A neural architecture for selective attention to speech features

Nika Jurov[1,2]*, William Idsardi*[1]*, Naomi H. Feldman*[1,2]

[1] Department of Linguistics, University of Maryland, USA
[2] UMIACS, University of Maryland, USA

njurov@umd.edu, idsardi@umd.edu, nhf@umd.edu

Abstract

Speech perception is complex and demands constant adaptations to the speaker and the environment (i.e. noisy speech, accent, etc.). To adapt, the listener relies on one speech feature more than another. This cognitive mechanism is called selective attention. We present a model that captures the idea of selective attention: we show that this dynamic adaptation process can be captured in a neural architecture by using a multiple encoder beta variational auto encoder (β-ME-VAE), which is based on rate distortion theory. This model implements the idea that optimal feature weighting looks different under different listening conditions and provides insight into how listeners can adapt their listening strategy on a moment-to-moment basis, even in listening situations they haven't experienced before.

Index Terms: speech perception, cognitive modeling, computational psycholinguistics, rate distortion theory

1. Introduction

Speech perception is an active process of extracting information from a highly complex and variable signal. When mapping the speech signal to phonetic categories, such as [p] or [b], listeners can make use of information that is spread across several aspects of the speech signal (*features*) [1]. For example, one commonly used feature that distinguishes [p] and [b] in English is voice onset time (VOT), which denotes a difference between the burst and the start of voicing. Another feature signaling the [p] and [b] distinction is pitch (F0). In English, [b] has shorter VOT and lower F0, whereas English [p] has longer VOT and higher F0. In typical listening situations, listeners rely more on VOT than on F0 [2]. Previous models have aimed to capture this unequal reliance on different speech features as optimal inference under uncertainty, and have hypothesized that long-term input statistics determine feature weighting [3, 4].

However, recent data suggest that listeners are extremely flexible and fast at reweighting features when encountering new speakers and new listening conditions [2, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Specifically, [2] showed that listeners primarily rely on VOT in a clear listening condition, but they rely primarily on F0 when VOT is obscured (noisy listening condition). In addition, after hearing miscorrelated features (i.e. VOT typical for [b] and F0 typical for [p]), listeners rely even more on their preferred feature (VOT in the clear condition, F0 in the noisy condition) and ignore the information given by the other feature.

Previous models cannot easily capture this pattern because they do not take into account that speech can be momentarily perturbed. Any flexibility in feature weighting in those models requires training a separate model for each listening condition. This clearly cannot account for cases where listeners quickly adapt to miscorrelated features (a set of input statistics that they have not previously experienced).

In this paper, we introduce a neural architecture and show that it can facilitate rapid changes in perceptual feature weighting. Our model instantiates the idea of selective attention [14], a flexible cognitive process prioritizing one feature over another that enables listeners to adapt to the speaker and situation without conscious control. Mathematically, our model is based on rate distortion theory (RDT) [15], implemented as a multiple encoder beta variational autoencoder (ME-β-VAE). We show how speech features can be weighted flexibly on a moment to moment basis within this neural architecture without needing to retrain the network for each individual situation. This flexible weighting allows the model to switch between different conditions and provides insight into how listeners can achieve the fast, flexible reweighting that has been observed empirically.

2. A model based on rate distortion theory

We propose a new neural architecture that can allow listeners to flexibly shift their attention. Our model is based on an idea from information theory known as rate distortion theory (RDT).

RDT is a probabilistic model of a system (often also called a channel) trying to maximize its performance with capacity constrained information processing [16, 15, 17]. Its objective is to minimize perceptual errors. Therefore, the encoding channel extracts as much information as possible that is relevant to the task: in this case, reconstructing the acoustics of the speech signal and mapping the speech signal to a category, like [p] or [b]. RDT assumes that the constrained capacity is the source of what is often called internal or sensory noise. The information that passes through the channel is chosen so that task performance is maximized, subject to the constraints on channel capacity.

Our neural architecture is based on the β-VAE [18], which has been shown mathematically to implement the RDT framework of efficient information processing [19]. The β-VAE is a probabilistic deep neural network model trained to optimize the loss containing the rate (forcing the encoder to learn meaningful latent representations) and the reconstruction terms (forcing the model to reconstruct the input as best possible):

$$\mathcal{L}(\theta, \phi; \boldsymbol{x_i}, \boldsymbol{y_i}) = \\ -\beta D_{KL}(q_\phi(\boldsymbol{z_i}|\boldsymbol{x_i})||p_\theta(\boldsymbol{z_i})) + MSE(\boldsymbol{x_i}, \boldsymbol{y_i}) \quad (1)$$

where bolded symbols are vectors and acronyms denote specific losses. In particular, $\boldsymbol{x}$ is the input, $\boldsymbol{y}$ is the output, $\boldsymbol{z}$ is the latent information found at the end of the encoder. Losses: D_{KL} denotes KL divergence, MSE denotes mean square error. θ, ϕ are parameters of the probability distributions; $q(), p()$ denote the probability distributions, where $q()$ is an approximation of $p()$. β is a parameter scaling the KL divergence proportionally to the MSE loss.

Our β-VAE incorporates several advances relative to the original architecture from [18]. First, properties of the optimal channel in RDT depend on the loss function, and the channel would attend more to a specific part of the input if deviations in reconstructing that part of the input were penalized more in its loss function. This idea has been implemented in research on visual attentional allocation [20] and we adopt it here to simulate asymmetric feature weighting (i.e., reliance on one primary feature). Specifically, we obtain this by scaling each dimension of the reconstruction loss (MSE) with a predefined asymmetric weight which we call ω, so that the second term in the loss function becomes $MSE(\boldsymbol{x_{i_j}}, \boldsymbol{y_{i_j}})\boldsymbol{\omega_{i_j}}$. Setting a high value of ω for the reconstruction of a particular feature, such as VOT, leads the network to prioritize that feature in its latent encoding.

Second, following [21], we add a supervised categorization model, enabling category mapping based on the features. This allows us to extract the information in such a way that it best serves the mapping of the input to one of the categories. Any information not contributing to the categorization is as such less important to send through the channel. The categorization model is trained jointly with the β-VAE on cross-binary (BCE) entropy loss (see full loss in Equation 2): the loss compares the output of the categorization model given latent information z_i with the ground truth binary label l_i (either [p] or [b]) fed to the model during training: $BCE(cat(\boldsymbol{z_{i_j}}), l_i)$.

Third, our model has multiple encoders as seen in [22]. Each encoder extracts different feature information. This is achieved with feature weighting as in [20]: we force each encoder-decoder combination to have higher reconstruction accuracy on one of the dimensions — VOT for encoder 1, and F0 for encoder 2 — and inversely, lower reconstruction accuracy on the other. This unequal accuracy reconstruction makes the model learn different information with each encoder. This models internal manipulations that we hypothesize listeners to be performing on the input when optimizing their perceptual system based on experience.

In other words, we show how listeners can build a perceptual system that allows them to allocate their focus more to one feature rather than another, given their needs in any given listening situation. The joint training of encoders is achieved by getting KL and MSE losses for each encoder and then averaging over them (see full loss in Equation 2). The model is jointly optimized

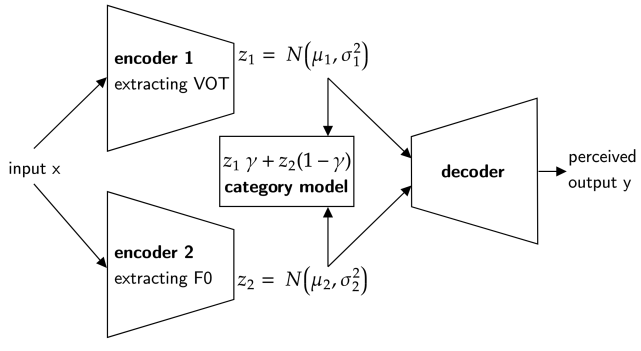

Figure 1: ***Multiple encoder β-VAE training*** *- Schematic architecture of training of multiple encoder β-VAE with an added categorization model. The number of encoders can be extended. Each encoder encodes the same input. A sample z_i from each encoder E_i is decoded separately in training. The category model receives a random proportion of data from each encoder.*

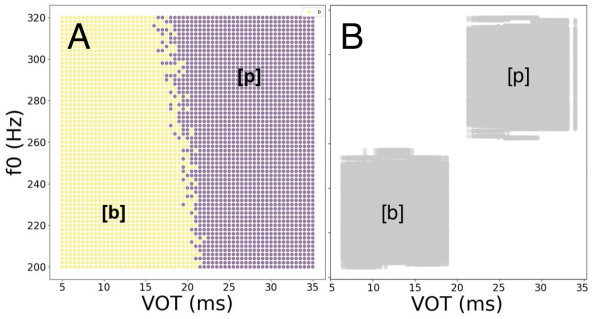

Figure 2: ***Feature weighting (single β-VAE) and training data*** *- **A:** Modeled decision boundary of a single channel β-VAE that is forced to reconstruct VOT better than F0. It cannot adapt to the moment-to-moment statistics unless retrained. **B:** Visualization of training data used for all simulations. For categorization, data in the lower left corner was labeled [b], and data in the upper right corner was labeled [p].*

to extract features, weight them by how reliable they are in the moment and map them to a category:

$$\begin{aligned}&\mathcal{L}(\theta, \phi; \boldsymbol{x_i}, \boldsymbol{y_i}) = \\ &\frac{1}{n}\sum_{j=1}^{n}\Big(- \beta D_{KL}(q_{\phi_j}(\boldsymbol{z_{i_j}}|\boldsymbol{x_{i_j}})||p_\theta(\boldsymbol{z_{i_j}})) \\ &+ MSE(\boldsymbol{x_{i_j}}, \boldsymbol{y_{i_j}})\boldsymbol{\omega_{i_j}}\Big) \\ &+ BCE(cat(\sum_{j=1}^{n}\gamma_j \boldsymbol{z_{i_j}}), l_i)\end{aligned} \tag{2}$$

Each training step consists of each encoder encoding the same input x. The last layer of each encoder is a sampling step where the latent variable z_i is obtained. Each z_i is pushed through the decoder to obtain the reconstructed input, which is in turn used to calculate the loss function. γ_j is a randomly sampled weight from $[0, 1]$ interval, such that $\sum_{j=1}^{n}\gamma_j = 1$. γ_j defines the proportion of the information taken from each encoder, such that their combination is pushed through the category model $(cat(\sum_{j=1}^{n}\gamma_j \boldsymbol{z_{i_j}}), l_i))$.

To motivate our architecture choice we also show results on a single channel β-VAE, that is one encoder one decoder architecture. Its loss is as stated in equation 2 with 1 encoder ($n = 1$). This model is able to asymmetrically extract a single feature at the expense of another, but does not have the power of adaptation like the β-ME-VAE does.

3. Model training and testing

We implemented a scenario with 2 features (VOT and F0) to model listeners' ability to use different primary features in different listning conditions, as in [2], where listeners used VOT in clear listening conditions and F0 in noisy conditions and also increased their reliance on the primary feature when hearing a reversed correlation between the two features. That experiment tested comparable scenarios on both consonants and vowels, but since experiments on vowels show qualitatively the same results as consonants, we simulate only the consonants. We show that the preference between the two features is flexible given the speaker and/or environment. This preference can be visualized as a decision boundary or a line on a 2D plot that shows which of

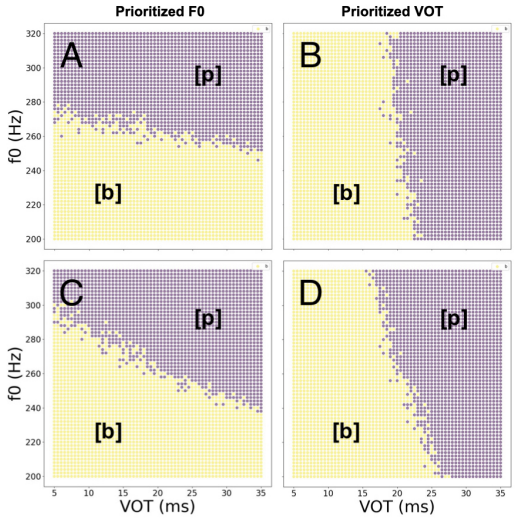

Figure 3: ***Flexible adaptation in perception*** *- Modeled decision boundaries due to different combinations of encoders. Encoder 1 is focused on VOT, encoder 2 is focused on F0. Depending on how much weight we give to each encoder, we arrive to different decision boundaries.* ***A:*** *all of the decision is based on F0.* ***B:*** *all of the decision is based on VOT.* ***C:*** *80% of the information is coming from encoder 2, 20% from encoder 1.* ***D:*** *80% of the information is coming from encoder 1, 20% from encoder 2.*

the features (each on one of the axes) contributes more towards the categorization of the stimulus (Figure 3).

Our experiments are conducted with Python version 3.8.12 and TensorFlow version 2.8.0 on macOS 13.2.1 system with a single GPU. The architecture and the parameters used for the simulations were the following: there were two encoders (one for the β-VAE), each consisting of two 2D convolutional layers, followed by a dense layer (2000 units) and a layer extracting the mean and the log variance (each 500 units) of the latent variable z, all having rectified linear activation functions. The model was trained with Adagrad optimizer with learning rate 0.001, batch size 32 and β of 0.0025. To make each encoder focus on one feature, the ω was set to 0.999 for the upweighted feature (VOT in encoder 1 and F0 in encoder 2) and to 0.001 for the downweighted feature (F0 in encoder 1 and VOT in encoder 2). The decoder was a one layer model with 2 units and a linear activation function. The category model consisted of two layers, the first with 100 units with rectified linear activation functions and the second with 2 units and sigmoid activation functions. This overparametrized architecture was already used by [20], except that here we increase the number of encoders. All parameters were chosen to give good performance but no attempt was made to systematically optimize them, as we were more interested in qualitative characteristics of the model. The code is available at: `https://github.com/n-ika/adapt2noise`.

The input to all simulations was 2-dimensional data, where the first value denotes VOT and the second F0. These values simulate those as reported in the 7-step continuum in [2]: VOT varied as 5-ms steps between 5ms and 35ms and F0 varied as 20-Hz steps between 200Hz to 320Hz. However, because of numeric differences between the scales (one between 5-35, the other between 200-320), the actual numbers simulating these two scales were both set to the same scale, centered around 0. In total, there were 180,000 data points. Each data point was created by sampling from a normal distribution with variance 1 and a mean either -3, -2, or -1 for [b] stimuli; or a mean 1, 2, or 3 for [p] stimuli. This is because shorter VOT and lower F0 are typical of [b] and longer VOT and higher F0 are typical of [p]. There were no values sampled around 0, just like in [2], since this is a value between typical [p] and [b] VOT/F0 values (i.e. not characteristic of either of the two categories). For visualization of the training data, see Figure 2.

During test time for the β-ME-VAE, the latent variables (z_1 and z_2) are joined with a different ratio, depending on how much information is wanted from each encoder:

$$\sum_j \gamma_j \times E_j(x_j) \tag{3}$$

Where γ_j denotes the ratio of information from each encoder $E_j()$ and $\sum_j^n \gamma_j = 1$. For example, to simulate total reliance on VOT, only the information coming from encoder encoding mostly VOT is taken and none from the encoder encoding mostly F0.

4. Experiments

Speech perception based on feature weighting has been described by prior research to be active and thus attention guided [23, 14, 24]. Listeners have a preferred feature they focus on more which does not seem to be a result of perceptual distances themselves [25]. Our first simulation illustrates one way in which listeners may learn to attend to a primary cue based on an internal weighting, rather than simply by matching external input statistics (as in [3]).

Our single channel (one encoder model) was trained to better reconstruct VOT information rather than F0, similar to encoder 1 of the β-ME-VAE. It was trained with the loss function seen in Equation 2 with $\omega = 0.999$ to upweight VOT information and $\omega = 0.001$ to downweight F0 information. This channel bases its categorization mapping almost exclusively on VOT information (Figure 2). This model has a similar learning outcome to the model in [3], in that it weights the two features to different degrees. The difference between the models is in the cause of that behavior: whereas the model in [3] relies on long term input statistics to determine feature weighting, our model's feature weighting depends on an internal feature weighting manipulation, encoded in the loss function as $\boldsymbol{\omega}$.

However, an important limitation to this kind of a channel is that it does not change its feature weighting unless it is retrained. To force it to rely more on F0 information, we would need to either give it different data or a different set of weights ω, which would result in catastrophic forgetting. This means that the model does not have adaptive power of switching between features, but rather it is always attentive to one feature's information. This is not what humans do, as they would categorize a set of stimuli differently based on the listening condition they are in. Even in listening situations that are new, such as when hearing miscorrelated features, the listeners can quickly adapt their feature weighting to resolve the conflict of the miscorrelated features that are never present together for a single English speech sound.

4.1. β-ME-VAE

The β ME-VAE is built to allow for the type of flexible adaptation seen in humans. Human listeners base their listening decisions on environment and/or speaker. Our model shows

similar behavior, depending on which encoder most of the information comes from. Encoder 1 was trained to better reconstruct VOT information rather than F0, similar to the single-channel β-VAE, whereas encoder 2 was trained to better reconstruct F0 information than VOT. To model reliance on features at test time as seen in [2], we encode all of the test stimuli with both trained encoders. We then probabilistically weight this information as described in Equation 3. For example, to simulate the fact that a listener biases their decision mostly on VOT, we take most of the information from encoder 1 (prioritizing VOT).

As seen in Figure 3, the probabilistic weighting at test time shows that the model can change the feature weighting that characterizes its decision boundary by manipulating how much information comes from each encoder. It can put nearly all its weight on one feature (top row), like humans do when they are listening to miscorrelated features. It can also attend to both features, with one of the two features being primary, and can flexibly switch its primary feature (bottom row).

This flexibility is evident in human listeners, who can switch their reliance on features given the speaker and the situation they are in. For example, if they find themselves in a situation that has VOT masked, they change the feature primacy to F0 by taking most of the information from the encoder 2 (prioritizing F0) (see lower left decision boundary). Then, with miscorrelated data, they use even more of the information coming from the encoder extracting mostly F0 and even less of the information coming from the encoder extracting mostly VOT (see upper left decision boundary).

5. Discussion

This paper has introduced a new neural architecture that captures selective attention and enables behavioral adaptation (shifting reliance between features) as seen in [2]. This has allowed us to propose a potential explanation for how the optimal response of a system can change on a moment-to-moment basis in perception, even if the relevant environmental statistics have not been observed through long-term experience.

Relying on features dynamically is a process seen in carefully designed laboratory experiments on humans. However, human listeners are extremely robust in perceiving speech in unforeseen conditions or with speakers with accents they never heard before in more naturalistic conditions as well [10, 13]. RDT is an ideal framework for modeling this type of behavior in humans, because it allows us to model the fact that not all information is encoded at every moment - only the features that are most informative for category mapping in the moment. The use of multiple encoders allows us to capture the fact that which specific features are most important changes given the situation and the speaker.

5.1. β VAE

In the future, it will be important to expand this model beyond just two features. For example, as many as 16 different features can be used in discriminating voiced versus voiceless stop intervocalically in English, not just VOT or F0 [1]. To expand our architecture, we could expand the number of encoders and extract more features, each with a separate encoder.

This idea is appealing because it has qualitative similarities with neuronal computations of auditory cortex seen experimentally in animals and humans, specifically spectro-temporal receptive fields (STRFs). STRFs are aggregated neuronal responses to acoustic features and usually reflect activity of several neurons [26, 27]. Recent work [14] has hypothesized that rapid modulation of STRFs in auditory cortex may play a role in listeners' rapid adaptation to different listening conditions. Changes in STRFs are thought to show facilitated sound detection in ferrets guided by attention [26] and in people guided by prior experience [27]. Selective attention then is a feature weighting process that may have markers in the brain and enables listeners to unequally weight feature information according to their needs.

Based on the parallels with STRFs described above, each encoder's output could hypothetically represent an STRF-like filter in primary auditory cortex (A1). To further evaluate this hypothesis, we would ideally want to see the weighting mechanisms present in addition to STRFs. Evidence for such weighting mechanisms would need to show how these "feature detector" neurons are pooled together from A1 to make a higher order sound representation, i.e., how information from A1 is treated by later processing areas like the superior temporal gyrus (STG). There is some evidence of category information [28, 29] in STG. Further research can determine how closely those neural mechanisms correspond to the weighting mechanism proposed here.

Prior research in vision shows that humans are sensitive to the noise that happens during encoding of the stimuli — the noisier the stimulus the more difficult it is to perceive. However, humans ignore the noise that may occur while integrating different cues with high confidence [30]. This suggests that flexible feature weighting may be a cognitive strategy beyond speech perception. In addition, neuroscience research in animal audition has shown active neuronal suppression of the non informative part of the stimulus in a given moment [31]. Ignoring noisy part of the stimulus would in our model result as simple down-weighting of information coming from a specific encoder that we would not want to contribute to the category mapping.

Finally, the choice of how to set attention weights associated with each encoder is in the present model manual. How listeners decide on this weighting in a particular listening situation remains an interesting question for future research. Future work can also expand this model beyond the stimuli that were used in a particular laboratory study, to account for perception of complex speech signals in naturalistic conditions with multiple speakers.

6. Conclusion

This research showed that results of selective attention, a strategy facilitating speech perception, can be implemented in a neural architecture. We presented a new model β-ME-VAE that is based on RDT. This framework explains that perceptual distortions happen as a byproduct of capacity constrained information extraction. This process enables us to extract the most important speech features for a particular speech perception task. The extensions of RDT to a multiple-encoder model enable us to model selective attention or dynamic feature weighting that changes on a moment-to-moment basis. When humans show perceptual flexibility, this may be the result of a neural architecture that can flexibly focus attention on the most important parts of the speech signal, rather than simply the result of accumulating the statistics of the environment.

7. Acknowledgments

We thank 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.

8. References

[1] L. Lisker, ""voicing" in english: A catalogue of acoustic features signaling/b/versus/p/in trochees," *Language and speech*, vol. 29, no. 1, pp. 3–11, 1986.

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

[3] J. C. Toscano and B. McMurray, "Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics," *Cognitive Science*, vol. 34, no. 3, pp. 434–464, 2010.

[4] D. F. Kleinschmidt and T. F. Jaeger, "Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel." *Psychological Review*, vol. 122, no. 2, p. 148, 2015.

[5] K. Idemaru and L. L. Holt, "Specificity of dimension-based statistical learning in word recognition." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 40, no. 3, p. 1009, 2014.

[6] M. Lehet and L. L. Holt, "Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing," *Cognition*, vol. 202, p. 104328, 2020.

[7] R. Liu and L. L. Holt, "Dimension-based statistical learning of vowels." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 41, no. 6, p. 1783, 2015.

[8] K. Idemaru and L. L. Holt, "Generalization of dimension-based statistical learning," *Attention, Perception, & Psychophysics*, vol. 82, no. 4, pp. 1744–1762, 2020.

[9] H. Zhang, S. Wiener, and L. L. Holt, "Adjustment of cue weighting in speech by speakers and listeners: Evidence from amplitude and duration modifications of mandarin chinese tone," *The Journal of the Acoustical Society of America*, vol. 151, no. 2, pp. 992–1005, 2022.

[10] C. M. Clarke and M. F. Garrett, "Rapid adaptation to foreign-accented english," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3647–3658, 2004.

[11] M. H. Davis, I. S. Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan, "Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences." *Journal of Experimental Psychology: General*, vol. 134, no. 2, p. 222, 2005.

[12] S. Guediche, J. A. Fiez, and L. L. Holt, "Adaptive plasticity in speech perception: Effects of external information and internal predictions." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 42, no. 7, p. 1048, 2016.

[13] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 2008.

[14] L. L. Holt, A. T. Tierney, G. Guerra, A. Laffere, and F. Dick, "Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing," *Hearing research*, vol. 366, pp. 50–64, 2018.

[15] C. R. Sims, "Rate–distortion theory and human perception," *Cognition*, vol. 152, pp. 181–198, 2016.

[16] H. B. Barlow *et al.*, "Possible principles underlying the transformation of sensory messages," *Sensory Communication*, vol. 1, no. 01, 1961.

[17] C. R. Sims, "Efficient coding explains the universal law of generalization in human perception," *Science*, vol. 360, no. 6389, pp. 652–656, 2018.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[19] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken ELBO," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 159–168. [Online]. Available: https://proceedings.mlr.press/v80/alemi18a.html

[20] C. J. Bates and R. A. Jacobs, "Optimal attentional allocation in the presence of capacity constraints in uncued and cued visual search," *Journal of Vision*, vol. 21, no. 5, pp. 3–3, 2021.

[21] C. J. Bates, R. A. Lerch, C. R. Sims, and R. A. Jacobs, "Adaptive allocation of human visual working memory capacity during statistical and categorical learning," *Journal of Vision*, vol. 19, no. 2, pp. 11–11, 2019.

[22] L. Ternes, M. Dane, S. Gross, M. Labrie, G. Mills, J. Gray, L. Heiser, and Y. H. Chang, "A multi-encoder variational autoencoder controls multiple transformational features in single-cell image analysis," *Communications Biology*, vol. 5, no. 1, pp. 1–10, 2022.

[23] S. L. Heald and H. C. Nusbaum, "Speech perception as an active cognitive process," *Frontiers in Systems Neuroscience*, vol. 8, p. 35, 2014.

[24] Z. Harmon, K. Idemaru, and V. Kapatsinski, "Learning mechanisms in cue reweighting," *Cognition*, vol. 189, pp. 76–88, 2019.

[25] A. L. Francis, N. Kaganovich, and C. Driscoll-Huber, "Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in english," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1234–1251, 2008.

[26] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nature Neuroscience*, vol. 6, no. 11, pp. 1216–1223, 2003.

[27] C. R. Holdgraf, W. De Heer, B. Pasley, J. Rieger, N. Crone, J. J. Lin, R. T. Knight, and F. E. Theunissen, "Rapid tuning shifts in human auditory cortex enhance speech intelligibility," *Nature Communications*, vol. 7, no. 1, pp. 1–15, 2016.

[28] I. Bhaya-Grossman and E. F. Chang, "Speech computations of the human superior temporal gyrus," *Annual Review of Psychology*, vol. 73, pp. 79–102, 2022.

[29] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[30] S. Herce Castañón, R. Moran, J. Ding, T. Egner, D. Bang, and C. Summerfield, "Human noise blindness drives suboptimal cognitive inference," *Nature communications*, vol. 10, no. 1, pp. 1–11, 2019.

[31] Z. P. Schwartz and S. V. David, "Focal suppression of distractor sounds by selective attention in auditory cortex," *Cerebral Cortex*, vol. 28, no. 1, pp. 323–339, 2018.

[32] P. K. Kuhl, "Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not," *Perception & Psychophysics*, vol. 50, no. 2, pp. 93–107, 1991.

[33] N. H. Feldman, T. L. Griffiths, and J. L. Morgan, "The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference." *Psychological Review*, vol. 116, no. 4, p. 752, 2009.

[34] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman, 1982.

[35] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[36] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*. PMLR, 2015, pp. 1083–1092.

[37] G. B. Keller and T. D. Mrsic-Flogel, "Predictive processing: a canonical cortical computation," *Neuron*, vol. 100, no. 2, pp. 424–435, 2018.