Towards a General-Purpose Model of Perceived Pragmatic Similarity

Nigel G. Ward[1], *Andres Segura*[1], *Alejandro Ceballos*[1], *Divette Marco*[1]

[1]University of Texas at El Paso, USA

nigelward@acm.org, andressegura915@gmail.com, aceballos4@icloud.com, divettemarco@outlook.com

Abstract

Models for estimating the similarity between two utterances are fundamental in speech technology. While fairly good automatic measures exist for semantic similarity, *pragmatic* similarity has not been previously explored. Using a new collection of thousands of human judgments of the pragmatic similarity between utterance pairs, we train and evaluate various predictive models. The best performing model, which uses 103 features selected from HuBert's 24th layer, correlates on average 0.74 with human judges for the highest-quality data subset, and it sometimes approaches human inter-annotator agreement. We also find evidence for some degree of generality across languages.

Index Terms: dialogue, utterance-level perceptions, English, Spanish, prosody, HuBert

1. Introduction

Pragmatics, the aspects of language use in which people convey information beyond the semantic content, is becoming more important for computational purposes, as applications increasingly target more natural spoken dialog and more embodied use cases.

Models of similarity underlie much of speech technology: in their guise as loss functions for training; as error measures for system performance evaluation; for analysis, as in clustering; and as system components, for example in nearest-neighbor-based classifiers. While many useful models of lexical, semantic, and prosodic similarity have been developed, modeling pragmatic similarity is a new and different challenge.

This paper contributes: 1) an overview of some needs that a model of pragmatic similarity could serve, 2) a new way to use features from a self-supervised learning (SSL) model for a downstream task, 3) a simple HuBert-based model that closely predicts human judgments, and 4) the finding that this model can have value even without re-tuning to a specific language.

2. Applications and Related Work

This section overviews how a model of pragmatic similarity could support progress in three areas of speech technology.

The first area is speech synthesis. Of the three main contributors to progress in this area — models, data, and loss functions — weaknesses in the latter may be the limiting factor [1]. Of course, there are ways to estimate perceived intonational similarity [2, 3, 4, 5, 6], and, more generally, prosodic similarity [7, 8, 9, 10], as surveyed in [11], and these may be adequate for read-style speech, where the task of the synthesizer may only be to convert a string of text to an intelligible sound. However it is increasingly noted that speech synthesizers whose output is prosodically neutral and pragmatically uninformative are not adequate for many other use cases, such as human-robot interaction [12, 13]. While cross-entropy can sometimes work well, in cases discrete-unit representations are sufficient [14], in general a continuous pragmatically-sensitive loss function can most directly support the broadening of synthesizer utility [15, 1].

One recent topic of interest is synthesis for speech-to-speech translation, where support for conversational uses will need elements of the source-language pragmatics to be faithfully conveyed in the target-language output [8, 16, 17, 18]. For this purpose it is clearly not adequate to just augment a synthesizer with options for a small finite set of emotions or of speaking styles. A good pragmatic-similarity metric could help these systems learn more fine-grained control, by enabling training to minimize the pragmatic gap between system output and human-generated reference translations. This use case inspired the two most relevant previous efforts: Barrault *et al.* [18] trained a model, AutoPCP, for estimating overall similarity, but unfortunately this has been evaluated only for its utility in systems-level comparisons, and not for its ability to model pragmatics specifically, much less for individual judgments of similarity. Avila and Ward [17] proposed a model using a Euclidean distance metric over 100 features designed to capture the main prosodic indications of pragmatic functions, but its performance was never properly evaluated. Both of these efforts were incidental to larger projects, and neither has yet, as far as we can determine, resulted in usable code, so for current purposes they are just sources of inspiration.

The second area is assessment of human speech and dialog behavior. This can be done to rate or to help people learning a new language, or to diagnose communication disorders or monitor the progress of those seeking to overcome them. For many people in many situations, the pragmatic aspects of language behavior may matter most [19], but existing assessments highlight only phonetic, lexical, syntactic or semantic aspects. Several use cases could be supported by a pragmatic similarity estimator, for example, the automatic comparison of a subject's dialog behavior samples to those of an exemplar speaker, or set of reference speakers, such that, if the behavior of the subject is pragmatically dissimilar, then correction or referral for intervention may be appropriate. Today building such assessment tools requires task-specific labeled training data [20], such as a set of matched typical/autistic utterances. However, a good pragmatic-similarity metric could avoid that need.

The third area is retrieval-based dialog systems. These rely on estimators for semantic similarity [21, 22], of which many exist [23, 11]. However, semantically-appropriate utterances are not always pragmatically appropriate [24], especially for speech. For example, the word *okay* can have very different functions, depending on the prosody. While constraining retrieval in various ad hoc ways can help, a general pragmatic-similarity metric could be a simpler solution.

Thus a model of pragmatic similarity could be widely useful. In psychological modeling, similarity has been noted to be "one of the most important relations humans perceive" [25] — as it underlies many aspects of learning, classification, and generalization — and the perception of similarities of various types have been well-studied. However there appear to have been no previous studies specifically of pragmatic similarity perceptions, or, indeed, much work at all on the perceptual space of pragmatic functions.

3. Data and Task

We started by collecting human judgments of the pragmatic similarity between pairs of utterances, as described in [11]. A total of 689 utterance pairs were each judged by 6 to 9 judges, on a scale from 1 to 5. The utterance pairs were crafted to be more similar than would be accomplished by random sampling. Four-fifths were based on original utterances from unstructured conversations among students, and the rest from task-oriented dialogs among students plus a handful of toddler utterances. Judgments were obtained in three four-hour sessions: the first with 220 American English pairs and 9 judges, the second with 234 American English pairs and 9 judges, and the third with 235 Mexican/American-border Spanish pairs and 6 judges.

To highlight some important properties of this dataset: First, the judgments are task-agnostic: judges were asked only to indicate "How pragmatically similar are the two clips, in terms of the overall feeling, tone, and intent?" Second, the judgments likely reflect unbiased perceptions, as the judges were not taught any theory or taxonomy of pragmatic functions. Third, the judgments are high-quality — thanks doubtless to the judges being hand-picked for having demonstrated sensitivity to and adeptness with the nuances of the languages, and being well-compensated and constantly supervised — with good inter-annotator agreement, at least after the initial session. Fourth, this data includes both stimulus pairs with different lexical content and pairs whose lexical content is the same, and thus differ mostly in the prosody, in a ratio of about 1 to 2. Having pairs of both types is helpful because, for many use cases, such as those discussed above, the ideal similarity metric would perform well regardless of whether the two utterances have identical word sequences.

Our modeling task is to predict the human similarity judgments from the audio for both utterances in the stimulus pair: thus, to build a Pragmatic Similarity Estimator. Our primary quality metric is the correlation between model predictions and human judgments. This is computed in two ways. First, as the main performance metric, we use the correlation between the model predictions and the average of the human judgments (Method M1a). Second, to enable direct comparison with human inter-annotator agreement, we also report the average of the correlations between the model and each judge (Method M1b). In addition we sometimes report the mean absolute error between model predictions (rescaled to best match the human judgments) and human judgments (Method M2). We also report some qualitative analyses.

We report results separately for each session. As will be seen, some of the results for the first and third sessions use information gleaned across all judgments in those sessions. However all results for the second session as test data are pristine, coming from models whose hyperparameter selection and training process was blind to that session's data.

models	correlation		
	Eng.1	Eng.2	Spa.
WavLM + cosine	.12	.17	.06
Wav2Vec2.0 + cosine	.31	.41	.24
HuBert + cosine	.45	.41	.40
selected HuBert + cosine	.69	.74	.53

Table 1: *Correlation with Human Judges' Averages (M1a)*

4. Models and Results

4.1. Modeling Approach and a Basic Model

Our strategy is to reduce each utterance to a set of features and estimate pragmatic similarity from the two sets of features. As our aim is a general-purpose model, we wanted to avoid overfitting in either the features or the model. Accordingly, in this subsection we use generic features, and a very simple model, namely cosine similarity (which we found to outperform Euclidean distance). We chose to try features taken from SSL models pretrained on generic prediction tasks [26, 27, 28]. While these are pretrained on audiobook data, and thus may not be expected to represent pragmatic information well, previous research has shown that SSL features in fact support many prosody- and pragmatics-related tasks [29].

Because SSL models produce a firehose of features, up to 1024 per layer and per 10-millisecond frame, we reduced these in two ways: for every experiment we used features from only one layer, and we average-pooled the features across time. Average-pooling is certainly not the most sophisticated way to exploit SSL features [29], and it may, on the one hand, risk discarding temporal information, but, on the other hand, the transformer layers may have already paid adequate attention to any temporally-localized informative features. Empirically, HuBert was better than Wav2Vec2.0 and WavLM, and the HuBert Large features outperformed the HuBert Small. For each SSL model, many layers did almost equally well. Table 1 reports the best results for each SSL model; for HuBert Large this was Layer 24, the last layer. For Session 2, the correlation was 0.41, and the correlation was significant at $p < 1e-10$. The runtime on a laptop was about 70 milliseconds for HuBert to compute features per second of audio, around 50 milliseconds each for the pooling operations, unoptimized, and 0.02 milliseconds for the cosine. Thus, for example, the total time to estimate the similarity between two 3-second clips was just over a half second.

To understand the weaknesses of this model we qualitatively examined a sampling of the utterance pairs for which its predictions were farthest from the average of the human judges. We found that there were many utterance pairs which this model estimated to be much more similar than the judges thought: including cases where both clips were from children, where both clips had similar background noise levels, or where a synthesized voice happened to closely match the original utterance in its pacing. Thus this basic model is often sensitive to factors not directly relevant to pragmatic similarity.

4.2. A Selected-Features Model

Since HuBert features are known to generically support many tasks, we hypothesized that feature selection could arrive at a subset more useful for modeling pragmatic similarity specifically. Again favoring simplicity, we chose to continue using the cosine, just with fewer features. As far as we know, feature se-

lection has not previously been used as a way to exploit SSL speech features.

With 1024 features to consider, considerations of speed led us to avoid standard methods and create our own feature-selection algorithm: We split the 1024 features into subsets of 10, and within each 10 did exhaustive search over all pairs of features to find the pair which gave best performance, evaluated, as always, in terms of the correlation between the cosine values and the human judgments. This gave us 102 candidate feature pairs, of which we retained the 50 pairs (100 features) which performed best. This procedure was fast: 160 seconds on a laptop for all the Session 1 data.

We evaluated this method by 10-fold cross validation on the Session 1 data, giving the result seen for "selected Hubert + cosine" in the first column of Table 1. (We experimented with various hyperparameters for this procedure, but there were only slight differences in performance, as long as the number of features selected was around 100.) Because the features found for each fold varied, we also developed a way to find a stable set: we selected those features that were retained in at least half of the folds. There were 103 such features, and we used this set to predict the Session 2 and Spanish data, giving the results seen in columns 2 and 3 of Table 1. Evaluating on the data from Session 2 using Method M2, the performance with these 103 features was better than with all 1024 features, with the mean absolute errors being 0.53 and 0.61, respectively, and the difference being significant by a matched-pairs one-sided t-test ($p < .0001$). Interestingly, feature selection improved performance on the lexically-identical pairs but hurt performance on the pairs with different lexical content. the table (table 3?)

As alternatives to feature selection, we also tried feeding the pairwise feature deltas into a single-layer network, and feeding the concatenated features of each pair into a single-layer network. Performance was much worse with these methods, whether we started with the 1024 features or a selected subset. Such side explorations aside, the good results across the three test conditions indicate that this approach can produce models with good generality.

Attempting to understand the weaknesses of this model, we examined a sampling of pairs for which its predictions were worst. The problems noted for the basic model were not seen, and we did not see any general patterns of error.

To better understand the power of this model, especially its ability to discriminate very similar versus moderately similar pairs, we randomly picked 8 seed utterances from the first session, and listened to the utterances which were most similar by the metric, for each. Apart from toddler utterances, in each case there was, perceptually, a strong similarity. These similarities were of different kinds, including: being a suggestion, being intended to persuade but also hesitant, sharing a certain strange pattern of pausing, and expressing mixed feelings about something. This suggests that this similarity model is performing well across different regions of the space of utterances. It also suggests that the model is capturing many dimensions of pragmatic similarity, including aspects not covered by [30] or, as far as we know, any other taxonomy of pragmatic functions.

5. Comparisons

5.1. Comparisons to Acoustic-Prosodic Similarity Models

While there are no previous models of pragmatic similarity, we can compare our results here to those of various classic models of acoustic and prosodic similarity, commonly used in evaluation of speech synthesis, as those may be capturing some of the same information.

	lexically		
	same	different	all
cepstral distance	.30	-.06	.24
F_0 DTW	.12	.12	.11
mel-cepstral DTW	.31	.03	.23
mel-cepstral independent DTWs	.31	.00	.24
duration	.05	.11	.05
BertSimilarity	–	.50	–
HuBert + cosine	.47	.33	.41
selected HuBert + cosine	.80	.20	.74

Table 2: *Correlations with Human Judges' Averages (M1a), for Session 2 and its subsets*

To provide some details on the models evaluated: Cepstral distance was computed using Sternkopf's implementation of Kubichek's Mel-Cepstral Distance Measure. Dynamic Time Warping (DTW) was done using F_0 computed using librosa [31] and Meert's DTW code [32]. To obtain everywhere-defined pitch, we patched regions where none was detected by using the most recent detected pitch value. We did no speaker normalization. The few utterances with no pitch detected at all were simply excluded when computing the correlations. Mel-Cepstral DTW was done 13 mel-cepstral features, using librosa's implementation [33]. We did this in two ways. First, in the normal way, we found the single best alignment for all features, using FastDTW [34]. Second, we found the best time alignment for each of the 13 features independently, using Meert's DTW, and then took the average of the distances for the 13. In addition, having noticed that judges seldom rated pairs as similar if they were greatly different in length, we built a trivial predictor that used the absolute difference in duration between the utterances in the pair. We note that all these models have no free parameters, being based on theories of what people will perceive as similar, and thus were used without training.

As seen in Table 2, these classic models had only modest success on the Session 2 data, especially for the lexically-different pairs. This was also true for the other sessions. This suggests that the HuBert features well-capture the pragmatically-relevant prosodic information, a result that aligns with the findings of [29]. We do not know specifically where the benefit comes from. It could come from gathering more prosodic information than just intonation features, from better normalizing, from better modeling temporal patterns, or some combination of these and other factors.

5.2. Comparison to a Word-based Similarity Estimate

We next tried a word-based similarity model. While the working assumption of this paper is that pragmatic similarity is different from semantic similarity, they are not unrelated. We investigated using a high-performing semantic similarity model, BertSimilarity [35].

As seen in Table 2, BertSimilarity does quite well for the lexically-distinct subset, greatly outperforming our model. However, it can, of course, provide no information for the lexically-identical pairs. From the table we also see that the HuBert-based models do relatively poorly for lexically-different pairs, perhaps in part because the training data (the Session 1 data) was mostly lexically-identical pairs.

	Eng.1	Eng.2	Spa.
models			
Wav2Vec2.0 + cosine	.26	.36	.20
HuBert model + cosine	.32	.36	.33
selected HuBert + cosine	.50	.64	.45
humans			
worst human	.29	.68	.62
average human	.45	.72	.66
best human	.53	.78	.70

Table 3: *Average of Correlations with Human Judges (M1b)*

To understand the relative merits of BertSimilarity and our model, we examined a few pairs where its estimates and our model's most differed. An illustrative case was the pair: *your payment has been processed* and *the payment went through*. BertSimilarity rated these two only modestly similar, but our model matched with the human judges in rating them very similar. Listening to the audio, we attribute this to the prosody of both conveying, in addition to the content, a business-like stance, confidence, reassurance, and topic closure. This suggests that our model might usefully complement a word-based similarity model for some purposes.

5.3. Comparison to Humans

To compare the models' performance to human performance, we looked at how well they correlate with each human judge (Evaluation Method M1b). This enables direct comparison to the performance of the human judges, unlike the values seen in Table 1, where the goal is to predict the average of the judges (M1a), an easier task because averaging removes hard-to-model noise. Thus, Table 3 shows the average of the correlations between the model and each human, and, for comparison, the average of the correlations between all pairs of human judges. As seen in the table, it outperformed the average judge in Session 1, but did relatively less well for the subsequent sessions, in which the judges were more experienced and agreed more [11].

Our best model did better for the lexically-identical stimulus pairs, but so did the human judges. Comparing for the lexically-identical Session 2 data, using Method M1b, the average of the correlations with human judges for the best model was 0.72, not far behind the average human judge's average correlation, 0.80.

6. Performance Across Languages and Genres

To explore the generality of the method and models across languages, we measured the performance across the three sessions of data of three models: the one with all features, the one with feature selection done using the English Session 1 data, and one with feature selection done on the Spanish data. Incidentally, of the 101 features in the latter model, only some two dozen overlapped the 103 found for English. The results, Table 4, indicate that tuning even on different-language similarity judgments is helpful, but also that top performance requires language-specific modeling.

We also explored whether the best model also has value for a different genre. We used the Switchboard corpus, since, unlike most of our test data, its speakers are strangers, mostly middle-aged, have an East Texas accent, and talk over the telephone. For a random sampling of 10 utterances as seeds, we examined the utterances that our model found as most similar to each. In 9 of the 10 cases the most-similar utterances were from a different speaker, and 9 of the 10 were perceptually, to us, highly similar. Many types of similarity were evident, many surprisingly specific, including reminiscing about something amusing, trying to evocatively describe a neighborhood scene, closing out a topic by summarizing, discussing negative experiences with telephone calls, attempting to persuade through logical argument, clarifying a point made in a previous statement, explaining something technical, and disparaging unselective behavior. While there were a few lexical similarities and occasional topic similarities, most similarities were pragmatics-related.

	Eng.1	Eng.2	Spa.
original HuBert	.45	.41	.40
English-tuned HuBert	.69	.74	.53
Spanish-tuned HuBert	.59	.63	.72

Table 4: *Cross-language Modeling: Correlations with the human judges averages (M1a)*

7. Summary, Implications, and Future Work

This paper has reported the first work on building models able to estimate the perceived pragmatic similarity between utterance pairs. We obtained good performance with a surprisingly simple model, which uses the cosine similarity between selected time-averaged HuBert features for the two utterances. This model outperformed existing prosodic and acoustic similarity measures, works where semantic similarity provides no information, and can come close to human performance. These things are true for both English and Spanish.

A major limitation of this work is that we were unable to properly examine the generality of the models for different genres of speech. Doing so will require new data resources. Future work should also explore how to build explainable models, for exampl by using meaningful and interpretable prosodic features, and also explore whether we can improve performance and robustness, including by conditioning the similarity estimates on additional factors, such as dialog activity, partner behavior, local context, and, for robot applications, the physical environment.

The prospects are bright for utility for the use cases mentioned in Section 2. Even though this was just an initial exploration, our models may already be useful. In particular, to expand on two aspects: For reference-based evaluation of the pragmatic abilities of speech synthesizers, since our model approaches the performance of even selected and supervised human judges, it is likely to far outperform the usual method of using crowdworkers. For the assessment of human speech, we have recently obtained promising results using this model for data-light classification of autistic versus neurotypical adolescents. We make our code available at `https://github.com/andysegura89/Pragmatic_Similarity_ISG` to support further investigations and applications.

8. Acknowledgments

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

9. References

[1] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander *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.

[2] 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)*, 2008, pp. 63–68.

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

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

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

[6] 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.

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

[8] 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.

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

[10] 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.

[11] N. G. Ward and D. Marco, "A collection of pragmatic-similarity judgments over spoken dialog utterances," in *Linguistic Resources and Evaluation Conference (LREC-COLING)*, 2024.

[12] 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.

[13] J. Miniota, S. Wang, J. Beskow, J. Gustafson, É. Székely, and A. Pereiral, "Hi robot, it's not what you say, it's how you say it," in *32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2023, pp. 307–314.

[14] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed *et al.*, "Generative spoken dialogue language modeling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.

[15] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors Affecting the Evaluation of Synthetic Speech in Context," in *Proc. 11th ISCA Speech Synthesis Workshop*, 2021, pp. 148–153.

[16] 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.

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

[18] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[19] K. Bottema-Beutel, "Glimpses into the blind spot: Social interaction and autism," *Journal of communication disorders*, vol. 68, pp. 24–34, 2017.

[20] T. Linzen, "How can we accelerate progress towards human-like linguistic generalization?" in *58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*. Association for Computational Linguistics (ACL), 2020, pp. 5210–5217.

[21] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *arXiv preprint arXiv:1603.08023*, 2016.

[22] C. Tao, J. Feng, R. Yan, W. Wu, and D. Jiang, "A survey on response selection for retrieval-based dialogues." in *IJCAI*, 2021, pp. 4619–4626.

[23] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity: A survey," *ACM Computing Surveys*, vol. 54, pp. 1–37, 2021.

[24] L. Pragst, "On the generation of pragmatic paraphrases for dialogue systems," Ph.D. dissertation, Ulm University, 2022.

[25] R. Richie and S. Bhatia, "Similarity judgment within and across categories: A comprehensive model comparison," *Cognitive Science*, vol. 45, no. 8, p. e13030, 2021.

[26] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in neural information processing systems*, 2020, pp. 12 449–12 460.

[28] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[29] 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, pp. 1104–1111.

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

[31] https://librosa.org/doc/main/generated/librosa.pyin.html.

[32] W. Meert, "dtaidistance.dtw," 2022, https://dtaidistance.readthedocs.io/en/latest/modules/dtw.html.

[33] https://librosa.org/doc/main/generated/librosa.feature.mfcc.html.

[34] S. Tanitter, "Implementation of Salvador and Chan's FastDTW," https://github.com/slaypni/fastdtw.

[35] A. Majumder, "A BERT embedding library for sentence semantic similarity measurement," 2020, https://github.com/abhilash1910/BERTSimilarity.