Interspeech 2024
1-5 September 2024, Kos, Greece

Pragmatically Similar Utterance Finder Demonstration

Nigel G. Ward[1], *Andres Segura*[1]

[1]University of Texas at El Paso, USA

nigelward@acm.org, andressegura915@gmail.com

Abstract

Models for estimating the similarity between two utterances are fundamental in speech technology. While fairly good automatic measures exist for semantic similarity, we only recently built a model of *pragmatic* similarity, the first. We propose to present this model by letting participants try out our Pragmatically Similar Utterance Finder. This system listens to one side of a live conversation, identifies the utterances, and for each retrieves the most similar utterances, according to our model, from a large corpus. Participants and viewers will then be able to hear these utterances, and judge for themselves the prospects for pragmatic-similarity modeling.

Index Terms: dialogue, perceptions, prosody, interactive demo

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.

We therefore built the world's first Pragmatic Similarity Estimator, able to predict the pragmatic similarity, as it would be judged by a human, given the audio for any two utterances.

While this is a core technology, not yet embodied any system, this demonstration exposes it so that Interspeech participants can appreciate the problem, judge the quality of the solution, and perhaps see how it relates to their own research or to application needs.

2. Model

The model that underlies the demo is fully described in our Interspeech 2024 submission, Towards a General-Purpose Model of Perceived Pragmatic Similarity.

In brief, this model was trained to approximate the human judgments in our PragSim corpus [1], which contains thousands of human judgments of the pragmatic similarity between utterance pairs. After much experimentation, we created a high-performing model. This uses 103 features selected from HuBert's 24th layer, and simply computes the cosine between the feature representations for each utterance. This correlates on average 0.74 with human judges for the highest-quality data subset, and sometimes approaches human inter-annotator agreement. It outperforms Mel-Cepstral Dynamic Time Warping and F_0 Dynamic Time Warping.

3. Demo System

The demo is interactive. We will ask for a volunteer to wear a microphone and have a short conversation with anyone for a minute or less. As they do, the system extracts his or her utterances. For each utterances, it finds utterances in the corpus that it computes as being very pragmatically similar, slightly less similar, etc. It then exposes these in an interface, Figure 1, where the volunteer, or the experimenter, can click any original utterance and the ones retrieved for various degrees of similarity.

As they listen, we expect they will discuss whether they think the utterances are in fact similar, and, when they are, in what ways they are similar.

The interface also offers settings for the corpus, Switchboard, DRAL, or NMSU [2, 3, 4], and if the former, to chose whether to limit retrieval to only male or only female utterances. The number of utterances varies from about 1500 to over 5000. The features for the corpus utterances are precomputed, so that the system runs and populates the interface in real time.

The model is currently released, at https://github.com/andysegura89/Pragmatic_Similarity_ISG, and the demo code will be released before the conference.

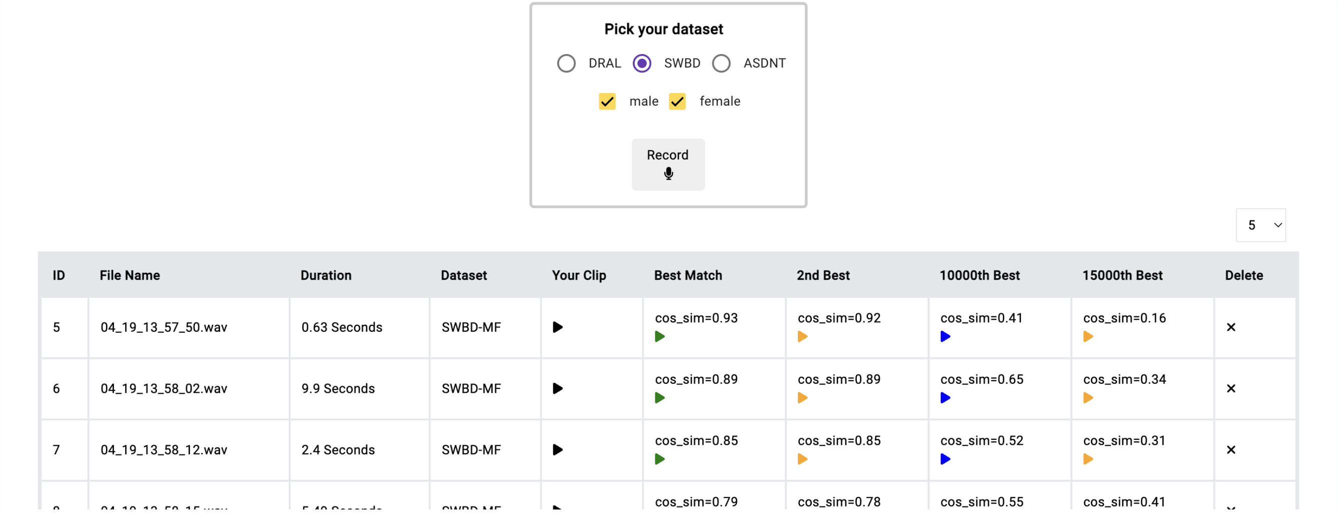
In terms of implementation, the demo adds the minimal functionality to the model that is needed to give users a clear understanding. Figure 2 shows how it works.

4. Logistics

The demo only requires a good-quality microphone, a laptop, and speakers. We will bring them all, plus a poster to explain how to use it and how it was built.

5. Expected Value for Interspeech Participants

The demo is not a product, but a proof of concept for a general-purpose module that could be built into many system, or used in a training process. We expect the demo to catalyze discussions about our model and how to improve it, and more generally, about the nature of pragmatic similarity, and how such models may be useful. We will not attempt to control such discussion, but may note the potential utility for 1) speech synthesis, where a pragmatically-sensitive loss function could support training and evaluation to support the widening of synthesizer utility for dialog applications [5, 6], 2) speech-to-speech translation, where support for conversational uses will need elements



Pick your dataset

○ DRAL ● SWBD ○ ASDNT

☑ male ☑ female

Record

5

ID	File Name	Duration	Dataset	Your Clip	Best Match	2nd Best	10000th Best	15000th Best	Delete
5	04_19_13_57_50.wav	0.63 Seconds	SWBD-MF	▶	cos_sim=0.93 ▶	cos_sim=0.92 ▶	cos_sim=0.41 ▶	cos_sim=0.16 ▶	×
6	04_19_13_58_02.wav	9.9 Seconds	SWBD-MF	▶	cos_sim=0.89 ▶	cos_sim=0.89 ▶	cos_sim=0.65 ▶	cos_sim=0.34 ▶	×
7	04_19_13_58_12.wav	2.4 Seconds	SWBD-MF	▶	cos_sim=0.85 ▶	cos_sim=0.85 ▶	cos_sim=0.52 ▶	cos_sim=0.31 ▶	×
8	04_19_13_58_15.wav	5.40 Seconds	SWBD-MF	▶	cos_sim=0.79	cos_sim=0.78	cos_sim=0.55	cos_sim=0.41	×

Figure 1: *Demo Screenshot*

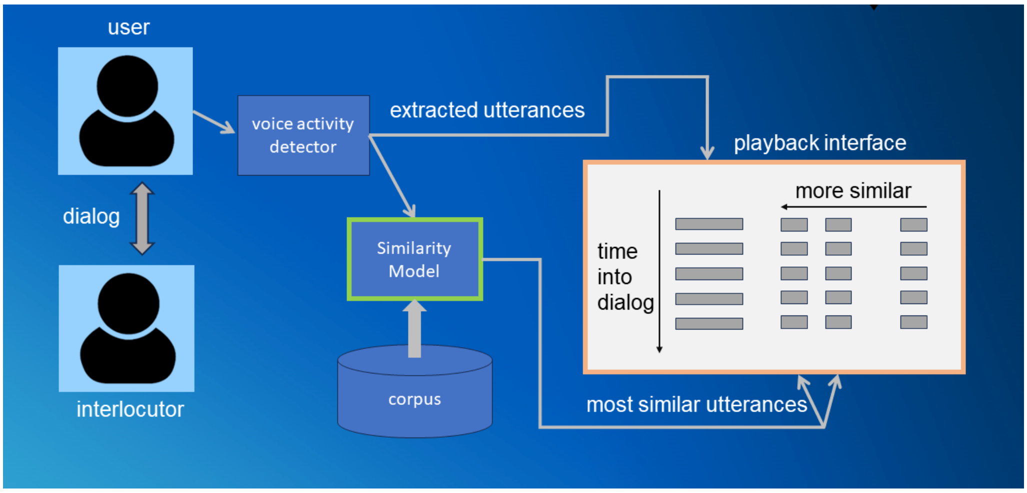


Figure 2: *Demo Components*

of the source-language pragmatics to be faithfully conveyed in the target-language output [7, 8], 3) assessment of human speech and dialog behavior, for example by k-Nearest Neighbors classification when only small training data is available, and 4) retrieval-based dialog systems.

6. Acknowledgments

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. AI Institute for support.

7. References

[1] 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.

[2] J. J. Godfrey, E. C. Holliman, and J. McDaniel, “Switchboard: Telephone speech corpus for research and development,” in *Proceedings of ICASSP*, 1992, pp. 517–520.

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

[4] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, “Prosodic entrainment in conversations of verbal children and teens on the autism spectrum,” *Frontiers in Psychology*, vol. 11, p. 2718, 2020.

[5] 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.

[6] 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.

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

[8] 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.