Towards a neurally, behaviorally, and developmentally plausible model of speech perception

Keywords: learning, speech, perception, rational model, statistics, sparse coding

Speech perception is the process by which a continuous, acoustic signal—the pressure waves produced by a speaker—is converted into a discrete, linguistic representation—words, sentences, and ultimately meanings. The fact that state-of-the-art automatic speech recognition systems perform as well as a four-year-old human, and only after extensive training and in tightly proscribed situation, demonstrates that this is a task of extreme complexity which normal humans accomplish with remarkable ease. Naturally, speech perception has intrigued and challenged psychologists, linguists, and engineers for as long as it has been remotely tractable. The key to understanding this process is to develop a model of the way that speech sounds are represented at the point where they are categorized in a linguistically-relevant way. I propose a framework for such a model that is based on well-known principles of efficient coding (which well-describe the functioning of early perceptual areas in the brain) and optimal cue-combination (which have been used to explain how different sources of information are combined to make inferences about underlying causes.

For some types of phonetic distinctions, phoneticians have identified relevant acoustic cues. For instance, *peach* and *beach* differ only in the voicing of their initial sounds—/p/ is called voiceless and /b/ voiced. One important cue to voicing is Voice-Onset Time (VOT), which is the amount of time between the opening of the lips and the beginning of voicing. As VOT increases, perception abruptly switches from a /b/ percept to a /p/ at around 40 ms VOT.

Based on the existence of acoustic cues such as VOT, it would be natural to think that phonetic category identification should be as simple as reading off a few acoustic measures, but in practice this picture is complicated by two facts. The first is that, due to systematic and random variation, categories are actually realized as overlapping distributions on any given cue dimension [1]. The second is that, in order to deal with the ambiguous nature of these cues, listeners combine information from a large number of cues in order to make phonetic judgements, and almost every cue is used in making multiple phonetic distinctions.

Furthermore, it's not clear that the acoustic cues identified so far are the ones actually used in human speech perception. My proposed model thus has two parts. The first part models how the auditory system extracts informative acoustic cue dimensions by adapting to the statistical structure of complex natural stimuli. The second part models how information from these cues is combined to form phonetic categories. Both of these models are domain-general statistical models of perception and categorization, supported by rigorous mathematics, and have been used to explain a variety of behavioral and neural data.

The first phase of the proposed project is a model which learns informative acoustic cue dimensions using sparse coding and related methods. As discussed in Previous Research Experience, my preliminary data [2] has demonstrated that this approach learns representations that match those used by the primary auditory cortex quite well, based on both speech and non-speech natural sounds.

The second phase will be the development of a rational model of how the different acoustic cues learned by the first phase can be best combined to effectively make phonetic distinctions. Work on rational cue-combination has shown that the best way to make a decision based on multiple cues is to pay more attention to more reliable cues. Such models have already been used to explain the trading relationships between VOT and other cues in determining voicing [4]. These models simultaneously learn the underlying number of categories and the optimal way of combining different cues. I plan to extend such models to learn the optimal weighting from the acoustic features extracted from speech in the first phase to known phonetic distinctions (e.g. voicing).

I will evaluate this model using the same behavioral dataset as [4], training the model on examples of voiced and voiceless consonants extracted from actual speech and represented using the features used in phase one. Using a common task allows a direct comparison between features identified by phonetics research (e.g., VOT and vowel length) and those learned by sparse coding, and is a strong test of the ability of features learned in this way to support speech perception.

Another phenomenon that could be explained using this two-stage model comes from studies on perceptual adaptation to non-standard pronunciation. These studies systematically manipulate the way that certain sounds are pronounced (e.g., making /s/ sound more like /f/), and find that the boundaries between phonetic categories shift noticeably after only a small amount of training [3]. A rational model of cue-combination like the one proposed here can account for these effects as Bayesian inference, where prior knowledge about the characteristics of phonetic categories is combined with observations to produce updated beliefs about phonetic categories. Moreover, there are puzzling differences in whether or not subjects generalize to different speakers and related phonetic distinctions. The proposed framework could explain this through differences in how strongly the manipulated features cue for different category distinctions or for speaker identity.

The third phase will investigate the effect of allowing these different learning processes to interact. While the features represented in A1 may be genetically specified, it is possible that they are learned during development, and learning phonetic categories simultaneously may change the nature of these low-level features. One strength of the statistical framework used for this project is that both stages can be combined into a single, two-stage model which simultaneously learns acoustic features, phonetic categories, and the weighting between them. The low-level acoustic cues learned by such a hybrid model can be compared to those learned by the phase one model alone, in order to test what top-down effect learning phonetic categories might have on basic acoustic representations.

The **intellectual merit** of this approach to modeling and understanding human speech perception lies in the confluence of methods and results linguistics, neuroscience, psychology, and machine learning. The models that I have proposed are general statistical techniques that have been applied broadly, but which are generally consistent with the current neural and behavioral data on speech perception, as well as being plausible from broader neural, developmental, and behavioral perspectives. This is a unique and novel approach to understanding speech perception.

The **broader impacts** of this project are twofold. First, because it brings together insights and methods from many fields, it will require broad collaboration and dissemination across disciplines. To language scientists, it potentially offers a way of integrating speech perception into the broader framework of cognitive science. For researchers in neuroscience, machine learning, and general cognitive psychology, speech perception can serve as a challenging model domain for their methods and theories. Second insights gleaned from this work on how the auditory system underlies speech perception have potential to advance automatic speech recognition technology as well, and more importantly can inform the development of better cochlear implants.

[1] Hillenbrand, J., et al., 1995. Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, 97(5.1) 3099–111.

[2] Kleinschmidt, D., 2010. Efficient coding of speech in the auditory cortex. *poster presented at the 2nd Annual Neurobiology of Language Conference*.

[3] Kraljic, T. et al., 2006. Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, 13(2) 262–8.

[4] Toscano, J.C. et al., 2008. Using the distributional statistics of speech sounds for weighting and integrating acoustic cues. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 433–438.