WHICH PROSODIC FEATURES MATTER MOST FOR PRAGMATICS?

Nigel G. Ward, Divette Marco, Olac Fuentes

Computer Science
University of Texas at El Paso
El Paso, Texas, 79912, USA

ABSTRACT

We investigate which prosodic features matter most in conveying pragmatic functions. We use the problem of predicting human perceptions of pragmatic similarity among utterance pairs to evaluate the utility of prosodic features of different types. We find evidence that the duration-related features are most important, that pitch-related features are much less important and less adequate, and that complete modeling will require additional acoustic and prosodic features, including nasality and phonetic reduction. These findings can guide future basic research in prosody, and suggest how to improve speech synthesis evaluation, among other applications.

Index Terms— speech synthesis evaluation, error metrics, pragmatic similarity, prosodic feature sets, feature-importance analysis, English, Spanish

1. MOTIVATION

We ask: What prosodic features matter most in the expression of pragmatic functions?

We choose to focus on pragmatic functions because of their importance in emerging scenarios for speech technology, such as dialog systems involving interpersonal sensitivity or deployed in situated robots [1]. Prosody is well-known to have important roles in conveying many pragmatic functions [2, 3, 4].

Our question may seem dated. Certainly it is now irrelevant for any short-term project for which there is adequate training data. In such cases, rather than agonize over which features to use, we can just use everything available, leaving it to the machine learning algorithm to exploit them effectively. This strategy is especially well-suited to the use of pretrained models, which can encode much prosodic information [5].

Nevertheless, attempting to answer this question could serve:

a) To support system development, when lacking sufficient training data to build a model from scratch. Developers can use knowledge of which prosodic features generally matter most, to build a starting-point architecture or model, which can then be refined.

b) To support prioritization of research directions and questions, for psycholinguists and others doing basic research; and to help applied researchers chose what to focus on when describing, for example, some understudied language, the nature of some communication disorder, or some aspect of sociolinguistic variation.

c) To support the design of control parameters to serve as "knobs" for human-in-the-loop specification or post-editing of the prosody of speech synthesizer output [6].

d) To enable better evaluation of the quality of the prosody output of a generative system, for practical needs, including:

d1) The evaluation of speech synthesizer output [7]. A quick survey of the papers on speech synthesis at Interspeech 2023 shows that, while most discuss prosody, the features considered were mostly pitch and duration, with only 10% mentioning intensity and only one voicing properties. Even when specifically targeting better expressivity, conversational style, and better prosody [8, 9, 10], designers of metrics, lacking true knowledge of what matters, may fall back on what's familiar, namely pitch and sometimes duration.

d2) Building better speech-to-speech translation systems, where there is increasing interest in faithfully conveying more than just the lexical content [11, 12]. Knowledge of which prosodic features matter most can support the design of better loss functions.

d3) Evaluating the power of discrete and other learned representations [13].

d4) The design of better speech codecs. Compression algorithms have been traditionally designed to maximize intelligibility and naturalness, but for use in interactive communication, preservation of certain prosodic features is also likely important.

d5) Designing interpretable feedback. This can be for people wishing to learn how to communicate more effectively in business or in relationships, or for special populations such as second language learners, those in rehabilitation after a stroke, and children with speech or language pathologies.

Over the decades, there have been numerous investigations of which prosodic features matter most for various specific purposes, mostly in classification and regression tasks. These include emotion recognition [14, 15, 16], classification of linguistic structures such as tones, boundaries, and accents [17, 18, 19], estimating judgments of language learners' accentedness, intelligibility, and other properties [20], prediction of turn-taking actions [21], language modeling [22], speaker identification [23], language identification [24], detecting clinical conditions [25], and speech synthesis [26]. Many interesting things have been found, and this body of work has greatly influenced the features included in prosodic-feature toolkits [15, 27]. However, none of this work has addressed pragmatics other than incidentally.

In this paper we use "pragmatics" in a broad sense, to include all aspects of interaction in dialog that go beyond the lexical semantic meaning. These are diverse: in dialog, people frequently show enthusiasm, make clarifications, cue action, clarify, criticize, praise, introduce a new topic, yield the turn to the other, and so on.

2. METHODS

Our interest is the prosody of pragmatic functions in general, not for any specific task. We accordingly chose to study prosody in the context of a general problem: estimating the perceived pragmatic similarity between pairs of utterances. (This can be seen as a generalization of attempts to model how humans perceive similarity for intonation [28, 29, 30] and similarity of expressivity [10]). Our premise is that a set of prosodic features that can support such estimates, across a wide variety of data, is likely to be useful for many applications involving pragmatic functions. Thus, we leverage this model-building problem to glean insight into what aspects of prosody matter.

2.1. Data

We exploit a dataset recently collected for another purpose [31]. This consists of pairs of utterances, each with an assessed pragmatic similarity value, on a continuous scale from 1 to 5, based on the average rating of 6 to 9 human judges. There are 458 pairs in American English, our main focus for this paper, and 235 Northern Mexican Spanish pairs.

Each utterance pair consists of a seed utterance extracted from a recorded dialog and a subsequent re-enactment of that utterance. Re-enactments were done under six conditions designed to create a variety of degrees of similarity, including two where the lexical content may differ from the seed and one with a synthesized voice. The dialogs were recorded with diverse scenarios to broaden the coverage of pragmatic functions [32], and within these, the seeds were also selected for diversity [31]. Thus the coverage is likely far broader than seen in any single-genre corpus.

In preliminary analysis we noted an interesting property of this data set: some feature distributions differ between the re-enactments and the seeds. In particular the re-enactments tend to have less variation in the pitch features, and to be louder and more creaky.

2.2. Features

We wanted a set of features that was broad in coverage, robust for dialog data, generally perceptually relevant, and simple, to enable easy interpretation of the results.

Specifically, we chose Avila's [33] adaptation of selected Midlevel Toolkit [34, 27] features to tile utterances. This set included 10 base features: intensity, lengthening, creakiness, speaking rate, peak disalignment (mostly late peak), cepstral peak prominence smoothed (CPPS), an inverse proxy for breathy voice, and four pitch features, namely measures of perceived pitch highness, pitch lowness, pitch wideness, and pitch narrowness. While this feature set is far from ideal, it is suitable for this exploration. Uniquely, it was designed to capture the prosody of pragmatic functions — unlike prosodic feature sets designed for paralinguistic properties, music, or general signal processing — and it was designed to be robust to microprosody and various phenomena of conversation. At the same time, it is flawed. Like other feature sets, none of its component features is simultaneously fully robust, fully corresponding to perception, and fully accurate. For example, the pitch features are not conditioned on sonority [35], CPPS correlates only roughly with perceptions of breathiness, and the speaking rate feature is based on spectral flux, and so is susceptible to diverse ancillary and confounding factors, including the presence of creaky voice. Nevertheless, when used for statistical and modeling purposes over sufficient data, the features in this set can be useful, as seen by their utility in numerous basic and applied studies [34, 36].

Each base feature is normalized per track to be roughly speaker-independent. We then use average values over each of ten non-overlapping windows spanning fixed percentages of its duration: 0–5%, 5–10%, 10–20%, 20–30%, 30–50%, and symmetrically out to 100%. This representation is not suited to syllable- or word-bound prosodic phenomena, but can roughly represent the sorts of overall levels, contours, and patterns that are often associated with pragmatic functions.

	Correlation
Euclidean Distance	–0.33
Linear Regression	0.44
KNN Regression	0.58
Random Forest Regression	0.70
cosine over selected HuBert	0.74

Table 1: Pearson's correlation between each models' predictions and the human judgments.

This set is simplistic, and in particular includes nothing relating to time-sequence modeling, notably no temporal deltas;,let alone functionals, but our working assumption is that we can still learn from it.

2.3. Models and Prediction Results

While our main aim in this paper is to analyze feature importance, modeling pragmatic similarity is a problem of importance in its own right [31, 37], for example, for use cases d1 and d2 above, so this subsection focuses on that perspective.

Our primary metrics for model quality are correlations between the systems' similarity estimates and the human judgments. We also computed MSE, and the results were consistent. Our primary train/test split was between judgments collected in Sessions 1 and 2, a month apart. We also did experiments using 10-fold cross-validation across all the data, and the results were similar.

The models used are as follows: Euclidean Distance is a re-implementation of [33]. In this all features are weighted equally, after z-normalization). For the next three models, the inputs were the 100 feature deltas, that is, the feature values for the seed minus the values for the reenactment. (Performance using instead the absolute differences was always somewhat lower, as one might expect for models accordingly blind to the seed-reenactment distinction.) For the KNN Regression Model, k was 50. For the Random Forest Regression there were 100 trees. The "selected HuBert" model uses the cosine similarity between feature representations consisting of 103 Hubert layer-24 features selected to maximize performance on the training data [37]. While none of these models is very sophisticated — lacking dynamic time warping or other alignment methods, average- or max-pooling, non-linear or configurational compositions of features, and so on — they may suffice for exploratory purposes.

The results are seen in Table 1. First, we see the usual trade-off between model simplicity/explainability and performance. More interestingly, we see that the best designed-features model, with random forest regression, is doing almost as well as the pretrained features model. This indicates that the penalty for using designed features is small; perhaps a more sophisticated decision model could close the gap. Further, examining the correlations separately for pairs which were lexically different and for pairs which were lexically-identical, performance on the former was near-random, but 0.80 for the latter, as good as with pretrained features [37].

3. FEATURE-IMPORTANCE ANALYSES

Given a set of features and a task, there are many ways to measure the importance of feature sets and subsets [18]. We investigated using three methods. First, we simply computed the Pearson's correlation between each of the 100 features and the target judgments. Second we examined how much each feature contributed to the performance

Feature	Importance	Correlation
speaking rate	43.7%	0.64
lengthening	20.9%	0.54
peak disalignment	7.8%	0.32
CPPS	5.9%	0.04
pitch highness	4.1%	0.12
pitch narrowness	4.0%	–0.06
pitch wideness	3.9%	0.00
creakiness	3.5%	0.14
pitch lowness	3.3%	0.13
intensity	3.1%	–0.03
4 pitch features	15.2%	0.16
all 10 features	100.0%	0.70

Table 2: Feature types, ordered by importance for the random forest regression model and also showing performance of a model using features of this type alone.

of the best model, using the impurity decrease for Random Forest Regression Models averaged across all folds. Third, we did subset and ablation studies, examining performance when including only, or when excluding only, various feature types.

The implications noted below are all multiply supported, so, to save space, we present only a selection of the evidence. However we note that there is no consistent ranking of features, as seen in Table 2. This is not surprising; rather, the existence of feature types with low correlations but relatively high importance indicates that the features are not independent, and instead, as often noted [34], specific configurations of features likely bear specific meanings. We also found that the features are highly redundant: ablating any specific type only slightly reduces the performance.

3.1. Most Important Feature Types

Table 2 shows results per feature type, and Figure 1 the correlations for five informative feature types. We draw three implications: 1) Duration features are important, with speaking rate the top feature by every measure. Interestingly, while lengthening is strongly anticorrelated with speaking rate, it still has some independent value, increasing the performance over speaking rate alone by 0.02 (correlation with human judgments). 2) The value of the pitch features is low. This was not entirely a surprise: the reasons that pitch is popular — being salient, easy to visualize, familiar from music, relatively easy to measure, and historically important — do not imply actual utility, and we suspected that the self-evident importance of pitch features for modeling read speech may not carry over to dialog. However the importance of the pitch features was surprisingly small. As this is the first study to actually measure the value of prosodic features for pragmatic functions, the last word is yet to be written, but we can conclude at least that pitch features do not deserve the exclusive respect that they often get. 3) The least informative features overall are intensity and pitch narrowness.

3.2. Most Important Feature Positions

We next examined evidence for which feature positions matter most. Figures 1 and 2 suggest that the prosody around 70–90% into utterances is relatively informative, especially for peak disalignment, pitch wideness, pitch highness, and lengthening. We suspect that this relates to the common occurrence of various types of pitch peak (such as nuclear accents) in this region in many utterances. Other interactions between feature type and position included: speaking rate being especially informative in the beginnings and middles of utterances, peak disalignment at the beginning, and the lengthening feature mostly toward the end. Incidentally the sharp drop in importance at the end may be an artefact of variation in where the labelers marked utterance ends, since in these dialogs the utterances often trailed off. Overall, the tendencies were weaker than we had expected; rather, it seems that informative prosody is widely distributed across utterances.

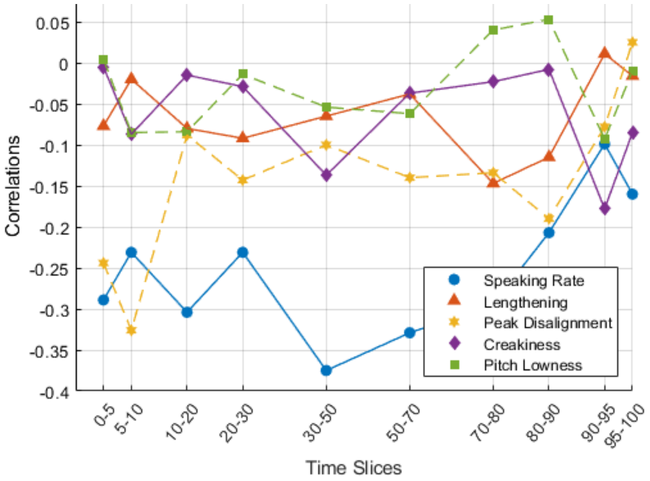

Fig. 1: Single-feature correlations between the judgments and the deltas for five of the most informative feature types, across both Session 1 and Session 2 data. The X-axis represents the regions, defined by fixed percentages of the utterance duration.

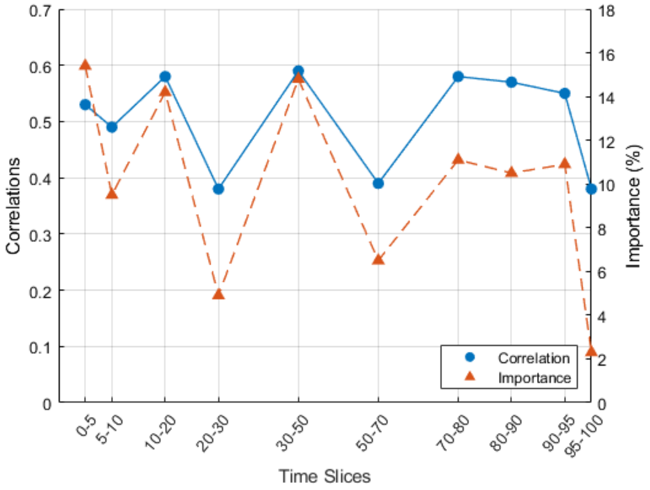

Fig. 2: Feature importance as a function of position (time slice): on the left axis, performance of a model using only features at that position; on the right axis, summed importance in the random forest regression model.

4. QUALITATIVE ANALYSES

Our first qualitative analysis was a brief exploration of why pitch-only model performed so much worse than the all-feature model. We examined a small sampling of pairs for which the predictions of the former were far more accurate than those of the latter. Several pragmatic functions were common in these pairs, mostly commonly positive/negative assessment, turn hold/yield, and correction of a misunderstanding. The most common prosodic-acoustic properties present in these pairs, which also seemed to be involved in conveying these meanings, included nasality and speaking rate variation. Saliently, all of these poorly-handled pairs had a synthesized-speech re-enactment. From this we infer that the synthesizer used, namely Amazon Polly, is not able to effectively control (or even much vary, it seems) many of the prosodic characteristics that are important to human perception. As far as we know, this may be true for all synthesizers, and we speculate that this is due in large part to the pitch-prioritizing loss functions that they are trained to.

Our second qualitative analysis explored the limitations of Avila's 100-feature set. Although designed to be widely inclusive, it did not quite support state-of-the-art performance, at least with the models tried. Again we did failure analysis, this time more thoroughly, examining the 30 pairs for which the predictions of our best model, the Random Forest Regression model, tested in 10-fold cross validation, had the highest divergences from human judgments.

First we examined pairs which the model rated much higher than did the judges, looking for differences that our ears could hear but that the model likely had missed. Almost all of these involved differences in nasality. Also common were differences in pause frequency, length, and location. Other factors we noticed include, in rough order of frequency, words said with or without laughing, the exact phonetic form of non-lexical utterances such as *oh*, phonetic reduction including devoicing, stressing of specific words, vibrato, falsetto, non-lexical sighs, uses of glottal stops, ejectives, and strong harmonicity. Speaking rate variations and breathiness were also common factors, even though the speaking rate and CPPS features were intended to cover for such perceptions.

Second, we examined pairs that the model rated much lower than the judges. One frequent factor was differences in pacing or pause placement that were not significant to our ears, but seemed to trip up the model. This suggests, unsurprisingly, that a model could do better with some kind of alignment or max-pooling. Another factor was apparent individual- or, often, gender-dependent variant prosodic forms for conveying the same meaning. For example, in *oh my god, it's working* (female) and *yo, it's working* (male), where, in addition to the lexical differences, the former used vibrato, breathy and falsetto voice, and the male creaky voice, both conveyed excitement and matched well in nuance. Thus, while most aspects of spoken English are amenable to gender-agnostic modeling, this suggests that this strategy may not work well for pragmatics-related prosody.

Third, we revisited these pairs and a sampling of pairs that were handled well, hoping to discover which pragmatic-function distinctions remain problematic, even with the full feature set. However, there were no clear patterns. We also found that all the functions identified as problematic for the pitch-only model were often handled well by the full model, and of course, the magnitude of the divergences was much less than for the pitch-only model.

Audio illustrating these points is available at https://www.cs.utep.edu/nigel/pros-prag/.

5. SPANISH

Wondering which of the findings above might apply beyond English, we repeated most of the analyses using the Spanish data from [31]. In brief, we found: 1) These features serve to predict pragmatic similarity fairly well for Spanish also (0.73 correlation, with 10-fold cross-validation). 2) The feature types with the highest correlations only partly overlapped those for English, with the top three being speaking rate, creakiness, and pitch wideness. 3) Modeling using pitch alone was again far inferior to using all features, but the penalty was less than for English (correlation of 0.41, with 10-fold cross-validation), 4) Utterance-final features were again the least informative, 5) Models trained on one language and tested on the other performed reasonably well (e.g. Spanish trained on English: correlation 0.68), but not as well as language-specific models. 6) Common features lacking from the model but important for human perception of differences again included nasality and devoicing. 7) Humans often discounted differences in creakiness, nasality, and breathiness, especially when these reflected different gender-dependent ways to convey the same pragmatic function.

6. SUMMARY AND LIMITATIONS

We have reported the results of the first systematic study of which prosodic features matter most for pragmatics. While we can, of course, provide no definitive answer to this question — the best feature set will always depend on the task, language, speaker population and so on — this exploration has contributed:

- a new method for evaluating the pragmatic adequacy of prosodic feature sets
- some explainable models for predicting human judgments of pragmatic similarity
- a new method for discovering important but lacking features
- indications of which features should be included in evaluation metrics for applications, notably not only pitch features but also duration-related and voicing features
- identification of pragmatic functions that are poorly handled by pitch-only feature sets, including making corrections, marking positive or negative feeling, and indicating turn hold/yield intentions
- identification of features that are understudied but important for pragmatics, and thereby deserving of further study, notably nasality, vibrato, and phonetic reduction

While we have broken new ground, we note that this study has numerous limitations, including the small data sizes, the simplicity of the features and modeling, the lack of coverage of all genres of dialog, and the focus on American English. Further work is needed.

Acknowledgments

This work was supported in part by National Science Foundation award 2348085 and by the AI Research Institutes program of the NSF and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 – National AI Institute for Exceptional Education. We also thank the anonymous reviewers.

7. REFERENCES

[1] Matthew Marge, Carol Espy-Wilson, et al., "Spoken language interaction with robots: Research issues and recommendations," *Computer Speech and Language*, vol. 71, 2022.

[2] Dagmar Barth-Weingarten, Nicole Dehé, and Anne Wichmann, *Where Prosody Meets Pragmatics*, Brill, 2009.

[3] Harm Lameris, Joakim Gustafsson, and Éva Székely, "Beyond style: Synthesizing speech with pragmatic functions," in *Interspeech*, 2023, pp. 3382–3386.

[4] Weiqin Li, Peiji Yang, et al., "Spontaneous style text-to-speech synthesis with controllable spontaneous behaviors based on language models," in *Interspeech*, 2024.

[5] Guan-Ting Lin, Chi-Luen Feng, et al., "On the utility of self-supervised models for prosody-related tasks," in *IEEE Works. on Spoken Language Technology (SLT)*, 2022, pp. 1104–1111.

[6] Dan Andrei Iliescu, Devang S Ram Mohan, Tian Huey Teh, and Zack Hodari, "Controllable prosody generation with partial inputs," in *IEEE ICASSP*, 2024, pp. 11916–11920.

[7] Petra Wagner, Jonas Beskow, et al., "Speech synthesis evaluation: State-of-the-art assessment and suggestion for a novel research program," in *Proceedings of the 10th Speech Synthesis Workshop (SSW10)*, 2019.

[8] Wen-Chin Huang, Benjamin Peloquin, et al., "A Holistic Cascade System, Benchmark, and Human Evaluation Protocol for Expressive Speech-to-Speech Translation," in *ICASSP*, 2023.

[9] Yayue Deng, Jinlong Xue, et al., "ConCSS: Contrastive-based context comprehension for dialogue-appropriate prosody in conversational speech synthesis," in *IEEE ICASSP*, 2024.

[10] Kevin Heffernan, Artyom Kozhevnikov, et al., "Aligning speech segments beyond pure semantics," in *Findings of the Assn. for Computational Linguistics*, 2024, pp. 3626–3635.

[11] Loïc Barrault, Yu-An Chung, et al., "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[12] Eliya Nachmani, Alon Levkovitch, et al., "Translatotron 3: Speech to speech translation with monolingual data," in *IEEE ICASSP*, 2024, pp. 10686–10690.

[13] Leyuan Qu, Taihao Li, et al., "Disentangling prosody representations with unsupervised speech reconstruction," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2023.

[14] Anton Batliner, Stefan Steidl, Bjorn Schuller, et al., "Whodunnit: Searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech and Language*, vol. 25, pp. 4–28, 2011.

[15] Florian Eyben, Klaus R. Scherer, 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.

[16] Bogdan Vlasenko, Sargam Vyas, et al., "Comparing data-driven and handcrafted features for dimensional emotion recognition," in *IEEE ICASSP*, 2024, pp. 11841–11845.

[17] Neville Ryant, Malcolm Slaney, et al., "Highly accurate Mandarin tone classification in the absence of pitch information," in *Proceedings of Speech Prosody*, 2014.

[18] Anton Batliner, Jan Buckow, et al., "Prosodic feature evaluation: Brute force or well designed," in *Proc. 14th Int. Congress of Phonetic Sciences*, 1999, vol. 3, pp. 2315–2318.

[19] Anton Batliner, Jan Buckow, et al., "Boiling down prosody for the classification of boundaries and accents in German and English," in *Eurospeech*, 2001, pp. 2781–2784.

[20] Eduardo Coutinho, Florian Hönig, et al., "Assessing the prosody of non-native speakers of English: Measures and feature sets," in *Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 1328–1332.

[21] Gabriel Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech & Language*, vol. 67, 2021.

[22] Nigel G. Ward, Alejandro Vega, and Timo Baumann, "Prosodic and temporal features for language modeling for dialog," *Speech Communication*, vol. 54, pp. 161–174, 2011.

[23] Luciana Ferrer, Nicolas Scheffer, and Elizabeth Shriberg, "A comparison of approaches for modeling prosodic features in speaker recognition," in *IEEE ICASSP*, 2010, pp. 4414–4417.

[24] Raymond W. M. Ng, Tan Lee, et al., "Analysis and selection of prosodic features for language identification," in *IEEE Int'l. Conf. on Asian Language Processing*, 2009, pp. 123–128.

[25] Ethan Weed and Riccardo Fusaroli, "Acoustic measures of prosody in right-hemisphere damage: A systematic review and meta-analysis," *Journal of Speech, Language, and Hearing Research*, vol. 63, no. 6, pp. 1762–1775, 2020.

[26] Ivan Bulyko, M Ostendorf, and P Price, "On the relative importance of different prosodic factors for improving speech synthesis," in *Proceedings of ICPhs*, 1999, vol. 99, pp. 81–84.

[27] Nigel G. Ward, "Midlevel prosodic features toolkit (2016-2023)," https://github.com/nigelgward/midlevel, 2023.

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

[29] Uwe D. Reichel, Felicitas Kleber, and Raphael Winkelmann, "Modelling similarity perception of intonation," in *Interspeech*, 2009, pp. 1711–1714.

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

[31] Nigel G. Ward and Divette Marco, "A collection of pragmatic-similarity judgments over spoken dialog utterances," in *Linguistic Resources and Evaluation Conference*, 2024.

[32] Nigel G. Ward, Jonathan E. Avila, Emilia Rivas, and Divette 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.

[33] Jonathan E. Avila and Nigel G. Ward, "Towards cross-language prosody transfer for dialog," in *Interspeech*, 2023.

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

[35] Jonathan Barnes, Alejna Brugos, Nanette Veilleux, and Stefanie Shattuck-Hufnagel, "Segmental influences on the perception of high pitch accent scaling in American English," *Language and Speech*, 2024.

[36] Nigel G. Ward, Ambika Kirkland, et al., "Two pragmatic functions of breathy voice in American English conversation," in *11th Conference on Speech Prosody*, 2022, pp. 82–86.

[37] Nigel G. Ward, Andres Segura, Alejandro Ceballos, and Divette Marco, "Towards a general-purpose model of perceived pragmatic similarity," in *Interspeech*, 2024.