Towards a Model of the Mapping Between English and Spanish Prosody

by Jonathan Avila

Motivation

- Speech-to-speech translation systems are already useful for short interactions but are less useful for conversations
- One reason for this is an inadequate translation of prosody – the stress, rhythm, and intonation of speech
- Prosody conveys many intents and stances

Example:
Yeah can have different interpretations based on its prosody

Research Objective

Improve the pragmatic fidelity of speech-to-speech translation for dialog by exploring cross-lingual prosody mappings

Outline

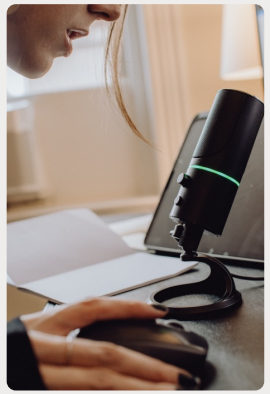
- Protocol & corpus
- Representation of utterance prosody
- English-Spanish prosodic relationships through correlations
- Metric for model evaluation
- Hypotheses & simple models
- Results, hypotheses validation & English-Spanish differences
- A better model

Outline

- Protocol & corpus
- Representation of utterance prosody
- English-Spanish prosodic relationships through correlations
- Metric for model evaluation
- Hypotheses & simple models
- Results, hypotheses validation & English-Spanish differences
- A better model

Multilingual Speech Corpora

- Machine learning approaches often heavily rely on data
- Prior multilingual speech corpora consist of **monologues** and/or **read**, **synthesized**, or **acted** speech, e.g.,
  - *CoVoST 2*
  - *mTEDx*
  - *Heroes*
- Consequently, such data lacks
  - Spontaneity
  - Nuances of interpersonal interactions
  - Pragmatic uses of prosody



The Dialogs Re-enacted Across Languages (DRAL) Protocol

- Bilingual participants engage in a 10-minute recorded conversation, mostly unscripted

- Under producer guidance, they listen to and re-enact utterances in their other language “with the same feeling”

The Dialogs Re-enacted Across Languages (DRAL) Corpus: Example

X: *You're going to have your own,*

Y: *Ah, that's right.*

X: *apartment.*

Y: *Already on Thursday.*

X: *On Thursday?*

Y: *On Thursday they're going to give it to me, on Thursday at three in the afternoon.*

X: *Are you parents gonna come, or?*

X: *Vas a tener tu propio,*

Y: *Ai, si cierto.*

X: *departamento.*

Y: *Ya el jueves.*

X: *¿El jueves?*

Y: *El jueves me lo van a dar, el jueves a las tres de la tarde.*

X: *¿Van a venir, venir tus papás para?*

The Dialogs Re-enacted Across Languages (DRAL) Corpus

- 3,816 pairs of English and Spanish utterances, each produced by the same speaker
- Topics include
  - Getting to know each other
  - Sharing personal experiences
  - Discussing hobbies and interests

- Protocol and corpus detailed in technical report UTEP-CS-23-27
- Accepted by the Linguistic Data Consortium

Outline

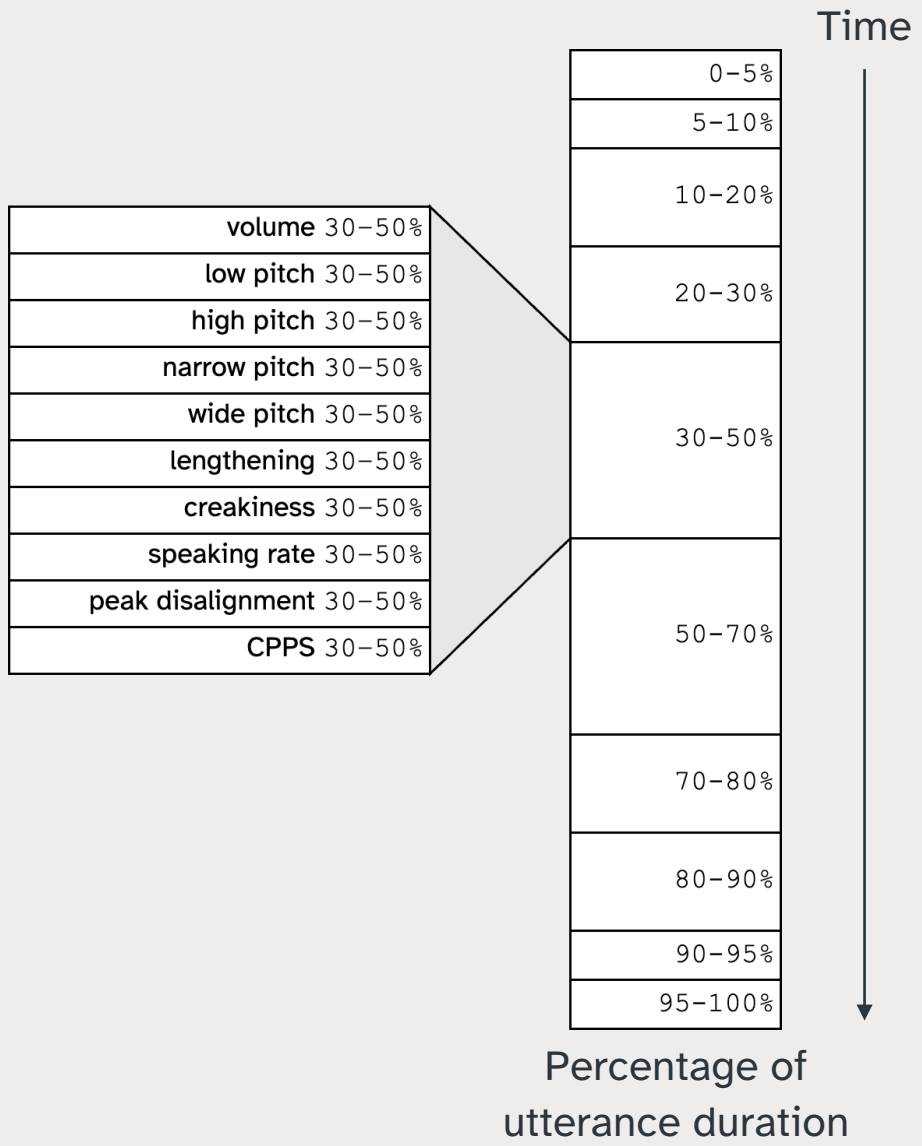
- Protocol & corpus
- Representation of utterance prosody
- English-Spanish prosodic relationships through correlations
- Metric for model evaluation
- Hypotheses & simple models
- Results, hypotheses validation & English-Spanish differences
- A better model

Speech Representations in General

- Encode the underlying factors of speech relevant to its application, chosen based on research objectives
  - Commonly used tools for feature extraction: *openSMILE*, *Kaldi*, *Praat*
- Gap: A representation with focus on prosody from non-read-speech

My Prosodic Feature Set

- Features are
  - robust for dialog data
  - generally perceptually relevant
  - normalized per speaker
- Ten base features computed over ten non-overlapping window proportional to utterance duration, spanning the utterance



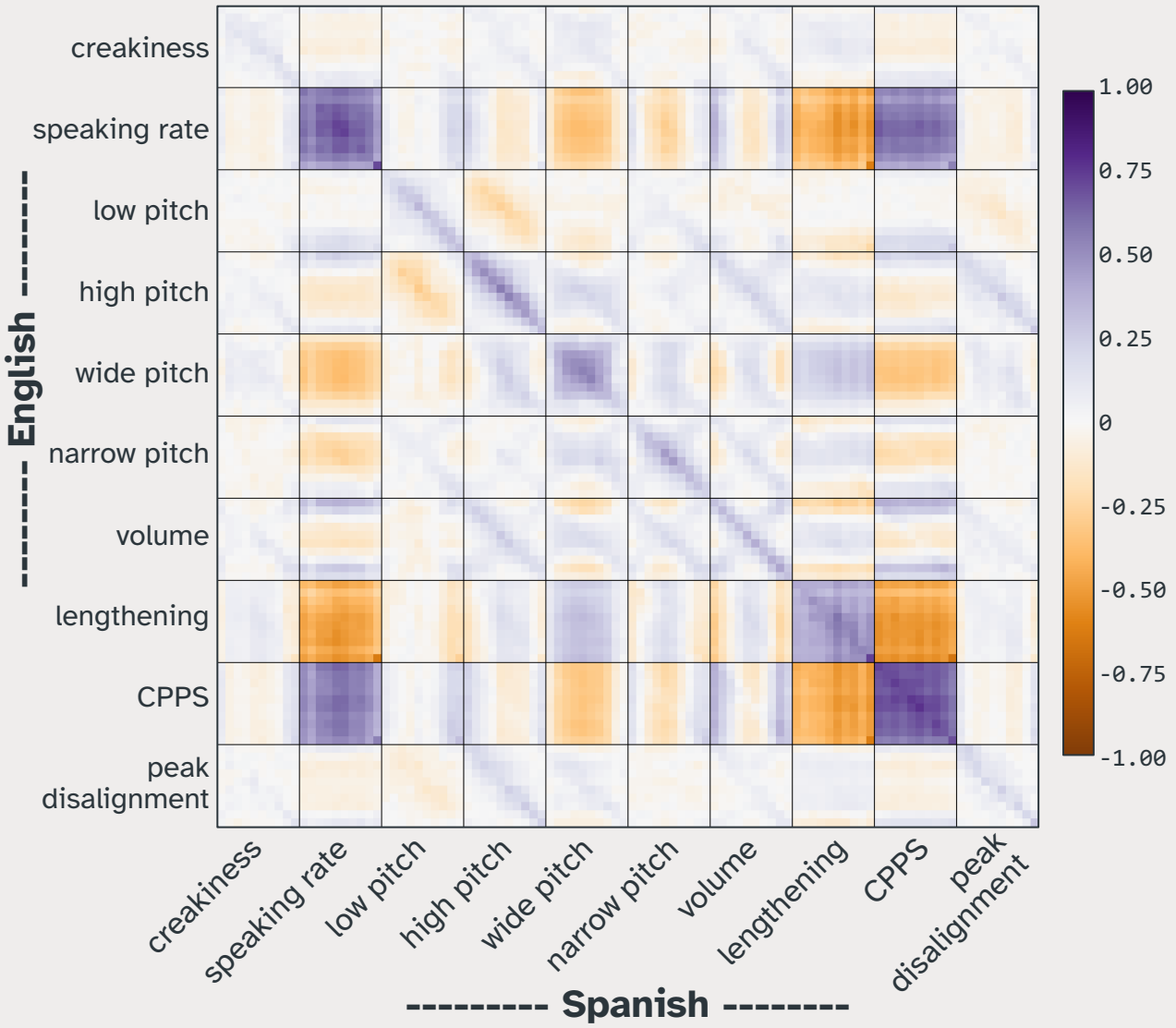
Outline

- Protocol & corpus
- Representation of utterance prosody
- English-Spanish prosodic relationships through correlations
- Metric for model evaluation
- Hypotheses & simple models
- Results, hypotheses validation & English-Spanish differences
- A better model

Prosodic Feature Correlations

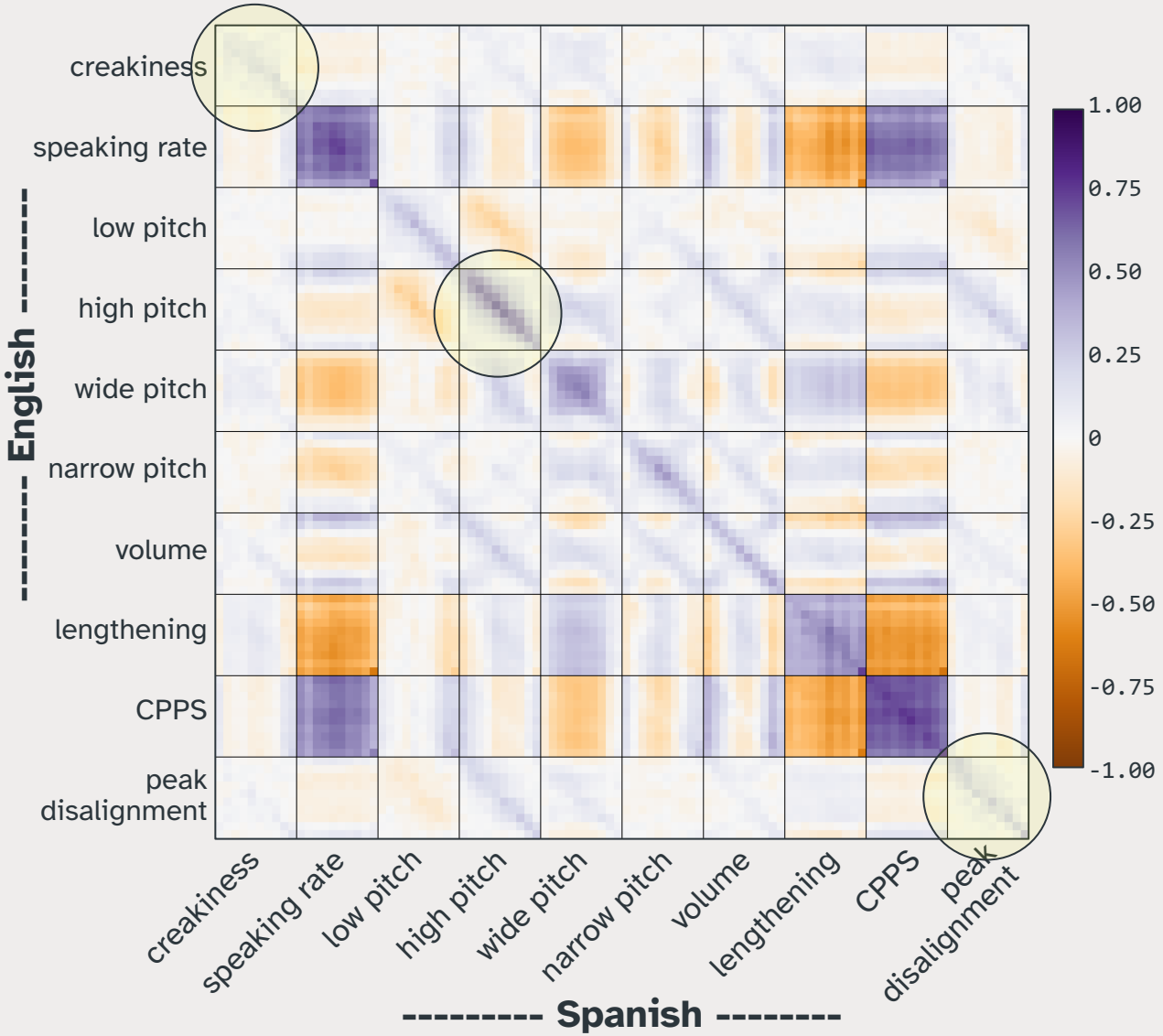
- **Purpose:** To gain a preliminary understanding of the relationship between English and Spanish prosody
- **Method:** Examined correlations between 100 prosodic features across all matched English and Spanish pairs in the DRAL corpus

English and Spanish Prosody: Similarities (1/2)



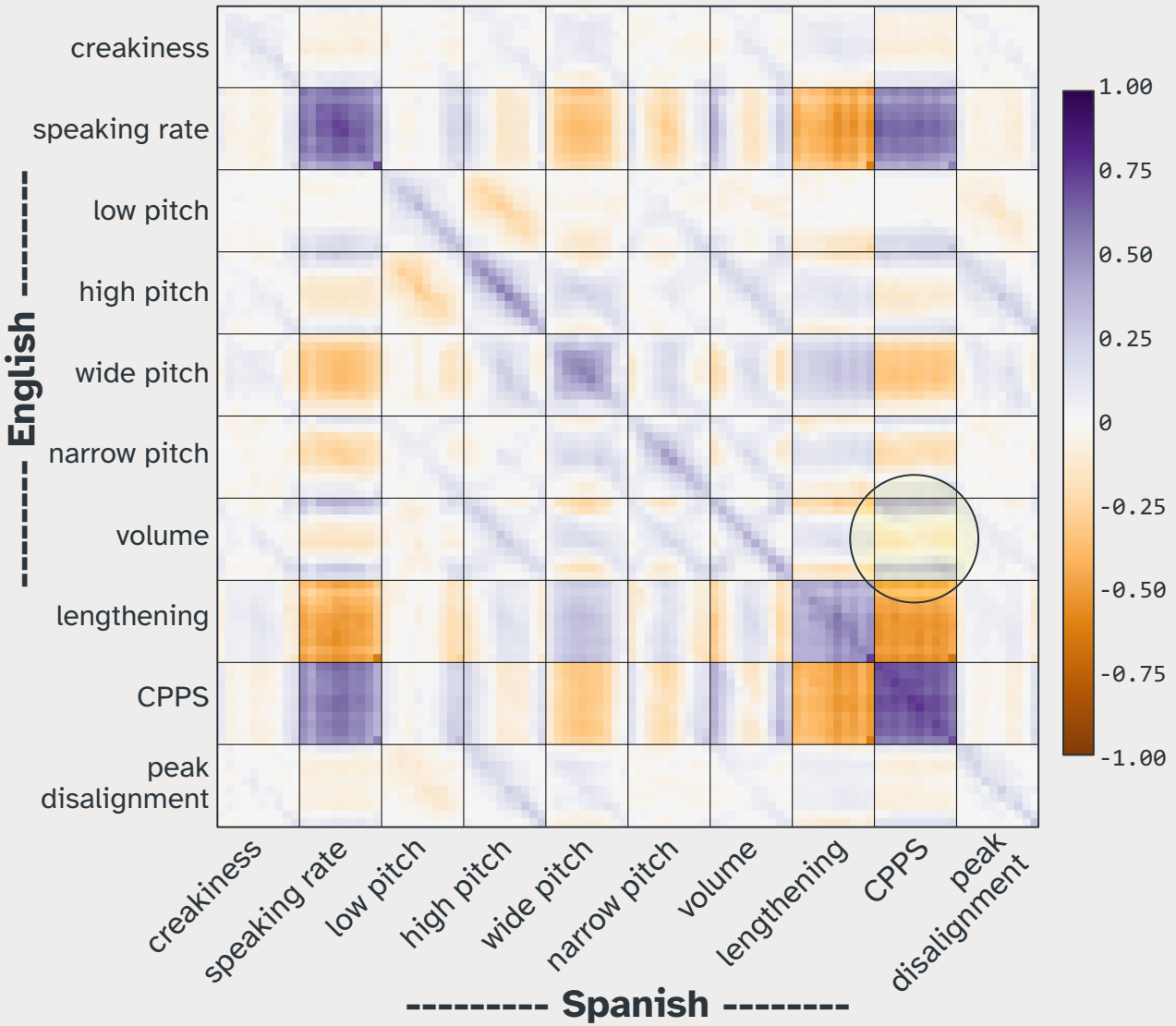
- Overall, English and Spanish prosody are similar
- Over half of the main-diagonal correlations are ≥ 0.3

English and Spanish Prosody: Similarities (2/2)



- Pitch highness is highly similar, particularly in the middle of utterances (e.g., 30-50%, $\rho = 0.56$)
- Creakiness and peak disalignment had the weakest cross-language correlations, suggesting different functions in the two languages

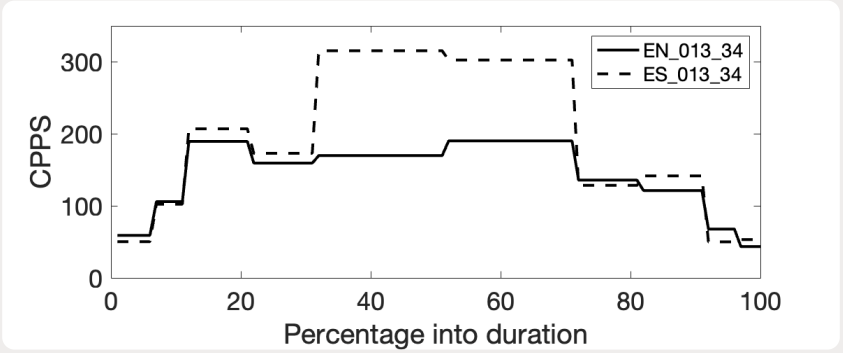
English and Spanish Prosody: Differences

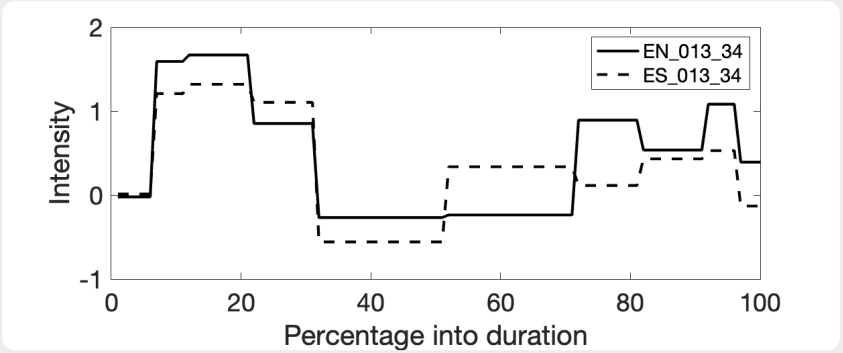


- Some off-diagonal correlations were expected (e.g., speaking rate and lengthening) but not all
- Example: English intensity and Spanish CPPS

English Intensity and Spanish CPPS: Observations

- Ten pairs strongly reflected this pattern: high English initial and final intensity and high Spanish CPPS (non-breathiness)
- In half of these pairs, the speaker was preparing for a follow-up utterance





Outline

- Protocol & corpus
- Representation of utterance prosody
- English-Spanish prosodic relationships through correlations
- Metric for model evaluation
- Hypotheses & simple models
- Results, hypotheses validation & English-Spanish differences
- A better model

Speech-to-Speech Translation Evaluation Metrics in General

- Current automatic evaluation metrics (e.g., *BLEU*, *COMET*, *BLASER)*
  - Rely on error-prone automatic transcriptions
  - Disregard prosody, or focus on a limited range of prosodic features
  - Estimate s*emantic* similarity, which is different from *pragmatic* similarity
- Use: Compare predicted target-language prosody with prosody of human-produced reference

My Metric for Prosodic Similarity Between Utterances

- Quantifies prosodic similarity between pairs of utterances
- Based on the inverse Euclidian distance of the utterance's prosody representations
- Values closer to zero indicate greater similarity

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

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

Where *d* and *q* are prosodic representations

Performance of Metric as a Proxy for Human Judgments

- The metric captures many aspects of pragmatic similarity, including:
  - speaker confidence
  - revisiting unpleasant experiences
  - describing sequences of events
- While some pairs shared lexical content, there was generally no correspondence between prosodic similarity and lexical similarity
- The metric performs better than chance in estimating the most similar and most dissimilar utterances (50 out of 56 estimates examined)

Outline

- Protocol & corpus
- Representation of utterance prosody
- English-Spanish prosodic relationships through correlations
- Metric for model evaluation
- Hypotheses & simple models
- Results, hypotheses validation & English-Spanish differences
- A better model

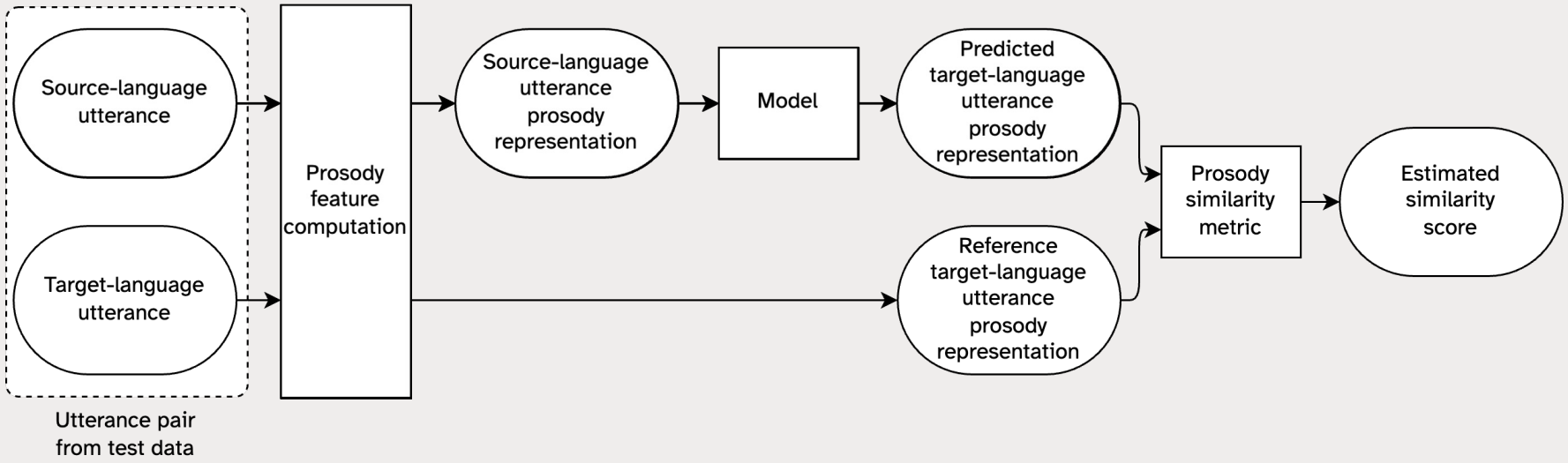
Hypotheses

Predicting the prosody representation of a target-language utterance from cross-language patterns will yield, on average, a higher similarity compared to predicting it

- as identical to that of the source-language utterance **(Hypothesis 1)**
- based solely on the lexical content of the source-language utterance **(Hypothesis 2)**

Prosody Translation Task Definition

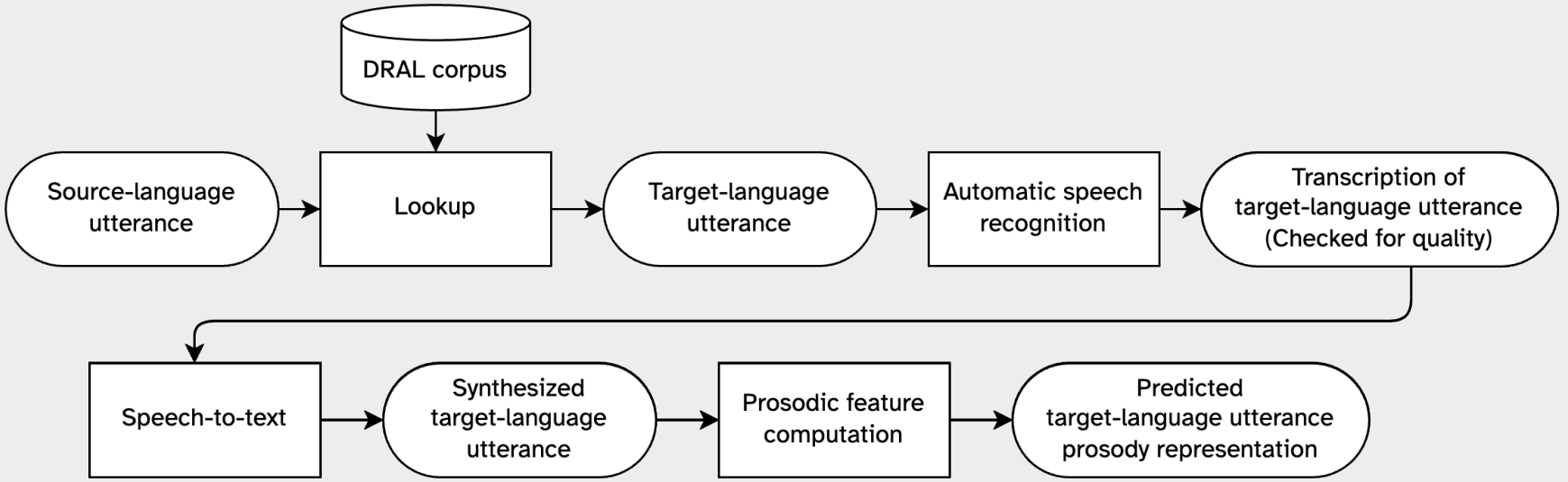
- Task: Predict the target-language prosody representation from that of a source-language utterance
- Evaluation: Average error, determined by its similarity with the prosody of a human reference utterance



Description of Models (1/3)

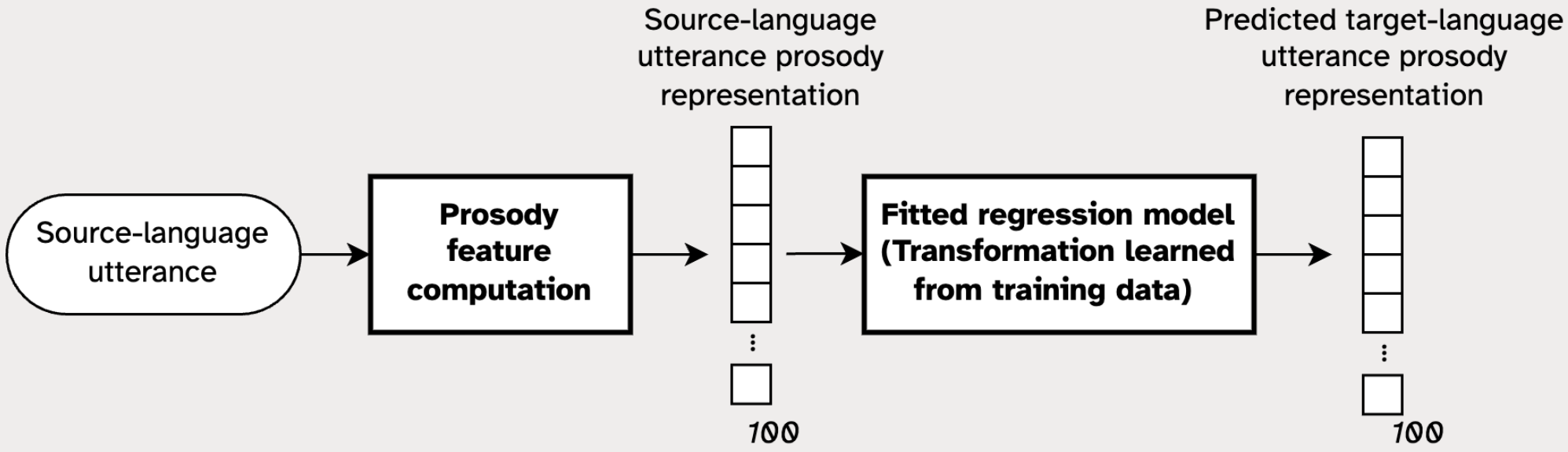
Direct-transfer baseline: Predicts target-language representation as identical to that of source-language utterance

Source-ignoring baseline: Predicts target-language representation based on content of source-language utterance



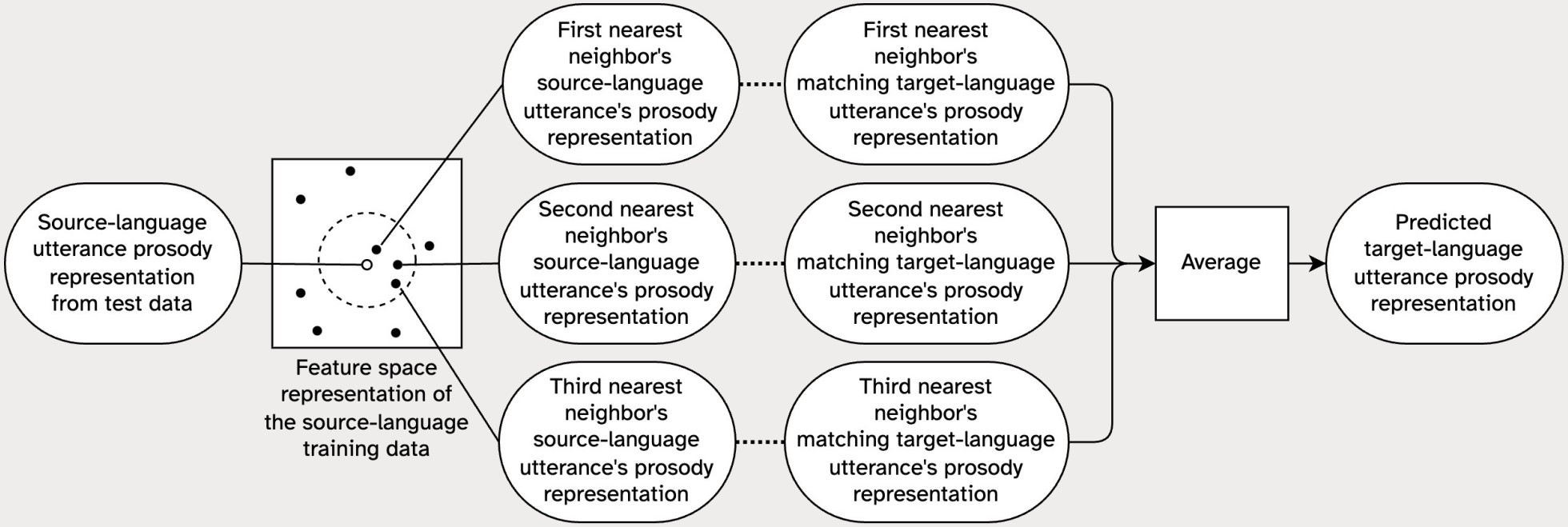
Description of Models (2/3)

Linear regression model: Parametric approach, mapping English and Spanish prosodic features as a linear function



Description of Models (3/3)

***k*-nearest neighbor regression model**: Local approach, predicts target-language representation based on proximity in feature space of source-language representation



Outline

- Protocol & corpus
- Representation of utterance prosody
- English-Spanish prosodic relationships through correlations
- Metric for model evaluation
- Hypotheses & simple models
- Results, hypotheses validation & English-Spanish differences
- A better model

Comparison of Model Performance

Table 6.2: Average error of prosody translation models.

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

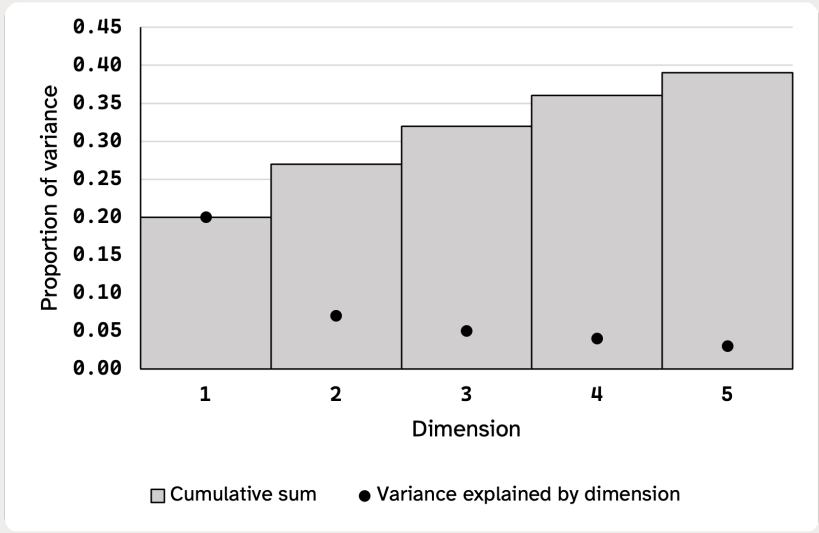
- The linear regression model outperformed both baseline models
- **Hypothesis 1 verified:** Modeling prosody as a linear relationship is a more effective strategy on average
- **Hypothesis 2 verified:** A simple model like linear regression can map certain aspects of prosody, so it can get better

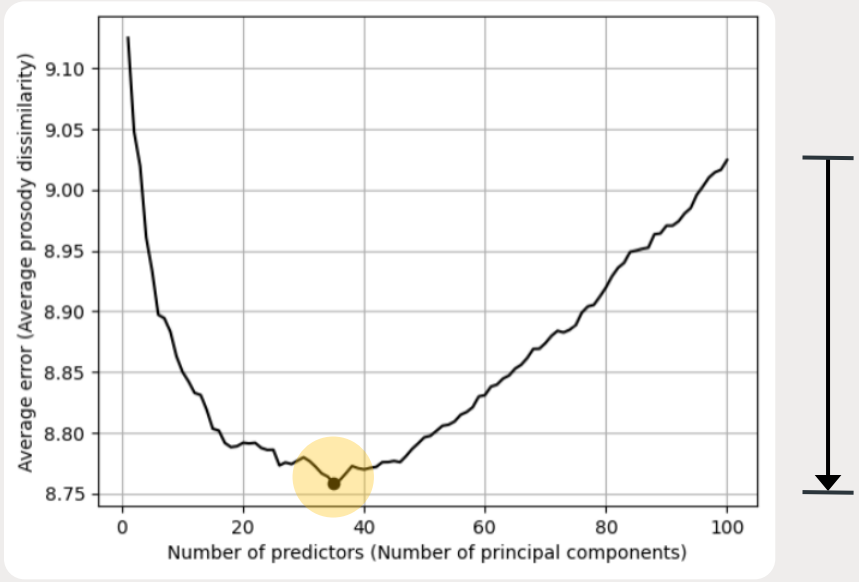
Outline

- Protocol & corpus
- Representation of utterance prosody
- English-Spanish prosodic relationships through correlations
- Metric for model evaluation
- Hypotheses & simple models
- Results, hypotheses validation & English-Spanish differences
- A better model

Improved Model: Linear Regression After Dimensionality Reduction

- Principal Component Analysis (PCA) for dimensionality reduction
- The first five principal components explain 39% of variance in both English and Spanish data
- Reducing the number of features has benefit compared to using the full feature set





Bonus: The Dimensions are Interpretable

Interpretation of principal components by examining the loadings and extremes for each dimensions

Pragmatic functions of English and Spanish dimensions

English	**Spanish**
1. Focus on speaker	1. Focus on speaker
2. Engaged/animated	2. Engaged/animated
3. Existence of shared understanding	3. Predictability
4. Intent to continue topic	4. Authority
5. Checking existence of shared understanding	5. Certainty

Failure Analysis: Baseline Models

- Source-ignoring baseline model’s errors included:
  - Failure to lengthen vowels or vary speaking rate during uncertainty
  - Failure to change pitch at turn ends
- Direct-transfer baseline model’s differences to reference included:
  - English utterances having more rising pitch endings
  - English being breathier in some areas
  - These differences may be due to English uptalk, which isn't common in in the Spanish data

Future Work

- Improvements, extensions to
  - Corpus
  - Representation of utterance prosody
  - Metric for similarity
  - Models for mapping cross-language prosody

Summary of Contributions

- A corpus with parallel utterances from dialog
- A representation of utterance prosody
- A metric for prosody-conveyed pragmatic similarity of utterances
- A reduced dimensionality representation of utterance prosody
- **Hypotheses verified, from analysis of cross-language mapping modeling strategies**
- **Findings on English and Spanish prosody**

Acknowledgments

- Professor Nigel Ward
- Emilia Rivas and Divette Marco
- Ann Lee, Benjamin Peloquin, and Justine Kao
- Meta, the National Science Foundation, and UTEP’s University Research Institute

Towards a Model of the Mapping Between English and Spanish Prosody

by Jonathan Avila