David Nguyen
dxn180015
Dr. Karen Mazidi
CS 4395.001

HW4 - ACL Paper Summary

The title of the ACL Paper that I will be summarizing is “Do self-supervised speech models develop human-like perception biases?” by Juliette Millet and Ewan Dunbar. Whose affiliations are with the CoML, ENS/CNRS/EHESS/INRIA/PSL, LLF, University of Paris, CNRS, CRI, FAN, IIFR and CoML, ENS/CNRS/EHESS/INRIA/PSL, University of Toronto respectively.

Many types of machine learning methods have a problem of over dependence on quality labeled data for training. This is a bottleneck in the training process as it is impractical to collect and label all sorts of varied data. Self-supervised learning aims to solve this problem by reducing the dependency on labeled data where the model will train itself to learn one part of the input from another part of the input (Shah). This method turns an unsupervised problem into a supervised problem by auto-generating the labels (Shah). This machine learning method is increasingly becoming popular as it may eliminate manual labeling. How do babies learn how to talk while growing up? As babies listen to others talk, they develop a specialized perception of sounds based on the listeners’ native languages. They will learn to discriminate sounds from their native language, but they will lose their ability to discriminate against non-native sounds. So often native speakers in one language may have trouble distinguishing certain sounds that are used in other languages. The problem addressed by the authors of this paper is to determine whether the same thing (developing a specialized perception bias of sounds) happens in self-supervised models. They want to see how models’ “representational spaces compare to the

perceptual spaces of human listeners, as inferred from behaviour on phone discrimination experiments" (Millet and Dunbar).

The authors said they were not the "first to compare speech models' representational space with human[s]" (Millet and Dunbar). They noted multiple prior works done by other researchers like 'Schatz,' "demonstrating that all three models preserve and enhance linguistically relevant speech sound contrasts in the language they are trained on" (Millet and Dunbar). As they listed previous work by other researchers, they explained what those researchers did and found such as how some speech models were better at better or more sensitive at speech/language discrimination. For example, the authors showed how "Schatz et al. (2021) showed, for example, that a simple self-supervised speech model reproduces the reduced sensitivity to the English [r]/[l] contrast when trained on Japanese speech recordings" (Millet and Dunbar). The authors also used their Perceptimatic database in the article and noted how it is one of the "many multiple datasets containing human behavioral data [that] have been collected and openly released to encourage comparison of models with humans" (Millet and Dunbar). They then explained how they are building on others work and how their article will inform others on "the kind of information speech models learn" and how their model comparisons to humans "can have a broader impact on our knowledge of how human[s] perceive speech, and how they learn to do so" (Millet and Dunbar).

The unique contribution(s) of this paper overall is recording quantitatively how well self-supervised models perform while also comparing them to humans and supervised learning models trained on speech and non-speech data. The authors found that "Self-supervised models trained on speech recordings are better than models trained on acoustic scenes" at speech discrimnation and "good at predicting human discrimination behaviour at the stimuli level" but

are “worse than neutral acoustic features” (Millet and Dunbar). They also found that contrary to prior results, the “supervised reference system is quite good at predicting human discrimination behaviour” and predicts “a native language effect” (Millet and Dunbar). Self-supervised models performed the same as or better “than the supervised reference and human listeners” (Millet and Dunbar). By training the models on both speech and acoustic scenes, they were able to “shown that training on speech data is essential to obtaining a human-like perceptual space” and that the “the benefits of self-supervised speech models comes from learning characteristics of human speech” not because the models are just better “general audio features” (Millet and Dunbar). They also learn that these models don’t typically learn about speech like how humans perceive it. They are not typically “not language-specific” as Wav2vec 2.0 and HuBERT showed little of the native language effect (Millet and Dunbar). Those models can be seen as modeling a “language-neutral or universal speech perception space” (Millet and Dunbar).

The authors of the paper evaluated their work by utilizing the “Perceptimatic benchmark datasets” which is “a collection of experimental speech perception data intended to facilitate comparison with machine representations of speech” (Millet and Dunbar). They then train 3 self-supervised models: wav2vec 2.0, HuBERT, and a CPC model in English and French as a way to “study speech models’ perception biases and compare them with humans” (Millet and Dunbar). The authors also compared the performance of these models to a supervised ASR model, DeepSpeech trained on the same data but with phonemic labels (Millet and Dunbar). To see how much a model’s representational space is impacted by the properties of speech, they also trained the models on acoustic scenes (non-human sounds) (Millet and Dunbar). The method they used to measure the discrimination was based on the “Human ABX test” which is used for human speech perception, in which “participants hear three speech extracts A, B, X” and the

subjects guess if A or B sounds closest to X. They measure the contrast ("correct (target)" p1 and the "distractor (other)" p2). They tested the models in the same way where they used a formula to get the Δ-value. To compare, they used "two metrics: the log-likelihood of a binary regression model on the experimental responses, and the Spearman's correlation between the average of the model's Δ-values and participants' accuracies averaged within each phone contrast" (Millet and Dunbar). They can now measure the predictions with granularity: "the discriminability of individual experimental items and the overall discriminability of pairs of phones" (Millet and Dunbar).

Millet was cited 98 times and Dunbar was cited 957 times based on Google Scholar. Generally the lead author is a PhD student while the advisor is the last author and this tracks true with this paper. Millet has far fewer citations compared to Dunbar, who has the most citations. Their work in this paper was important because they test how different types of models perform at sound/speech discrimination (when trained on both speech and acoustic senses data) and compare them to humans which is crucial to learn the strength and weaknesses of models and thinking of how to use that information to improve machine learning methods in regards to speech perception. They did the experiments and were then able to discuss a "different approach to developing more human-like representational spaces in self-supervised models" (Millet and Dunbar).

Works Cited

Millet, Juliette, and Ewan Dunbar. “Do Self-Supervised Speech Models Develop Human-like Perception Biases?” Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, https://doi.org/10.18653/v1/2022.acl-long.523.

Shah, Deval. “Self-Supervised Learning and Its Applications.” Neptune.ai, 24 Jan. 2023, https://neptune.ai/blog/self-supervised-learning.