RI: Small: Modeling Prosody for Speech-to-Speech Translation

Project Description

Speech-to-speech translation (S2ST) systems are valuable tools for enabling cross-language communication. While very useful already for short, transactional interactions, they are less so for richer exchanges. Indeed, the utility of these systems will always be limited unless they can translate more than words. In particular, prosody transfer is needed to reliably convey many intents and stances. Unfortunately, current translation systems generally aim only to produce prosody that sounds natural, which is not always sufficient.

The construction of speech-to-speech translations systems that can translate prosody is today beyond the state of the art: no such systems exist, nor does the field know how to build them. Our aim is to advance scientific understanding and technical knowledge to enable the construction of such systems.

1 The Richness of Dialog and the Importance of Prosody

Consider the Spanish utterance

> *¿Van a venir, venir tus papás para ...*
> (are-they-going to come, come your parents to [help you move in]?)

and its translation:

> *Are your parents gonna come, or ...*

The original utterance occurred in the context of the interlocutor talking about being excited into moving into his own apartment in a couple of days. The words alone are only modestly informative, but as this was speech, there is in addition the prosody. (In this proposal we'll use prosody in the broadest sense, to include all aspects of the speech signal not explained by the word sequence nor the speaker identity, including phenomena of pitch, intensity, duration, timing, and features that pattern with them.) In this example the prosody conveys: that this is a question, that this is eliciting not a yes/no response but some elaboration on the situation and the interlocutor's feelings, that the speaker invites him to infer what she's leaving unsaid (*to help you move in*) and expects that he can infer it, and that she doesn't mean to pry, in case the reasons for moving out of his parents' home are something he doesn't want to go into. In a cross-language interaction, if a translation system could carry these nuances through into the English[1], the conversation could go smoothly. If not, there could be confusion, awkwardness, and misunderstandings. In some cases this might not matter much — the excerpt above was from a casual conversation — but the same sort of issues arise in high-stakes dialogs.

Imagine a scenario where a patrol team who don't share a language are moving through a complex urban environment. Through a translation device one might ask the others to *go check behind the truck* and receive as confirmation *behind the truck, okay*, but if all that is communicated is agreement on the location, this will often be suboptimal. Speakers sharing a language might use these same words, but, in addition, use prosody to convey things like whether each is stressed, semi-distracted, frustrated, tired, unsure of the reason for this action, confident that this is the right thing to do, and so on. Beyond such speaker-state indications, appropriate nonverbals might also convey information about the planned action, including the anticipated difficulty of moving to the target location, and the likelihood of finding something behind the truck. Speech features can also

[1]Since it is hard to convey the added value of speech over text in a written document, I will mention that the Spanish and English audio for this example are available via [a URL findable via the query UTEP DRAL].

English

holding the turn
showing reluctance
showing enthusiasm
taking the turn
yielding the turn
backchanneling
cuing a backchannel response
topic closing
topic continuation
topic elaboration
positive assessment
empathy bid
indifference
marking a transient disfluency
marking topic-comment structure
politeness
long floor yield
off-topic interpolation
marking a significant point to consider
cuing action
grounding *vs* grounded
difficult *vs* easy
comfortable *vs* awkward

Mandarin

holding the turn
taking the turn
yielding the turn
negative evaluation
brief pause while thinking
new information receipt
cue for new information receipt
reiterating a point
topic elaboration
enthusiasm
engagement
empathy bid
display of empathy
topic exhaustion
cue for a short response
contrast

Figure 1: Some of the most common pragmatic functions of prosody in English and Mandarin. Based on analyses of American English conversations among strangers (Switchboard) [1], Scottish English task-oriented dialogs (MapTask) [2], and Mandarin conversations among friends or family members (CallHome) [3].

convey dialog-contextual factors such as whether the hearer is confident he or she understands the meaning, whether this demands urgent action or is just setting up a plan, or whether it is intended as, or understood as, an immutable directive or a contingent suggestion.

The same phenomena can be seen for a single word. For example, *okay* can mean many things, such as "I don't understand what you're asking," or "I understand what you mean," or "I'm convinced and will do what you ask," or "I'm starting to execute a tricky maneuver and need to concentrate," or "I've succeeded," or "I'm ready for more instructions," and so on. When translated with neutral prosody, the hearer can only guess the intended meaning.

There are many other aspects of meaning conveyed with prosody, as suggested by Figure 1. It is true that not all dialog situations involve communication of intents like these, for example, simple answering of factual questions may not. It is also true that, in some dialog situations, users sometimes can read the intention directly from their interlocutor's source-language behavior (although this can be risky: for example, an Arabic speaker who ends an utterance with a sharp pitch downdash is often merely cueing the listener to backchannel to show that they are paying attention [4], but English speakers may misperceive this as an accusation or curt dismissal [5]). Further, it is true that speakers can back off to a slower, more explicit style, where more of the meaning is conveyed through words, to reduce the risk of prosodic misunderstandings.

In general, however, there are many dimensions of interaction that are present in human dialog, that speech-to-speech translation should support, and that can be facilitated with appropriate prosody.

2 The Research Landscape: Speech-to-Speech Translation

Systems like Google Translate and Skype Translator are technical marvels. They are also very useful in many situations. Thanks largely to recent advances in deep learning and the development of enormous corpora, progress in S2ST has been very rapid, towards higher quality translations, more languages, smaller footprints, lower latency, textless learning, smaller training-data requirements, and so on [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. However, currently, the vector of progress is not pointing towards better support for dialog interactions: indeed, at present, no other research team seems to be even contemplating work on this. This is not because of a lack of awareness of the need, but due to other factors, including three huge barriers.

2.1 Barrier 1: Suitable Data

Today translation systems are trained on data. No progress is made without data, and the field has accordingly assembled many impressive resources for text translation and speech-to-text translation. Corpora for speech-to-speech translation have lagged, but some impressive offerings exist [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31], as we have surveyed elsewhere [32, 33].

However, none are adequate to advance translation for *dialog* purposes. These corpora mostly are based on monologues, and in particular are derived from readings [22, 34, 35, 21], political discussions [18], or informative talks [36, 20, 27]. For general use, we really do not want to produce speech-to-speech translations systems that are useful only for polished one-way communication, yet that is what these datasets support. The few corpora that exhibit dialog elements were derived from television show or movie dubs [27, 37, 31] or lectures and press conferences [26], or created via (generic) speech synthesis [13, 24]. The speech in these collections lacks interactivity, spontaneity, and most of the pragmatic and prosodic variation common in real dialog.

2.2 Barrier 2: Suitable Quality Measures

To judge the quality of prosody transfer, the field will need a measure of pragmatic similarty between utterances. This will enable estimates of how closely a model's predicted prosody matches the observed prosody in the human reference translation. Lacking such a metric, there is currently no way to automatically estimate whether, for example, two instances of *okay* mean roughly the same thing, or different things entirely. Similarly, there is no way to determine whether speech saying *that's really interesting* and *that reminds me of a story I heard* may be functionally closely similar, given suitable prosody, despite the different lexical content, for conveying a polite topic closure intent.

As the field has recently made great progress in metrics for related tasks, there is a lot we can learn from.

For semantic similarity there exist fairly good metrics, developed mostly by the natural language generation and machine translation communities. Beyond classic edit-distance metrics and BLEU, there are embedding-based distance metrics, MoverScore, and metrics trained to match human judgments [38, 39, 40]. For speech, textless similarity metrics are desirable [41], and here Blaser is an inspiring new metric [42], but it may work well only for fairly large quality differences,

and models only semantic similarity. Cross-lingual embedding models have great promise, but none yet have been trained on or tested for pragmatic similarity.

Prosodic similarity models are less developed. Still popular are *a priori* metrics such as the mel-cepstral distortion, which uses a crude representation of everything the human ear can hear, and, prosody-specific measures, focusing mostly on intonation (F_0) alone [43, 36, 44, 45, 46]. More recent work adds duration features [47, 48, 49], and more [50, 51], but never all aspects (pitch, intensity, timing rate, phonetic reduction, and voice quality features). While various reasonably comprehensive utterance-level representations of prosody exist [52, 1], almost all have been designed to support supervised learning, not attempting to place utterances in a space in which distances have meaning. No prosodic similarity metrics have been tuned for or tested for ability to detect pragmatic similarity, except very crudely in our own preliminary work, described below.

Metrics for speech synthesis quality are mostly for naturalness, which is almost orthogonal to pragmatic appropriateness, but do suggest ways to train to match human judgments, and architectures of models for estimating them [53, 54, 55]. Ways to measure synthesizer output quality in terms of any specific dimension of pragmatic function or style can be built, as has been done for clarity of turn-hold/yield intentions [56]. However, the number of important pragmatic functions is large, such as those relating to the marking of topic shifts and connections, the coordination of action and the managing of turn taking, and the expression of engagement, stance, and attitude. No existing taxonomy includes all functions that are important in dialog [57, 58, 59], and anyway, one-by-one modeling of quality in terms of specific single functions is not scaleable. A general pragmatic similarity model is needed.

2.3 Barrier 3: Basic Models of Cross-Language Prosody Correspondences

Prosody is the key difference between spoken and written communication. (Nonlexical utterances, such as *uh-hmmm*, are probably the (distant) second.) People use prosody to convey many important aspects of meaning, feeling, and intent, as illustrated in Section 1. This section overviews four relevant research strands and explains why the problem of cross-language prosody mappings remains open.

First, for speech-to-speech translation, a few researchers have, over the decades, sporadically modeled aspects of prosody (e.g. [60, 61, 62, 63, 37]). To date, these target only a few functions of prosody, notably its roles in conveying paralinguistic/emotional state, emphasis, and syntactic structure, and target only a few prosodic features, notably F_0, pausing, and word duration. Very recent work has shown that this can measurably and significantly improve perceived translation quality [37], but also that these techniques so far only close less than half of the perceived gap between default prosody and the human reference, and only for one melodramatic corpus. Perhaps because none of this work has addressed most of the functions that are most common and most significant in actual dialogs, there has been no real uptake or impact.

A radically different strategy involves the rejection of modular architectures, based in part on the critique that classic, "cascaded" (modular) architectures for S2ST, which combine a speech recognizer, a machine translation model, and a speech synthesizer [64], and thus rely on intermediate representations that are just text, lose everything that text does not represent, notably prosodic information. In this newer "direct" strategy, models are trained or tuned end-to-end. An initial motivation for this strategy was the hope of learning some prosody transfer ability implicitly, from the data [6]. Despite much progress with direct models [11, 9, 12, 10, 13], as yet the only evidence they can handle prosody is in a spate of papers this summer on "style transfer." While interesting and certainly related to prosody, such styles (variously including sets such as {broadcast, storytelling ...}, {polite, vibrant, cute, honest, gentle ...}, and various emotions [65, 49, 16, 15]), are

coarse-grained and chosen to support one-way communication, so the relevance of this work to prosody translation for dialog is unclear.

Second, across speech technology research more generally, prosody is a popular topic, but there is a lack of suitable knowledge and methods for the *pragmatic* aspects of prosody. This is in contrast to prosody applications for intelligible speech synthesis and for paralinguistic inference (in which a system infers a speaker's identity, health conditions, emotion or other attributes from speech samples). Given sufficient training data, it is possible to train a system to perform any of these tasks very well. However, these two applications areas relate only weakly to the pragmatics-related functions that are so important in dialog. Improving synthesizers to do better for pragmatic functions and for prosody in general is, however, a very active area of research [66, 67, 68, 69].

Third, in the linguistics sphere, recent advances in the prosody of pragmatic functions have taken us far beyond earlier focus merely on final pitch contours that distinguish sentence modalities. In particular, much recent work considers prosodic features across entire utterances, diverse pragmatic functions, and the role of pragmatic features beyond pitch. To expand on the last point, it is now known that pitch features alone often fail to capture most of the details of meaningful prosody [70, 71, 72, 73]. For example, English has four constructions which are difficult to distinguish using pitch alone, as all involve a pitch downstep, but if timing, intensity, and voicing properties are considered, the four patterns — for apologizing, cursing, praising, and cuing action — are clearly different. However the speech technology field mostly still treats prosody as a matter of intonation (pitch contours) and durations alone.

Fourth, comparative work on prosody across languages has a long history (although there is a countervailing tendency in linguistics to avoid ethnocentric bias by describing each language purely in its own terms). Unfortunately, this work focuses mostly on syllable-level, lexical, and syntactic prosody [74, 75, 76, 77]. In particular, there is relatively little work on differences in how prosody conveys pragmatic functions. Even for English and Spanish, a well-studied pair, with a great deal of work on their prosody (surveyed for English in [1] and for Spanish in [78, 79, 80]), comparative work has been rare. Documented knowledge is limited to a few topics such as turn-taking [81], questions and declaratives [82, 83, 84], and expression of certainty [75]. However, these certainly do not exhaust the prosodic meanings important for dialog. Further, these studies have been mostly limited to differences in intonation and duration, leaving out most prosodic features.

Overall, our understanding of pragmatics-related prosody is as yet inadequate to support dialog-effective speech-to-speech translation.

3 Goals

Our overall aim is making progress towards conversational speech-to-speech translation. We plan a multi-pronged strategy, targeting four specific goals.

G1 We aim to characterize the issues in cross-language prosody transfer, to explore what is easy with current technology, what is difficult, and what further advances we need.

G2 We aim to develop models that map prosody between English and Spanish, as a constructive way to approach G1. Specifically, these models will take as input an utterance of one language and predict appropriate prosody for the translation in the other language. Aiming to build these mappings gives us a specific goal with both practical utility and driver for progress on fundamental issues of representation and modeling.

G3 We aim to produce sufficient bilingual conversation data, closely matched, utterance by utterance, to support G1, G2, and other investigations by other research groups.

G4 We aim to develop evaluation methods and metrics for modeling human perceptions of pragmatic similarity between spoken utterances, to support G2, and also applications beyond translation.

All of these goals are novel. Each has value in itself, yet they are strongly synergistic.

To be clear, we do *not* aim to produce or improve a full working system. (While we hope to do this eventually, we cannot realistically commit to this within the scope of a Small project. In any case, because S2ST is a very active research area, we expect that other groups will use our results to also build working systems, probably sooner than we can.) Thus the predictions of our models will *not* be of actual audio output, but of specifications for the prosody of the audio output. The advantage of backing off from systems-building is that it enables us, and other groups, to focus attention on one key problem. The disadvantage lies in complicating evaluation, but this problem can be overcome, as discussed below.

4 Preparation

To prepare to do the proposed research, we collected a small corpus, implemented a first-pass metric, and made some small discoveries:

4.1 Preliminary Corpus Development

To overcome the limitations noted in Section 2.1, last year we developed the Dialogs Re-enacted Across Languages (DRAL) data collection protocol [33]. This involves pairs of nonprofessional, bilingual participants. They first have a ten-minute conversation, which we record. These conversations are unscripted, although we sometimes suggest topics or activities to boost the pragmatic diversity. Depending on their relationship, the participants mostly get to know each other, catch up on recent happenings, and/or share personal experiences. Subsequently, under the direction of a producer, they select an utterance or exchange and closely re-enact it in their other language, which may take several attempts until the naturalness and fidelity satisfies everyone: the speaker, the interlocutor, and the producer. They then re-enact another utterance, and iterate. The yield is typically a few dozen matched pairs per one-hour session. Our design choices are discussed further in a technical report [33]. This protocol enables the collection of data that 1) exhibits a wide range of communicative functions, 2) is closely parallel at the utterance level, and 3) reflects real language use. Indeed, bilinguals informally listening to randomly sampled utterances were rarely able to distinguish re-enacted from original utterances better than chance.

Following this protocol we have so far collected 3816 matched English-Spanish utterance pairs, with an average duration of 2.7 seconds, from a total of 129 speakers [33]. Approximately 24% of this data is the designated test set. The latest release, including examples, pairs, source recordings and metadata, is available at [a URL findable via the query "UTEP DRAL"]. In the explorations described below, we used the first 1139 matched "short" pairs.

While this small dataset has already been useful to us, extensions and enhancements are needed.

4.2 Preliminary Metric Development

Needing a metric of pragmatic similarity, this year we started by trying the Euclidean distance between the prosodic representations of two utterances, with all features given equal weight, where each utterance is represented with 100 features: 10 base features (specifically: intensity, lengthening, creakiness, speaking rate, pitch highness, pitch lowness, pitch wideness, pitch narrowness, peak disalignment (mostly late peak), and cepstral peak prominence smoothed (CPPS), the latter

Model	English →Spanish	Spanish →English
Source-Prosody-Ignoring	12.65	12.32
Naive Prosody Transfer	11.35	11.35
Linear Regression	9.23	9.37

Table 1: Average Error (Euclidean distance between prediction and reference)

an inverse proxy for breathy voice), each computed over ten non-overlapping windows, together spanning the whole utterance. There was a fair bit of cleverness in designing these features to be robust, well-normalized and perceptually relevant, but they still have significant limitations [1, 32, 85].

Nevertheless, a small-scale validation study found three gratifying things [32]. First, the metric captures many aspects of pragmatic similarity, including speaker confidence, revisiting unpleasant experiences, discussing plans, describing sequences of events, and describing personal feelings. Thus prosodic similarity can be, as expected, a useful proxy for pragmatic similarity. Second, the similarities found were not generally lexically governed: while some words and syntactic structures have characteristic prosody, a metric that ignores lexical alignments can have value. Third, at least one of the causes for the metric sometimes failing to match perceptions may have a simple fix [32].

Thus, while this metric has already been useful, it needs to be greatly improved or replaced.

4.3 Preliminary Findings

We used the data and metric described above in a preliminary investigation of four basic questions [32]. We did this by comparing the performance of models representing different ways to produce prosody in a translation context. While we describe these below for the English→Spanish direction, the conclusions held for both directions. In each case, the model's task was to infer the proper values for the 100-feature representation of the target-language prosody.

First, we wanted to test our basic assumption that source-language prosody is informative regarding the appropriate target-language prosody. To test this, we compared the performance of two models. The first synthesized Spanish utterances based only on the words, thus entirely ignoring the information in the source-language prosody. The prosody of these synthesized utterances was then computed and compared to the prosody of the human-generated translations, using the similarity metric described above. The second naively proposed the prosody of the source-language utterance as the appropriate prosody for the target-language utterance. As seen in the first two lines of Table 1, the second method did better. Clearly, as expected, ignoring the source-language prosody is not a great strategy.

Second, we wanted to see how far we could get with a very simplistic method. We built a linear regression model to predict the prosodic feature values for Spanish from those for English. As seen in the last line of Table 1, this outperformed the naive baseline, showing that even a simple model can learn some aspects of the mapping between English and Spanish prosody. The task is certainly not intractable.

Third, we wanted to see how hard the problem was. From the table we see that, although the simple linear model shows a benefit, its prediction error is still very high: there is still a lot to do.

Fourth, we wondered where the prosody of English and Spanish most differ. We examined cross-language feature correlations and found, for example, that high CPPS in Spanish across utterances

correlates with intensity at the beginning and end of English utterances, while no correlations exist within either language. Examining sample pairs that mostly closely embodied this correspondence indicated that that they were utterances that were setting the stage for follow-up explanation. These connections seem all to be unknown in the literature. We also examined a sampling of cases where the naive model and the linear regression models performed worst, and found, for example, that the common final breathy pitch rise in English (frequently used in grounding referents), was lacking in the corresponding Spanish utterances. We later learned that this difference is well-known among scholars of American Spanish. We also tentatively noted differences in the uses of creaky voice, and differences common in the functions of moving the topic in a more personal direction, of leading into something, of marking contrast, and of taking the turn.

These findings are based on a mere 1139 matched utterances, an old-school small hand-crafted set of prosodic features, an ad hoc quality metric, and very simplistic modeling. Thus all are tentative and should be revisited. Still, they suggest that our basic assumptions are correct, and demonstrate that we are prepared to do the work proposed in the next section.

5 Approach

As our aim is novel and our goals also novel, we will not seek to break new ground in methodology, preferring well-known, reliable techniques as much as possible.

5.1 Thrust 1: Metrics for Pragmatic Similarity

The creation of a metric of quality, in terms of pragmatic similarity, will support comparison of different approaches, serve the shared task (discussed below), and enable evaluation of overall project success. A ideal metric will support not only system-level comparisons, but also utterance-by-utterance quality judgments to support fine-grained analysis of where more work needs to be done, and perhaps even serve as a loss function for model training.

As noted earlier, the output of our models will be specifications for the prosody of utterances. This brings a challenge. If there existed a synthesizer capable of realizing arbitrary prosodic specifications, we could just use it to create audio, and then measure quality using human perceptions of the pragmatic similarity between the synthesized and reference speech. However, no existing synthesizer is capable of this for the rich set of prosodic features that are important for dialog. Thus we need a method that can estimate this using prosodic specifications, rather than using fleshed-out speech signals. We will consider many possible approaches, but at the moment are inclining towards two.

First, we plan an empirical approach, aiming to build a model that works directly from the acoustic signal, and that directly models human similarity judgments. Accordingly, we will collect such judgments for both English and Spanish (which we will release as a general resource), and train a model to approximate them. We will exploit all prosodic properties, and favor models that are mostly explainable, probably including hand-designed features and hand-crafted architecture elements. We will start by simply learning per-feature weights that improve on the Euclidean model, and then iterate through cycles of failure analysis and redesign. Among other things, these may lead us to improve the features, add new features to the set, or shift to an existing or newly developed pretrained feature set. The result will be an utterance-prosody representation that is better in the sense of being well suited for similarity estimation.

Second, we will try precision-at-k modeling, where the similarity model is trained together with the prosody mapping model to maximize the probability of selecting the correct, human-generated translation, from among all the possible translations (our entire Spanish utterance set), or at least

finding it in the top k matches. That is, we will predict the target-language utterance prosody, and then select the utterance which best matches that prediction.

This thrust will not only produce a useful metric, but provide answers to some general research questions: What is the added value of including prosodic features beyond pitch and intonation for pragmatic meanings? Do prosodic features and normalizations designed specifically for the pragmatics-related aspects of prosody do better than generic prosodic feature sets, such as OpenSmile [86]? Alternatively, how much leverage can pre-trained models provide? How well does an English-optimized similarity model work for Spanish, and conversely? How far can we get using prosodic features alone? To what extent can prosody support pragmatic similarity judgments between utterances with different lexical content? The answers should be useful beyond S2ST, for those needing pragmatic similarity metrics for language proficiency assessment, screening for communicative disorders, and dialog systems development.

We will evaluate the metrics primarily in terms of Spearman correlation with average normed human judgments.

5.2 Thrust 2: A Shared Task

A good way to catalyze progress on speech technology problems like this is to create a shared task and challenge all comers to do their best. The speech technology community is particularly enthusiastic about such evaluations. Speech-to-speech translation research specifically has, in recent years, benefited greatly from many shared tasks, such as, most recently, [87]. However, so far all such evaluations have been done using read speech, monologue data, or synthesized speech.

Having data and, soon, a metric, we will be able to support a shared task, likely named Translation of Prosody in Dialog Utterances. There is a community very likely to engage on this: speech translation is a very active area of research. One core group favors the annual International Workshop on Spoken Language Translation (IWSLT), with 60+ attendees in 2023. Other research teams favor larger venues such as Interspeech. There is active industry involvement, notably by Meta, Amazon, Google, Apple, Microsoft, Zoom and Bytedance, with involvement also by Huawei, Xiaomi, and Naver, and smaller companies such as Papercup, AppTek and Interactions. University involvement includes JHU, CMU, GMU, UCSB, Maryland, Oregon State, and Northeastern in the US, and universities abroad in China, Germany, Japan, Czechia, Israel, and France. ARL and AFRL also have projects in this area, although with applied aims and low profiles.

From my review of the literature, various individual conversations, and, especially, attendance at the IWSLT 2023 panel discussion, it is clear that there is a nascent consensus that supporting more user needs is important. Further, the community broadly wants to expand beyond working on the same few standard tasks, but lacks suitable data to do so. This is as true for the top-tier tech companies as for the smallest research teams.

Our steps in organizing a shared task will include: 1) Affiliating it either with Interspeech, which would reach a broad research community, tend to highlight the scientific questions, and tap wide expertise in prosody, speech synthesis, and machine learning, or with IWSLT, which would engage a focused group of researchers and would likely support a multi-year shared task. 2) Forming an advisory board and determining the scope of the task: with our data and metric it would be possible to broaden the task beyond prosody to cover all aspects of dialog, but doing so would have both pros and cons. 3) Adopting the right competition-supporting platform, perhaps CodaLabs. 4) Fixing the details of evaluation. In particular, as we expect some teams will build full-fledged systems, in addition to automatic-metric performance results, we will need to provide summative evaluations using human judgments. Among the various dimensions of machine translation quality (semantic fidelity, naturalness, etc.), pragmatic equivalence will be our sole

focus. (We will also likely not consider the role of source-language prosody in target-language word choice.) We will mostly likely do MUSHRA-like evaluations [88] by banks of human judges convened in Saturday sessions on-campus, again leveraging our bilingual population. Judges will evaluate the appropriateness of target-language utterances produced by various systems (or under various conditions), likely in terms of "overall appropriateness," and perhaps "utility for someone attempting to communicate but not knowing the source language." However the details will need to be determined in consutation with the advisory board.

In terms of effort level, since we already plan to develop the metric and prepare adequate data, the added time needed to organize a shared task will be modest, and the benefits should far exceed that cost. Even purely selfishly, the feedback that other teams will give on our metrics, our data, and our modeling ideas will be invaluable. We also accept that participating teams may explore approaches or aspects that would never have occurred to us, or show performance levels that leave ours in the shade. In fact, we welcome such outcomes. Overall, it seems very likely that well-designed infrastructure here would attract wide community participation and be a critical enabler for the field.

We will evaluate this thrust based on the level of participation and the quality and impact of the resulting systems and findings.

5.3 Thrust 3: Mappings/Models

In this, the main thrust, our target will be a module that accurately maps the prosody from a dialog utterance in one language to the equivalent prosody in the other. While the resulting software artifact may be useful, the true value of this thrust will be in insights it gives on numerous fundamental issues.

The first class of research questions relates to modeling and algorithms. We plan to learn generally how to best build such mappings, and also: What is the value of models pretrained on standard tasks? How much leverage do multilingual semantic embeddings provide, and where do they fall short? How can we build models that learn to disentangle the pragmatics-related aspects of prosody, which should be transferred, from its other functions, which mostly should not? How much value is in the notion of prosodic constructions (temporal configurations of multiple streams of prosodic functions, which can be present to greater or lesser extents, each associated with a meaning or function) and the associated mechanism of superposition?

The second class of research questions involve fundamental properties of human languages: How large is the actual role of prosodic features other than pitch? To what extent can we model pragmatics-related prosody without regard to the word stream? To what extent to different languages convey the same meanings with prosody, instead of diversely using lexical content, gesture, or context, or leaving them implicit? To what extent is there free variation in prosody, without implications for meaning or function? How much does conditioning interpretations on the individual speaker, beyond simple normalizations, help? To what extent is the meaning of prosodic configurations context-dependent? (Pervasively we feel that, when listening to an utterance out of context, we can infer the intent, the previous context, and some likely possible continuations. If this is true, then the target-language prosody will be mostly determinable from the source-language utterance alone, without considering the context. This is not only a practical question, but a theoretical one. In particular, the interactional linguistics literature frequently stresses how prosody (or at least intonation) is only meaningful in context. We will measure the extent to which prosody is adequately translatable without reference to the larger context, and thus provide evidence regarding the extent of truth of the general claim.) We will address these questions in part by ablation studies and in part by analysis of where simple models fail.

The third class of research questions involves representations. These are important for two reasons.

The first reason is the data quantity issue. If we could somehow obtain thousands of hours of suitable data, we could probably use machine learning to create models able to handle everything, including the prosody mappings, with no need for prosodic representations or perhaps even prosody-specific research at all. This is seen by the success of models like AudioPaLM and Voicebox [15, 17]. However such volumes of bilingual parallel dialog data are unlikely to become available for decades, if ever, so we need to design and learn representations that can support data-efficient learning of the cross-language mappings [13]. Conceptually, we envisage exploiting monolingual conversation data to build initial language-specific models, and subsequently using our bilingual data as a kind of Rosetta Stone to enable mapping them into the same space, but we will see.

The second reason is the potential utility for many other applications. Low-dimensional embeddings for the semantics of words and word sequences have revolutionized natural language processing, and similar representations of prosodically conveyed pragmatic meanings should also have a broad impact. While pragmatic functions have traditionally been represented using discrete labels, we will develop vector-space models. A priori, it seems that prosody is even more suited to low-dimensional representation than words: instead of hundreds of thousands of words, there are, at most, a dozen relevant prosodic feature streams, and likely only tens of meaningful temporal configurations. A vector-space model is also natural in terms of the nature of the meanings that prosody conveys: moving to a vector-space model should not be limiting; for example, traditional categorical notions, such as "confirmation question" can be mostly represented as combinations of more basic elements, such as "assume agreement + factual focus + expect short answer." While vector-space representations of prosodic meanings seem like a natural choice, this possibility has not been previously explored.

The ideal representation would be universal yet map simply to the surface-level prosody of every language, and, as noted above, directly support simple computation of accurate pragmatic similarity estimates. Research questions include how to develop or learn such representations. At the moment we think that max pooling (over an utterance) of the counts and strengths of various language-specific prosodic configurations would be a good next step beyond our 100-features surface-level represention described above, but there is a wide space of possibilities to explore. These include extending multilingual semantic embeddings [89] to include pragmatic information. General pretrained models (a.k.a. self-supervised models, the equivalent of large language models for speech [90]), may also capture useful information [91], although so far such models have never been trained or tested on conversational speech. Designed-by-hand representations, for example perhaps based on those that have been useful for linguistic description, or for describing the pragmatic functions of utterances [92], may also have utility.

Specific representation-related research questions include: How many dimensions are needed to adequately describe pragmatic meanings? How do pragmatic meaning representations or spaces relate to lexical embedding spaces? Are the prosodic representations best support cross-language mappings also the best in terms of supporting similarity estimates? and What are the merits of utterance-wide versus word-by-word and moment-by-moment representations of prosodic meaning?

We will evaluate this thrust based on performance according to the metric and based on impact on the field. While we do not expect to produce the ultimate model, nor to provide the final answer to any of these questions, we will produce a proof-of-concept and informative first-pass answers that developers can rely on.

5.4 Thrust 4: Data

The most suitable data set for the investigations described above, namely our own DRAL, needs enhancements and extensions.

We plan to address two critical limitations of our existing data set.

First, since there is more than one way to validly translate an utterance, additional references are needed, to support more meaningful evaluations of system performance. There are at least two ways we might collect these. One possibility would be to have the original speakers produce alternatives, prompting them to "now do it again, but with different words or in a slightly different style." Another possibility would be to have a third-party speaker produce the translations after the fact. The latter could provide much higher throughput.

Second, we need an additional language pair to support evaluation of generality, both for our own methods and for the shared task.

We would also, of course, like to increase the size of our Spanish-English collection, both for our own purposes and for the teams in the shared evaluation. Further, we would like to collect data in a new genre, perhaps problem-solving dialogs. Specific planning will happen after we learn from progress on the first three thrusts and after receiving requests and suggestions from the shared-task teams. We will choose based on what is most likely to drive progress towards insight, utility and generality. While we have outsourcing options lined up if necessary for the new-language pair, in all other respects we plan to collect the data at our institution, leveraging our largely bilingual student body.

Whatever the details, the outcome will be a data collection appropriate for the shared task, and also able to support other research endeavors.

We will evaluate these data collections based on intrinsic quality measures [33], utility for the other thrusts, and uptake by other teams.

5.5 Thrust 5: Descriptive Case Studies

While advancing the science of linguistics is not a direct aim of this project, we will generate some descriptive knowledge about the prosody of Spanish and English. These efforts will help identify challenges, inform our modeling work, and hopefully inspire other teams to do research in this area, not only technologists but also linguists. The direct outcome will be a publication that presents an overview of the most common ways in which Spanish and English most differ in their use of prosody to convey pragmatic functions, and great detail on one or two functions as case studies.

Thanks to our novel corpus and our willingness to apply mixed methods, we expect to rapidly make significant findings. We will not scorn observational methods, but the main power will come by using our computational framework to evaluate the performance of the naive model, that assumes that the prosody of English and Spanish are the same. The utterances for which this baseline does worst will, we expect, be mostly associated with a handful of pragmatic functions, and we will identify them using qualitative-inductive methods. This might enable us to, for example, characterize the common problems as relating to specific pragmatic functions or intents (perhaps sarcasm), or to specific contexts (perhaps topic change regions), or specific speakers, or interactions with lexical content, and so on.

5.6 Thrust 6: User Studies

For further insight into the issues, beyond what we can obtain from corpus studies and systems building, we plan a small human-subjects investigation, to explore the in-use value of prosody-aware translation. While definitive findings will have to wait for experiments with future systems

in actual use, we will work towards early estimates of how much value prosody adds and which aspects of prosody are of most value to users. The details remain to be worked out, but we may do in-the-community Wizard-of-Oz studies in which human interpreters facilitate conversations in two conditions: translating only the words or additionally adding the prosodic meanings. We may then have participants evaluate the resulting interactions in terms of temporal efficiency, quality and quantity of information transferred, perceived likability of the interlocutor, cognitive load, and overall preference, and also provide free-form comments and suggestions. This data will also tell us which prosodic features and which functions the interpreters strive most to retain, for example, likely not most aspects of turn-taking intentions, but probably most expressions of topic structure and connections, stance, and attitude. A nice side-effect of these experiments will be a new set of translation pairs that we can release as a supplementary corpus.

The results will inform not only our own work, but also help others to roadmap the path to broadly valuable translation systems. This understanding will be useful not only to researchers and developers, but also to potential customers of future systems, ranging in size from small businesses on the border to the US Army, enabling them to understand the components of the value proposition, and to be able to better specify what they really need from technology.

6 Expected Technical Impacts

While we may not see success for every goal, our multi-pronged strategy will certainly advance knowledge of prosody modeling and prosody translation. While prosody is not the only thing needed for the field to move beyond content-only to dialog-suitable speech-to-speech translation (non-lexical utterances such as *uh-hmm* come to mind), we expect that it will take us a long way.

We do not see prosody translation as an eternal question. After three years we expect that we will have good answers to most of the research questions. Builders of speech-to-speech translation systems will then know a lot about what works and what doesn't: hopefully enough to create and deploy prosody-capable systems. This will be true for those who favor modular (cascaded) systems, enabling them to build few-parameter prosody-specific modules. This will also be true for those who favor end-to-end systems, probably relying on foundation models in the form of new multilingual embedding spaces that are pretrained to also represent the prosodic and pragmatic aspects of utterances. Realistically, only the top-tier tech companies have the resources to build such foundation models, but the work we propose is essential for informing how they design, train, and evaluate them, and for informing the field how to use them to build systems.

Advancing our knowledge of prosody and pragmatic functions will also help the field of speech technology more broadly. Our answers to the research questions, will, for example, also transform the possibilities for human-agent and human-robot interactions. Improved interaction capabilities are essential for reaping the full power of AI, a national priority. Referring to the report of a recent NSF-sponsored workshop on the future of spoken language interaction with robots [93], our proposal aligns with Recommendation 10, "better exploit prosodic information," and Recommendation 13, "extend the pragmatic repertoire of speech synthesizers." Today speech synthesizer output tends to be prosodically neutral and uninformative, and our findings will help to advance beyond this. In general, we need our AI-enabled devices to better understand our intents, in order to communicate faster and more reliably. For example, our "behind the truck" scenario from Section 1 is equally relevant if one of the partners is a robot. In particular, we need ways to make synthesizers able not only to produce intelligible and natural output, but output that conveys specified communicative functions, and that can therefore be informative and effective in dialog applications. The representations that we develop may meet the need for a specification suitable for such meaningful control of synthesizer output, and the metric will support evaluation

and improvement.

7 Broader Impacts

The first category of broader impacts will stem from the improved speech-to-speech translation systems this work will enable.

This will support many use cases, some well-known, such as enabling businesspeople to have more effective conversations with more people and get more things done, helping diplomats communicate better and reduce conflict among nations, and helping soldiers avoid tragic misunderstandings and better coordinate with allies and civilians. Adding pragmatic competence to these systems will increase the bandwidth, improve understanding, and enable better outcomes.

Beyond such high-prestige genres, our work will also benefit ordinary people using language for humble purposes [94, 95, 96, 97]. The primary purpose of many real-world conversations is not to convey information, but to establish connections and rapport. Use of prosody, for example in conveying micro-intents, is an important mechanism for establishing rapport [98, 99]. Society could benefit from more rapport across people with different language abilities, even within families: we know people who can't really communicate with some cousins, or with their grandchildren. Stronger family connections across extended immigrant families can help reduce loneliness, and enable family members to share wisdom, support each other, understand their roots, and feel more love. Beyond the family, there can also be benefits for society. For example, many immigrants today end up doing jobs where only limited communication skills suffice, as maids, gardeners, taxi drivers, and so on. One study of immigrant populations found that "participants felt that the [existing] translation tools had limitations that severely impacted their life" [95]. Enabling such individuals to interact more effectively would have an economic benefit, enormously increasing the effective "human capital" of our nation, and a huge social-inclusion benefit.

The second category of broader impacts will come by helping people learn to communicate better, unaided. I admit that improved translation systems might reduce the incentive to learn foreign languages and to learn to adapt to other conversation cultures, but they may also bring more people into a learning zone, where talking to someone across languages is no longer intimidating but rather a challenge to rise to, fostering growth opportunities. In addition, corpus and out descriptive case studies will enable and hopefully inspire linguists to pile in on questions of cross-language prosody. The findings they generate could eventually help teachers of English (respectively, Spanish) to be more specific and directed as they help their learners go beyond the acquisition of lexical and grammar skill to achieve true communicative competence. I also foresee utility for language practice systems. In the long term, advancing our explicit knowledge of how people convey pragmatic intents in dialog may also help those people with autism [100] and others who, for whatever reason, haven't mastered the prosodic aspects of communication in the usual subconscious way.

The third category of broader impacts will arise from student involvement in the work itself. This will be substantial and, very likely, involve mostly Hispanic students, reflecting the demographics of our institution, our department, and our research group, although Hispanics are, nationally, an underrepresented minority.

8 Results from Prior NSF Support

We are part of the recently awarded "AI Institute for Transforming Education for Children with Speech and Language Processing Challenges" (National Science Foundation and the Institute of Education Sciences, U.S. Department of Education Award # 2229873, 1-15-2023 $\sim$ 1-14-2028,

UTEP allocation $475,309). The overall project aims include improved workflows for screening for early diagnosis of language disorders for children. Our goals include building software to model perceived pragmatic similarity for children's utterances, so there is good synergy with the current proposal. As yet our only results are the explorations described in Section 4.2 and [32].

9 Notes on Fairness

Ultimately our work will help democratize speech-to-speech translation, by supporting more dimensions of interaction, in more scenarios, and for more people. However we acknowledge many limitations in our plan. In particular, our data collection will not be representative of all languages or even all dialects of English and Spanish. We do not expect it to be used directly for training of deployable systems, but if they are used in this way, the resulting systems will be less useful for speakers of other dialects. Both our data collections and our human-studies work will rely on our undergraduate population, again limiting generality. Ultimately fairness will require future follow-on work with more diverse populations.

10 Team Qualifications

The PI has a long record of contributions in speech processing, especially involving dialog and prosody, mostly for English [76, 1, 101, 102, 103, 104], but also for other languages [105, 3, 106]. He also has experience with pretrained models [91, 107], and doing human-subjects experiments [108, 5], including by crowdsourcing [109, 110]. He has released software tools [85] and collaborated widely across diverse academic disciplines and with industry partners. He has experience planning and managing a shared task [111], which attracted 10 teams, 4 of whom persisted to completion, and led to findings reported in multiple publications, including a journal article [112]. He is a well-connected member of the speech technology community, with specific connections already with several of the main players in speech-to-speech translation.

The CoPI brings broad expertise in machine learning and its applications across problems in vision and language. Basic research themes include algorithms for deep network initialization, attribute selection and creation, and transfer learning. Applications include emotion detection, prosodic analysis, brain mapping, analysis of astronomical and solar data, earthquake prediction, and ecological monitoring. He is Spanish-English bilingual.

Together with our students, we are well-prepared and eager to do what we propose.

11 Rough Timeline and Effort Allocation

Thrust		Timespan			Effort	Supports
1	Metrics	Y1	Y2		23%	Goals 4, 1, 2; Thrusts 2, 3, 4
2	Models	Y1	Y2	Y3	43%	Goals 1, 2; Thrust 5
3	Shared Task	Y1	Y2	Y3	13%	Goals 2, 1, 4; Thrust 2
4	Data		Y2		7%	Goals 3, 1, 2; Thrusts 2, 3
5	Case Studies			Y3	7%	Goal 1
6	User Studies		Y2		7%	Goal 1

References

[1] N. G. Ward, *Prosodic Patterns in English Conversation*. Cambridge University Press, 2019.

[2] N. G. Ward, "Automatic discovery of simply-composable prosodic elements," in *Speech Prosody*, pp. 915–919, 2014.

[3] N. G. Ward, Y. Li, T. Zhao, and T. Kawahara, "Interactional and pragmatics-related prosodic patterns in Mandarin dialog," in *Speech Prosody*, 2016.

[4] N. Ward and Y. Al Bayyari, "A prosodic feature that invites back-channels in Egyptian Arabic," in *Perspectives on Arabic Linguistics XX* (M. Mughazy, ed.), pp. 186–206, John Benjamins, 2007.

[5] N. G. Ward and Y. Al Bayyari, "American and Arab perceptions of an Arabic turn-taking cue," *Journal of Cross-Cultural Psychology*, vol. 41, pp. 270–275, 2010.

[6] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," *arXiv preprint arXiv:1904.06037*, 2019.

[7] R. Fukuda, S. Novitasari, Y. Oka, Y. Kano, Y. Yano, Y. Ko, H. Tokuyama, K. Doi, T. Yanagita, S. Sakti, K. Sudoh, and S. Nakamura, "Simultaneous speech-to-speech translation system with transformer-based incremental ASR, MT, and TTS," in *24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques*, pp. 186–192, IEEE, 2021.

[8] T. Kano, S. Sakti, and S. Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 958–965, IEEE, 2021.

[9] A. Lee, P.-J. Chen, C. Wang, J. Gu, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, *et al.*, "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[10] Q. Dong, F. Yue, T. Ko, M. Wang, Q. Bai, and Y. Zhang, "Leveraging pseudo-labeled data to improve direct speech-to-speech translation," *arXiv preprint arXiv:2205.08993*, 2022.

[11] S. Popuri, P.-J. Chen, C. Wang, J. Pino, Y. Adi, J. Gu, W.-N. Hsu, and A. Lee, "Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation," *arXiv preprint arXiv:2204.02967*, 2022.

[12] X. Li, Y. Jia, and C.-C. Chiu, "Textless direct speech-to-speech translation with discrete speech representation," *arXiv preprint arXiv:2211.00115*, 2022.

[13] Y. Jia, Y. Ding, A. Bapna, C. Cherry, Y. Zhang, A. Conneau, and N. Morioka, "Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation," in *Interspeech*, pp. 1721 –1725, 2022.

[14] K. Wei, L. Zhou, Z. Zhang, L. Chen, S. Liu, L. He, J. Li, and F. Wei, "Joint pre-training with speech and bilingual text for direct speech to speech translation," *arXiv preprint arXiv:2210.17027*, 2022.

[15] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, *et al.*, "AudioPaLM: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[16] J. Swiatkowski, D. Wang, M. Babianski, P. L. Tobing, R. Vipperla, and V. Pollet, "Cross-lingual prosody transfer for expressive machine dubbing," in *Interspeech*, 2023.

[17] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, and W.-N. Hsu, "Voicebox: Text-guided multilingual universal speech generation at scale," *arXiv preprint arXiv:2306.15687*, 2023.

[18] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, pp. 993–1003, 2021.

[19] P.-A. Duquenne, H. Gong, N. Dong, J. Du, A. Lee, V. Goswani, C. Wang, J. Pino, B. Sagot, and H. Schwenk, "SpeechMatrix: A large-scale mined corpus of multilingual speech-to-speech translations," *arXiv preprint arXiv:2211.04508*, 2022.

[20] R. Cattoni, M. A. Di Gangi, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: A multilingual corpus for end-to-end speech translation," *Computer Speech & Language*, vol. 66, p. 101155, 2021.

[21] M. Z. Boito, W. N. Havard, M. Garnerin, É. L. Ferrand, and L. Besacier, "MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible," in *LREC*, 2020.

[22] C. Wang, A. Wu, and J. Pino, "CoVoST 2 and massively multilingual speech-to-text translation," *arXiv preprint arXiv:2007.10310*, 2020.

[23] Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen, "CVSS corpus and massively multilingual speech-to-speech translation," in *LREC*, 2022. arXiv preprint arXiv:2201.03713.

[24] C. Zhang, X. Tan, Y. Ren, T. Qin, K. Zhang, and T.-Y. Liu, "UWspeech: Speech to speech translation for unwritten languages," in *AAAI 2021*, vol. 59, p. 132, 2020.

[25] T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki, "A speech and language database for speech translation research," in *Third International Conference on Spoken Language Processing*, pp. 1791 – 1974, 1994.

[26] K. Doi, K. Sudoh, and S. Nakamura, "Large-scale English-Japanese simultaneous interpretation corpus: Construction and analyses with sentence-aligned data," in *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pp. 226–235, 2021.

[27] A. Öktem, M. Farrús, and A. Bonafonte, "Corpora compilation for prosody-informed speech processing," *Language Resources and Evaluation*, vol. 55, pp. 925–946, 2021.

[28] M. Baali, W. El-Hajj, and A. Ali, "Creating speech-to-speech corpus from dubbed series," *arXiv preprint arXiv:2203.03601*, 2022.

[29] P. Jeuris and J. Niehues, "LibriS2S: A german-english speech-to-speech translation corpus," in *LREC*, 2022.

[30] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2023.

[31] W. Brannon, Y. Virkar, and B. Thompson, "Dubbing in practice: A large scale study of human localization with insights for automatic dubbing," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 419–435, 2023.

[32] J. E. Avila and N. G. Ward, "Towards cross-language prosody transfer for dialog," in *Interspeech*, 2023.

[33] N. G. Ward, J. E. Avila, E. Rivas, and D. Marco, "Dialogs re-enacted across languages, version 2," Tech. Rep. UTEP-CS-23-27, University of Texas at El Paso, Department of Computer Science, 2023.

[34] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *LREC (12th)*, pp. 4218–4222, 2019.

[35] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[36] E. Salesky, J. Mäder, and S. Klinger, "Assessing evaluation metrics for speech-to-speech translation," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 733–740, 2021.

[37] W.-C. Huang, B. Peloquin, J. Kao, C. Wang, H. Gong, E. Salesky, Y. Adi, A. Lee, and P.-J. Chen, "A Holistic Cascade System, Benchmark, and Human Evaluation Protocol for Expressive Speech-to-Speech Translation," in *ICASSP*, 2023.

[38] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance," in *ENMLP*, 2019.

[39] J. Wieting, T. Berg-Kirkpatrick, K. Gimpel, and G. Neubig, "Beyond BLEU: training neural machine translation with semantic similarity," in *ACL*, 2019.

[40] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," *Journal of Artificial Intelligence Research*, vol. 77, pp. 103–166, 2023.

[41] L. Besacier, S. Ribeiro, O. Galibert, and I. Calapodescu, "A textless metric for speech-to-speech comparison," *arXiv preprint arXiv:2210.11835*, 2022.

[42] M. Chen, P.-A. Duquenne, P. Andrews, J. Kao, A. Mourachko, H. Schwenk, and M. R. Costa-jussà, "Blaser: A text-free speech-to-speech translation evaluation metric," in *ACL*, 2023.

[43] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, pp. 63–68, 2008.

[44] D. J. Hermes, "Auditory and visual similarity of pitch contours," *Journal of Speech, Language, and Hearing Research*, vol. 41, pp. 63–72, 1998.

[45] U. D. Reichel, F. Kleber, and R. Winkelmann, "Modelling similarity perception of intonation," in *Interspeech*, pp. 1711–1714, 2009.

[46] O. Nocaudie and C. Astésano, "Evaluating prosodic similarity as a means towards L2 teacher's prosodic control training," *Proceedings of Speech Prosody 2016*, pp. 26–30, 2016.

[47] H. Mixdorff, J. Cole, and S. Shattuck-Hufnagel, "Prosodic similarity: Evidence from an imitation study," in *Speech Prosody*, pp. 571–574, 2012.

[48] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T.-A. Nguyen, M. Rivière, A. Mohamed, E. Dupoux, *et al.*, "Text-free prosody-aware generative spoken language modeling," in *ACL*, 2022.

[49] W.-C. Huang, B. Peloquin, J. Kao, C. Wang, H. Gong, E. Salesky, Y. Adi, A. Lee, and P.-J. Chen, "A holistic cascade system, benchmark, and human evaluation protocol for expressive speech-to-speech translation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[50] A. Rilliard, A. Allauzen, and P. Boula de Mareuil, "Using dynamic time warping to compute prosodic similarity measures," in *Interspeech*, 2011.

[51] N. G. Ward, S. D. Werner, F. Garcia, and E. Sanchis, "A prosody-based vector-space model of dialog activity for information retrieval," *Speech Communication*, vol. 68, pp. 85–96, 2015.

[52] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 190–202, 2016.

[53] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The voiceMOS challenge 2022," in *Interspeech*, 2022.

[54] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, "SpeechLMScore: Evaluating speech generation using speech language model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[55] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, IEEE, 2022.

[56] E. Ekstedt, S. Wang, É. Székely, J. Gustafson, and G. Skantze, "Automatic evaluation of turn-taking cues in conversational speech synthesis," in *Interspeech*, 2023.

[57] R. Fernandez, D. Haws, G. Lorberbom, S. Shechtman, and A. Sorin, "Transplantation of conversational speaking style with interjections in sequence-to-sequence speech synthesis," in *Interspeech*, 2022.

[58] H. Bunt and V. Petukhova, "Semantic and pragmatic precision in conversational AI systems," *Frontiers in Artificial Intelligence*, vol. 6, p. 896729, 2023.

[59] F. Seebauer, M. Kuhlmann, R. Haeb-Umbach, and P. Wagner, "Re-examining the quality dimensions of synthetic speech," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[60] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *Verbmobil: Foundations of speech-to-speech translation*, pp. 106–121, Springer, 2000.

[61] P. D. Aguero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I, IEEE, 2006.

[62] Q. T. Do, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Improving translation of emphasis with pause prediction in speech-to-speech translation systems.," in *IWSLT*, 2015.

[63] T. Kano, S. Sakti, S. Takamichi, G. Neubig, T. Toda, and S. Nakamura, "A method for translation of paralinguistic information," in *Proceedings of the 9th International Workshop on Spoken Language Translation: Papers*, pp. 158–163, 2012.

[64] G. Kumar, M. Post, D. Povey, and S. Khudanpur, "Some insights from translating conversational telephone speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3231–3235, IEEE, 2014.

[65] K. Song, Y. Ren, Y. Lei, C. Wang, K. Wei, L. Xie, X. Yin, and Z. Ma, "StyleS2ST: Zero-shot style transfer for direct speech-to-speech translation," in *Interspeech*, 2023.

[66] A. Abbas, S. Karlapati, B. Schnell, P. Karanasou, M. G. Moya, A. Nagaraj, A. Boustati, N. Peinelt, A. Moinet, and T. Drugman, "ecat: An end-to-end model for multi-speaker tts & many-to-many fine-grained prosody transfer," in *Interspeech*, 2023.

[67] M. Lenglet, O. Perrotin, and G. Bailly, "Local style tokens: Fine-grained prosodic representations for tts expressive control," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[68] S. Wang, G. E. Henter, J. Gustafson, and E. Szekely, "On the use of self-supervised speech representations in spontaneous speech synthesis," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[69] Y. Hu, C. Zhang, J. Shi, J. Lian, M. Ostendorf, and D. Yu, "Prosodybert: Self-supervised prosody representation for style-controllable tts," in *ICLR*, 2023, submitted.

[70] O. Niebuhr and H. Pfitzinger, "On pitch-accent identification: The role of syllable duration and intensity," in *Speech Prosody*, 2010.

[71] O. Niebuhr, "On the phonetics of intensifying emphasis in German," *Phonetica*, vol. 67, pp. 170–198, 2010.

[72] O. Niebuhr, K. Gors, and E. Graupe, "Speech reduction, intensity, and F0 shape are cues to turn-taking," in *SigDial*, pp. 261–269, 2013.

[73] O. Niebuhr, "Stepped intonation contours: A new field of complexity," in *Tackling the Complexity of Speech* (R. Skarnitzl and O. Niebuhr, eds.), pp. 39–74, Charles University Press, 2015.

[74] J. Romero-Trillo, *Pragmatics and prosody in English language teaching*. Springer, 2012.

[75] D. Ramirez Verdugo, "A study of intonation awareness and learning in non-native speakers of English," *Language Awareness*, vol. 15, pp. 141–159, 2006.

[76] N. G. Ward and P. Gallardo, "Non-native differences in prosodic construction use," *Dialogue and Discourse*, vol. 8, pp. 1–31, 2017.

[77] M. Ortega-Llebaria and L. Colantoni, "L2 English intonation: Relations between form-meaning associations, access to meaning, and L1 transfer," *Studies in Second Language Acquisition*, vol. 36, pp. 331–353, 2014.

[78] J. I. Hualde, *The Sounds of Spanish, Chapter 14: Intonation*. Cambridge University Press, 2005.

[79] C. de la Mota, P. M. Butragueño, and P. Prieto, "Mexican Spanish intonation," in *Transcription of Intonation of the Spanish Language* (P. Prieto and P. Roseano, eds.), pp. 319–350, Lincom Europa, 2010.

[80] V. Escandell-Vidal and P. Prieto, "Pragmatics and prosody in research on Spanish," in *The Routledge Handbook of Spanish Pragmatics* (D. A. Koike and J. C. Felix-Brasdefer, eds.), pp. 149–166, Routledge, 2020.

[81] A. Berry, "Spanish and American turn-taking styles: A comparative study," in *Pragmatics and Language Learning Monograph Series, Volume 5* (L. F. Boulton, ed.), pp. 180–190, University of Illinois, Urbana-Champaign: Division of English as an International Language, 1994.

[82] J. D. Bowen, "A comparison of the intonation patterns of English and Spanish," *Hispania*, vol. 39, pp. 30–35, 1956.

[83] M. G. V. Farias, "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*, vol. 13, pp. 1061–1083, 2013.

[84] G. Zárate-Sández, "Production of final boundary tones in declarative utterances by English-speaking learners of Spanish," in *9th International Conference on Speech Prosody*, pp. 927–31, 2018.

[85] N. G. Ward, "Midlevel prosodic features toolkit (2016-2023)." https://github.com/nigelg ward/midlevel, 2023.

[86] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings, International Conference on Multimedia*, pp. 1459–1462, 2010.

[87] S. Agrawal, A. Anastasopoulos, L. Bentivogli, O. Bojar, C. Borg, M. Carpuat, R. Cattoni, M. Cettolo, M. Chen, W. Chen, *et al.*, "Findings of the IWSLT 2023 evaluation campaign," in *Proceedings of the 20th International Conference on Spoken Language Translation*, pp. 1–61, 2023.

[88] ITU, "Method for the subjective assessment of intermediate quality level of audio systems," Tech. Rep. ITU-R BS.1524.3, International Telecommunication Union Radiocommunication Assembly, 2015.

[89] A. Conneau, A. Bapna, Y. Zhang, M. Ma, P. von Platen, A. Lozhkov, C. Cherry, Y. Jia, C. Rivera, M. Kale, *et al.*, "XTREME-S: Evaluating cross-lingual speech representations," in *Interspeech*, pp. 3248 – 3252, 2022.

[90] A. Mohamed, H. yi Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, *et al.*, "Self-supervised speech representation learning: A review," *arXiv preprint arXiv:2205.10643*, 2022.

[91] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *IEEE Workshop on Spoken Language Technology (SLT)*, 2022.

[92] H. Bunt, "The multifunctionality of utterances in interactive discourse," in *Multifunctionality in English* (Z. Yin and E. Vine, eds.), pp. 11–29, Routledge, 2022.

[93] M. Marge, C. Espy-Wilson, N. G. Ward, *et al.*, "Spoken language interaction with robots: Research issues and recommendations," *Computer Speech and Language*, vol. 71, 2022.

[94] X. Wang and K. Evanini, "Empirical evaluation of the communicative effectiveness of an automatic speech-to-speech translation system," in *SLaTE*, pp. 150–155, 2017.

[95] D. J. Liebling, M. Lahav, A. Evans, A. Donsbach, J. Holbrook, B. Smus, and L. Boran, "Unmet needs and opportunities for mobile translation AI," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[96] L. Nunes Vieira, C. O'Sullivan, X. Zhang, and M. O'Hagan, "Machine translation in society: Insights from UK users," *Language Resources and Evaluation*, pp. 1–22, 2022.

[97] S. Santy, K. Bali, M. Choudhury, S. Dandapat, T. Ganu, A. Shukla, J. Shah, and V. Seshadri, "Language translation as a socio-technical system: Case-studies of mixed-initiative interactions," in *ACM SIGCAS Conference on Computing and Sustainable Societies*, pp. 156–172, 2021.

[98] G.-A. Levow, S. Duncan, and E. T. King, "Cross-cultural investigation of prosody in verbal feedback in interactional rapport," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[99] C. De Pasquale, C. Cullen, and B. Vaughan, "An investigation of therapeutic rapport through prosody in brief psychodynamic psychotherapy," in *Interspeech*, pp. 3043–3047, 2019.

[100] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, and S. B. Gaigg, "Is voice a marker for autism spectrum disorder? a systematic review and meta-analysis," *Autism Research*, vol. 10, no. 3, pp. 384–407, 2017.

[101] N. G. Ward and J. E. Avlia, "A dimensional model of interaction style variation in spoken dialog," *Speech Communication*, vol. 149, pp. 47–62, 2023.

[102] N. G. Ward and S. Abu, "Action-coordinating prosody," in *Speech Prosody*, 2016.

[103] N. G. Ward, J. C. Carlson, and O. Fuentes, "Inferring stance in news broadcasts from prosodic feature configurations," *Computer Speech and Language*, vol. 50, pp. 85–104, 2018.

[104] N. G. Ward and A. Vega, "A bottom-up exploration of the dimensions of dialog state in spoken interaction," in *13th Annual SIGdial Meeting on Discourse and Dialogue*, 2012.

[105] N. G. Ward, "Ten prosodic patterns of turn-taking in Japanese conversation," in *Proc. 10th International Conference on Speech Prosody 2020*, pp. 764–768, 2020.

[106] N. G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes, "Turn-taking predictions across languages and genres using an LSTM recurrent neural network," in *IEEE Spoken Language Technology Conference*, pp. 831–837, 2018.

[107] D. Aguirre, N. G. Ward, J. E. Avila, and H. Lehnert-LeHouillier, "Comparison of models for detecting off-putting speaking styles," in *Interspeech*, 2022.

[108] N. G. Ward and J. A. Jodoin, "A prosodic configuration that conveys positive assessment in American English," in *International Congress of the Phonetic Sciences*, pp. 3368 – 3372, 2019.

[109] N. G. Ward, A. Kirkland, M. Włodarczak, and E. Székely, "Two pragmatic functions of breathy voice in American English conversation," in *11th International Conference on Speech Prosody*, pp. 82–86, 2022.

[110] N. G. Ward, J. E. Avila, and A. M. Alarcon, "Towards continuous estimation of dissatisfaction in spoken dialog," in *SigDial*, pp. 13–20, 2021.

[111] N. G. Ward, S. D. Werner, D. G. Novick, T. Kawahara, E. E. Shriberg, L.-P. Morency, and C. Oertel, "The similar segments in social speech task," in *MediaEval Workshop*, 2013.

[112] N. G. Ward, S. D. Werner, F. Garcia, and E. Sanchis, "A prosody-based vector-space model of dialog activity for information retrieval," *Speech Communication*, vol. 68, pp. 86–96, 2015.