Neural network vs. HMM speech recognition systems as models of human cross-linguistic phonetic perception

Thomas Schatz[1,2] **(thomas.schatz.1986@gmail.com)**
Naomi H. Feldman[1,2] **(nhf@umd.edu)**
[1]Department of Linguistics & UMIACS, University of Maryland, College Park, USA
[2]Department of Linguistics, Massachusetts Institute of Technology, Cambridge, USA

Abstract

The way listeners perceive speech sounds is largely determined by the language(s) they were exposed to as a child. For example, native speakers of Japanese have a hard time discriminating between American English /ɹ/ and /l/, a phonetic contrast that has no equivalent in Japanese. Such effects are typically attributed to knowledge of sounds in the native language, but quantitative models of how these effects arise from linguistic knowledge are lacking. One possible source for such models is Automatic Speech Recognition (ASR) technology. We implement models based on two types of systems from the ASR literature—hidden Markov models (HMMs) and the more recent, and more accurate, neural network systems—and ask whether, in addition to showing better performance, the neural network systems also provide better models of human perception. We find that while both types of systems can account for Japanese natives' difficulty with American English /ɹ/ and /l/, only the neural network system successfully accounts for Japanese natives' facility with Japanese vowel length contrasts. Our work provides a new example, in the domain of speech perception, of an often observed correlation between task performance and similarity to human behavior.

Keywords: phonetic perception; ASR; neural networks; HMM

Introduction

Humans experience the external world through complex, high-dimensional sensory interfaces, such as the retina or the cochlea. Recent progress in machine learning has led to the development of artificial systems that can handle such complex sensory inputs and yield performance that sometimes rivals that of humans (e.g. Xiong et al., 2016; Mnih et al., 2015; He, Zhang, Ren, & Sun, 2015). Does that mean that these systems operate in a way that is similar to humans? While the similarity between computational systems and humans is not a logical necessity, several studies have reported a correlation between normative performance of the systems and their ability to predict human behavior or brain activity (e.g. Yamins et al., 2014; Banino et al., 2018).

We focus in this paper on the case of speech perception and ask whether neural network systems, which are more accurate for speech recognition than hidden Markov model (HMM) systems, are also better predictors of how humans will confuse foreign speech sounds. Existing theories of human cross-linguistic phonetic perception propose that foreign sounds are mapped onto native categories, and that these categories act as a language-specific filter affecting how non-native phonetic contrasts are perceived (e.g. Best, 1995; Flege, 1995). To model this category filter quantitatively, we train an ASR system on 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 posteriorgram). We then supply these 'native' representations to a simple model of a discrimination task: the machine ABX evaluation metric (Schatz et al., 2013; Schatz, 2016). This allows us to measure patterns of confusion predicted by each model for contrasts of interest and compare these to human perceptual judgments.

In a previous study (Schatz, Bach, & Dupoux, 2018), we used this approach to show that HMM ASR systems can correctly account for some effects that have been empirically observed in human speech perception, including for the difficulty of distinguishing American English /ɹ/-/l/ for native listeners of Japanese (Goto, 1971). However, HMMs are known to be structurally limited in their ability to model segment duration (Pylkkönen & Kurimo, 2004), and this may limit their utility as models of human speech perception. Vowel length contrasts are common cross-linguistically and are cued by segment duration, making them an ideal test case for determining whether neural networks are better than HMMs at modeling duration-based aspects of human speech perception.

In this paper we show that neural network models can predict the same /ɹ/-/l/ effect that had been previously captured by HMM models and that, unlike HMM models, they can correctly predict that vowel length contrasts in Japanese are easier to perceive for Japanese native listeners than for American English native listeners. This provides empirical evidence that neural network ASR systems are not only better at recognizing speech, but also at modeling human speech perception.

Methods

We train models on four different corpora of continuous speech, two in American English—the Wall-Street Journal corpus (WSJ), consisting of read news articles (Paul & Baker, 1992), and the BUCKEYE corpus (BUC), consisting of casual spontaneous conversations (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005)—and two in Japanese—the GlobalPhone Japanese corpus (GPJ), also consisting of read news articles (Schultz, 2002), and the Corpus of Spontaneous Japanese (CSJ), consisting of spontaneous relations of personal stories in front of an audience (Maekawa, 2003). Each corpus is di-

Table 1: Language, training and test set duration, speech register, and number of speakers for each corpus, as well as word error rates (WER) obtained with each HMM and neural network (NN) ASR systems on the test set of their training corpus.

Corpus	Language	Train	Test	Register	No. speakers	WER HMM	WER NN
WSJ	American English	19h30	9h39	Read	143	10.7%	8.12%
GPJ	Japanese	19h33	9h40	Read	143	23.19%	19.77%
BUC	American English	9h13	9h01	Spontaneous	40	63.4%	58.5%
CSJ	Japanese	9h11	8h57	Spontaneous	40	39.1%	33.6%

vided into a training and test set; only the training set is used to train models. The main properties of the corpora for each system are reported in Table 1, along with the word error rates (WER) obtained with each system on the test set of its training corpus. As expected, for each corpus, the neural network system has a lower WER than the corresponding HMM system.

We train HMM and neural network ASR systems with the Kaldi speech recognition toolkit (Povey et al., 2011). All instances of each type of model (HMM or neural network) are trained with the same recipe, adapted from the Wall Street Journal recipe, using the same default parameter values. We trained diagonal covariance word-position-dependent triphone Gaussian mixture model acoustic models with global semi-tied covariance transforms, linear discriminant analysis features, and feature-space maximum likelihood linear regression (fMLLR) speaker adaptation. We trained deep belief network acoustic models using the `nnet1` kaldi recipe, with unsupervised pre-training, followed by frame-level cross-entropy optimization and sequence-discriminative training. The neural networks acoustic models are initialized using the HMM acoustic models and take as input linear discriminant analysis features that are fMLLR speaker adapted using the HMM acoustic models. We refer the reader to the Kaldi documentation for further technical detail.[1] To compute word error rates, we train a word-level bigram language model on the training set of each corpus and combine it with the HMM or neural network acoustic model trained on the same corpus to perform Viterbi decoding of the test set. To extract frame-by-frame phone-level posteriors (see next section), we train a phone-level bigram language model on the training set of each corpus and combine it with the HMM or neural network acoustic model trained on the same corpus to obtain Viterbi-smoothed posteriors on the test set of any of the corpora.

Machine ABX 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 (Schatz et al., 2013; Schatz, 2016). 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's representation for X is closer to the model's representation for A than it is to the model's representation for B. The probability of an error (i.e. X is closer to B than to A) 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 zero, the two categories are discriminated perfectly. If it is equal to 0.5, discrimination is at chance level.

The model's representation of a vowel or consonant is obtained as a sequence of phone-level posteriorgrams taken every 10 ms. Posteriorgrams are vectors that indicate the posterior probability, under a particular model, that a time frame of the speech signal corresponds to each of the possible phonetic categories in the language that the model was trained on. To quantify how close two model representations are to each other, we use dynamic time warping (DTW) (Müller, 2007) which allows us to compute similarities between variable-length speech segments. We use Kullback-Leibler divergence as the underlying dissimilarity metric in the DTW algorithm.

In the specific ABX task considered here, we only evaluate triplets where A, B and X occur in the same phonetic context (same preceding phone and same following phone) and are uttered by the same speaker. For each phonetic contrast an aggregated ABX error rate is calculated by averaging over stimulus order, context and speaker.

Results

American English /ɹ/-/l/

American English /ɹ/ and /l/ are much harder to distinguish for Japanese than for American English native speakers (Miyawaki et al., 1975). Figure 1 shows that both HMM and neural network models correctly predict this effect. For both HMM systems (left panel) and neural network systems (right panel) the two 'Japanese native' models (in blue) have a much higher error rate for discriminating American English /ɹ/-/l/ than either of the 'American English native' models (in red). Two controls show that similar to humans, this deficit of Japanese models is particularly strong for the American English /ɹ/-/l/ contrast: although Japanese models are worse than American English models at discriminating American English consonants on average, and the American English /w/-/j/ contrast in particular, the decrease of performance in those cases is much more moderate. These results were obtained by testing on stimuli from the WSJ corpus; similar results are obtained when using test stimuli from the BUC corpus (not shown).

[1] See `http://kaldi-asr.org/`.

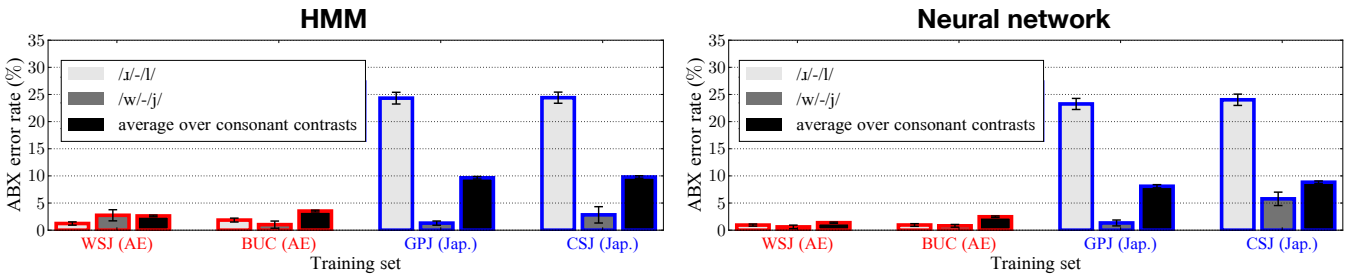

Figure 1: ABX error rates for HMM systems (left) and neural network systems (right) for the American English /ɹ/-/l/ contrast and two controls (using test stimuli from the WSJ corpus). 'American English native' (AE) models are in red and 'Japanese native' models (Jap.) are in blue. A specific deficit of Japanese models on American English /ɹ/-/l/ is clearly visible. Error bars indicate mean plus and minus one standard deviation and were obtained by resampling the ABX errors at the level of speakers.

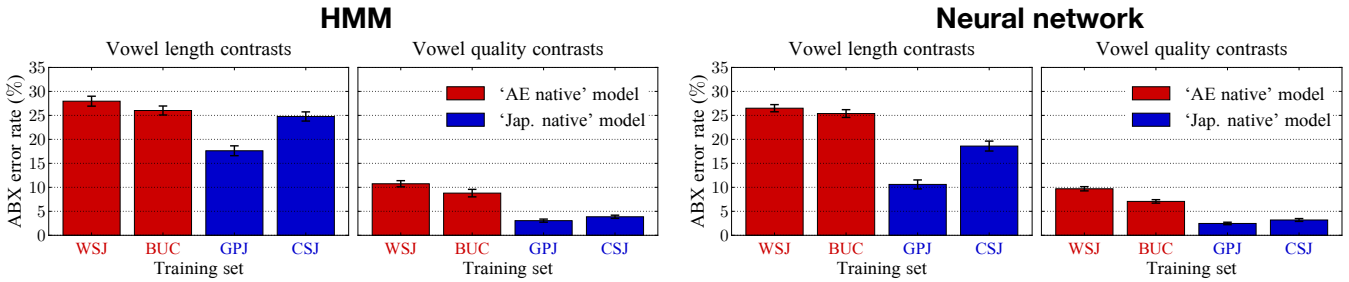

Figure 2: Average ABX error rate over all Japanese vowel length contrasts and all vowel quality contrasts for HMM systems (left) and neural network systems (right) using test stimuli from the GPJ corpus. 'American English native' models (AE) are in red and 'Japanese native' models in blue. The x axis indicates the corpus on which the model was trained. While both HMM and neural network 'Japanese' models outperform 'American English' models on vowel quality contrasts, for vowel length contrasts the 'Japanese' HMM model trained on the CSJ corpus does not have a clear edge over 'American English' models. Error bars as in Figure 1.

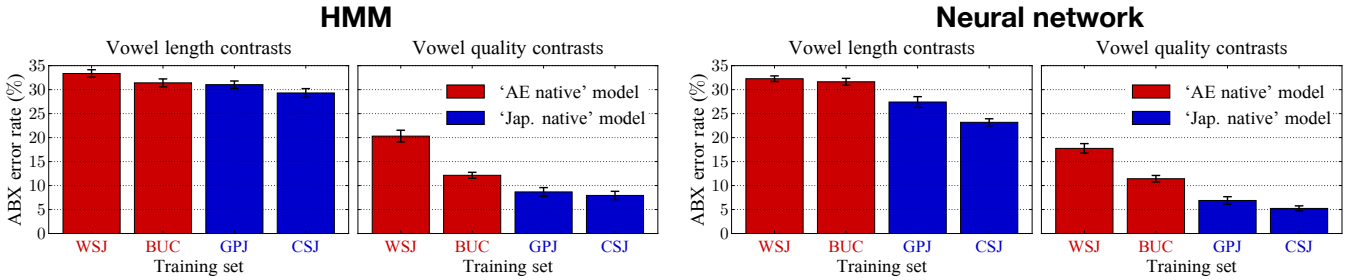

Figure 3: As in Figure 2, but using test stimuli from the CSJ corpus. Both of the 'Japanese' HMM models do not appear much better than the 'American English' models at discriminating Japanese vowel length contrasts.

Japanese vowel length

Each of the five Japanese vowels comes in a long and a short version which are phonemically contrastive (i.e. they need to be distinguished to properly identify certain words in Japanese). Long and short vowels are believed to differ primarily in their duration, making this an ideal test case for distinguishing models' ability to capture human-like perception of duration. Japanese vowel length contrasts are easier to perceive for Japanese natives than for American English natives (Hisagi, Shafer, Strange, & Sussman, 2010). In Figure 2, stimuli from the GPJ corpus are used to test the ability of HMM (left) and neural network (right) models to discriminate Japanese vowel length. In Figure 3, test stimuli from the CSJ corpus are used. Discrimination scores for Japanese vowel quality contrasts (contrasts between two different short vowels or two different long vowels) are also reported, as a control.

For HMM models, 'Japanese native' models (in red) appear clearly better than their 'American English native' counterparts (in blue) at discriminating Japanese vowel length in just one case: when testing the model trained on GPJ with test stimuli from that same corpus. The result does not generalize when the same model is tested with stimuli from the CSJ corpus

instead. The HMM model trained on the CSJ corpus does not appear better than the 'American English native' models, irrespective of whether it is tested with stimuli from the CSJ or GPJ corpus. This indicates that HMMs are—at best—inconsistent in their ability to capture the relevant duration cues for distinguishing phonemic vowel length in Japanese. They can learn to represent duration categories only in a corpus-specific way, and have trouble learning even corpus-specific representations when trained on a corpus of spontaneous speech.

In contrast, both of the 'Japanese native' neural network models are much better at discriminating vowel length than either of the 'American English native' neural network models when tested with either GPJ or CSJ stimuli. In particular, the 'Japanese native' neural network models are better than the 'American English native' models at discriminating vowel length, even when tested with stimuli from a different corpus than the one on which they were trained. This generalization across corpora provides strong evidence that the 'native benefit' effect observed with neural network models on Japanese vowel length contrasts reflects a genuine language-specificity of the learned representations that, unlike for HMMs, cannot be explained away by channel effects associated with specific recording conditions. Finally, let us emphasize that, when trained on American English, even neural network models do poorly at discriminating Japanese vowel length. This is important, because it shows that neural networks are not just better overall at processing duration, independently of training conditions (one could imagine, for example, that neural networks trained on American English would pick up on informative duration cues from American English tense/lax vowel contrasts, which involve duration). Rather, similar to humans, a facility for Japanese vowel length discrimination is observed only for the 'Japanese native' neural network models.

Conclusion

While both HMM and neural network models correctly predict that American English /ɹ/ and /l/ are very hard to discriminate for Japanese native listeners, only neural network models appear to robustly predict that Japanese vowel length contrasts are easier to perceive for Japanese natives than for American English natives. These results suggest that, in addition to being better from a normative point of view (i.e. better at recognizing speech), neural networks also constitute better predictors of human cross-linguistic speech perception patterns.

More generally, we have introduced a method for generating quantitative predictions regarding the discriminability of any foreign phonetic contrast for native listeners of any language (given a suitable training corpus). This makes it straightforward to compare the empirical adequacy of a wide range of speech processing systems with respect to the many effects reported in the empirical literature on cross-linguistic phonetic perception. It would be interesting, in particular, to look for effects that both HMM and neural network models systematically fail to capture, as these could point toward better models, both from the point of view of modeling human perception and, potentially, for practical applications, such as speech recognition.

Acknowledgments

This research was supported by NSF grant BCS-1734245. We thank Jon Gauthier and Marianne Duyck for helpful comments and discussion.

References

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., ... others (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204).

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277).

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia*, *9*(3), 317–323.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).

Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. (2010). Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and electrophysiological measures. *Brain Research*, *1360*, 89–105.

Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE workshop on spontaneous speech processing and recognition*.

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.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Müller, M. (2007). Dynamic time warping. *Information retrieval for music and motion*, 69–84.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proc. workshop on speech and natural language* (pp. 357–362).

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... others (2011). The Kaldi speech recognition toolkit. In *Proc. workshop on automatic speech recognition and understanding*.

Pylkkönen, J., & Kurimo, M. (2004). Duration modeling techniques for continuous speech recognition. In *Proc. INTERSPEECH*.

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

Schatz, T., Bach, F., & Dupoux, E. (2018). Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception. *The Journal of the Acoustical Society of America*, *143*(5), EL372–EL378.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proc. INTERSPEECH*.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at Karlsruhe university. In *Proc. INTERSPEECH*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2016). Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.