## B.1 English Dimension 1: High

1. EN_021_13 *And the beach is really strange because it's like a, you see like, the beach is not like straight, like, it was like a doughnut*

2. EN_020_12 *Umm, hmm, I like, walking*

3. EN_025_47 *Yeah, so that's like, I don't know, all, it's not high BPM, or it's not super uplifting*

4. EN_029_26 *I think if-if I ever found that out, like, about, like, my dad, like if he was involved in that stuff, I don't think, like, I could ever look at him, like, the same*

5. EN_029_17 *I think my grandma was a great person with me, but my mom thinks she was like the horrible- the horrible person*

6. EN_029_30 *Ah no that's-that's actually, like, that's really sad because like, the gang violence down there is terrible*

7. EN_050_11 *I like working outside, and then I just ask, and they are like "Okay," and then they buy it and that's cool*

8. EN_020_5 *People say it so often they're like "Oh I-I don't think there's anything to do, or it's, like, boring"*

9. EN_029_13 *The bad guy is like his father figure, so it's, like, really emotional when you get that plot twist that he's actually the bad guy*

10. EN_025_2 *Yeah so, I was gonna have a hard time, like, describing the type of music that I listen to*

11. EN_032_2 *Everything is becoming more expensive, but they are not paying us the price we need to live*

12. EN_020_13 *And I like, exercising but not out for a hike*

13. EN_025_29 *Find out what that movie's called, but, yeah, they put artificial intelligence into this woman and*

14. EN_040_17 *I want in engineering and I don't see myself studying, uh, nursing or something like that*

15. EN_013_29 *But I feel like I still haven't had enough time to even explore, like, the subfields of computer science to know what I want to do*

16. EN_028_1 *Uh, I heard that you were gonna take twelve hours in the fall*

## B.2 English Dimension 1: Low

1. EN_009_48 *That's weird!*

2. EN_011_35 *How old are you?*

3. EN_005_30 *I have an older brother*

4. EN_002_2 *No*

5. EN_009_13 *Oh really?*

6. EN_011_42 *And you?* interlocutor by asking the same question back

7. EN_005_35 *It's easier*
8. EN_006_40 *Yeah, yeah*
9. EN_001_34 *Yeah dude*
10. EN_027_27 *Yeah*
11. EN_025_34 *Robot?*
12. EN_016_28 *Oh wow*
13. EN_005_34 *I'm ninteen*
14. EN_009_26 *Nice*
15. EN_005_16 *Well, yeah*
16. EN_025_28 *In that genre*

## B.3 English Dimension 2: High

1. EN_006_40 *Yeah, yeah*
2. EN_009_48 *That's weird!*
3. EN_006_21 *Psychology undergrad and then I'm*
4. EN_024_27 *We were just doing our thing*
5. EN_021_46 *You were that far?*
6. EN_021_52 *My shoulders are, like, killing [me]*
7. EN_006_38 *I remember that I got out of the cell*
8. EN_021_47 *Yeah, I-I, we didn't, like, know how long we were there*

9. EN_015_15 *Yeah, yeah*

10. EN_024_43 *Then not even because like everyone, everyone was like leaving, right?*

11. EN_050_9 *It's not as easy anymore*

12. EN_039_9 *I kind of agree with you*

13. EN_001_23 *One day I went on my own*

14. EN_005_3 *Uh, no but actually what I was gonna tell you was*

15. EN_027_10 *Yeah, if I could, I would*

16. EN_045_10 *Oh, oh! I-I have another blouse*

## B.4 English Dimension 2: Low

1. EN_009_45 *Then it's really easy, because*

2. EN_001_16 *Your adventures*

3. EN_028_34 *Faster*

4. EN_004_22 *Yeah*

5. EN_016_18 *It's better to ask questions and learn than to stay confused*

6. EN_011_24 *Mechatronics*

7. EN_009_16 *Yeah*

8. EN_009_12 *No, grandpa*

9. EN_026_12 *If you're working in that type of job Yes.*

10. EN_009_6 *Uh, we hang out*

11. EN_011_15 *And aspects that I couldn't, like, see back then*

12. EN_003_19 *Sad, face*

13. EN_016_36 *So yeah*

14. EN_006_2 *Whatever you want*

15. EN_003_7 *Interrupt?*

16. EN_026_7 *It's like, when you go to a place that serves, like, fast food*

## B.5 English Dimension 3: High

1. EN_015_5 *We haven't played online, but*

2. EN_019_4 *Okay, okay*

3. EN_010_19 *Oh my god, okay*

4. EN_019_14 *No, just kidding, we watched like one a day*

5. EN_019_48 *Oh, you get dizzy? Or what?*

6. EN_004_8 *It happens*

7. EN_019_13 *But, we watched like one every week*

8. EN_006_25 *I actually worked at the*

9. EN_006_34 *Barely on February*

10. EN_024_23 *No it was actually, there's-there's some in, um, El Paso*

11. EN_024_10 *Yeah, definitely, yeah*

12. EN_011_41 *And I really like Mejia because he is the one that is always like telling me “Hey you should apply to this, you should apply to this” so*

13. EN_006_1 *Do you want to start in English or Spanish?*

14. EN_020_20 *They probably need more space*

15. EN_023_12 *I always think I could rest a little more, but*

16. EN_023_6 *That would be*

## B.6 English Dimension 3: Low

1. EN_006_13 *Have their babies here*

2. EN_006_40 *Yeah, yeah*

3. EN_004_7 *So*

4. EN_006_21 *Psychology undergrad and then I’m*

5. EN_002_10 *You save money*

6. EN_006_38 *I remember that I got out of the cell*

7. EN_021_39 *I’m-I’m getting tired*

8. EN_007_8 *Mario Kart wasn’t even a thing back then*

9. EN_005_3 *Uh, no but actually what I was gonna tell you was*

10. EN_006_35 *Oh wow*

11. EN_025_14 *That the people like, and that, was also appropriate for a wedding*

12. EN_011_22 *So, I don’t really remember*

13. EN_003_12 *To put an order*

14. EN_025_57 *Beach House. They have pretty good music*

15. EN_024_27 *We were just doing our thing*

16. EN_003_10 *I still needed to study two*

## B.7 English Dimension 4: High

1. EN_006_16 *My, my grandpa and my grandma*

2. EN_015_28 *Lab or something*

3. EN_006_37 *Yeah, what it's gonna be like*

4. EN_040_4 *Well, I have a desktop*

5. EN_053_23 *Very competetive*

6. EN_010_11 *Go over there*

7. EN_053_6 *Yeah it's, it's kinda difficult to learn a new language*

8. EN_024_28 *I was sixteen, we were partying, you know*

9. EN_009_6 *Uh, we hang out*

10. EN_039_4 *Time that you can also be using to study*

11. EN_008_56 *Like in warehouses and all that*

12. EN_008_19 *Yeah, just my dad*

13. EN_024_8 *I have nothing, like, to do*

14. EN_006_13 *Have their babies here*

15. EN_028_15 *And then they work towards your degree*
16. EN_040_8 *I got some from UTEP*

## B.8 English Dimension 4: Low

1. EN_022_34 *Oh, yeah yeah*
2. EN_013_13 *Yeah, you know*
3. EN_003_25 *Why?*
4. EN_021_9 *It's, it was really cool*
5. EN_050_21 *This Friday*
6. EN_025_23 *Oh that's good, I like it*
7. EN_021_41 *I stop*
8. EN_011_42 *And you?*
9. EN_004_37 *Two?*
10. EN_008_48 *You can't?*
11. EN_005_16 *Well, yeah*
12. EN_039_5 *I kind of agree with you*
13. EN_027_27 *Yeah*
14. EN_011_11 *I mean*
15. EN_001_5 *Is, sorry, is Montse's boyfriend*
16. EN_039_9 *I kind of agree with you*

## B.9 English Dimension 5: High

1. EN_021_39 *I'm-I'm getting tired*
2. EN_011_5 *I started college there*
3. EN_050_21 *This Friday*
4. EN_022_18 *So yeah, yeah, like, I go out*
5. EN_024_47 *After like a couple hours*
6. EN_028_26 *Do you know which ones after these you're taking?*
7. EN_026_33 *Confront the people*
8. EN_023_5 *I don't know if those count, like, the three*
9. EN_023_34 *Turn it in?*
10. EN_011_23 *Studied mechatronics engineer*
11. EN_006_23 *And you want to work with, kids?*
12. EN_035_2 *So cool*
13. EN_027_19 *I feel like that could be a bucket list item.*
14. EN_021_10 *Well the thing is that it was really alone*
15. EN_036_2 *You shower here?*
16. EN_021_2 *I went with my dad, my mom, and my sister*

## B.10 English Dimension 5: Low

1. EN_006_21 *Psychology undergrad and then I'm*
2. EN_053_5 *Professors are not on top of you and nothing like that, and*
3. EN_003_12 *To put an order*
4. EN_005_3 *Uh, no but actually what I was gonna tell you was*
5. EN_006_40 *Yeah, yeah*
6. EN_004_26 *And you didn't like it?*
7. EN_017_1 *I don't know, I've always been pretty bad at placing blame*
8. EN_002_41 *They have a lot more athletic clothes, clothes that's more*
9. EN_018_8 *But like, how many times a semester?*
10. EN_024_29 *You know because of the, the area there was also, like, some crazy people there, not even gonna lie*
11. EN_006_13 *Have their babies here*
12. EN_004_7 *So*
13. EN_029_27 *I think it would completely, like, mess up, like, my perception of him, you know*
14. EN_011_38 *I'm nineteen, I'm turning twenty in January*
15. EN_003_10 *I still needed to study two*
16. EN_024_22 *Kind of downtown? Or*

## B.11 Spanish Dimension 1: High

1. ES_029_26 *Yo creo que si me entere que mi papá estaba haciendo esas cosas no creo que lo pod- lo podría ver lo mismo otra vez* (EN_029_26 *I think if-if I ever found that out, like, about, like, my dad, like if he was involved in that stuff, I don't think, like, I could ever look at him, like, the same*)

2. ES_013_29 *Siento que no he tenido suficiente tiempo para explorar las disciplinas de ciensas computacionales para saber lo que yo quiero hacer* (EN_013_29 *But I feel like I still haven't had enough time to even explore, like, the subfields of computer science to know what I want to do*)

3. ES_039_26 *Tiene ese miedo de que se va a romper la amistad que ya nunca me va a poder hablar como antes nos hablamos* (EN_039_26 *Have that fear of breaking the relationship and not be able to talk like they used to*)

4. ES_028_12 *Y otros tipos de electives no has considerado como* (EN_028_12 *And other types of electives you haven't considered like uh*)

5. ES_050_18 *Pero ya que trabajo aquí, ta más, me gusta significativamente aquí porque aquí me dan comida gratis* (EN_050_18 *Well now that I work here, I like working here significantly more because they give me free food*)

6. ES_026_14 *Si como que, solamente eres bueno para eso y* (EN_026_14 *Yeah it's like, you're only good at doing that and*)

7. ES_001_10 *¿En qué época está, pues si, en que- en que época sucede o sea, después el imperio? ¿Durante el imperio?* (EN_001_10 *So in what time is it? Like, yeah, what time period is it in? Uh, is it, after the empire? During the empire?*)

8. ES_029_17 *Yo pienso que mi abuela era una buena persona, pero mi mamá piensa que es una horrible, digo, horrible persona* (EN_029_17 *I think my grandma was a great person with me, but my mom thinks she was like the horrible- the horrible person*)

9. ES_026_9 *Pues están allí para, solamente para hacer la comida, hacen la comida y nada más* (EN_026_9 *Well they're there to serve you the food, to make you the food and nothing else*)

10. ES_029_30 *La verdad es- está muy triste porque la violencia de pandillas está muy malo allá abajo* (EN_029_30 *Ah no that's-that's actually, like, that's really sad because like, the gang violence down there is terrible*)

11. ES_054_12 *Si Beto para allá vale verga, estás como, okay, bueno, está horrible, pero atentamos algo nuevo* (EN_054_12 *If Beto's still ass, then you're like okay, he's still terrible, but we tried something new*)

12. ES_029_12 *Su papá era como un explorador, pero él lo abandono a él y su madre* (EN_029_12 *His dad was also like a, explorer or something like that, but he abandoned him and his mother*)

13. ES_039_7 *Es la etapa donde no tiene que explorar, tiene que conocer a gente* (EN_039_7 *But it's also the time to get to explore and get to meet more people*)

14. ES_023_30 *Y si, no tienes como una manera de pensar de que vas a hacer y cuando lo vas a hacer* (EN_023_30 *And if you don't have like a way of thinking about what you're going to do and when you're going to do it*)

15. ES_029_23 *Te está dando algo, pero, pues, al mismo tiempo te está haciendo, cosas ilegales, muy malas* (EN_029_23 *Give you something, but like, at the same time he may doing illegal, really bad stuff*)

16. ES_052_18 *Sí, y pues los puentes son, ya son puentes, ya lo hicieron con dos carriles, no sé cómo lo van a hacer con tres* (EN_052_18 *Yeah, and so the bridges, well they're bridges, and they made them with two lanes, I don't know how they're going to make them with three*)

## B.12 Spanish Dimension 1: Low

1. ES_005_34 *Tengo diecinueve* (EN_005_34 *I'm ninteen*)

2. ES_019_4 *Ah ok, ok* (EN_019_4 *Okay, okay*)

3. ES_016_28 *Oh wow* (EN_016_28 *Oh wow*)

4. ES_006_5 *Soy de* (EN_006_5 *I'm from*)

5. ES_005_17 *Tiene que ser* (EN_005_17 *It's gotta be*)

6. ES_021_51 *Puedo ir allí* (EN_021_51 *I can go there*)

7. ES_002_37 *Yo no* (EN_002_37 *I don't*)

8. ES_015_1 *¿Juegas videojuegos?* (EN_015_1 *Do you play video games?*)

9. ES_004_15 *Que interesante* (EN_004_15 *That's interesting*)

10. ES_009_48 *¡Qué raro!* (EN_009_48 *That's weird!*)

11. ES_010_37 *¡Ándale!* (EN_010_37 *Oh my gosh*)

12. ES_006_9 *¿Me oyes bien?* (EN_006_9 *Can you hear me fine?*)

13. ES_004_26 *¿Y no te gusto?* (EN_004_26 *And you didn't like it?*)

14. ES_001_34 *Simón bato* (EN_001_34 *Yeah dude*)

15. ES_015_28 *Lab o algo* (EN_015_28 *Lab or something*)

16. ES_002_10 *Te ahorras* (EN_002_10 *You save money*)

## B.13 Spanish Dimension 2: High

1. ES_015_30 *Casi me dijo, "Hazlo otra vez"* (EN_015_30 *She almost told me like, "Do it again"*)

2. ES_011_11 *Um, bueno e* (EN_011_11 *I mean*)

3. ES_022_4 *Pues me gustaba estar en el rollo* (EN_022_4 *Well I like to go on the roll*)

4. ES_001_6 *Primeramente yo estaba todo estresado* (EN_001_6 *First I was stressed out on the first day*)

5. ES_002_18 *No es un deporte* (EN_002_18 *It's not a sport*)

6. ES_012_23 *Pues sí, sabe mejor* (EN_012_23 *But yeah, I like the flavor*)

7. ES_045_11 *En serio sí, es que están chidas* (EN_045_11 *Really, it's because they're cool*)

8. ES_002_2 *No* (EN_002_2 *No*)

9. ES_005_35 *Está más fácil* (EN_005_35 *It's easier*)

10. ES_007_12 *Es la que posee de mi familia* (EN_007_12 *It's the one that my family owns*)

11. ES_004_17 *No, no fue por eso* (EN_004_17 *Mm, mm, no it wasn't because of that*)

12. ES_012_11 *Tienes que cocer el tocinito primero* (EN_012_11 *You have to cook the bacon first*)

13. ES_019_15 *Porque, era la pandemia* (EN_019_15 *Because, we we're in the pandemic*)

14. ES_043_13 *Una pieza es real* (EN_043_13 *One piece is real*) (Note: While *One Piece* is a Japanese manga series, it is not referred to as *Una Pieza* in Spanish, so I think the speaker means it literally.)

15. ES_016_22 *Me terminé uniendo a MAES y SHPE* (EN_016_22 *So I ended up joining MAES and SHPE*) (Note: MAES and SHPE both organizations.)

16. ES_026_24 *No es mi amigo, es amigo de mi amigo* (EN_026_24 *He's not my friend, he's my friend's friend*)

## B.14 Spanish Dimension 2: Low

1. ES_015_37 *Ah* (EN_015_37 *Oh, okay*)
2. ES_010_11 *Ice para allá* (EN_010_11 *Go over there*)
3. ES_015_33 *Y acabe como a las diez* (EN_015_33 *And I ended like at ten*)
4. ES_003_20 *Que, no me la supe* (EN_003_20 *Because I didn't know*)
5. ES_006_11 *Ah, ok* (EN_006_11 *Okay*)
6. ES_006_13 *Tener a sus bebes aquí* (EN_006_13 *Have their babies here*)
7. ES_025_41 *¿Rápido?* (EN_025_41 *Upbeat?*)
8. ES_005_15 *Muy importante* (EN_005_15 *Very important*)
9. ES_036_30 *¿A tu casa?* (EN_036_30 *To your house?*)
10. ES_009_26 *Padre* (EN_009_26 *Nice*)
11. ES_011_20 *Pues, yo diría que Monterrey* (EN_011_20 *Um, I think Monterrey*)
12. ES_011_24 *Mecatrónica* (EN_011_24 *Mechatronics*)
13. ES_003_34 *Porque si no me iban a poner un cero en el examen* (EN_003_34 *Because I would get a zero on the exam*)
14. ES_017_3 *Sí* (EN_017_3 *Yes*)

15. ES_004_22 *Sí, sí, sí* (EN_004_22 *Yeah*)

16. ES_020_32 *Enormes* (EN_020_32 *Massive*)

## B.15 Spanish Dimension 3: High

1. ES_001_15 *Mhmm* (EN_001_15 *Mhmm*)

2. ES_012_19 *Porque, no sé* (EN_012_19 *Because, I don't know*)

3. ES_005_35 *Está más fácil* (EN_005_35 *It's easier*)

4. ES_022_4 *Pues me gustaba estar en el rollo* (EN_022_4 *Well I like to go on the roll*)

5. ES_028_2 *Yo le había dicho que podía hacer más, quería hacer más* (EN_028_2 *I told her that I could do more, I want to do more*)

6. ES_002_2 *No* (EN_002_2 *No*)

7. ES_001_37 *Para ya comprar el boleto* (EN_001_37 *So I can buy the ticket*)

8. ES_002_20 *Es una actividad* (EN_002_20 *It's an activity*)

9. ES_040_11 *Tiene muchas máquinas para cortar madera* (EN_040_11 *He has a lot of machines to cut wood*)

10. ES_023_28 *Entonces el tiempo que estás trabajando* (EN_023_28 *So then the time that you're working*)

11. ES_050_17 *Si no te conviene, pues no* (EN_050_17 *If it doesn't benefit you, then nah*)

12. ES_006_16 *Mi, mi abuelo y abuela* (EN_006_16 *My, my grandpa and my grandma*)

13. ES_006_21 *La licenciatura, sicología* (EN_006_21 *Psychology undergrad and then I'm*)

14. ES_023_42 *Sí, sí me imagino, ay ya me da miedo* (EN_023_42 *Yeah, I can imagine, oh now I'm scared*)

15. ES_008_33 *Sí, sí, o sea* (EN_008_33 *It's just that, I don't know, like*)

16. ES_028_54 *Es ya donde estás ocupado* (EN_028_54 *It's when they're really busy*)

## B.16 Spanish Dimension 3: Low

1. ES_009_48 *¡Qué raro!* (EN_009_48 *That's weird!*)

2. ES_009_13 *¿Oh, en serio?* (EN_009_13 *Oh really?*)

3. ES_024_21 *Era-era un, un barrio sospechoso* (EN_024_21 *So it was like-it was like a sketchy, like, neighborhood*)

4. ES_004_48 *Yo digo como cincuenta y tres horas* (EN_004_48 *I have like fifty three hours*)

5. ES_022_44 *Cuando yo, yo empecé trabajando desde los, desde los dieciséis* (EN_022_44 *When I, when I started working uh, I was sixteen*)

6. ES_010_19 *¡Ándale! Ok* (EN_010_19 *Oh my god, okay*)

7. ES_007_8 *Mario Kart ni existía antes* (EN_007_8 *Mario Kart wasn't even a thing back then*)

8. ES_043_22 *¿Todo de Juárez? Riquísimo* (EN_043_22 *Everything from Juárez? Peak*)

9. ES_013_28 *Ah, okay* (EN_013_28 *Oh, okay*)

10. ES_003_26 *No pasa nada* (EN_003_26 *Nothing happens*)

11. ES_020_3 *Pero regresan* (EN_020_3 *But they come back*)

12. ES_011_41 *Eh la verdad es me agrada Mejía porque siempre me está diciendo “Ay, aplica a esto, aplica a esto”* (EN_011_41 *And I really like Mejia because he is the one that is always like telling me “Hey you should apply to this, you should apply to this” so*)

13. ES_013_22 *Yo digo que las clases de seguridad cibernéticas son muy difíciles* (EN_013_22 *I’ve heard cybersecurity classes are really hard*)

14. ES_016_30 *Motivar a los nuevos estudiantes de CS para que busquen oportunidades y internships* (EN_016_30 *Encourage people who are new to CS to look for new opportunities and internships*)

15. ES_019_14 *No, no te creas, vimos como una cada día* (EN_019_14 *No, just kidding, we watched like one a day*)

16. ES_025_23 *No, está bien, me gusta* (EN_025_23 *Oh that’s good, I like it*)

## B.17 Spanish Dimension 4: High

1. ES_015_28 *Lab o algo* (EN_015_28 *Lab or something*)

2. ES_002_43 *En Ross* (EN_002_43 *In Ross*)

3. ES_019_42 *Ah, sí* (EN_019_42 *Oh, yeah*)

4. ES_004_28 *No era lo que yo esperaba* (EN_004_28 *It wasn’t what I expected*)

5. ES_015_15 *Sí, sí* (EN_015_15 *Yeah, yeah*)

6. ES_025_4 *Para una boda que voy a ir* (EN_025_4 *For a wedding that I’m going to*)

7. ES_009_47 *Wow, qué loco* (EN_009_47 *Wow, that’s crazy*)

8. ES_027_14 *Quedando allí por unos años* (EN_027_14 *Staying there for a couple years*)

9. ES_005_19 *Es más importante saber* (EN_005_19 *It’s more important to know*)

10. ES_005_23 *Ah no te creas, perdón* (EN_005_23 *Oh no, just kidding*)

11. ES_005_36 *No tienes que preocuparte de renta* (EN_005_36 *You don’t have to worry about rent*)

12. ES_006_16 *Mi, mi abuelo y abuela* (EN_006_16 *My, my grandpa and my grandma*)

13. ES_019_14 *No, no te creas, vimos como una cada día* (EN_019_14 *No, just kidding, we watched like one a day*)

14. ES_040_12 *Mi papá le gusta mucho la carpintería* (EN_040_12 *He like a lot, carpentry*)

15. ES_015_13 *Uh, como todo el día* (EN_015_13 *Uh, like all day*)

16. ES_022_55 *Pero lo mandaron adentro a la concina* (EN_022_55 *But they sent him inside to the kitchen*)

## B.18 Spanish Dimension 4: Low

1. ES_002_2 *No* (EN_002_2 *No*)

2. ES_029_9 *Es que, no me acuerdo como después de eso como lo tomo* (EN_029_9 *I really don’t remember after that how he handled it*)

3. ES_006_40 *Sí, sí* (EN_006_40 *Yeah, yeah*)

4. ES_001_5 *Es, entonces es* (EN_001_5 *Is, sorry, is Montse’s boyfriend*)

5. ES_004_29 *Me aburrió, honestamente, como* (EN_004_29 *Bored me, honestly*)

6. ES_015_30 *Casi me dijo, “Hazlo otra vez”* (EN_015_30 *She almost told me like, “Do it again”*)

7. ES_008_33 *Sí, sí, o sea* (EN_008_33 *It's just that, I don't know, like*)

8. ES_050_15 *No más se estaba quejando como "No más me dijiste que el sábado"* (EN_050_15 *She's complaining like "You only told me it was on Saturday"*)

9. ES_012_23 *Pues sí, sabe mejor* (EN_012_23 *But yeah, I like the flavor*)

10. ES_004_39 *¿Aquí a Juárez?* (EN_004_39 *Here in Juárez?*)

11. ES_005_35 *Está más fácil* (EN_005_35 *It's easier*)

12. ES_001_4 *¿Fue? ¿Ya no es?* (EN_001_4 *Was? He's not anymore?*)

13. ES_001_2 *Sí* (EN_001_2 *Yes*)

14. ES_022_46 *Siempre nos decía "Ah, que no están haciendo esto bien" o que nos gritaba* (EN_022_46 *He would always tell us "Ah, you guys aren't doing this right" or he would scream at us*)

15. ES_013_4 *Ah, es- es- está padre* (EN_013_4 *That-that's cool*)

16. ES_009_35 *¿Cómo las casas verdad?* (EN_009_35 *Like the houses right?*)

## B.19 Spanish Dimension 5: High

1. ES_004_27 *Mm, no* (EN_004_27 *Mm, no*)

2. ES_005_4 *Cuando* (EN_005_4 *When*)

3. ES_016_20 *Te uniste a una* (EN_016_20 *Did you join a, um*)

4. ES_025_28 *En ese género* (EN_025_28 *In that genre*)

5. ES_021_51 *Puedo ir allí* (EN_021_51 *I can go there*)

6. ES_027_13 *Mudando a un lugar* (EN_027_13 *Moving one place*)

7. ES_009_9 *¿Ah, rentaste la cabina? ¿O era como* (EN_009_9 *Oh did you rented the cabin? Or like*)

8. ES_025_37 *Entonces eso era la primera* (EN_025_37 *So that was the first one*)

9. ES_053_5 *Los profes no están encima de ti, ni nada de eso, y* (EN_053_5 *Professors are not on top of you and nothing like that, and*)

10. ES_026_16 *Se enojan demasiado* (EN_026_16 *They get way too angry*)

11. ES_003_28 *¿Yo sola? No* (EN_003_28 *Me alone? No*)

12. ES_009_32 *Uh, no* (EN_009_32 *Uh, no*)

13. ES_011_38 *Tengo diecinueve. Voy a cumplir veinte en enero* (EN_011_38 *I'm nineteen, I'm turning twenty in January*)

14. ES_009_10 *Como una cabina de un amigo* (EN_009_10 *It was a friend's, like a friend's house*)

15. ES_029_11 *He visto Bambi antes, pero no me acuerdo de la historia ni nada* (EN_029_11 *I've watched Bambi before, but, like, a long time ago, like, I don't remember the plot or anything*)

16. ES_016_28 *Oh wow* (EN_016_28 *Oh wow*)

## B.20 Spanish Dimension 5: Low

1. ES_013_4 *Ah, es- es- está padre* (EN_013_4 *That-that's cool*)

2. ES_002_17 *Los dos* (EN_002_17 *Both*)

3. ES_002_43 *En Ross* (EN_002_43 *In Ross*)

4. ES_028_54 *Es ya donde estás ocupado* (EN_028_54 *It's when they're really busy*)

5. ES_022_19 *Y me dice de qué "¿Oyes, donde andas?"* (EN_022_19 *Uh, so he told me "Hey where are you?"*)

6. ES_002_18 *No es un deporte* (EN_002_18 *It's not a sport*)

7. ES_054_5 *¿No e-, eres de Horizon? ¿Votaste?* (EN_054_5 *No, you're from Horizon right? Did you vote?*)

8. ES_001_6 *Primeramente yo estaba todo estresado* (EN_001_6 *First I was stressed out on the first day*)

9. ES_005_26 *Lo voy hacer otra vez* (EN_005_26 *Ah so you moved a lot then*)

10. ES_005_23 *Ah no te creas, perdón* (EN_005_23 *Oh no, just kidding*)

11. ES_001_20 *No tocamos, por culpa del otro guitarrista* (EN_001_20 *We didn't play because of the other guitarist's fault*)

12. ES_032_6 *Pero también pienso que, el trabajo que hago* (EN_032_6 *But I also think, the work I do*)

13. ES_028_2 *Yo le había dicho que podía hacer más, quería hacer más* (EN_028_2 *I told her that I could do more, I want to do more*)

14. ES_022_9 *Pero le decimos ticket* (EN_022_9 *Well we call it ticket*)

15. ES_033_18 *Para desahogar, como si tienes un, día malo* (EN_033_18 *To release your feeling, like if you had like a rough day*)

16. ES_043_19 *Muchas personas se lo saltan como "Ay, está muy lento"* (EN_043_19 *A lot of people skip it because they're like "Oh, it's so slow"*)

# Appendix C

# Utterance Prosody Similarity Metric Observation Notes

- EN_016_16
  - **Content:** *I would be kind of scared to ask questions to the professor or...*
  - **Notes:** sharing something personal (they'd be afraid of doing asking a professor a question), extends a word while they try to come up with the right description of how they feel
  - Close utterances
    - EN_034_20
      - **Content:** *It's like, I would do meds, but in a lotion form.*
      - **Notes:** [truly close] sharing something personal (about the form of their medication), takes a pause after mentioning medication and before mentioning its form
    - EN_018_12
      - **Content:** *What have been like, some challenges for you in your career?*
      - **Notes:** [truly close] asking a question that might be personal (about challenges in the other person's career), takes a pause while they think of how to ask
    - EN_025_1
      - **Content:** *So overall, what what music do you prefer to listen to?*

- **Notes:** [truly close] asking a question that might be personal (about the other person's music preference), repeats a word to add a bit of delay before asking

- EN_025_7
  - **Content:** *So I have to pick music that I like, but also that people. . .*
  - **Notes:** [falsely close] (marginal) talking about something that might be personal (about the music they listen to), has a pause but sounds more like they're thinking of what to say next

- Far utterances
  - EN_011_41
    - **Content:** *And I really like Mejia because he's the one always like telling me 'Hey, you should apply to this, you should apply to this'*
    - **Notes:** [truly far] showing appreciation for a person and quoting them, uses *like* but in the different sense (and not to take a pause)
  - EN_024_1
    - **Content:** *So uh yesterday you were telling me about, like, a weird, like, experience you had with the cops in Mexico, right?*
    - **Notes:** [falsely far] (marginal) wants the other person to retell a story they had previously talked about, asking about something that might be personal (the story involves the police), uses *like* twice and might be delaying asking
  - EN_021_13
    - **Content:** *And the beach is really strange because it's like a, you see, like the beach is not like a straight line. It was like a doughnut.*
    - **Notes:** [truly far] incredulous, or trying to get the other person to find it incredulous
  - EN_019_19

- **Content:** *But do you think that someone who hasn't seen a Marvel move can just watch any movie? Or is there any specific movies they have to watch?*
- **Notes:** [truly far] asking for the other's opinion, starting a topic, asking a question, but not something that might be personal

- EN_018_16
  - **Notes:** giving their own opinion, summarizing an idea they were just talking about, closing the topic, flat, not much going on
  - Close utterances
    - EN_026_3
      - **Notes:** [truly close] summarizing an idea they were just talking about
    - EN_029_9
      - **Notes:** [truly close] asking for the other person's opinion (they'd like the other person to answer because the answer might be helpful to them)
    - EN_018_15
      - **Notes:** [truly close] mentions a third party's opinion, probably will continue
    - EN_028_3
      - **Notes:** [truly close] mentions a third party's opinion, puts on a voice to indicate someone else said this
  - Far utterances
    - EN_011_41
      - **Notes:** [truly far] showing appreciation for a person and quoting them, mentioning a third party's suggestion
    - EN_024_1

            - **Notes:** [truly far] wants the other person to retell a story they had previously talked about, starting a topic, is interested in what the other person has to say
        - EN_021_13
            - **Notes:** [truly far] incredulous, or trying to get the other person to find it incredulous, no opinions
        - EN_019_19
            - **Notes:** [falsely far] (marginal) asking for the other's opinion, starting a topic, asking for an opinion

- EN_019_27
    - **Notes:** annoyed about how an outcome was predictable, uses an undulating pace like meaning to say *yada yada*, negative mood towards the whole thing
    - Close utterances
        - EN_026_20
            - **Notes:** [truly close] talking about the past, negative because they are annoyed by other people
        - EN_051_16
            - **Notes:** [truly close] talking about the past, annoyed with them themselves (because they overslept)
        - EN_021_36
            - **Notes:** [truly close] talking about a past unpleasant experience, annoyed about their experience and uses the same undulating pace like meaning to say *and it was annoying, and it was annoying*
        - EN_051_12
            - **Notes:** [truly close] talking about an experience, how they would've had an unpleasant experience (had they made a different choice)

  - Far utterances
    - EN_011_41
      - **Notes:** [truly far] showing appreciation for a person and quoting them, is happy and not dismissive
    - EN_024_1
      - **Notes:** [truly far] wants the other person to retell a story they had previously talked about
    - EN_021_13
      - **Notes:** [truly far] incredulous, or trying to get the other person to find it incredulous, the outcome here (the shape of the beach) was not predictable/expected
    - EN_019_19
      - **Notes:** [truly far] asking for the other's opinion, starting a topic

- EN_033_12
  - **Notes:** leading up to something more interesting, trying to recall the right number (number of months they've been practicing), sounds flat and maybe unenthusiastic
  - Close utterances
    - EN_025_30
      - **Notes:** [truly close] leading up to something more interesting (about to reveal a twist in a movie or book)
    - EN_023_30
      - **Notes:** [truly close] leading up to something more interesting (the consequence if you don't follow these steps)
    - EN_011_14

      - **Notes:** [truly close] leading up to something more interesting (how them *understanding a lot of views* is important), sounds flat and maybe unenthusiastic
    - EN_032_11
      - **Notes:** [falsely close] trying to come up with the number (how much weight they lost), already got to the interesting part
  - Far utterances
    - EN_013_9
      - **Notes:** [truly far] might be answering a simple question
    - EN_021_13
      - **Notes:** [truly far] incredulous, or trying to get the other person to find it incredulous
    - EN_024_1
      - **Notes:** [truly far] wants the other person to retell a story they had previously talked about, could be leading to something more interesting, but this would be coming from the other person
    - EN_019_19
      - **Notes:** [truly far] asking for the other's opinion, starting a topic
- EN_037_1
  - **Notes:** telling an order of events, has a pause before the last word
  - Close utterances
    - EN_036_18
      - **Notes:** [truly close] telling an order of events
    - EN_056_18
      - **Notes:** [truly close] telling an order of events

    - EN_045_15
        - **Notes:** [falsely close] explaining their setup (notebook for writing notes)
    - EN_025_44
        - **Notes:** [truly close] pause before the last word, possibly finishing the telling of events
  - Far utterances
    - EN_013_4
        - **Notes:** [truly far] showing interest in what the other person has to say but doesn't contribute, just shows approval
    - EN_011_24
        - **Notes:** [truly far] completing the other person's sentence OR trying to be funny and making a choice for the other person, a single word
    - EN_013_9
        - **Notes:** [truly far] might be answering a simple question, no order of events but possibly ending one
    - EN_002_16
        - **Notes:** [truly far] showing interest in what the other person has to say but doesn't contribute, just shows approval

- EN_037_20
  - **Notes:** unenthusiastic, downplaying what they're saying to let the other person know it's not very important (what they did: eat pizza), talking about a past event
  - **Content:** Uh, we went to go eat pizzas.
  - Close utterances
    - EN_036_4

            - **Notes:** [truly close] closing the topic, downplaying what they're saying to let the other person know it's not very important (the variety in their exercise regimen)
        - EN_056_1
            - **Notes:** [truly close] downplaying what they're saying to let the other person know it's not very important (what they did: made it to Portugal), unenthusiastic
        - EN_055_10
            - **Notes:** [truly close] unenthusiastic (explicit negative emotion saying *I also didn't like*), might be talking about a past event saying *I also didn't like,* closing the topic
        - EN_044_2
            - **Notes:** [truly close] talking about a past event (but doesn't mention the event yet saying *last year with like-*), possibly disappointed about this event
    - Far utterances
        - EN_011_41
            - **Notes:** [truly far] showing appreciation for a person and quoting them
        - EN_024_1
            - **Notes:** [truly far] wants the other person to retell a story they had previously talked about, asking the other person to talk about a past event
        - EN_019_19
            - **Notes:** [truly far] asking for the other's opinion, starting a topic
        - EN_021_13
            - **Notes:** [truly far] incredulous, or trying to get the other person to find it incredulous, talking about a past event but is excited about this

- EN_050_13
  - **Notes:** sharing their future plans, sharing about themselves but not expecting too much of a reaction, their plans are unknown to the other person
  - Close utterances
    - EN_036_19
      - **Notes:** [truly close] sharing their future plans (coincidentally, about the upcoming Friday) but not expecting too much of a reaction
    - EN_025_42
      - **Notes:** [falsely close] explaining what they were thinking during a decision, why they changed their mind, talking about past and sharing about themselves, like *here's something about me*
    - EN_006_41
      - **Notes:** [truly close] sharing something about themselves the other person doesn't know
    - EN_011_36
      - **Notes:** [truly close] sharing about themselves, what will happen to them in the future (will turn 21 years old)
  - Far utterances
    - EN_011_41
      - **Notes:** [truly far] showing appreciation for a person and quoting them
    - EN_019_19
      - **Notes:** [truly far] asking for the other's opinion, starting a topic
    - EN_021_13
      - **Notes:** [truly far] incredulous, or trying to get the other person to find it incredulous, talking about a past event and expecting a reaction
    - EN_024_1

· **Notes:** [truly far] wants the other person to retell a story they had previously talked about, about a past event

# Curriculum Vitae

Jonathan E. Avila is a Ph.D. candidate in Computer Science at the University of Texas at El Paso, where he has conducted research under the guidance of Professor Nigel G. Ward.

In previous work, Jonathan has worked with models for the continuous estimation of dissatisfaction in dialog, and interaction styles. At his internship at the University of Pennsylvania, he modeled Twitter users' ages from posts related to prescription medicine dosages. He has assisted at the university with Computer Organization and Operating Systems courses.

Jonathan is seeking roles that involve applied research in machine learning with a specific focus on speech system applications. For those interested in knowing more about Jonathan's work or reaching out for professional opportunities, his resume is available on his home page: `www.jonavila.dev`.