HUBERT-BASED MODELS AND EVALUATION STRATEGIES FOR PRAGMATICALLY-FAITHFUL SPEECH TO SPEECH TRANSLATION

by

JAVIER VAZQUEZ

THESIS

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

MASTER OF SCIENCE

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2025

Chapter 1

Introduction

1.1 Speech Translation

Communication is a key component of society, allowing us to build relationships and solve problems. Nevertheless, it can be tricky to do these things when communicating across languages. To facilitate this communication process, translation systems have been developed.

Translation systems have come a long way, from simply replacing each word with their target-language correspondent to understanding the context behind the translation. They have been built to support various formats including text-to-text (T2T), text-to-speech (TTS), speech-to-text (S2T), and speech-to-speech (S2S). These systems have achieved notable results, being able to translate across 100s of languages, and do well in the area of semantics [16]. This is often an adequate solution for many transactional dialogs; however, it is lacking for deeper or more casual dialog where meaning is based not only on what the users say but also how they say it.

1.2 Prosody & Pragmatics

The "how we say things" when people communicate is critical for the recipient to discern the correct interpretation of the message from the range of possible interpretations that words and context allow. The intent of the speaker, or the function of the message, is known as pragmatics. Pragmatics has a vast array of functions, including things such as positive assessment, sarcasm, and turn-yield. Among the main indicators for pragmatics are the acoustic elements known as prosody. Prosody can encompass intonation, rhythm, stress, and much more, quantifying how things were said. These prosodic elements can combine in various ways to form prosodic constructs that each convey a pragmatic function. One could see it as pragmatics being the "destination" to which we arrive using the "road" of prosody.

For a faithful translation, S2S systems are becoming increasingly popular since they now have the potential to leverage the acoustic elements from the source speech to provide a translation with acoustic elements that carry the same intent. However, S2S systems still struggle at leveraging these aspects and problems in the real world may arise due to incorrect translations. These inconsistencies in translations can lead to misunderstandings, potentially escalating a problem. For example, in [6] there were issues in translation where intonations that indicated a question were mistranslated as statements. The errors found in the translations can often be attributed to an inaccuracy in pragmatics. Being able to translate the pragmatics can be a challenging task as prosodic features can participate in many pragmatic functions and languages do not share a direct mapping for their corresponding pragmatic functions. For example, [1] found that follow-up explanations tend to have high near-final intensity in English while in Spanish they would have high cepstral peak prominence smoothed, a proxy for less breathy voice. Thus, being able to learn the different relationships in prosody across languages could enable translation systems to improve by reducing the rate of pragmatically inaccurate translations.

1.3 Pragmatically-Faithful Translation

Given that pragmatics and prosody are closely interconnected, we aim to improve the mapping of prosody to obtain more pragmatically-faithful translations. We define cross-lingual prosody transfer to be the use of a prosody representation from a source language to synthesize speech in a target language that conveys the same pragmatic functions, using whatever target-language prosody is appropriate to do so. This work focuses on advancements towards improved cross-lingual prosody transfer through the use of HuBERT features as a prosodic speech representation, specifically tested between English and Spanish.

Chapter 2

Related Work

Although prosody transfer is starting to receive more attention as a research topic, the focus has mainly been on intra-lingual prosody transfer, the generation of target speech with the same prosodic features as a given source utterance in the same language. Cross-lingual prosody transfer, a more complex problem than intra-lingual transfer, is still largely understudied. Nevertheless, two key advancements have served as the inspiration for this thesis, the Avila model [1], and the Seamless system [16]. This thesis leverages the use of dialog-based parallel speech like the Avila model, and has a foundational BERT-based algorithm for speech encoding similar to the Seamless system.

2.1 Avila's Model

Avila made the initial attempt of mapping prosody between English and Spanish by training models based on features extracted from the then-novel parallel dataset, Dialogs Reenacted Across Languages (DRAL) [23].

The DRAL dataset consists of recordings of conversations between pairs of speakers from the greater El Paso area. To focus on natural speech, the speakers picked most topics, and only some were suggested by the producers. These conversations contain many interesting prosodic patterns that are essential for translating dialog.

With the data at hand, Avila created a prosodic representation based on hand-crafted features. These were a set of 10 features containing: intensity, lengthening, creakiness, speaking rate, pitch highness, pitch lowness, pitch wideness, pitch narrowness, peak disalignment, and cepstral peak prominence method as a proxy for breathiness. These 10 acoustic features were each computed for 10 fixed-percentage non-overlapping regions of different durations tiling any given utterance. With the 100 generated features to represent prosody of a given source utterance, different models could be implemented to predict the corresponding prosody of a target utterance.

The model that saw the most success with these features was a Linear Regression model. The idea was to compute each target feature from a linear combination of the source features. The Linear Regression model was able to perform better than baseline models such as speech from a TTS with no prosodic representation and direct transfer of the source prosody. While this straightforward approach worked to some extent, the remaining large error rates suggest that non-linear relationships exist between English and Spanish [1]. Another weakness was that the evaluation was performed with an unverified metric based on the Euclidean distance of the computed features. The metric was argued to be a decent starting point for objective prosodic evaluation, as there was agreement with overall human judgments. Nevertheless, there were certain aspects where the metric strongly deviated from human judges. These were sometimes related to nasality and energy contours, information that is not represented by the ten features.

2.2 The Seamless System

In contrast to Avila's simple, focused approach with a novel dataset, the Seamless Communication team at Meta produced a massively multilingual foundational model, SeamlessM4T v2, for different applications such as expressiveness and real-time translation [16]. The relevance here of this work is mainly for the SeamlessExpressive model which, as the name suggests, prioritizes translating the expressivity of the source speech, i.e aspects of the prosody of the source.

While the details are not critical for the current thesis, it is still beneficial to understand the changes made to the baseline model that resulted in improved performance with regard to prosody. To begin with, the baseline Seamless model is the finetuned combination of two pretrained models: wav2vec-BERT 2.0 for speech encodings, and the No Language Left Behind (NLLB) encoder for text encodings. Thus, the latent space encodings given by the wav2vec-BERT and NLLB models serve as the acoustic and lexical representation, respectively. Both of these encodings are fed into an NLLB decoder, which has been expanded to accept both, to produce a sequence of translated text. These decoded artifacts are converted into acoustic units through a Non-AutoRegressive Text-to-Unit (NAR T2U) model. The NAR T2U model takes in text, encodes it into a latent space where the model predicts the duration of the acoustic units, and decodes the acoustic units from the latent space. These three modules comprise the UnitY2 architecture which is the core of the baseline Seamless model. The acoustic units that are produced by the UnitY2 architecture are then passed into a conventional HiFi-GAN unit vocoder that converts the acoustic units into an audio waveform. Figure 2.1 below provides a simplified overview of the baseline Seamless model.

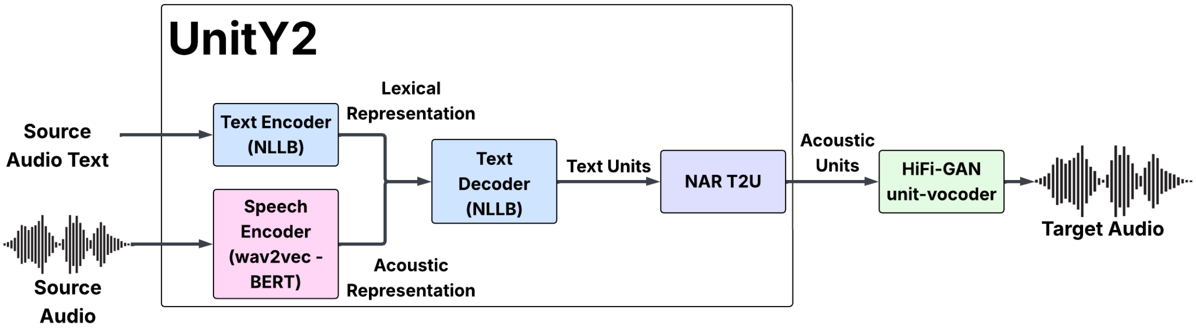

Figure 2.1: A module diagram of the foundational Seamless model.

For the Expressive version, an alteration was made to UnitY2 to account for the lack of prosodic representations, labeled as Prosody UnitY2. Essentially, a new expressivity representation was concatenated to the encoded output of the NAR T2U model before the alignment and decoding process. The expressivity representation was generated through a modified version of the ECAPA-TDNN architecture. With this new representation, Prosody UnitY2 can encode different acoustic aspects to guide unit generation with proper rhythm, speaking rate, and pauses. Additionally, a new module, called PRETTSEL, was incorporated as a step between the Prosody UnitY2 and the HiFi-GAN vocoder. This PRETTSEL module was introduced to ensure that the generated speech retains the encoded expressivity. The PRETTSEL module takes as input the acoustic units produced by Prosody UnitY2 and the source expressivity representation, processes them through a textless acoustic model, and produces Mel-filterbank features for a HiFi-GAN mel-vocoder. The textless acoustic model generates the Mel-filterbank features by predicting high-level context features and local-level prosody features and upsampling them from unit-level to frame-level timescale. Figure 2.2 below provides a simplified overview of the SeamlessExpressive model.

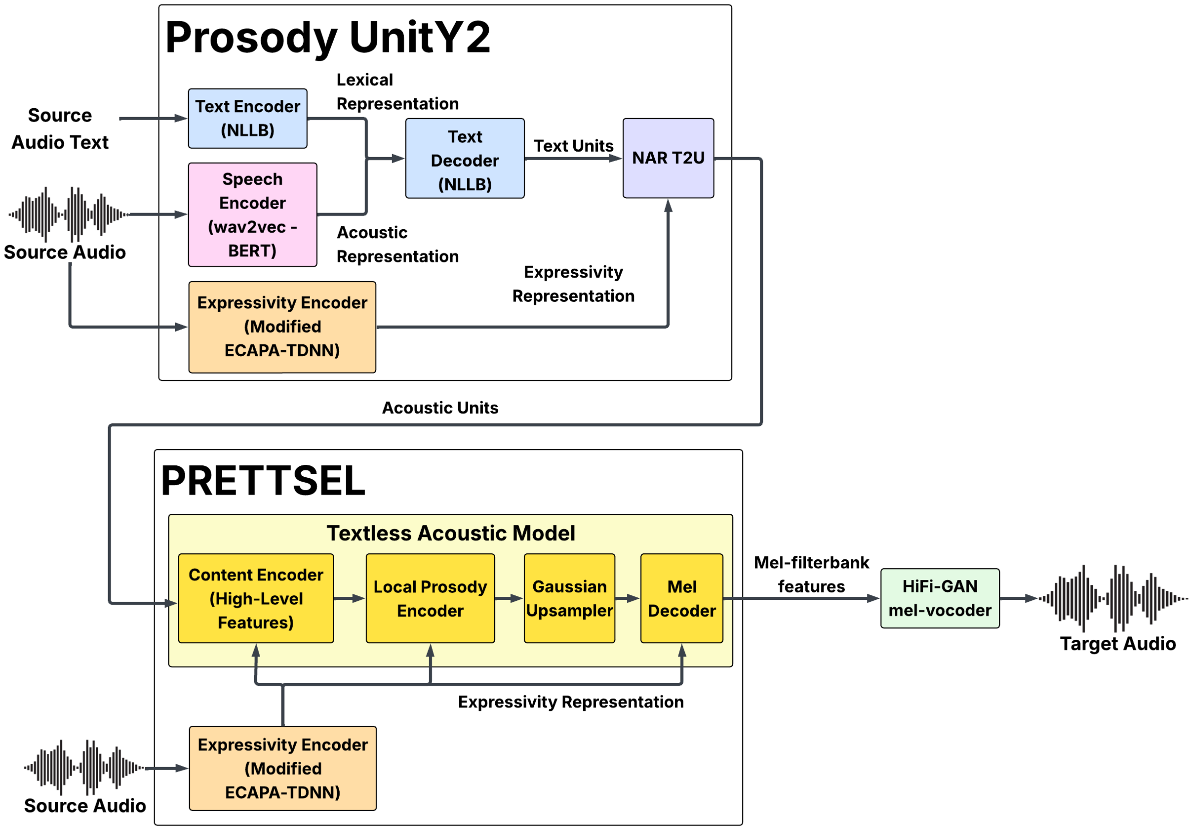

Figure 2.2: A graphical representation of the expressive Seamless model.

These prosody-encoding modules were trained on SeamlessAlignedExpressive, a semantically and prosodically aligned dataset comprised both of manually- and computationally-aligned data. The manually-aligned data are known as mDRAL and mExpresso; the former is an extension of DRAL [1] and the latter is an extension of Expresso [14]. These initial datasets were extended by incorporating other languages such as French and Mandarin. The computationally-aligned portion of the dataset was based on different non-parallel datasets. The semantic alignment was based on the embeddings of a model trained on Automatic Speech Recognition data. These embeddings carry semantic meaning, thus for each instance in the dataset, the closest embeddings would be considered for possible prosodic alignment. The prosodic alignment was performed based on the predicted similarity score from AutoPCP. The AutoPCP module is a connected neural network that predicts the PCP score, a protocol for obtaining human judgments of selected aspects of prosodic similarity. Thus, for the prosodic alignment, the AutoPCP score was predicted by feeding an embedding vector for each of the two considered audios to the neural network.

The SeamlessExpressive model performed well in preserving the intent conveyed by the source prosody, improving upon the foundational Seamless model, according to metrics such as AutoPCP and rate and pause correlations.

2.3 Other Notable Trends

Although Seamless had positive results with their approach, other research teams have tackled prosody transfer from different angles, attempting to solve different problems. The rest of the section overviews this other research in terms of 4 trends: data efficiency in non-parallel datasets, local and global scales for prosodic representations, disentanglement of speaker information, and other generated parallel datasets.

2.3.1 Data Efficiency with Non-Parallel Datasets

To circumvent the need for parallel data, other approaches have looked at prosodic representations generated from non-parallel data. [2] utilizes a novel implementation of the

Source-Filter theory alongside preprocessed acoustic features to represent the prosodic information in speech. It was able to achieve comparable results in Mean Opinion Score (MOS)-based evaluations to other models with only 1/60 of the data. Although this is a great improvement in data efficiency, the architecture would still be limited by the amount of preprocessing needed for the acoustic features. [19] utilizes an unsupervised approach of variational autoencoders (VAEs) for prosodic representation. This creates a need for a large amount of data, which they supplied with noisy multimedia content. To deal with the noise, they developed a noise modeling module that was able to separate the noise and prosodic elements to distinct embeddings after passing the original audio through an external denoiser. Thus, armed with the denoiser to use their in-house multimedia content, they were able to achieve more natural sounding speech compared to their baseline model, although performance was still behind a human-made translation. Another idea that emerged to get around the lack of data was to use transfer learning for modeling speech characteristics. [4] pairs a pretrained voice cloning module with a speaker encoder to preserve the prosody across a Cascade-S2ST system. It was able to do comparatively well to other landmark models such as Translatotron 2 [7].

2.3.2 Prosody Representation at Local and Global Scales

One of the key trends that improved various models was prosodic representation on both a local and a global scale. [19] had an initial model that improved upon their baseline by only considering features that were extracted at a global scale. In [20], they further improved their model by introducing the extraction of local phenomena. [12, 26] used fine and coarse grained emotion representation, mapping to the local and global scales, respectively. The local representations would capture language specific features while the global representation focused on language agnostic features. While emotion does not have a 1:1 relationship to prosody, there is high correlation between certain emotions across languages that can aid our search for cross-lingual prosody transfer. Lastly, [3] uses a similar idea by having f0 representations at different timescales to improve the naturalness.

These features cover 10 timescales that can be led to the generation of local, speaker related attributes and global, language related attributes. Something to note is that for [3], the global scales were language-related while in [12, 26] the global scales were language-agnostic. However, they both coincide in that local representations have speaker-specific features.

2.3.3 Disentanglement of Speaker Information

A key issue in the transfer of prosody is the *foreign accent* problem, where speaker information from a reference audio leaks into the synthesized speech. This makes the disentanglement of prosody and speaker information imperative. Adversarial training to discriminate between speaker and prosody seems to be an effective approach that enables uncorrelated representations. [2] uses Source-Filter theory to disentangle the speech attributes into a source and filter component, the filter having the semantic and speaker specific elements of the audio while the source having the prosodic elements. [11] separated the speech signal into language, speaker, and prosody representations. [12] explicitly focused on separating the speaker information from an emotion embedding. Once again, this was done through adversarial training to ensure that the learned representation of prosody was not directly correlated to the source speaker. Overall, disentanglement is critical for a generalizable model that can sound natural for a variety of languages.

2.3.4 Parallel Datasets

While DRAL is the focus of this work, there are other notable parallel datasets such as SP2 [17] and CVSS [8]. SP2 has speakers reenact chosen sentences, with some reenactments having variation in which word received emphasis. The difference in emphasis can provide different meanings for the same sentence, allowing for the training of more robust systems due to the creation of training data exhibiting more variation in pragmatics. Additionally, the dataset also has a protocol similar to DRAL that can be used to expand the amount of speakers and instances covered. Although this streamlines the collection process, it still remains costly to have a variety of speakers with a wide range of prosody. Another noteworthy point regarding these types of datasets is that they depend on the proficiency of the speakers selected for the prosodic replication. CVSS is a massively multilingual-to-English dataset that has sentences from 21 languages translated into English. While CVSS boasts thousands of hours of translated speech, it is lacking in the natural prosody found in dialog as it is based on read speech. Thus, it can be beneficial for the semantic evaluation of a translation system but it is poor for the evaluation of the varying natural prosody found in dialog.

Chapter 3

Task

3.1 Aim of the Work

The current work aims to improve on the prosodic mapping between English and Spanish, and vice-versa. The hypothesis is that HuBERT features are a better representation for prosodic elements which can thus be utilized to better learn the relationships between English and Spanish. While this may sacrifice the human interpretability of the model compared to hand-crafted features, the improved mapping could generate more pragmatically faithful translations. To evaluate the mapping of the prosody, we use the scores generated by the Segura model [24], which had a better reported correlation specifically tailored for dialog than other metrics.

3.2 Problem Definition

Formally, the general problem to solve is the following:

Given any audio clip, $\boldsymbol{X}$, in a specified source language, predict the prosodic speech representation, $\boldsymbol{y}$, for a corresponding audio translation in a target language with the highest pragmatic similarity.

While a general solution could provide significant contributions to S2ST, in this work we aim to achieve one small step in that direction by addressing the following:

Given a dialog utterance, $\boldsymbol{X}$, in English or Spanish, predict the prosodic speech representation, $\boldsymbol{y}$, in the target language as a set of select HuBERT features that are indicative of pragmatic similarity.

3.3 Hypothesis & Expected Results

The research questions to be explored in this thesis are:

R1: How effective is a HuBERT speech representation for cross-lingual prosody transfer?

R2: How effective is the Segura metric as a way to evaluate translation models in pragmatic similarity?

To answer the first question, we have a main hypothesis, which will be further explored through following sub-hypotheses:

1. HuBERT features enable better predictions, i.e. enable us to better control prosody.

- 1*a*) The new model outperforms previous Avila hand-crafted features in prosody transfer.
- 1*b*) The new model can match or outperform human generated prosody transfer.
- 1*c*) The new model can match or outperform the prosody transfer from SOTA models such as ElevenLabs Dubbing and SeamlessExpressive.

The second question will be investigated through an analysis of the predicted speech representations. We are mainly interested in observing the following aspects of the Segura metric: indication of pragmatic similarity, sensitivity to lexical content, and sensitivity to speaker differences. The Segura metric was already found to be informative of pragmatic similarity in [24], but the question of sensitivity to lexical and speaker differences still remains as does the application to translation evaluation.

Overall, the expected result is that this targeted approach will be able to compute a predicted prosody in the HuBERT embedding space (more details in 4.3.1) that is a closer approximation to the original target utterance compared to other models, corroborated by the performance according to the Segura metric.

Chapter 4

Approach

To determine whether the thesis models can better predict prosodic representations, each produced representation by the thesis models and the baselines will be compared to the ground-truth representation using the Segura model [24], a model that estimates pragmatic similarity. This process will be repeated through every utterance of the test set from the DRAL dataset. According to the average of the Segura scores, we will rank the models. There are two cases: for the baselines that produce audio, the HuBERT representation will be extracted from the translation-output audio, and for the thesis models and baselines that do not generate audio, the direct output will be taken as the predicted prosodic representation. The following diagram depicts the described approach:

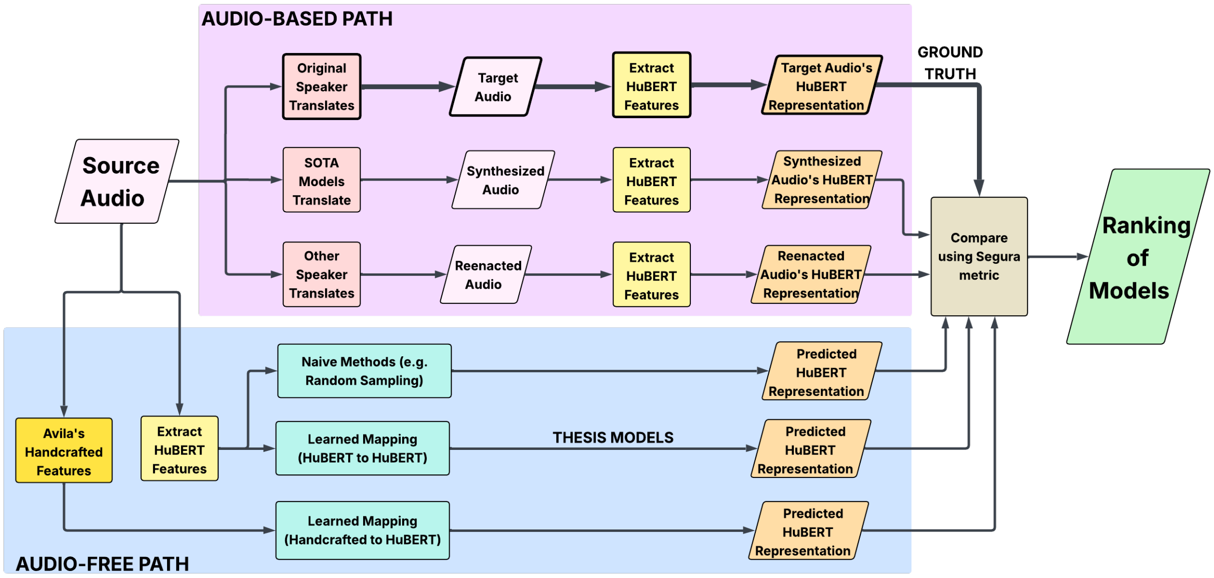
AUDIO-BASED PATH
Original Speaker Translates
Target Audio
Extract HuBERT Features
Target Audio's HuBERT Representation
GROUND TRUTH
Source Audio
SOTA Models Translate
Synthesized Audio
Extract HuBERT Features
Synthesized Audio's HuBERT Representation
Other Speaker Translates
Reenacted Audio
Extract HuBERT Features
Reenacted Audio's HuBERT Representation
Compare using Segura metric
Ranking of Models
Naive Methods (e.g. Random Sampling)
Predicted HuBERT Representation
Avila's Handcrafted Features
Extract HuBERT Features
Learned Mapping (HuBERT to HuBERT)
THESIS MODELS
Predicted HuBERT Representation
Learned Mapping (Handcrafted to HuBERT)
Predicted HuBERT Representation
AUDIO-FREE PATH

Figure 4.1: A graphical representation of the workflow of the thesis.

Corresponding to the two cases, Figure 4.1 has two types of path, Audio-Based and Audio-Free. The Audio-Based path, above in the figure, first translates the audio through different methods, and then computes the HuBERT representation. The Audio-Free path, below in the figure, extracts different feature sets for the source audio and, through a learned mapping or naive methods, predicts the target HuBERT representation, without ever explicitly generating target audio. These representations will all be ranked based on how similar they are to the target audio's representation, giving a sense of how well each method transfers pragmatic meanings. While the ideal comparison would be between synthesized translation audios through human judges evaluating pragmatic faithfulness, there is no current synthesizer that directly synthesizes from HuBERT features. Thus, we compromise and rely on this indirect comparison through objective metrics, leaving the synthesis for future work. The rest of this section goes into more detail about each element of the diagram, including data, baselines, and models.

4.1 Evaluation

The ranking will be done by evaluating the predicted representations for a test set through the Segura model.

4.1.1 The Segura Model

The Segura Model uses the cosine similarity of a language-specific subset of HuBERT features to predict pragmatic similarity [24]. The Segura model had high correlation to human judgments, indicating that it agreed well with the perceived pragmatic similarity. This is beneficial for our end goal of having a higher perceived prosody transfer for applications in dialog systems. Specifically, the language-specific subsets for English and Spanish had correlations with human judgments of 0.74 and 0.72, respectively. Thus, we will take the average Segura score across utterances of the test DRAL dataset to rank each of the models. This *Mean Segura Score* is defined as the average of the cosines, that is, of the dot products of the predicted and actual feature representations divided by the product of their magnitudes. The metric is defined by:

$$\text{Mean Segura Score} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i \cdot \hat{y}_i}{\|y_i\|\|\hat{y}_i\|} \tag{4.1}$$

where n is the number of utterances in the DRAL test set, y_i is the actual language-specific representation, and $\hat{y}_i$ is the predicted representation.

4.1.2 Other Metrics

While the Mean Segura Score is the only metric taken into account when ranking the models, other metrics are included for the evaluation of the thesis models. These metrics mainly serve as a tiebreaker between similarly performing thesis models because it is preferable to have the most accurate representation of the target prosody. Since it is a regression problem, the metrics of choice are the L1 loss and L2 loss. The L1 loss or Mean Absolute Error is defined as the average of the absolute differences between the ground truth and the predicted output. The loss can be represented by:

$$L_1 = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.2}$$

where the terms retain the definitions from the Mean Segura Score.

The L2 loss or Mean Squared Error is defined as the averaged of the squared differences between the ground truth and the predicted output. This can be represented by:

$$L_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.3}$$

where the terms retain the same definition from the L1 loss.

4.1.3 Data

Datasets for the task of cross-lingual prosody transfer are scarce. Thus, for this work, DRAL will be the sole dataset for training and evaluation. The DRAL dataset involves pairs of nonprofessional, bilingual participants from the greater El Paso area. A ten-minute conversation was recorded between each pair, with the dialog being completely natural to include natural prosodic patterns. Subsequently, "interesting" utterances were selected to be re-enacted in their other language, as described in [23]. The official release includes 2893 short Spanish-English utterances pairs constituting more than 2 hours of mono audio per language. These utterances were taken from 104 conversations with 70 unique participants. While the dataset is small in size, especially when compared to those for other speech related tasks, the richness in prosody within the conversations could allow the models to learn language-specific patterns and mappings.

To create the test set, we randomly selected 578 short utterances for each language from the initial 2893 utterances. The test set for the ranking equals 20% of the data, which was fixed to ensure an equal sample size across all models. Importantly, since the selection was utterance-based, not conversation-based, all speakers are seen both in the training and testing data.

4.2 Baselines

To gauge the quality of the predicted output of the models, they will be compared to six baselines. The baselines are:

1. SeamlessExpressive

2. ElevenLabs Dubbing

3. Linear Regression Model with Avila's Hand-crafted Features

4. Human Recreations from DRAL-Multiple References Supplement (DRAL-MRS) Features

5. Naive Implementations: Direct-Copy Transfer and Random Sampling

These baselines were selected to compare the new models to the SOTA, to previous work, and to people's ability to predict prosodic representations. As previously mentioned, the HuBERT representation produced by each baseline and the thesis models will be ranked on how similar they are to the ground-truth audio representation.

4.2.1 SeamlessExpressive & ElevenLabs Dubbing

The first two baselines, ElevenLabs Dubbing and Meta's SeamlessExpressive, are currently accessible SOTA models. These translation systems will be used out-of-the-box, with no adjustments being made to the models. Every short utterance from the DRAL dataset will be fed into these systems so they are translated into the target language. After the translation, the prosodic representation will be extracted through HuBERT by feeding it the synthesized speech.

The SeamlessExpressive model is downloaded through Meta's public repository, and run through an instance of Google Colab. Once again, I would like to extend my appreciation to Vanessa Bolado for her help with the workflow to obtain the translated audios.

The ElevenLabs Dubbing system is accessed through their API service. By specifying the target language, source audios could be dubbed into the target language. One thing to note of this approach is ElevenLabs enforces outputs to have the same duration as the source audio which can affect characteristics such as speaking rate. ElevenLabs has options to modify the timing and transcription but such post-hoc hand-crafting was avoided to retain a fair comparison between systems.

4.2.2 Avila's Features

The next baseline, Avila's Linear Regression, will be utilizing the hand-crafted features developed in [1]. The goal of this baseline is to provide a comparison between hand-crafted features and self-supervised features for the prediction of the prosodic representation. With this in mind, no audio will be produced through this baseline as the aim is only to evaluate the quality of the mapping from one feature space to another.

4.2.3 DRAL-Multiple References Supplement Reenactments

The fourth baseline, human recreations from DRAL-MRS, will provide a comparison to human's capabilities of recreating pragmatic functions in the target language. To go into more detail about DRAL-MRS, the new speakers did not go through the same process of having a conversation and re-enacting selected clips. Instead, the new speakers re-enacted utterances from the original dataset. This supplement covers 2168 of the original utterances. The speakers were from the same greater El Paso region with ages ranging from 19-25. Most speakers were of the same sex as the original speaker of the utterance they were reenacting; however, there were times were the sex would differ.

Each reenacted audio will be passed through HuBERT to obtain the selected set of features. This extracted feature set will be treated as the predicted prosodic representation in the baseline ranking.

4.2.4 Naive Implementations

The last baseline are naive approaches to predicting the target prosodic representation, of which we have included two: Direct-Copy Transfer and Random Sampling. The first is to "directly transfer" features from English to Spanish and vice-versa. This is done by swapping the features that are normally considered for each language. Thus, for the English inputs we compute a HuBERT representation using the Spanish subset and for the Spanish inputs we use the English subset. This switch is possible because the language subsets are retrieved from the last layer of the same model. For a simplified example, imagine that the English subset considers the 1st and 2nd features and the Spanish subset considers the 3rd and 4th features, such that, from an English layer of $[a,b,c,d,e]$, we normally retrieve a and b but with Direct-Copy Transfer we retrieve c and d. Given the same subsets, from a Spanish layer of $[u,v,\ x,\ y,\ z]$, we normally retrieve x and y but with Direct-Copy Transfer we retrieve u and v. The diagram below illustrates the simplified example::

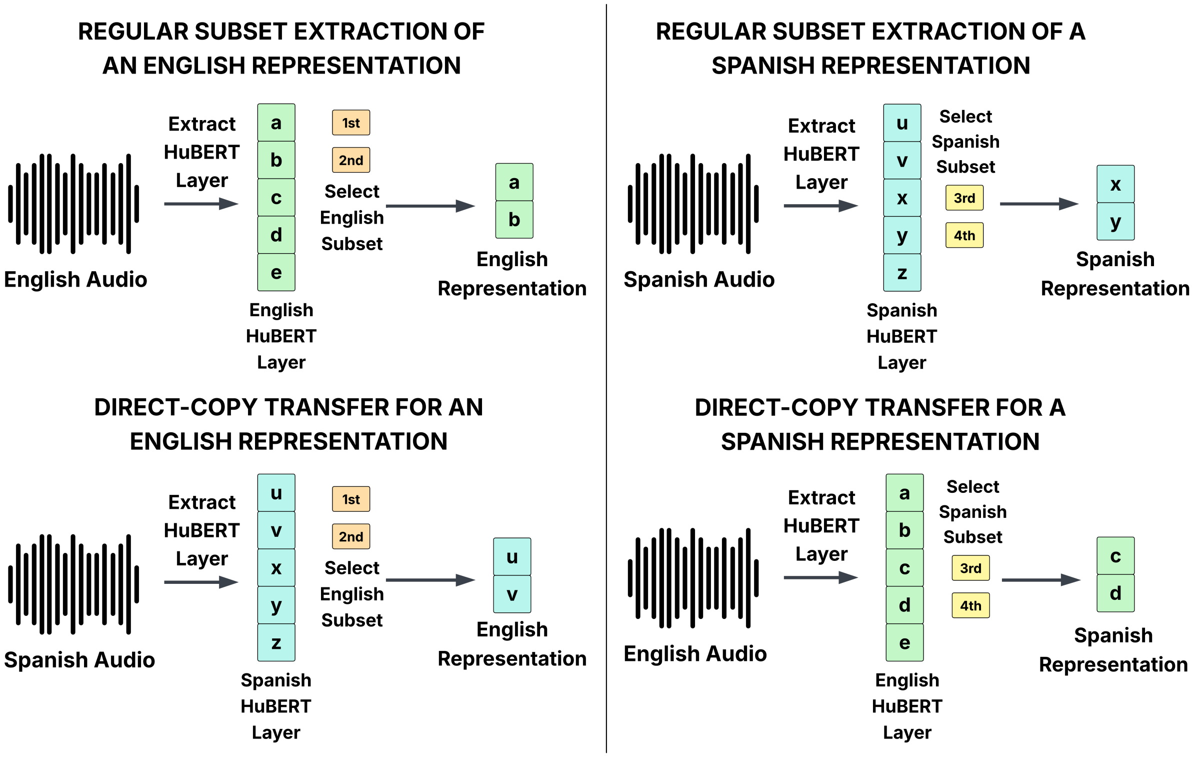

Figure 4.2: A graphical representation of Direct-Copy Transfer operation.

The second is Random Sampling where we take the prosodic representation of a randomly-chosen English utterance B as the prediction for our target English utterance A, ensuring that each utterance is only picked once. These naive implementations serve as the bottom baselines.

4.3 HuBERT-Based Models

To compute the mapping, three models will be evaluated, a Multilinear Regression model, a Locally Weighted Linear Regression model, and a Multilayer Perceptron model. All three models will take as input HuBERT features extracted from utterances of DRAL. The novelty of this approach lies in using HuBERT as a speech representation for transfer of prosodic qualities. As [24, 13, 15] have explored, HuBERT seems to be well-equipped for retaining information relevant to pragmatic similarity, thus there is hope for the transfer of prosodic characteristics.

4.3.1 HuBERT

HuBERT is the application of the BERT model on discrete hidden units derived from small audio segments [5]. The idea is to convert the audio into hidden units to have the speech data be more "language-like" and enable us to apply successful language models like BERT. For the scope of this thesis, the HuBERT model can be used as a black box where features representing an audio can be extracted. There are three sizes available for HuBERT: BASE which produces 768 features, LARGE which produces 1024 features, and X-LARGE which produces 1280 features. For this work, the features will be extracted using the LARGE model on each of the audios of interest. From the 1024 features, a subset of 103 features for English and a subset of 101 features for Spanish are of special interest due to [24] finding these subsets to be among the most pragmatically significant for their respective language.

Thus, each utterance of the DRAL dataset will have a prosodic representation computed by HuBERT LARGE. These will serve to represent both the input for the models and the ground-truth output, namely the representation from the reenacted audio. When computing the features, the version used is provided by the Torch Audio pipelines library which was trained on 60,000 hours of unlabeled audio from Libri-Light dataset [9]. Additionally, 1.5 seconds of silence are added to the beginning and end of audios before passing them into the model to prevent the noise discovered in [25].

4.3.2 HuBERT-Based Input Features

Three sets of inputs, that is, representations of the source-language utterance will be used with the models of this thesis.

1. Full Layer: The 1024 features from the last layer extracted by HuBERT averaged

over the entire utterance.

2. Selected Subset: The 103 or 101 features found by Segura to be the most informative regarding of pragmatic functions in English and Spanish, again time-averaged.

3. Time-Window Features: The selected language subset features averaged over 10 non-overlapping windows as described in [1].

Input 1 was selected to explore the full layer of HuBERT, to include information not captured in the select subset. Input 2 was chosen to see if these subsets, that were found to be pragmatically informative, were adequate. Input 3 was included to increase the granularity for the beginning and end of utterances, which are generally thought to be more prosodically informative. However, Input 3 could not be computed over utterances that were shorter than 400 ms, which excluded 3 instances from the English test set and 9 instances from the Spanish test set. We considered this to be a small enough difference to not require excluding this model from the ranking.

4.3.3 Multilinear Regression Model

The Multilinear Regression model consists of a fixed set of parameters to map the relationship between English and Spanish prosodic features as a linear function. Each feature in the target-language set, i.e. 103 for English and 101 for Spanish, is predicted as a linear combination of the source-language feature set using a set of estimated coefficients. Thus, each feature of the target-language prosody representation is predicted as the best fitting linear function of the n features of the source-language prosody representation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{4.4}$$

Each specific predicted target-language prosodic feature value is represented by y. The source-language prosodic feature values are represented by $x_1, \ . \ . \ . \ , x_n$. The regression coefficients, i.e the estimated contributions of the source-language features, are represented by $\beta_1, \ . \ . \ . \ , \beta_n$, and a constant β_0. Essentially, this model will provide a direct comparison to the Avila model and provide a baseline to determine the added value of HuBERT features as a prosodic representation.

4.3.4 Locally Weighted Multilinear Regression Model

The Locally Weighted Multilinear Regression model is a non-parametric approach that focuses on constructing various linear regression models based on the proximity and distribution of the data points. This allows us to capture more interesting behaviors due to the importance a kernel function gives to closer data points, which decreases non-linearly as the distance increases. The model can be represented by the following equations [10, 22]:

$$w^{(i)} = K(\mathbf{x}^{(i)}, \mathbf{x}_q) \tag{4.5}$$

$$K(\mathbf{x}^{(i)}, \mathbf{x}) = \exp\left(-\frac{(\mathbf{x}^{(i)} - \mathbf{x}_q)^T(\mathbf{x}^{(i)} - \mathbf{x}_q)}{2\tau^2}\right) \tag{4.6}$$

$$J(\beta(\mathbf{x}_q)) = \sum_{i=1}^{m} w^{(i)}(\mathbf{x}_q)(y^{(i)} - (\mathbf{x}^{(i)})^T\beta(\mathbf{x}_q))^2 \tag{4.7}$$

$$\beta(\mathbf{x}_q) = (\mathbf{X}^T\mathbf{W}(\mathbf{x}_q)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{x}_q)\mathbf{y} \tag{4.8}$$

$$\hat{y}(\mathbf{x}_q) = \mathbf{x}_q^T\beta(\mathbf{x}_q) \tag{4.9}$$

The vector of HuBERT features from the i-th training data point in the DRAL dataset is $x^{(i)}$. The vector of HuBERT features that we are taking in as input is x_q, which is used to predict the target features, labeled as $\hat{y}(\mathbf{x}_q)$. The weight assigned to the i-th training data point from DRAL based on a given kernel function is $w^{(i)}$. Our kernel function, K, will be the Gaussian kernel with τ being the bandwidth parameter controlling the size of the local neighborhood considered. The local regression coefficients $\beta(x)$ are then estimated by minimizing our cost, $J(\beta(\mathbf{x}_q))$, the weighted sum of squared errors. Ultimately, the locally weighted aspect allows us to draw focus to more specific behaviors that could be lost in the vast data of general speech.

4.3.5 Multilayer Perceptron Model

The Multilayer Perceptron model is a feedforward artificial neural network that consists of multiple layers of interconnected nodes. These nodes pass data between the layers and transform it using non-linear activation functions such as ReLU. The activation functions allow the model to learn non-linear relationships in the data. The model will take a vector $\mathbf{x} = [x_1, x_2, ..., x_p]^T$ as input, which will be our HuBERT prosodic representation. Each layer will pass on the information as determined by $z = f(W * a + b)$ where f is the activation function, W are the given weights for the node, a is the output of the previous layer activation function, and b is the bias for the given node. These layers would then output predictions for the features of the target language's HuBERT representation, i.e. 103 features for English and 101 for Spanish. Overall, the inclusion of this type of model is in order to be able to map non-linear relationships that are not representable by the basic Linear Regression model.

4.4 Model Implementations

This section covers the implementation of the models detailing things such as libraries and final architecture. All models were coded in Python and run on a laptop with 16GB RAM. All models were implemented using the same 80/20 split, leading to a training set of 2314 utterances and a test set of 578 utterances.

4.4.1 Linear Regression Model

The Linear Regression Model used the implementation found in the sklearn Python library. The fit function of the LinearRegression class was utilized for training and the predict function was used at inference time.

4.4.2 Locally Weighted Linear Regression Model

The Locally Weighted Linear Regression Model used was the implementation provided by [10]. The model was tested with different τ ranging from 0.1 to 100. Conventionally, τ used are around 1, but due to the large number of features, predictions with $\tau < 10$ would lead to an output of only zeros, meaning a vector containing only zeros would be produced. This is due to the size of the considered neighborhood being too small, retrieving no neighbors and thus leaving all vectors to have no weight in the prediction. To counteract this, larger τ were explored, allowing us to reach non-zero results. Additionally, the model was also tested with data z-normalized across the dataset in hopes of decreasing the value of τ needed for non-zero results.

4.4.3 Multilayer Perceptron

The Multilayer Perceptron was implemented in Python using the Pytorch library. Different architectures and parameters were tried during an initial pilot study, which led to the configuration described below:

- Hidden Layers:
    - Full Layer: 500 neurons, 250 neurons, 125 neurons (669,649 parameters for both English and Spanish).
    - Selected Subset: 100 neurons, 50 neurons (15,450 parameters for English and 15,250 parameters for Spanish).

  - Time Window Features: 500 neurons, 250 neurons, 125 neurons (672,125 parameters for English and 662,125 parameters for Spanish).

- Adam Optimizer
- Initial Learning Rate: 0.001
- Epochs: 40
- Batch Size: 500
- ReLU Activation Function
- L2 Regularization: 0.0001

While many different combinations are possible, we only used 2 or 3 layers to reduce the risk of overfitting due to the small size of the data. Nevertheless, other modifications were made to the initial architecture to see how the prediction were affected. The modifications attempted were:

- Leaky ReLU Activation Function: Avoids the “Dying ReLU” problem by having negative inputs be multiplied by a small constant instead of becoming 0.
- Stronger L2 regularization: Gives a bigger penalty for larger weights to avoid overfitting. A regularization of 0.05 was utilized.
- Dropout Layer: Randomly assigns the values of its input to be 0 based on a given percentage to avoid overfitting. The percentage used was 25%.
- Standardized Input Data Across Training Data: Standardized the data by z-normalization with respect to the entire training data to avoid a small number of features being the sole contributors to the prediction.

- Standardized Input Data by Speaker: Standardized the data with respect to each speaker by applying z-normalization to all HuBERT frames generated per speaker for each language. The reasoning behind applying this standardization is that many speech related tasks can be negatively impacted by speaker differences.

- Cosine Dissimilarity Based Loss Function: Set the loss function to be:

$$Loss = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{y_i \cdot \hat{y}_i}{\|y_i\| \|\hat{y}_i\|} \tag{4.10}$$

These modifications were selected in order to avoid overfitting, e.g. Dropout and Regularization, and to tailor the model to Segura-based ranking, e.g. Cosine Dissimilarity Based Loss Function .

Chapter 5

Results & Discussion

5.1 Evaluation Results

This section covers the results obtained from the ranking of the different systems with respect to Segura model on the ground truth utterances, and also the results of the hyper-parameter tuning for the MLP.

5.1.1 Ranking of Different Models

	Mean Segura Score		
Model	**English**	**Spanish**	**Notes**
Multilayer Perceptron	0.82	0.88	Thesis Model, Audio-Free
Locally Weighted Linear Regression	0.82	0.87	Thesis Model, Audio-Free
Linear Regression	0.81	0.87	Thesis Model, Audio-Free
Avila-Based Linear Regression	0.80	0.86	Previous work, Audio-Free
DRAL-MRS	0.75	0.78	Human-produced audio
ElevenLabs Dubbing	0.71	0.78	Full SOTA system, Audio-Based
SeamlessExpressive	0.71	0.81	Full SOTA system, Audio-Based
Direct-Copy Transfer	0.77	0.73	Naive Method, Audio-Free
Random Sampling	0.66	0.72	Naive Method, Audio-Free

Table 5.1: The Mean Segura Score achieved by each system for English and Spanish.

Each of the models of this thesis outperformed the baselines, with the best thesis model having a 0.02 improvement over the Avila hand-crafted features in both English and Spanish, and a larger 0.07 improvement over the best audio-output systems in both English and Spanish. While these are positive results, there is a confound, which was not foreseen before doing the experiment and seeing the results: the audio-free models are performing an easier task than the other systems. Baselines such as the human recreations in DRAL-MRS have to create the translated audio, while the audio-free systems do not. This complicates the task because those systems have to accomplish other goals such as a correct lexical translation. The additional goals can constrain the possible prosodic mappings. Thus, the audio-free systems might be able to create a beautiful prosodic specification that is not lexically feasible.

One observation from the results is that the naive method of Direct-Copy Transfer in English was able to outperform the SOTA models and even DRAL-MRS. This supports that there are some shared similarities between English and Spanish, although of course it is not a direct mapping, as discussed in [1]. Another observation is that SOTA models were able to match or outperform human recreations in Spanish, in contrast to English.

Understanding the Relationship between English and Spanish Scores

In Table 5.1, Spanish predictions performed better than English predictions for every model except Direct-Copy Transfer. Below we explore three possible explanations.

1. **Complexity of Pragmatic Functions in English vs Spanish:** The first explanation is that Spanish has simpler prosodic constructs for its pragmatic functions compared to English. This implies an easier prediction of these constructs for Spanish. Additionally, the higher performance of the Random Sampling naive implementation, by 0.07 relative to English, suggests that Spanish patterns are less diverse. [21] introduced the notion of Spanish having less complex prosodic constructs, such as a more narrow pitch contour. They proposed this was possible because Spanish relies more on lexical content for conveying pragmatics.

2. **Focus of HuBERT Training Data:** The second explanation is that HuBERT isn't able to capture the nuances of the Spanish language because of its English-only training data. Without being able to capture Spanish nuances, the model might be less informative in its extracted features. This makes the task easier by having to predict only a general construct instead of the specific nuances that would need to be predicted for English. However, the explanation is weakened by the fact that in [24], Spanish's correlation with human judges was almost the same as English's correlation with human judges, at 0.72 and 0.74, respectively. This indicates that the features captured by the HuBERT model are pragmatically informative for Spanish despite it never being trained on anything but English.

3. **Length of Feature Subsets:** The third explanation could be attributed to the length of the feature subsets per language. Prediction tasks become more difficult as the number of features increases, thus Spanish has an immediate advantage at its 101 features compared to the 103 from English. However, the increase in the number of features is minimal, adding only 2 more features when the number is already over 100. More importantly, the baselines that were trying to predict a translation instead of these language feature sets also followed the same trend.

The cause for the higher performance in Spanish could be a combination of these three reasons. Nevertheless, more investigation is needed on the relationship of pragmatic function complexity between English and Spanish.

5.1.2 Thesis Models Exploration Results

The following subsections cover the results obtained by each of the thesis models with respect to each set of input features, as discussed in Section 4.3.2. Additionally, the table results also cover the effects from changes to τ for Locally Weighted Regression, and all the modifications mentioned in Section 4.4.3 to the Multilayer Perceptron.

Multilinear Regression Results

	Feature Set								
	Full Layer			**Selected Subset**			**Time-Window Features**		
Metrics	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**
English Features	9.65	165.37	0.721	7.45	96.83	0.812	9.85	167.74	0.702
Spanish Features	8.50	125.67	0.801	6.71	82.36	0.867	8.89	134.37	0.776

Table 5.2: The scores achieved by variants of the Multilinear Regression model for predicting English and Spanish.

As seen in Table 5.2, the best performing input for Multilinear Regression was the Selected Subset. A strong performance by the Selected Subset input was expected since the subsets were selected to be pragmatically indicative, allowing the model to learn patterns in translation related to pragmatics. The stark difference in performance from the Selected Subset to the other inputs could also be explained by the nature of the mapping model. Simple linear regression models tend to struggle with larger number of inputs. Thus, the smaller size of the subsets gave it an advantage over the other two inputs.

Locally Weighted Multilinear Regression Results

	Feature Set								
Predictions for English Features	Full Layer			Selected Subset			Time-Window Features		
Tau:	MAE	MSE	Cosine	MAE	MSE	Cosine	MAE	MSE	Cosine
$\tau = 0.1$	13.08	275.39	0.0	13.08	275.39	0.0	13.02	271.76	0.0
$\tau = 1$	13.08	275.39	0.0	13.08	275.39	0.0	13.02	271.76	0.0
$\tau = 10$	NaN	NaN	NaN	15.50	439.85	0.521	13.02	271.76	0.0
$\tau = 50$	21.06	777.10	0.424	7.45	99.38	0.813	8.93	137.89	0.751
$\tau = 100$	13.15	309.75	0.601	7.36	95.22	0.816	23.49	952.18	0.374

Table 5.3: The scores achieved by adjusting τ for predicting English features.

	Feature Set								
Predictions for Spanish Features	Full Layer			Selected Subset			Time-Window Features		
Tau:	MAE	MSE	Cosine	MAE	MSE	Cosine	MAE	MSE	Cosine
$\tau = 0.1$	13.21	291.46	0.0	13.21	291.46	0.0	13.04	280.99	0.0
$\tau = 1$	13.21	291.46	0.0	13.21	291.46	0.0	13.04	280.99	0.0
$\tau = 10$	NaN	NaN	NaN	15.41	411.47	0.591	13.04	280.99	0.0
$\tau = 50$	18.00	539.85	0.514	6.84	88.17	0.865	7.67	103.26	0.834
$\tau = 100$	11.07	209.94	0.707	6.64	80.90	0.870	22.95	879.84	0.421

Table 5.4: The scores achieved by adjusting τ for predicting Spanish features.

From Table 5.3 and Table 5.4, we can see that Selected Subset was once again the highest performing input. The similar nature of Locally Weighted Multilinear Regression to Multilinear Regression could explain the lower performance of the other two inputs. A noticeable difference to the results of Multilinear Regression is that Time-Window features performed closer to Selected Subset features at $\tau = 50$. This could indicate that the considered neighborhood for the Time-Window features at $\tau = 50$ allowed it to focus on local, meaningful patterns; even so, it did not outperform the Selected Subset.

Another observation made was that a $\tau > 10$ was needed to achieve any non-zero vectors. This is due to the large number of features increasing the minimum distance needed for instances to be considered in the prediction. A larger τ considers a larger neighborhood which allows instances to have non-zero weights, meaning the instances contribute to the prediction. As a minor observation, the Time-Window results for zero-vectors are slightly different than the other two inputs due to the missing 3 instances for English and 9 instances for Spanish. Additionally, there are two Not a Number (NaN) results at $\tau = 10$, these are due to an overflow in division error caused by the weights calculated by the kernel function with the given τ for the large number of features. While there are different techniques to counteract this result, they were not explored here as no additional insight seems likely to be obtained.

	Feature Set								
Predictions for English Features	**Full Layer**			**Selected Subset**			**Time-Window Features**		
Tau:	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**
$\tau = 0.1$	13.08	275.39	0.0	13.08	275.39	0.0	13.02	271.76	0.0
$\tau = 1$	NaN	NaN	NaN	16.16	489.64	0.506	NaN	NaN	NaN
$\tau = 10$	11.96	254.68	0.638	7.36	95.19	0.816	11.48	227.86	0.637
$\tau = 50$	9.57	162.75	0.725	7.44	96.67	0.813	9.77	165.25	0.706
$\tau = 100$	9.62	164.49	0.722	7.44	96.79	0.812	9.83	167.05	0.703

Table 5.5: The scores achieved by adjusting τ for predicting English features with z-normalized data.

	Feature Set								
Predictions for Spanish Features	**Full Layer**			**Selected Subset**			**Time-Window Features**		
Tau:	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**
τ = 0.1	13.21	291.46	0.0	13.21	291.46	0.0	13.04	280.99	0.0
τ = 1	NaN	NaN	NaN	14.76	374.70	0.614	NaN	NaN	NaN
τ = 10	10.27	181.65	0.736	6.64	80.91	0.870	10.33	179.23	0.727
τ = 50	8.39	122.75	0.805	6.71	82.22	0.867	8.81	132.29	0.779
τ = 100	8.47	124.76	0.802	6.71	82.32	0.867	8.87	133.79	0.777

Table 5.6: The scores achieved by adjusting τ for predicting Spanish features with z-normalized data.

Further, from Table 5.5 and Table 5.6, we can see that z-normalizing the data did not much affect the outcome as Selected Subset outperforms the other two inputs by a similar margin as regular data. While we were able to lower the τ needed for a non-zero vector, there was no effect on the best performing mean cosine score, which is the metric considered for the ranking. Additionally, NaN results were present, now with a τ of 1, but again we did not investigate further, as they occurred in the lower performing models.

MLP Results

	Feature Set								
Predictions for English Features	Full Layer			Selected Subset			Time-Window Features		
Metrics	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**
Base MLP	7.34	95.35	0.816	7.48	97.31	0.809	7.81	108.15	0.791
Leaky ReLU Activation	7.45	98.20	0.811	7.45	96.56	0.811	7.82	108.22	0.792
L2 Regularization ($\alpha = 0.05$)	7.29	93.32	0.818	7.42	95.86	0.811	7.77	107.02	0.795
Dropout ($p = 0.25$)	7.39	95.90	0.815	7.50	98.41	0.808	7.60	102.28	0.805
Standardized Dataset	7.45	97.24	0.813	7.66	103.88	0.804	7.61	101.78	0.802
Standardized per Speaker	0.19	0.06	0.253	0.18	0.06	0.253	0.20	0.07	0.190
Cosine Dissimilarity Loss	8.21	121.15	0.814	8.40	123.79	0.810	8.97	141.29	0.792

Table 5.7: The scores achieved by applying various modification to the best performing MLP model for English features.

	Feature Set								
Predictions for Spanish Features	Full Layer			Selected Subset			Time-Window Features		
Metrics	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**	**MAE**	**MSE**	**Cosine**
Base MLP	6.56	77.86	0.875	6.78	83.54	0.866	7.05	88.55	0.852
Leaky ReLU Activation	6.65	79.34	0.874	6.78	84.11	0.865	7.11	91.49	0.848
L2 Regularization ($\alpha = 0.05$)	6.65	79.35	0.872	6.76	83.68	0.866	6.98	87.93	0.854
Dropout ($p = 0.25$)	6.65	80.60	0.873	6.83	85.95	0.864	6.82	85.13	0.865
Standardized Dataset	6.65	80.35	0.872	7.04	91.29	0.861	6.85	84.49	0.862
Standardized per Speaker	0.17	0.05	0.282	0.17	0.05	0.264	0.18	0.06	0.189
Cosine Dissimilarity Loss	7.74	111.13	0.870	7.59	105.25	0.861	8.23	121.48	0.857

Table 5.8: The scores achieved by applying various modification to the best performing MLP model for Spanish features.

As seen in Table 5.8 and Table 5.7, none of the modifications created any meaningful change for the best model. This suggests that the model is not overfitting, since techniques that reduce overfitting such as Dropout and Standardization had no benefit, with the only exception being a stronger L2 Regularization yielding a small improvement of 0.002 in English.

One interesting thing to note is that for the Time-Window features (Input 3), Dropout yielded an improvement of 0.014 for English and 0.013 for Spanish, and Standardization yielded an improvement of 0.011 for English and 0.010 for Spanish. While these might not be massive increases, these were the largest improvements observed from any of the explored modifications; however, it still did not surpass the Full Layer (Input 1) MLP. While Dropout and Standardization are techniques used to reduce overfitting, the Time-Window features did not benefit as much from a stronger regularization. Future work could possibly explore the cause of this increase, potentially relating it to the shorter duration of prosodic constructs observed at the start and end of utterances as opposed to the middle [1].

Lastly, the speaker-normalized data performed surprisingly poor, although speaker normalized data tends to improve performance in speech related tasks. We did not further investigate this change as standardization across the entire dataset had minimal impact on the performance of the base MLP. However, one speculation is that the HuBERT features are entangled. If HuBERT features are entangled, meaning some features represent both speaker identity and prosodic qualities, the changes related to speaker identity could affect the prosodic information. While the subsets were selected to be speaker-agnostic [24], meaning the Selected Subset input should not be affected by standardization, other research teams have suggested that current embeddings truly aren't able to separate speaker identity and prosodic information [18].

5.2 Qualitative Analysis of Successes & Failures

Identifying pragmatic functions or prosodic features that are commonly present in successes and failures can guide future investigations. It can help research teams focus on improving where the current models lack. To this end, a qualitative analysis was conducted on the best and worst translated audios for human recreations and for SeamlessExpressive, representing the SOTA models. For human recreations, we aim to see how much variety there is in possible translations and how well does the Segura metric correspond to human perceptions. For SeamlessExpressive, we want to see the general quality of the translations, how consistent they are with retaining pragmatic intent, and potential improvements for the model such as inappropriately generated prosodic qualities. The process for the qualitative analysis was to listen to 5-10 instances in the test set, which includes the source audio, the ground-truth target audio, and the translated audio. Additionally, we also listened to the ground-truth utterance pairs of the best and worst predicted instances of the MLP model. While it would be ideal to have translated audio for the analysis, these best and worst predictions can illustrate the types of pragmatic functions the MLP model seem to have learned from the data. We acknowledge that the analysis on the MLP model is highly speculative, relying on the performance of the Segura metric to suggest what pragmatic functions were or were not learned by the model.

5.2.1 Human Recreations (DRAL-MRS)

Once again, the goal of the human recreations analysis is to find how well humans agree in their percept of pragmatics. We want to observe how much variation in prosody exists in translations that kept the pragmatic intent, and if the Segura metric was able identify these similar pragmatic intents with differing prosodic forms. Additionally, we want to see what type of utterances are translated most and least consistently, as this could indicate the areas in which machine learning mappings should focus.

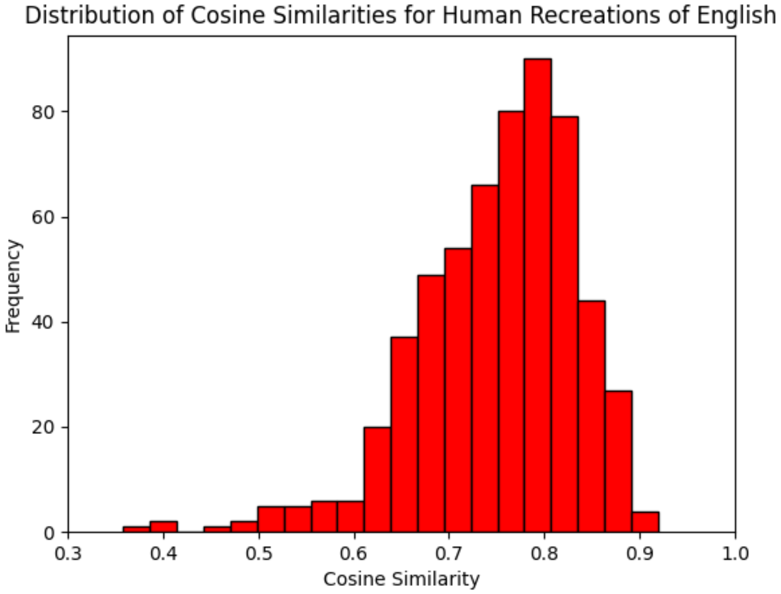

Figure 5.1: A histogram of the English human recreations.

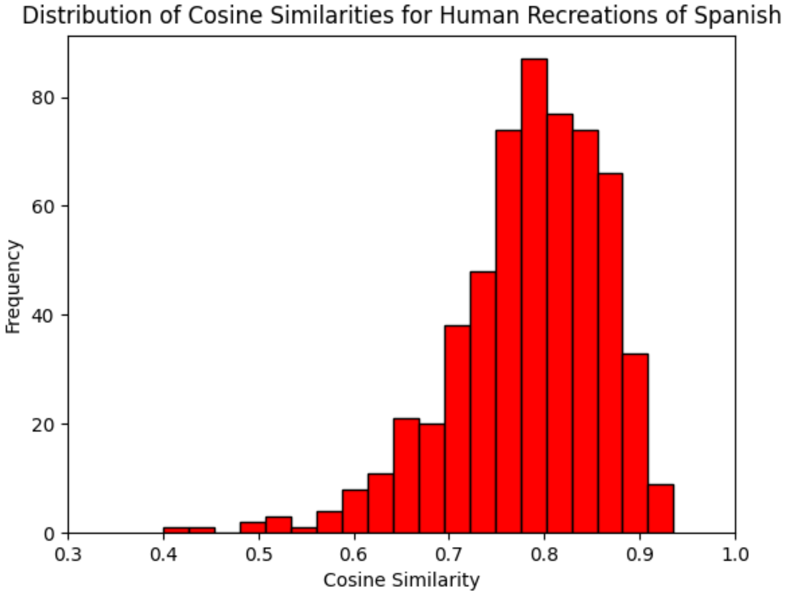

Figure 5.2: A histogram of the Spanish human recreations.

Best Performing for Predicting English

For the top 5 English utterances, similarity scores ranged from 0.89 to 0.92. These are high performances but it's also noteworthy that there were no perfectly recreated utterances (Cosine Score of 1.00).

Of these five audios, three were short, ranging from less than a second to 1 second. These were short utterances with a rising pitch saying small comments such as *"Oh"* or *"What's that?"*. The pragmatic functions for these were simple and straightforward.

The other two were longer audios of 3 and 7 seconds. The longer audios were non-commital, avoiding strong opinions, and had some degree of lengthening. For instance, the shorter utterance 012_12 said *"you have to cut it"* in both the ground-truth and the reenactment, with *"it"* being extremely lengthened in both. For a longer utterance, 019_19 had a ground-truth of *"But do you think that someone who hasn't seen a Marvel movie, can just watch any movie or is there any specific movies they have to watch?"* with a lengthened *"any movie"*. The presence of lengthening is maintained in the reenacted *"Do you think that someone who has never watched any Marvel movie, can watch any of them or like they have to watch a specific movie?"* with a lengthened *"that"* and *"or"*. Additionally, both utterances had three pauses present, albeit with the reenactment pauses being more salient. The ground truth had pauses after *"watch"*, *"there"*, and *"to"*. The reenactment had stronger pauses after *"movie"*, *"like"*, and *"watch"*. The degree to which lengthening is applied and the number of pauses present in both the ground-truth and the reenactments assist in conveying the original pragmatic intent, as corroborated by the Segura metric.

Best Performing for Predicting Spanish

For the top 5 Spanish utterances, scores ranged from 0.92 to 0.94. This is expected as the mean score for Spanish was higher than the mean English score, and it can be seen with the higher frequency towards the right edge in Figure 5.2. These utterances were on average longer compared to the best English utterances, with all being from 2 to 5 seconds. A prosodic construct observed on all utterances was an emphasized word through lengthening, rising pitch, or both. While the emphasized word would remain the same in the utterance pairs, the pragmatic function each pair conveyed could vary. Some words were emphasized to ask a question such as *"más"* in the ground-truth *"Ella fue la que te rompió el corazón, así más?"* and the reenacted *"No fue una que te rompió el corazón más?"*. The source audio for this utterance pair had a similar prosodic structure with an emphasized *"most"* in *"She was the one that broke your heart like the most?"*. Other emphasized words were used to convey a reaction such as *"Ahh"* in the ground-truth *"Ahh, se los estan copiando"* and *"Ay"* in the reenacted *"Ay, copiones"*. This carries the same intent from the source audio saying *"Ohh, you're just a copycat"*. An additional common pattern found in 3 of the utterances was a similar pause structure, which was relatively simple, featuring one or two pauses in the utterance. Having pauses and lengthening be present in both the English and Spanish best performers suggests that these are characteristics valued by both HuBERT subsets.

Worst Performing for Predicting English

The least-well predicted English utterances that were analyzed had similarity scores ranging from 0.36 to 0.51. They can be seen at the edge of the left tail on Figure 5.1. The worst performing utterance was 065_6 where the audio had a *"Yes"* followed by a "thump" noise which wasn't in the reenacted audio. These instances with data quality issues were not considered further since the purpose of this analysis is focused on prosody and pragmatics.

In the other utterance pairs, most were around 2 seconds long and had deviations in the conveyed pragmatic intent. For instance, utterance 010_17 had a different pragmatic function where the ground-truth stated *"...didn't have a cover"* with emphasis on the *"didn't"* while the reenacted audio had *"it didn't have a cover"* with no emphasis and a rising pitch on *"a cover"*. Thus, the pragmatic function of the ground-truth seemed to be an exclamation while the reenacted seemed to be a question. However, the most common differing pragmatic function was one in which the original speaker seemed thoughtful, while the re-enactor seemed more direct in their statement. These changes in pragmatics were seen in differences in pauses and lengthening. Either the ground-truth or the reenacted audio would contain pauses or lengthening that the other was lacking in the same degree. For example, utterance 052_14 had a ground-truth of *"How much time is it from here to your house?"* with lengthening and pauses at *"it"*,*"from"*, and *"here"*, while the reenactment said *"How much time does it take you from here to your house?"* without any pauses or lengthens. This was in contrast to many utterance pairs with higher similarity scores, where pauses and lengthening would appear more in both. Utterance 038_2, the instance with a score of 0.51, had a ground-truth of *"So, then, there's two beds?"* and a reenactment of *"So, there's two beds?"*. The ground-truth had strong lengthening in *"So"* and *"then"*, while the reenactment had lengthening only in *"So"*. That extra pause and lengthening in the original made the speaker seem more timid and thoughtful, while the singular lengthened word in the reenactment made the speaker seem more direct. These changes are subtle cues that can be missed easily by people such as apparently the re-enactors, but they appear to be valued highly by the Segura metric.

Worst Performing for Predicting Spanish

The scores of the worst Spanish utterances analyzed ranged from 0.40 to 0.57 while the length ranged from less than a second to 5 seconds. Once again, there were instances of data quality issues that, we didn't examine further, such as reenacted audio with echo or labeling errors, and they mainly encompassed the edge of the left tail in Figure 5.2. From the remaining utterances, the prominent pattern in the utterance pairs was hesitant versus confident statements. Either the ground-truth would show hesitance and the reenacted audio would show confidence or vice-versa.

Incidentally, to digress, utterances could show hesitance through three distinct manners. The first was through false starts such as in 055_4 with a ground-truth of *“El-el gordito”* and a reenactment of *“El gordo”*. In this utterance, the ground-truth conserved the pragmatic intent of the source audio having hesitance in saying *“The-the fat guy”*. The second way was through the use of filler words such as *“umm”* in the reenactment *”Si, umm si”* which wasn’t present in the ground-truth *“Si, claro”* or the source audio *“Yeah, definitely”*. The last method was differing placement of emphasis such as in utterance 006_5 where the ground-truth said *“Soy de”* with emphasis on the *“de”* while the reenactment said *“Yo soy”* with emphasis on *“Yo”*. The emphasis from the ground-truth kept the hesitance of the source audio *“I’m from”*, whereas the reenactment sounded more sure of themselves. The variety in prosody to convey hesitance showcases that there is no simple one-to-one relationship between prosodic constructs and pragmatic intent. There are many prosodic constructs that can map to the same function so people might find difficult to choose an adequate corresponding pattern as they translate from one language to the other.

Through the qualitative analysis, we were able to see that humans best transferred pragmatic intents with simple prosodic constructs. These constructs were mostly related to pauses and lengthening, perhaps being the features most salient to untrained listeners such as the re-enactors for DRAL-MRS. Excluding instances of data quality issues in DRAL, we saw that the majority of the worst performers had hesitance lost in translation. This could be through a lack of pauses, lengthening, or other prosodic features. Perhaps this could be attributed to some re-enactors being influenced by first hearing what they are going to say. It could be difficult for some people to generate hesitance when they know the words they will say. Overall, the Segura metric was able to correctly identify utterance pairs with similar intent, and penalize the pairs where the variation in prosody damaged the pragmatics.

5.2.2 SeamlessExpressive

For SeamlessExpressive, we are trying to identify the factors affecting the quality of the translations generated. The main focus of the analysis is on pragmatic intent, but other aspects such as quality of audio or changes to speaker identity will also be considered. Ideally, the system is able to capture the lexical content and pragmatic intent of any source audio and generate the corresponding target audio with a corresponding prosodic construct to convey the intent. If there are any weaknesses identified in this process, highlighting them would allow us to improve future translation tools. While the focus of this study is prosody and pragmatics, here we see that lexical content is also relevant for the analysis due to these three factors not being completely isolated. These elements of speech are not in competition, rather they are complements of one another, allowing for better communication. Thus, a translation system should be capable across all these factors to properly serve the people.

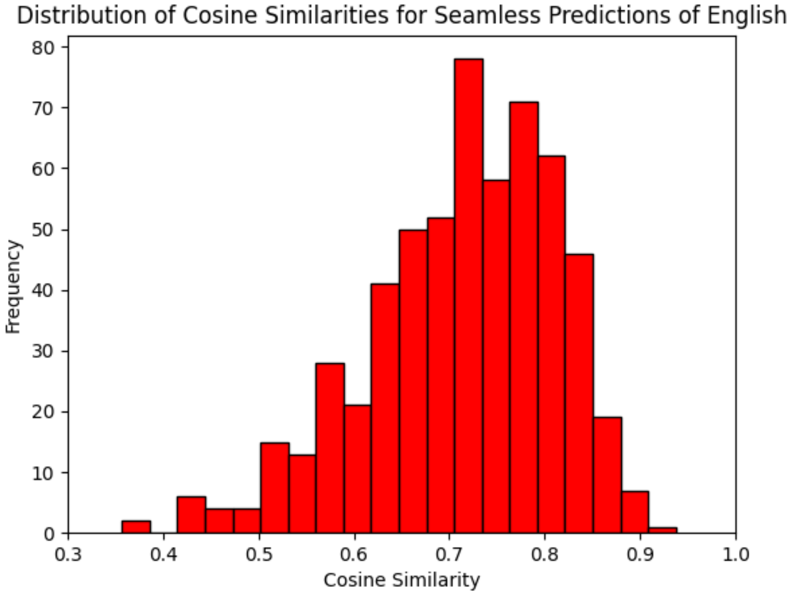

Figure 5.3: A histogram of the similarities for the SeamlessExpressive translations targeting English.

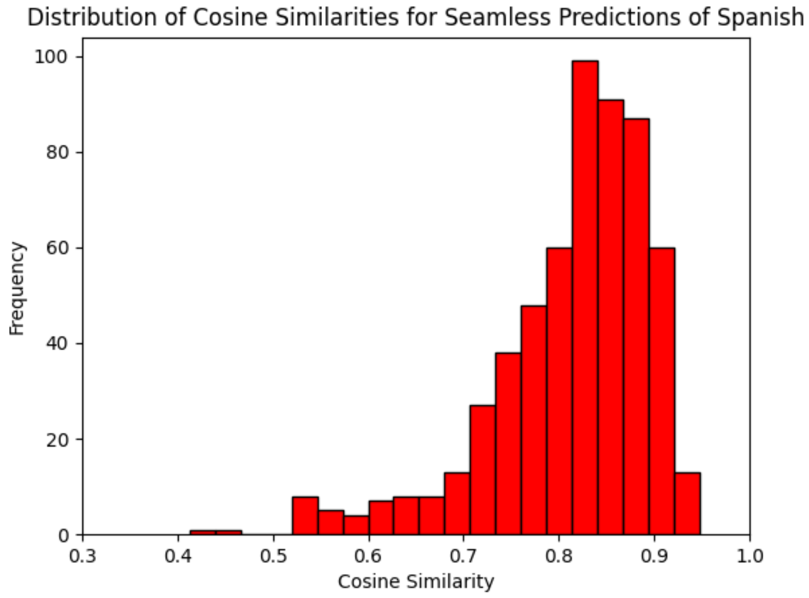

Figure 5.4: A histogram of the similarities for the SeamlessExpressive translations predicting Spanish.

Best Performing for Predicting English

For the best 10 predicting English utterances, scores were between 0.88 and 0.94. The utterances were short, with most being a second or less, and only two utterances being 2 seconds long. Two main patterns were observed in the top performers: backchannels and simple prosodic qualities. The best utterance was a backchannel that said *"Mhm"*, being the small sample on the right edge of Figure 5.3. As an observation, the source audio seemed to have a descending pitch, both the ground-truth target audio and the predicted audio seemed to have a higher pitch towards the center. Another backchannel was a *"Wow"* in utterance 019_16. In both the source and ground-truth target audio, the backchannel expressed a relaxed, surprised reaction. As a minor point, the predicted audio by Seamless seemed to lack the relaxed sentiment by exacerbating the surprise factor. These backchannels differed slightly from the target audio but did not appear to have inappropriate pragmatics. The other scenario in which Seamless did well was utterances with simple prosodic constructs. These were utterances that had no prominent features such as utterance 040_2 saying *"My bed is gray"* with a slight lengthening on the *"is"* which was present in both the ground-truth and the prediction. Utterance 002_26 would have a slightly higher pitch towards the end, having both the ground-truth and the predicted audio saying *"I don't drink coffee"* with the higher pitch being in *"coffee"*. As a side note for this utterance, Seamless appeared to make the voice of the male speaker more gender ambiguous, and the high rating regardless might indicate that the Segura metric is not highly sensitive to speaker differences. Other utterances had an emphasized word, such as utterance 010_17 with a ground-truth of *"didn't have a cover"* and a predicted *"I didn't have the coverage"*, having both an emphasized *"didn't"*. While the semantic content differs, the focus is on pragmatic fidelity, which the utterance appears to have. This showcases a strength of the Segura metric, that it is not highly lexically dependent. Something to note is that utterance 010_17 was part of the worst performers of the human reenactments due to the lack of the matching emphasis. Seamless seems to perform well by the Segura metric for short and straightforward utterances. However, there were also limitations of the Segura metric since some audios did not score too low, despite having noise interwoven between the lexical content. This noise damages the model's utility as a natural communication tool, and would most likely score lower in a MOS-type evaluation.

Best Performing for Predicting Spanish

For the 8 best predicted Spanish, scores ranged from 0.93 and 0.95 with durations ranging from 1 to 4 seconds. The prominent characteristic observed was the timing of the utterance, affecting the speaking rate or pauses to express a sentiment. For instance, utterance 068_15 had a source audio of *"Beto, I like Beto"* with a pause after the first *"Beto"* and speeding up after to show how much the person is liked. Similarly, the ground-truth and the predicted audio both say *"Beto, me gusta Beto"* with the same pause and follow up emphasis. Additionally, utterance 016_17 has a transition from a pause to emphasized speech. The ground-truth said *"...decirles que, no tengan miedo de hacer preguntas"* where *"decirles que"* was said slowly while *"no tengan "* was said with more emphasis. Its predicted audio said *"para decirles que no tengan miedo de hacer preguntas"* with a slight pause after *"para decirles"* and a more emphatic *"que no tengan miedo de"*. This conveys the pragmatics of the source audio saying *"tell them like, don't be afraid to ask questions"*, where the speaker wants to express reassurance. While slowing down was the most prominent way observed to highlight a word, there were other methods through which a word could be emphasized such as lengthening and rising pitch. For instance, *"Ahh"* is emphasized through lengthening, being from the same utterance mentioned in Section 5.2.1. For rising pitch, utterance 009_1 said both in the ground truth and the predicted audio *"tus vacaciones de diciembre"*, with a rising pitch in *"diciembre"*. This conveys the same intent of inquiry from the source audio *"your december vacations"*. Between the shared traits of the best performers for both Seamless and human recreations, it seems that the Spanish HuBERT subset values emphasized words being located in a similar region of the utterance.

Worst Performing for Predicting English

The worst predicted phrases of English had Segura scores ranging from 0.36 to 0.43. These audios were typically longer than the best performers, being around 2 to 4 seconds long, with one exception being less than a second long. This utterance's ground-truth was a *"So"* which was lengthened with a falling pitch to prompt a response, but Seamless was unable to capture the words, and produced a short and emphatic *"Don't say that"*, completely changing the pragmatic function. Additionally, there was another instance where Seamless was not able to capture the correct lexical content. In utterance 050_30, expression *"Y empecé a leer luego luego...Ah caray"* in the source audio, mistaking it for *"And I started reading to Kara"* in the generated audio. This change eliminates the intent of conveying that they had been surprised expressed through a sudden change in speaking style. In contrast, the ground-truth was able to keep this element by saying *"And I started reading and it was like... Oh what the-"* with a similar high intensity towards the end.

Beyond these one-off observations, the most common characteristic found was that translated audios by Seamless appeared to have a loss of sentiment, adopting a more neutral tone of voice. This loss in prosody negatively impacts the pragmatics, conveying a weak signal. For instance, utterance 039_6, expressed excitement through high intensity and a fast speaking rate, stating *"It-its a time where you're supposed to dedicate most of your time."*. On the other hand, Seamless outputted *"Cause it's the time when you have to be dedicated"*, with a neutral voice that still conveyed the exhortation to work hard, but the impact was lost. Another instance was 056_17 which had a ground-truth of *"Yeah, but that was like the next day, no?"* with lengthening on *"the"* as if recalling an event, and rising pitch on *"no?"* to ask for confirmation. Meanwhile, Seamless had *"Yes, but that was like the next day, wasn't it?"* with a somber, neutral voice and a slight rising pitch on the *"right?"*. One could infer the intent of confirmation from the lexical content, but the more neutral tone made Seamless appear apathetic to the conversation. This happened in other pragmatic functions as well, such as in the expression of frustration and the sharing of gossip, where a semblance of the intent could be inferred, but the clarity and effect suffered.

Worst Performing for Predicting Spanish

For the worst predicted Spanish utterances, we had to consider the top 20 utterances to avoid analysis being dominated by apparent post-processing errors in conversation EN_029, filling out more than half of the spots. Thus, for the remaining utterances, their scores ranged from 0.41 to 0.60; however, with one utterance in the 40s while the rest were in the 50s. This can be seen in the slight plateau on the left side of Figure 5.4, with the small "island" on the far left. There are two main patterns observed: failure to capture lexical content on short utterances that are breathy or lengthened, and failing to transfer hesitance from the source audio. For the worst performing instance, utterance 092_20 displays mistranslated lexical content with a short predicted *"Huh"* versus the ground-truth's lengthened *"Si"* and the source audio's lengthened *"Yeah"*. This completely eliminates the affirmation pragmatics that should have been maintained. Additionally, there was a breathy and lengthened *"gap"* in utterance 017_10 that was not understood by Seamless, returning a noisier version of the source audio. This contrasted highly with the soft and delicate pragmatics in the reenactment *"espacio"*. For cases of hesitance that wasn't transferred, Seamless would not transfer the pragmatics displayed in the source audio through filler words and false starts. For utterance 069_2, the source audio in English said *"I have, umm, that's the first one"* where *"have"* and *"umm"* were lengthened to express doubt before the correction. The ground-truth reenacted audio said *"Tengo...es la primera"* where the lengthening in *"Tengo"* was shortened and supplemented with a pause to convey doubt. However, Seamless said *"Esa es la primera"* with an extremely short pause between *"es"* and *"la"*, which changed the pragmatics from self-correcting to appearing confident in what they're saying. Another instance with eliminated doubt was utterance 086_9 which had a source audio of *"And then like it-it's like things happened"* and a reenactment of *"Las cosas pasan y-y"*. Both of these showed doubt through false starts but Seamless eliminated this by saying *"Y luego como que pasan cosas"* without pauses or any other prosody to indicate doubt.

Through the analysis, we were able to see that Seamless does well for short utterances or utterances with simple prosodic constructs. The simple utterances could have one or no salient prosodic properties such as an emphasized word, Additionally, through these best performers, we saw instances which HuBERT did not heavily penalize for changes in speaker identity or slight deviations in lexical content. Despite its competent performance, there were three areas observed in which Seamless can improve: more accurate lexical content, less neutral tone of voice, and less noisy output. First, the worst performers in English and Spanish had instances in which Seamless did not understand what was said.

Lexical content can directly influence what prosodic constructs are possible so it is crucial that it is correctly understood. Another reason to stress this point is that Seamless seemed to fail on instances with extreme prosodic qualities such as the intense *"Ah caray"* and the breathy *"gap"*. In the context of dialog, these instances are critical as they dominate the pragmatics of the message, so it is imperative that our translation tools are able to handle these occurrences. Second, the worst performers in English showed a variety of pragmatic functions which Seamless could not replicate due to it producing a more neutral tone of voice. Thus, by retaining the sentiment found in the original audio, it could become better at generating this wide array of pragmatics. Third, Seamless would produce some instances that were rated high despite them being noisy audios. This is a problem because the noisy audios tend to perform worse in MOS-type evaluations. It is possible that Seamless was only competent in this context due to the type of evaluation we performed. For translation systems, the less noise they produce the more useful they can be. However, this was a minor point when compared to the other two areas. Ultimately, through the analysis we were able to highlight some of the strengths and weaknesses of the Segura metric and SeamlessExpressive.

5.2.3 Multilayer Perceptron

As previously mentioned, the MLP model only predicts the expected speech representation, meaning no generated audio. Without any generated audio to analyze, we must rely on the source and ground-truth pairs to observe patterns in prosody and pragmatics; however, this makes the analysis even more tentative than for the previous two models.

Best Performing for Predicting English

For the best predicted English utterances, the scores ranged from 0.92 to 0.93 and the durations ranged from less than a second to 6 seconds. The source and reenacted audio pairs showed a wide variety of prosodic constructs. Some utterances are

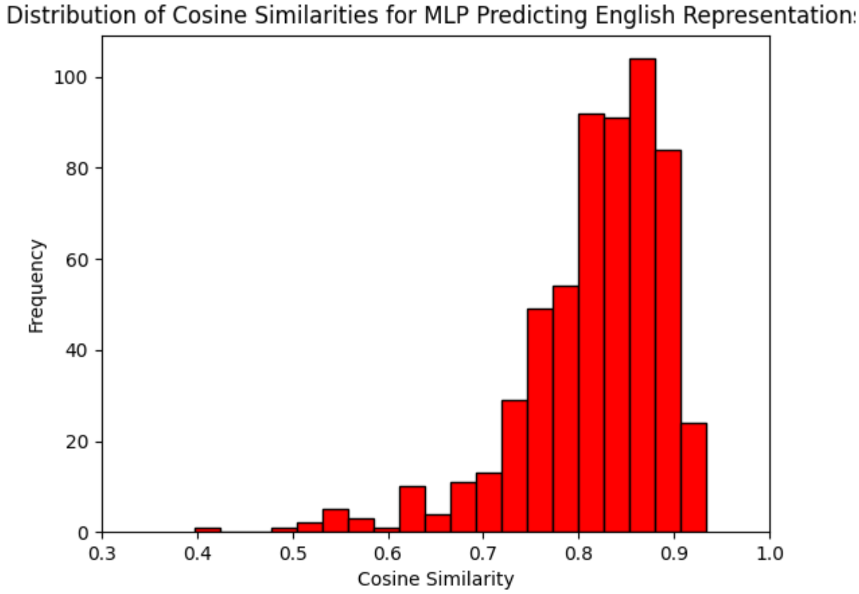

Figure 5.5: A histogram of the similarities for the MLP predictions for English.

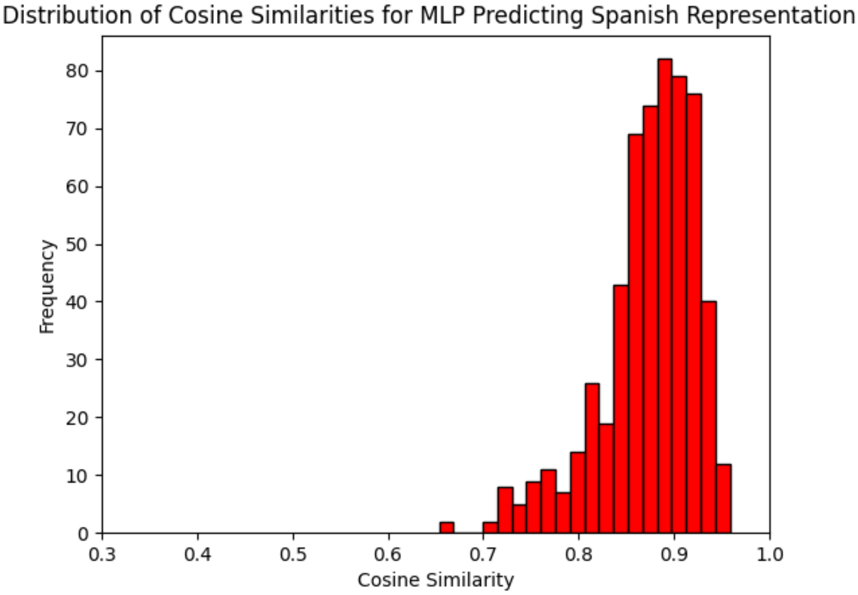

Figure 5.6: A histogram of the similarities for the MLP predictions for Spanish.

non-commital statements with no salient prosodic qualities while others are complex constructs showing sections of fast speaking rate, pauses, filler words, and emphasis through rising pitch. Nevertheless, the majority fell somewhere in between, having a pause and one other feature such as filler words, laughter, or rising pitch. This successful mapping of a large variety of utterances highlights the strength of the model, where it could map such meaningful constructs.

Best Performing for Predicting Spanish

The best predicted Spanish utterances had scores ranging from 0.95 to 0.96, with durations between 4 and 5 seconds. These utterance pairs had three things appear frequently: prominent pause structure, sections of fast speaking rate, and emphasized words. Most of them would share two out of the three items, and once more, the exact location in the utterance did not match, but the general area and order was retained. In terms of pragmatics, there was a variety of intentions displayed, some speakers looking for agreement while others were sharing their frustration.

Worst Performing for Predicting English

The worst performing English utterances had similarity scores ranging from 0.39 to 0.55 with all being 2 seconds or less. Only two were in the 0.40s, with the rest being above 0.50. This can be seen by the gap in the left tail in Figure 5.5. In terms of prosody, the most observed pattern was an additional lengthened word in the reenacted audio not found in the source audio. For instance, utterance 019_54 said in the source *"La primera de Iron Man"*, without any lengthening, and the reenactment said *"The first one of Iron Man"*, with lengthening in *"Man"*. Moreover, if there was lengthening already present in the source, an additional lengthened word would be found in the reenactment. Utterance 038_2 had *"Entonces, hay dos camas?"* had a lengthened *"Entonces"*, but the reenactment said *"So, then, there's two beds?"*, with lengthening in *"So"* and in *"then"*. It is important to note that the extra lengthens did not largely affect the pragmatic intent conveyed for some utterances. This highlights that the Segura metric is not perfect, being sensitive to alternative prosodic qualities which are still in the realm of appropriate pragmatics. Other than this observation, the worst performing utterances had rare behaviors in either the source or the reenactment that are not commonly found across the data, such as a person inhaling or silence after a person stopped speaking.

Worst Performing for Predicting Spanish

The worst predicted utterances for Spanish that were analyzed had scores between 0.67 and 0.73. In Figure 5.6, we can see the MLP having the strongest right skew between all the models, explaining the higher minimum for the worst performers. Once again, the erroneous conversation mentioned in Section 5.2.2 had to be filtered out since the conversation had a couple of spots in the worst performers. The remaining utterances seem to have differing prosodic qualities by substituting one with another type or changing the degree at which the quality is present. For instance, utterance 025_16 had a slight sad expressions with the ground-truth being *"heartbreak"* said in a breathy voice while the reenactment said *"corazones rotos"* with lengthening. Another substitution that was observed was an emphasized *"Ay"* in *"Ay, dios"* through high intensity while the source audio had a lengthened *"God"* in *"Oh my God"*. For the utterances that changed the degree at which prosodic qualities were present, they would add one more pause or lower the intensity at which it was said. These changes in prosody did not affect the conveyed pragmatic function, which showcases the nuanced relationship between the two and suggests that the MLP may have been doing acceptably in fact. With appropriate translations having a wide range of possible expressions, perhaps a multi-reference evaluation would be appropriate in future work. Since we have other human reenactments from DRAL-MRS, we could see if the representations predicted by the MLP matches the reenactments more than the ground-truth, testing the ability of the model to approximate at least one adequate correspondent.

Overall, from the qualitative analysis, we were able to see that pragmatic intent was conserved in the best performers of the Segura metric. Additionally, certain instances exemplified that the Segura metric does not heavily penalize on differences of lexical content such as in utterance 019_19. The Segura metric also did not heavily penalize changes in speaker identity as seen by the highest similarity scores for DRAL-MRS, which was reenactments from speakers different to the original speaker. Additionally, Seamless also had some high performers where the voice was slightly changed, such as the instance with a more gender ambiguous generated voice. Nevertheless, the Segura metric is not perfect: many instances of low-quality, noisy audio produced by Seamless were high performers, which most likely would not do as well in a MOS-type evaluation. Thus, for future work, the Segura metric should be taken into account, but it should not be the sole judge of quality.

By analyzing Seamless, we were able to see potential weaknesses in SOTA models that can be improved in future work. Seamless many times failed to capture the lexical content of short utterances that contained strong prosodic qualities such as utterances with a large amount of breathiness or lengthening. Another weakness was that Seamless would generate a more neutral voice than the one needed to convey the pragmatic intent, damaging the function of the message.

Lastly, for human recreations, we saw that they were able to correctly map pause structures, lengthening, and emphasis, producing a similar pragmatic intent. There were also instances where humans failed to produce adequate correspondent in prosodic constructs to convey a given intent, mainly around conveying doubt and hesitance. This highlights the complex nature of prosodic constructs to pragmatic intent, where some can be straightforward and others can be subtle.

Chapter 6

Concluding Remarks

6.1 Significance of the Results

Through the work of this thesis, we draw 8 tentative conclusions, although all need further investigations.

For conclusions relating to Research Question 1:

1. **Hypothesis 1a:** The thesis models outperformed Avila's hand-crafted features, as seen in Table 5.1, signifying that self-supervised features are more informative than hand-crafted features. While the phenomenon of self-supervised features performing better than hand-crafted features is well known, this is the first direct comparison in the cross-lingual prosody transfer space. It also should be noted that the slight improvement in performance comes at the cost of the model's interpretability, which could have been used to understand the mapping in future work.

2. **Hypothesis 1b:** The thesis models outperformed human performance as seen in Table 5.1. This suggests that a S2ST system with superhuman performance in cross-lingual prosody transfer is possible. However, this finding is only very tentative: when the audio is able to be generated, this should be corroborated by other evaluation metrics such as the results from a MOS evaluation.

3. **Hypothesis 1c:** The thesis models outperformed SOTA models as seen in Table 5.1. This implies that improvement in cross-lingual prosody transfer could be readily achievable. Similarly to the finding about human performance, there is the caveat

that the finding should be corroborated with other metrics, especially those that capture the percept of pragmatic fidelity.

For conclusions relating to Research Question 2:

4. The Segura metric serves well as a pragmatic similarity score. As discussed in Section 5.2, the generated audios that performed well seemed to retain the same pragmatic intent we would perceive from the utterance-pairs. Additionally, several utterances that were rated high by Segura metric conserved the intent while having slight differences in lexical content or speaker identity. These properties are useful, but there were only small deviations in the predicted audio from the ground-truth. In this work, we did not quantify how much of an impact the lexical and speaker differences have on the Segura metric. Moreover, the Segura metric displayed two other imperfections: penalizing audios with changed prosodic qualities that are still in the realm of appropriate pragmatics and having noisy audios in the best performers. These two scenarios likely wouldn't occur in evaluations based on human judgments. Towards the end goal of better communicative agents, human judgments should still be the gold standard as the agents are aimed to serve people, but the Segura metric has its place as an objective pragmatically-focused evaluation.

5. A proof of concept for a pragmatic similarity evaluation pipeline that allows for the comparison of audio-based and audio-free systems, as described in Figure 4.1. This pipeline can be useful to support development of new methods in which teams can contribute to advances even if they lack the resources to develop a full S2ST system.

6. Audio-free systems have an easier task than audio-based systems by only having to predict a prosodic representation. Non-naive audio-free systems always outperform audio-based systems as seen in Table 5.1. This affects the evaluation of methods since audio-free systems have an inherit advantage which should be kept in mind when performing future comparisons with this evaluation pipeline.

Additional, more speculative conclusions:

7. SOTA models are performing not too far below human-level in terms of pragmatic-faithfulness. This can be seen in the close similarity scores of SOTA models to human re-enactors in Table 5.1. However, this does not paint the full picture as SOTA models sometimes produce noisy output or fail to capture the correct lexical content. Thus, the current techniques available for prosody allow the SOTA models to be competent, but there is still clear room for growth.

8. Potential evidence suggesting that English is more pragmatically-complex than Spanish. The higher similarity scores for Spanish as compared to the scores for English in Table 5.1 could indicate that Spanish is less pragmatically complex than English. While the described relationship between English and Spanish could help explain this, it is not certain if this is the case or if it is the sole reason for the trend as discussed in Section 5.1.1.

These findings related to HuBERT and cross-lingual prosody transfer help shed more light to the evaluation of translation systems in the context of dialog. This is a critical step to help shift the paradigm from intelligible speech to natural speech in the speech synthesis field. This in turn is hoped to bring about better AI communication agents that can help connect people effectively despite the language barrier, avoiding common miscommunication issues present today.

6.2 Qualitative Indications

Through our qualitative analysis in Section 5.2, we were able to find initial observations with regards to human recreations, SeamlessExpressive, the HuBERT feature subsets, our models, and the DRAL dataset. These could each be followed up in future work as we continue to improve our understanding and modeling of prosody and pragmatics.

1. Humans do well in simple prosodic constructs, especially with constructs relating to pauses or emphasis. Perhaps this is due to these features being the most salient to untrained listeners. Some humans may find it difficult to convey hesitance or doubt despite there being many options for how to express it, such as filler words or false starts.

2. Seamless does well in messages with one salient property such as an emphasized word or no salient property. This benefit of no salient properties could be from it defaulting to a more neutral voice, which comes at the cost of a wider array of pragmatic functions that rely on more engaged voices. Seamless also struggles with understanding the lexical content of utterances with extreme prosodic properties, which is necessary for a translation tool in a dialog context.

3. The thesis models seemed to handle a range from simple to moderately complex prosodic constructs based on its high performing utterance pairs. This is with the caveat that the models did not synthesize audio so we cannot determine how much of the predicted representation would be kept in the final audio.

4. The HuBERT feature subsets seem to value properties such as pauses, lengthening, and intensity. This could potentially be due to how the embeddings were trained through a masked language modeling objective. For instance, pauses and high intensities could appear as low or high signals that are predicted through the masks. Additionally, lengthening could be a continuous signal over many frames, thus leading the model to learn it during training. However, this is highly speculative as there are currently no reliable ways to interpret self-supervised features.

5. Constructing pragmatically similar datasets can be tricky. Here we found a common weakness in several DRAL pairs where the re-enactor was not able to convey the hesitance or doubt found in the source audio. Potentially, re-enactors are influenced by the source audio telling them what to say instead of them having to think about

the lexical content from scratch. The effort needed to think about what to say can cause hesitance in speech, thus when the effort is removed it might be more difficult to express hesitance.

6.3 Next Steps

Future work can explore the impact of substituting different items of the approach. The following list highlights the initial items one might consider in future work.

1. **Feature Extractors:** The higher scores in Spanish have multiple possible explanations, one being the HuBERT model utilized. Thus, to quantify how much of an impact the feature extractor training data has on these translation tasks, a new study could be conducted with a HuBERT trained on multilingual data. The multilingual data would need a comparable amount of Spanish and English data to potentially affect the trend. If the trend remains, the complexity difference in English and Spanish pragmatic functions would be more likely a real effect.

2. **Input Aggregations:** In the thesis, we explored three different HuBERT-based feature sets with varying degrees of success; however, other forms of aggregations could be attempted. Dialog is highly interconnected in its sequence so more sophisticated modeling could help improve the mappings between languages. For instance, to take into account how some prosodic constructs have different scales of duration, one might model the audio by getting averages hierarchically: across the entire utterance, then across two windows, four windows, and so on. This varying degree of granularity would allow the features to represent the utterance at different timescales.

3. **Predictive Models:** Carrying on with the idea of more sophisticated modeling, different deep learning models could be evaluated with HuBERT-based features, particularly models that excel at sequential data such as Recurrent Neural Networks or Long Short-Term Memory Networks. These sequential models were mainly avoided

because the focus was to see if HuBERT is an appropriate speech representation, assuming that, if HuBERT is effective, even basic models should be able to perform reasonably well.

4. **Systematic Evaluation of the Segura Metric:** In this thesis, we assumed the Segura metric's pragmatic similarity information was appropriate in a translation context since only utterances of the same language were compared, as in [24]. While it did seem to perform adequately in terms of pragmatics, translation is a unique context in which utterances may be expressed differently to express the same function such as through differing lexical content. We observed in Section 5.2 that high performance by the Segura metric was possible despite slight variations in lexical content, but are unsure of the impact lexical differences have on the metric. With a systematic evaluation of the Segura metric, we would be able to quantify the impact lexical differences has on the Segura metric. This allows us to determine the best use of the Segura metric, and with it, a better evaluation of translation systems.

5. **Multi-reference Evaluation:** In our evaluation pipeline, we only considered one ground truth from the original DRAL dataset. This could potentially be improved by considering other references in the evaluation such as the ones provided by DRAL-MRS. Since a pragmatic function can have many adequate corresponding prosodic constructs, a multi-reference evaluation would allow us to determine if models choose at least one appropriate correspondent.

6.4 New Research Questions

The thesis had findings relating to cross-lingual prosody transfer, but it also brought into light more research questions that future work can answer. Future questions to answer stemming from this work are:

1. How does the complexity of prosodic constructs vary in English as compared to in

Spanish?

2. How much does the lexical content of a message affect the possible prosodic constructs in each language?
3. Are multilingual datasets necessary to obtain self-supervised features that are prosodically informative for multiple languages?

Answering these questions would help our understanding on how languages relate to one another and how we can better model them for improved communicative agents.

6.5 Towards Pragmatically Focused Speech Synthesis

Research questions aside, it would be easy to apply the thesis models for two purposes. First, for pragmatically focused speech synthesis, allowing us to build better communicative agents. The main application of the thesis models in speech synthesizers would be to have an additional penalty in the loss function for the difference between the output's representation and the target representation. Second, the thesis models would fit naturally in a cascaded S2ST system, where the predicted prosody representation can be used to condition the speech synthesizer. With such prosody-conditioned audio, new avenues for research and evaluation would be opened. We could begin to explore the limitations of the self-supervised features through qualitative analysis of the translated audios instead of relying on listening to the source and target pairs. Direct evaluations could also be made with other baselines through Mean Opinion Score experiments or other prosody related metrics such as AutoPCP [16]. For all these reasons, the construction of a S2ST system based on the thesis models would be the highest priority for future work.

References

[1] Jonathan E Avila and Nigel G Ward. Towards cross-language prosody transfer for dialog. In *Interspeech*, pages 2143–2147, 2023.

[2] Taejun Bak, Youngsik Eom, SeungJae Choi, and Young-Sun Joo. MultiVerse: Efficient and Expressive Zero-Shot Multi-Task Text-to-Speech. *arXiv preprint arXiv:2410.03192*, 2024.

[3] Zongyang Du, Kun Zhou, Berrak Sisman, and Haizhou Li. Spectrum and prosody conversion for cross-lingual voice conversion with cycleGAN. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 507–513. IEEE, 2020.

[4] Medha Hira, Arnav Goel, and Anubha Gupta. CrossVoice: Crosslingual prosody preserving cascade-S2ST using transfer learning. In *ICLR*, 2024.

[5] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, October 2021.

[6] Patricia Hudelson and François Chappuis. Using voice-to-voice machine translation to overcome language barriers in clinical communication: An exploratory study. *Journal of General Internal Medicine*, 39(7):1095–1102, Feb 2024.

[7] Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*, pages 10120–10134. PMLR, 2022.

[8] Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. CVSS corpus and massively multilingual speech-to-speech translation, 2022.

[9] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-Light: A Benchmark for ASR with Limited or No Supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673, 2020. `https://github.com/facebookresearch/libri-light`.

[10] Savya Khosla. ML | Locally weighted Linear Regression, 2023. `https://www.geeksforgeeks.org/ml-locally-weighted-linear-regression/`, Last accessed on 2025-04-24.

[11] Min-Kyung Kim and Joon-Hyuk Chang. Adversarial and Sequential Training for Cross-lingual Prosody Transfer TTS. In *Interspeech*, pages 4556–4560, 2022.

[12] Tao Li, Chenxu Hu, Jian Cong, Xinfa Zhu, Jingbei Li, Qiao Tian, Yuping Wang, and Lei Xie. DiCLET-TTS: Diffusion model based cross-lingual emotion transfer for text-to-speech—a study between English and Mandarin. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3418–3430, 2023.

[13] Guan-Ting Lin, Chiyu Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung yi Lee, and Nigel G. Ward. On the Utility of Self-Supervised Models for Prosody-Related Tasks. *IEEE Spoken Language Technology Workshop (SLT)*, pages 1104–1111, 2022.

[14] Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Interspeech*, pages 4823–4827, 2023.

[15] Livia Qian, Carol Figueroa, and Gabriel Skantze. Representation of perceived prosodic similarity of conversational feedback. In *Interspeech*, pages 374–378, 2025.

[16] Seamless Communication et al. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint arXiv:2312.05187*, 2023.

[17] Milan Sečujski, Branislav Gerazov, Tamás Gábor Csapó, Vlado Delić, Philip N Garner, Aleksandar Gjoreski, David Guennec, Zoran Ivanovski, Aleksandar Melov, Géza Németh, et al. Design of a speech corpus for research on cross-lingual prosody transfer. In *Speech and Computer: 18th International Conference, SPECOM, Budapest, Hungary, August 23-27*, pages 199–206. Springer, 2016.

[18] Atli Sigurgeirsson and Simon King. Do prosody transfer models transfer prosody? In *ICASSP 2023*, pages 1–5, 2023.

[19] Jakub Swiatkowski, Duo Wang, Mikolaj Babianski, Giuseppe Coccia, Patrick Lumban Tobing, Ravichander Vipperla, Viacheslav Klimkov, and Vincent Pollet. Expressive machine dubbing through phrase-level cross-lingual prosody transfer. In *Interspeech*, pages 5546–5550, 2023.

[20] Jakub Swiatkowski, Duo Wang, Mikolaj Babianski, Patrick Lumban Tobing, Ravichander Vipperla, and Vincent Pollet. Cross-lingual prosody transfer for expressive machine dubbing. In *Interspeech*, pages 4838–4842, 2023.

[21] M. Valenzuela Farías. A comparative analysis of intonation between Spanish and English speakers in tag questions, wh-questions, inverted questions, and repetition questions. *Revista Brasileira de Linguística Aplicada*, 13:1061–1083, 12 2013.

[22] Suraj Verma. Locally Weighted Linear Regression in Python, 2021. https://towardsdatascience.com/locally-weighted-linear-regression-in-python-3d324108efbf/#:~:text=In%20Locally%20weighted%20linear%20regression,(i)%20’s%20data., Last accessed on 2025-04-24.

[23] Nigel G. Ward, Jonathan E. Avila, Emilia Rivas, and Divette Marco. Dialogs Reenacted Across Languages, Version 2. Technical Report UTEP-CS-23-27, University of Texas at El Paso, 2023.

[24] Nigel G. Ward, Andres Segura, Alejandro Ceballos, and Divette Marco. Towards a general-purpose model of perceived pragmatic similarity. In *Interspeech*, pages 4918–4922, 2024.

[25] Nigel G. Ward, Javier Vazquez, and Emma Boushka. Tools for Estimating the Perceived Level of Phonetic Reduction. Language Resources and Evaluation Conference (LREC) 2026, submitted.

[26] Xinfa Zhu, Yi Lei, Tao Li, Yongmao Zhang, Hongbin Zhou, Heng Lu, and Lei Xie. METTS: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1506–1518, 2024.