Speech features are weighted by selective attention

Nika Jurov[1,2] (njurov@umd.edu), Grayson Wolf[1,2] (graywolf@umd.edu),
William Idsardi[1] (idsardi@umd.edu), Naomi H. Feldman[1,2] (nhf@umd.edu)
Department of Linguistics[1] & University of Maryland Institute for Advanced Computer Studies (UMIACS)[2], College Park, Maryland, USA

Abstract

Listeners typically rely more on one aspect of the speech signal than another when categorizing speech sounds. This is known as *feature weighting*. We present a rate distortion theory model of feature weighting and use it to ask whether human listeners select feature weights simply by mirroring the feature reliabilities that are present in their input. We show that there is an additional component (*selective attention*) listeners appear to use that is not reflected by the input statistics. This suggests that an internal mechanism is at play in governing listeners' weighting of different aspects of the speech signal, in addition to tracking statistics.

Keywords: speech perception, rate distortion theory, VAE, feature weighting, selective attention

People rely on aspects of the speech signal (features) in different proportions when mapping speech signal to categories. For example, when distinguishing [ɛ] (as in *bet*) from [æ] (as in *bat*), English listeners use the first formant (F1) as a primary feature and duration as a secondary feature (Wu & Holt, 2022; Liu & Holt, 2015). This has previously been hypothesized to directly reflect listeners' long-term experience with feature statistics in the input – i.e., how reliably a feature separates one sound category from the other (Toscano & McMurray, 2010). More recently, however, Holt, Tierney, Guerra, Laffere, and Dick (2018) proposed an alternative hypothesis known as *selective attention*, in which listeners choose to focus on one feature rather than another in their perception. Their proposal was motivated by listeners' reweighting of features in laboratory tasks (Wu & Holt, 2022; Liu & Holt, 2015; Lehet & Holt, 2020; Clarke & Garrett, 2004; Guediche, Fiez, & Holt, 2016; Idemaru & Holt, 2020, 2014).

In this paper we present a new argument in favor of selective attention, based on the fact that listeners' feature weighting does not match the long-term input statistics they hear. We collect corpus statistics from three corpora of English and then train a number of models that are optimal perceivers, defined in information theoretic terms through rate distortion theory (RDT). We create models that directly reflect the input statistics and compare them to models that incorporate selective attention. We show that the latter is qualitatively more similar to behavioral data on feature weighting from Wu and Holt (2022).

Methods

We model how listeners categorize standard American English [æ] and [ɛ], focusing on features akin to those manipulated by Wu and Holt (2022) – the first formant (F1)[1] and vowel duration (Liu & Holt, 2015). Smaller values of both typically correspond to [ɛ] and higher values correspond to [æ]. F1 is thought of as a primary feature, whereas duration is secondary (Wu & Holt, 2022; Liu & Holt, 2015).

Our simulations are conducted with a β-VAE architecture adopted from Bates and Jacobs (2021) used in cued visual attentional allocation research. We trained 2 versions of the model: one baseline model based only on input statistics and one incorporating selective attention. The selective attention model has an additional parameter $\vec{\omega}$ forcing the model to focus more on F1 (see below). For each of these two versions of the model, 10 models were trained on each of the three speech corpora of standard American English: TIMIT (Garofolo, Lamel, Fisher, Fiscus, & Pallett, 1993), Buckeye (BUC; Pitt, Johnson, Hume, Kiesling, & Raymond, 2005) and Wall Street Journal (WSJ; Paul & Baker, 1992). We compared these to listeners' feature weights from Wu and Holt (2022).

Our models are an implementation of RDT, which is an information theoretic model of a system trying to maximize its performance with capacity constrained information processing (Barlow et al., 1961; Sims, 2016, 2018). It is built out of an encoder that projects the information to the latent representation which is then both expanded back in the decoder, and also used to categorize the sound. Its objective is a tradeoff between minimizing perceptual errors for reconstruction and categorization, and developing parsimonious latent representations. We use this model of perception to numerically estimate optimal feature weights.

β-VAE is a neural implementation of RDT (Alemi et al., 2018). It is a probabilistic deep neural network model trained to optimize a loss containing the rate (forcing the encoder to learn meaningful latent representations) and the reconstruction of an input (forcing the model to reconstruct the input as best possible). In addition to the base VAE architecture we add a supervised category model mapping the latent information to one of the categories, similar to Bates, Lerch, Sims, and Jacobs (2019). Following Bates and Jacobs (2021), to create models with selective attention, we add a term $\vec{\omega}$ that forces the model to consistently reconstruct one of the features more accurately than another. The loss is:

$$\mathcal{L}(\theta,\phi;\vec{x_i},\vec{y_i}) = -\beta D_{KL}(q_{\phi_j}(\vec{z_{i_j}}|\vec{x_{i_j}})||p_\theta(\vec{z_{i_j}})) + MSE(\vec{x_{i_j}},\vec{y_{i_j}})\vec{\omega}_{i_j} + BCE(cat(\sum_{j=1}^{n}\gamma_j\vec{z_{i_j}}), l_i) \quad (1)$$

where $\vec{x}$ is input; $\vec{y}$ is output; $\vec{z}$ is latent information extracted

[1]The first formant is chosen instead of all 5 as in Wu and Holt (2022), since it is most different in the two vowels.

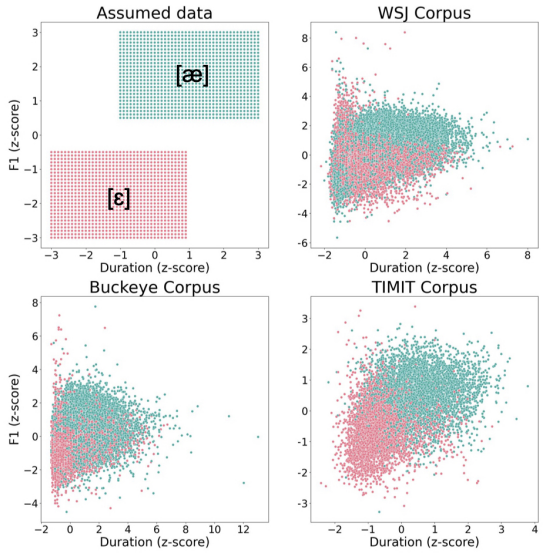

Figure 1: **Corpus data.** [æ] is green and [ɛ] is red. Assumed data is schematic given assumptions by prior research such as Toscano and McMurray (2010).

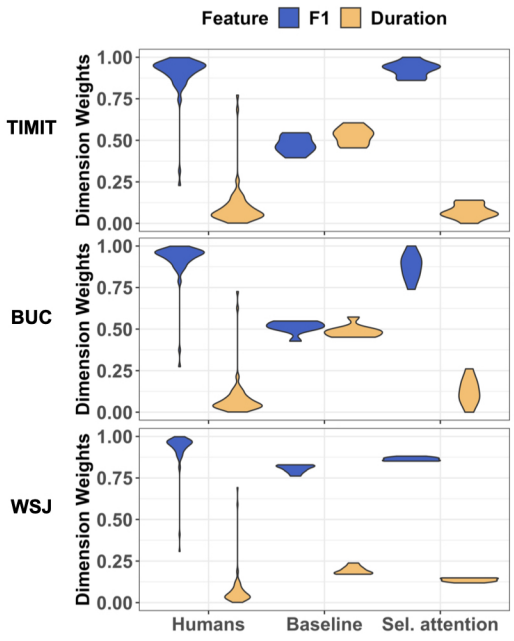

Figure 2: **Feature weights** in humans, models without selective attention, and models with selective attention.

by the encoder; D_{KL} is KL divergence; MSE is mean squared error; BCE is binary cross entropy; $q(), p()$ are probability distributions with θ, ϕ parameters where $q()$ is an approximation of $p()$; β is a parameter scaling the KL divergence proportionally to the MSE loss; $cat()$ is the categorization model; and $\vec{\omega}_{i_j}$ is the feature weighting. Parameters followed Bates and Jacobs (2021), except for $\vec{\omega}$, which was set at 0.1 for the downweighted feature (duration) and 0.9 for the upweighted feature (F1).

Simulations

We trained 10 models on each corpus separately and averaged results across the 10 replications.[2] The input to all simulations were [æ] and [ɛ] vowels encoded as 2-dimensional data (F1 and duration). These values were automatically extracted from each corpus using Praat and then normalized by z-scoring by speaker. Because of errors stemming from the automated feature extraction or from the automated alignments in the WSJ and BUC annotations, we excluded any vowel with an F1 greater than 1200Hz or a duration less than 30ms. Corpus data can be seen in Figure 1.

To obtain perceptual feature weights, following Wu and Holt (2022), we used linear regression[3] with features as predictors and category as the dependent variable. We tested models' categorization along the entire span of possible combinations of duration and F1 for a specific corpus, with a step of 0.1. To obtain comparable human feature weights, we used human data and experimental stimuli from Wu and Holt (2022). For each corpus, we first z-scored the experimental stimuli using the average mean and average variance of all female speakers of the corpus. We then used the step sizes in these z-scored duration and F1 values to recompute the human feature weights based on raw data from Wu and Holt (2022).[4]

Results are shown in Figure 2. Our baseline models that learn feature weights solely based on corpus statistics rely more on F1 than on duration, similar to humans. However, whereas the WSJ model matches human feature weighting relatively closely, TIMIT and BUC corpora show a more equal feature weighting that humans do. This difference cannot solely be due to the fact that WSJ contains well-articulated read speech, because TIMIT does as well (in contrast to BUC, which contains conversational speech). Our selective attention models ($\vec{\omega} = [0.1, 0.9]$) yield human-like feature weighting for all three corpora, as seen in Figure 2.

Discussion

This paper compared feature weighting in human speech perception to the input statistics that listeners hear. Our baseline models, whose feature weights were based solely on input statistics, trended toward relying more on F1 than duration, like humans do. Thus, which feature is primary does not appear to be random. However, our selective attention model showed qualitatively more human-like feature weighting than the baseline model did. This suggests that asymmetries in human feature weighting may not be solely determined by input

[2]Code is available at: https://github.com/n-ika/adapt2noise

[3]Logistic regression yielded qualitatively similar results.

[4]The lab stimuli were words produced in isolation, whereas the corpus data were extracted from continuous speech. This resulted in large acoustic differences between the two settings: the stimuli did not fall into the same acoustic range as the corpus, particularly along the duration dimension. Thus, the models trained on corpus data could not be used to directly predict human behavior on the lab stimuli. This mismatch also prevented us from conducting statistical tests to quantitatively compare the models' fit to human data.

statistics, as suggested by Toscano and McMurray (2010), but instead may be amplified through selective attention.

The parameter $\tilde{\omega}$ has been used to model attention in previous work on visual perception and has a natural interpretation within the theory of selective attention for speech perception (Holt et al., 2018), which hypothesizes that listeners selectively adjust how much different features contribute to perception. An interesting question for future research is why listeners control their attention to speech features in a way that deviates from the long-term statistics in their input.

Selective attention has also been hypothesized to generate neuronal shifts observed in humans and animals with STRF recordings in auditory cortex (Fritz, Shamma, Elhilali, & Klein, 2003; Holdgraf et al., 2016; Holt et al., 2018), and in the future, our models can potentially be extended to account for neural data using an architecture that implements selective attention.

Acknowledgments

We thank Charles Wu and Lori Holt for sharing their data with us, Christopher Bates for sharing his model code, and Philip Resnik and Thomas Schatz for helpful comments and discussion. This research was supported by NSF grant BCS-2120834.

References

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., & Murphy, K. (2018, 10–15 Jul). Fixing a broken ELBO. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 159–168). PMLR. Retrieved from https://proceedings.mlr.press/v80/alemi18a.html

Barlow, H. B., et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*(01).

Bates, C. J., & Jacobs, R. A. (2021). Optimal attentional allocation in the presence of capacity constraints in uncued and cued visual search. *Journal of Vision*, *21*(5), 3–3.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of vision*, *19*(2), 11–11.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658.

Fritz, J., Shamma, S., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature neuroscience*, *6*(11), 1216–1223.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, *93*, 27403.

Guediche, S., Fiez, J. A., & Holt, L. L. (2016). Adaptive plasticity in speech perception: Effects of external information and internal predictions. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(7), 1048.

Holdgraf, C. R., De Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J. J., ... Theunissen, F. E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature communications*, *7*(1), 1–15.

Holt, L. L., Tierney, A. T., Guerra, G., Laffere, A., & Dick, F. (2018). Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing. *Hearing research*, *366*, 50–64.

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009.

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. *Attention, Perception, & Psychophysics*, *82*(4), 1744–1762.

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*, 104328.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783.

Paul, D. B., & Baker, J. (1992). The design for the wall street journal-based csr corpus. In *Speech and natural language: Proceedings of a workshop held at harriman, new york, february 23-26, 1992.*

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.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, *34*(3), 434–464.

Wu, Y. C., & Holt, L. L. (2022). Phonetic category activation predicts the direction and magnitude of perceptual adaptation to accented speech. *Journal of Experimental Psychology: Human Perception and Performance*.