ASR Systems as Models of Phonetic Category Perception

Thomas Schatz[1] **(thomas.schatz@laposte.net)**
Francis Bach[2] **(francis.bach@ens.fr)**
Emmanuel Dupoux[1] **(emmanuel.dupoux@gmail.com)**
[1]LSCP (ENS/EHESS/CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, France
[2]SIERRA Project-Team (ENS/INRIA/CNRS), Département d'Informatique, Ecole Normale Supérieure, PSL Research University, France

Abstract

Recent advances in machine learning raised the prospect of developing quantitative models of human perception that can handle realistic sensory input. In this paper, we focus on the question of phonetic category perception, i.e. the way we perceive basic speech sounds (roughly consonants and vowels), which is largely determined by the language(s) to which we were exposed as a child. For example, native speakers of Japanese have a hard time discriminating between American English /r/ and /l/, a phonetic contrast that has no equivalent in Japanese. We show that typical GMM-HMM Automatic Speech Recognition (ASR) systems trained on large corpora of continuous speech correctly predict several perceptual effects observed in humans. Our work illustrates the value of considering large-scale machine learning systems in the context of modeling human perception.

Keywords: perception; phonetic categories; modeling; ASR

Humans are connected to the external world through sensory interfaces such as the retina, the cochlea, etc., which produce complex high-dimensional representations of the external world. Progress in machine learning lead to the development of artificial systems that can process such complex sensory inputs and yield a performance that rivals that of humans in specific tasks. Can these systems be used as quantitative models of human perception? Existing theories of perception remain largely descriptive, and quantitative models are needed to move toward truly predictive approaches. In this paper, we focus on speech-processing systems and phonetic category perception. The prevalent view is that, as infants, we learn our language's phonetic categories and, as adults, we perceive foreign phonetic contrasts through mapping these sounds onto our native phonetic categories. Existing theories, however, do not specify the mappings in question and thus cannot make predictions. In this paper, we use ASR systems as explicit models of the mappings from foreign sounds to native categories. We train an ASR system in a "native" language and then present it with speech in a "foreign" language, which the system transcribes in terms of a probability distribution over the phonetic inventory of the "native" language (phone-level posteriorgrams). We specifically test if our models can account for three well-documented effects in phonetic category perception. First, sounds from our native language are globally easier to process than non-native sounds (Gottfried, 1984); second, people with different native languages do not confuse the same foreign sounds (Strange, 1995); third, Japanese native speakers have difficulties in perceiving the distinction between American English (AE) /r/ and /l/ (Miyawaki et al., 1975).

Models

The same Gaussian-Mixture based Hidden-Markov-Model (GMM-HMM) architecture is trained independently on 5 different corpora of continuous speech, yielding 5 distinct ASR systems. Input features consists of 13 MFCC coefficients plus 3 pitch-related features and their delta and delta-deltas coefficients. Properties of the corpora for each system are reported in Table 1. Each corpus is randomly divided into a training and test set with the same number of speakers.

Table 1: Language, total duration, speech register and number of speakers for each corpus.

Corpus	Language	Time	Register	Spk
WSJ	AE	143h	Read	338
BUC	AE	19h	Spontaneous	40
CSJ	Japanese	15h	Spontaneous	75
GPM	Mandarin	30h	Read	132
GPV	Vietnamese	20h	Read	129

Model Evaluation

To quantify how easy it is to distinguish two phonetic categories based on representations produced by one of our models, we use a machine version of the ABX discrimination task. The basic idea is to take two acoustic realizations A and X from one of the phonetic categories and one acoustic realization B from the other category and to test whether the model representation for X is closer to the model representation for A than to the model representation for B. The probability for this to be false for A, B and X randomly chosen in a corpus is defined as the *ABX error rate* for the two phonetic categories according to the model. If it is equal to 0, the two categories are perfectly discriminated. If it is equal to $.5$, discrimination is at chance level. See (Schatz, 2016) for more details.

The model representation of a vowel or consonant is obtained as a sequence of phone-level posteriorgrams taken every 10 ms (i.e. vectors containing the probability that the portion of signal considered correspond to each of the possible phonetic categories). As a control, we also use directly the input features common to all systems as a model representation. To quantify how close two model representations are to

each other, we use Dynamic Time Warping (DTW) based on an underlying Kullback-Leibler divergence for posteriorgrams and an underlying cosine distance for input features.

Native vs Non-Native Contrasts

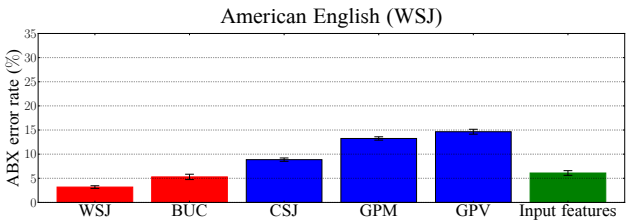

Figure 1: ABX error-rates averaged over all consonant contrasts of AE (using stimuli taken from the WSJ corpus).

Native phonetic categories are easier to distinguish than non-native categories (Gottfried, 1984). This is consistent with the predictions of our models shown in Figure 1. The discriminability of AE consonants obtained using stimuli from the WSJ corpus (read news) is higher for both models trained in AE (in red) than for models trained in a different language (in blue). This is true even though the BUC corpus contains speech in a different register (spontaneous speech) than the WSJ, GPV and GPM models (read speech).

Native-Language-Specific Confusion Patterns

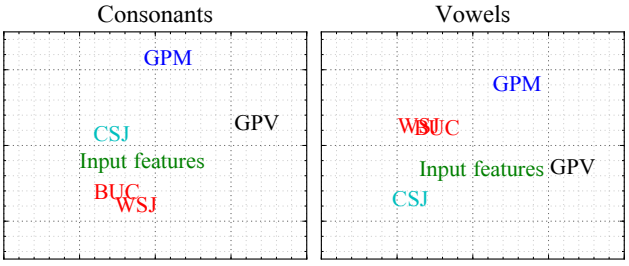

Figure 2: Two-dimensional embeddings of the different models based on the average cosine similarity between their patterns of ABX errors across the five test corpora. Left: for consonant contrasts. Right: for vowel contrasts.

The specific confusions we make between sounds of a foreign language are determined by our native language (Strange, 1995). Consistent with this effect, Figure 2 shows that the confusion patterns obtained with the two AE models over the different corpora are more similar to each other than to the confusion patterns obtained with models trained on other languages. In this figure, the distance between two points directly reflects the observed dissimilarity between the confusion patterns of the associated models. Importantly our measure of dissimilarity between confusion patterns does not depend on the average ABX errors studied in the previous section. See (Schatz, 2016) for more details.

Japanese Listeners and American English /r/-/l/

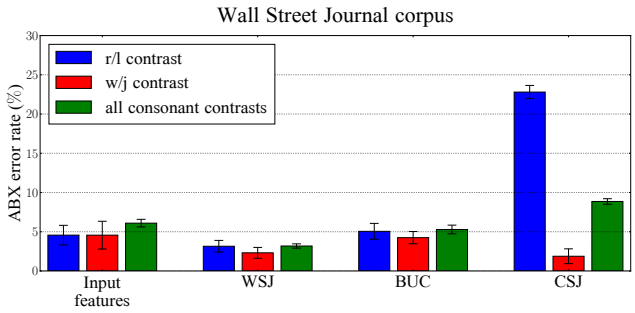

Figure 3: ABX error-rates obtained for the AE /r/-/l/ contrast and two controls (using stimuli from the WSJ corpus).

AE /r/ and /l/ are much harder to perceive for Japanese than for AE native speakers (Miyawaki et al., 1975). Figure 3 shows that our models' predictions are fully consistent with this effect: when comparing the Japanese model to both AE models and to the input features, the /r/-/l/ discriminability drops spectacularly, much more than the discriminability of comparable controls.

Conclusion

GMM-HMM ASR models can predict several effects in human phonetic category perception, showing the promise of using large-scale machine learning systems as quantitative models of human perception. GMM-HMM ASR systems have known limits however: they incorrectly model segment duration and while they are interesting as *perception* models, they are not plausible as *acquisition* models, because the speech transcriptions used to train them are not available to the learning infant. Further work will test more perception effects and will compare GMM-HMM to more powerful Neural-Network models as well as to more plausible *acquisition* models.

Acknowledgments

This research was supported by the European Research Council (grant ERC-2011-AdG 295810 BOOTPHON), the Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC).

References

Gottfried, T. L. (1984). Effects of consonant context on the perception of french vowels. *Journal of Phonetics*, *12*(2), 91–114.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of japanese and english. *Perception & Psychophysics*, *18*(5), 331–340.

Schatz, T. (2016). *ABX-Discriminability Measures and Applications*. Doctoral dissertation, Université Paris 6 (UPMC).

Strange, W. (1995). *Speech perception and linguistic experience: Issues in cross-language research*. York Press.