The dorsal stream in speech processing: Model and theory

James L. Keidel, Stephen R. Welbourne and Matthew A. Lambon Ralph
School of Psychological Sciences, University of Manchester
Manchester, M13 9PL United Kingdom

Abstract

The ability to produce and comprehend spoken language requires an internal understanding of the complex relations between articulatory gestures and their acoustic consequences. Recent theories of speech processing propose a division between the ventral stream, which involves the mapping of acoustic signals to lexical/semantic representations, and the dorsal stream, which mediates the mapping between incoming auditory signals and articulatory output. We present a connectionist model of the dorsal stream of speech processing that utilizes a novel schematic representation of time-varying acoustics and a featural mapping of articulation. The model successfully learns a large training vocabulary, accurately produces novel items and demonstrates patterns of perceptual errors highly similar to those observed in human subjects.

Keywords: speech perception; speech production; dorsal stream; neural networks

Introduction

In few instances do our intuitions more thoroughly deceive us than in the 'common sense' picture of speech perception. Phenomenologically, the experience of hearing speech is similar to that of reading text; empty gaps set off each word, and these words consist of individual stretches of sound corresponding more or less perfectly to the written character s. In reality, things are not nearly so clean-cut; the silent stretches that do appear in the signal (for instance, during voiceless stop closures) do not typically correspond to word boundaries, and one would be hard pressed to point to the instant in the word 'deed' where the signal switches from /d/ to /i/. To the extent that phonetic segments do exist, they are highly blurred together by the effects of the preceding and following articulatory gestures. In some cases the same stretch of sound induces different percepts in different contexts; thus the same burst of noise will produce a /p/ percept before a back vowel but a /k/ percept preceding a front vowel. In addition, each segment can be signaled by a constellation of cues, none of which is guaranteed to be present in a given utterance.

This apparent lack of acoustic invariants for phonetic identification led Liberman and his colleagues to propose the Motor Theory (MT) of speech perception, which holds that the true objects of speech perception are not acoustic but instead articulatory. Under this view, listeners employ a potentially innate understanding of vocal tract physics to recover the speaker's intended gestures from the acoustic waveform. The lack of invariance at the acoustic level is thereby resolved through reference to the putatively invariant underlying articulations.

Whether or not speech perception is mediated by neural representations of intended gestures, the many findings inspired by MT provide important insight into the nature of phonetic categorization. While there exist myriad cues to phonetic identity, these cues necessarily covary due to the constraints imposed on the signal by the physics of the articulatory apparatus. Thus while a low F1 onset frequency and a lack of F1 cutback are signals to voicing in word-initial stops, these two features tend to trade off as the supraglottal articulations in voiced and voiceless stops are highly similar. When voicing commences at the release of the stop closure, F1 will be low because, as an index of jaw position, it will reflect the fact that it is still relatively closed. Additionally, due to the presence of energy (from the voicing source) in the vicinity of the first formant, F1 will be audible and thus there will be little to no F1 cutback.

In spite of this extensive variability, the speech perception system displays an astonishing ability to recover the speaker's intended phonetic message and thereby her meaning. Many studies have demonstrated just how difficult it is to render a speech signal unintelligible. Remez, Rubin, Pisoni and Carrell (1981) showed that participants could understand speech signals composed only of sine waves tracking the movement of the first three formants of a sentence. Listeners in Shannon et al.'s (1995) study quickly learned to transcribe signals composed of four bands of white noise modulated by amplitude within each filtered band, a manipulation which all but removes temporal cues to phonetic identity. At the other end of the spectrum, Saberi and Perrott (1999) split speech waveforms into 50-100 ms chunks which were then each reversed in time; again, participants quickly adapted to this manipulation and could reproduce the distorted sentences.

Given the complexity of the stimulus, then, we should not be surprised that the corresponding neural activation is so extensive and difficult to pin down. The perception of a single word activates an extensive bilateral network centered around the primary auditory cortices and extending both anteriorly and posteriorly along the superior temporal gyri (Binder et al., 2000). Similarly, speech production activates a left-lateralized network consisting of IFG (Broca's area), insula, and primary and supplementary motor cortices (Hickok & Poeppel, 2007). Unsurprisingly, there is also a significant amount of overlap between the two systems, both in terms of the underlying representations that mediate the sound to articulation mapping, as well as the perceptuo-motor systems that provide online monitoring of one's own utterances and the ability to shadow heard speech at extremely short latencies.

While most researchers would accept primary auditory cortex as an appropriate starting point for analysis of the speech processing system, the consensus appears to end there as well. The advent of functional neuroimaging has rendered even the most general questions concerning speech processing open for debate. On the basis of lesion data, the classical Wernicke-Geschwind model posited a few very basic tenets about the organization of language in the brain: 1) Language in right-handed individuals is generally left-lateralized; 2) The posterior portion of the left superior temporal lobe ('Wernicke's area') is primarily responsible for language comprehension; 3) The left IFG ('Broca's area') is primarily responsible for guiding language production (Geschwind, 1970). All of these statements have been subject to some reevaluation; we focus here on the sensorimotor transformations required to compute the mapping between audition and articulation. As such, we will first examine receptive processing of the speech signal, and then integrate this discussion into an understanding of the mechanism of speech production. In this discussion we will follow the terminology of Hickok and Poeppel (2007), who pose a distinction between speech perception and speech comprehension. Speech comprehension involves recovering the speaker's intended message from the acoustic signal, and more or less aligns with our everyday use of speech. Speech perception, on the other hand, refers primarily to the sorts of behaviors beloved by speech researchers, such as identification and discrimination of speech sounds and other explicitly phonological or phonetic tasks. When no distinction need be made; viz., when a statement is true of both perception and comprehension, we refer to both under the aegis of speech processing.

Hickok and Poeppel (2007) propose a dual-stream model of speech processing, in which a ventral stream projecting bilaterally from A1 to posterior STG and MTG mediates the mapping between signal and lexicon, while a left-lateralized dorsal stream involving the left temporo-parietal junction (area Spt) and IFG links speech sounds to articulations. To a first approximation, then, this model fits well with the classical picture. However, on the basis of a number of neuroimaging studies, these authors propose some important elaborations to the Wernicke-Geschwind model. Key among these is the proposal that area Spt plays a central role in computing the mapping between incoming auditory information and articulatory gestures.

That there might be some region of cortex keyed into the relation between acoustics and articulation seems quite likely; Hickok and Poeppel offer two strong motivations for the necessity of such an area. First, when a child is learning to speak, the disparity between the intended and actual output provides the learning signal that drives organization of the articulatory output system. The importance of this function is not limited to development, however; the adult speaker must also monitor her output for errors. The quick and efficient operation of this system is perhaps best exemplified by the work of Houde and Jordan (1998), who manipulated the acoustic feedback to participants in such a way as to make them believe they had produced an incorrect vowel when reading single words aloud (e.g., when a participant produced 'head' she heard herself saying 'heed'). Participants in this experiment quickly adjusted their articulations to reflect the altered feedback concerning vowel height; a function which Hickok and Poeppel ascribe to the dorsal stream.

While the neuroimaging findings discussed above have certainly increased our understanding of the anatomical substrates of speech processing, there remain a number of open questions concerning the nature of the underlying conceptual representations. That is, while we may agree that area Spt is involved in the mapping between representations of sensory data (e.g., heard speech) and those underlying production (e.g., articulatory sequences), we would also like to know what the representations in Spt actually ***look like***. What stimuli are considered similar/equivalent in this area? Are there nonlinearities in the encoding of stimuli along perceptual or articulatory dimensions known to be relevant in speech processing? While imaging studies certainly can contribute to this goal (perhaps especially through the use of fMRI adaptation paradigms), it is likely that work in other methodologies will provide key insights as well.

Parallel Distributed Processing (PDP) provides an ideal framework for the exploration of the internal representations that guide behavior. This is perhaps especially true in speech perception, as the signal consists of multiple interacting probabilistic cues that likely require highly nonlinear weightings for correct stimulus classification. An important early precursor to the work presented here is the TRACE model of speech perception (McClelland & Elman, 1986), which was designed to account for a number of key phenomena in speech perception and lexical access. One of the key insights of this model was the highly interactive nature of speech processing: processing of the early part of the signal strongly constrains the interpretation of the latter part. However, there exist key differences between TRACE and the model presented here: while the input to TRACE was an acoustic featural description of the signal, in our model we use schematic spectrograms derived from actual recordings of English. In addition, the weights in TRACE were set by hand, while in our simulations the weights were adjusted as a function of the mismatch between the target and actual output.

A closer precedent to our simulations can be found in the work of Kello and Plaut, who explored phonological development in the context of neural network models. Plaut and Kello (1999) trained a multi-layer PDP network on the mappings between a schematic, feature-based acoustic input and both articulation and semantics, demonstrating the feasibility of the approach presented here. In addition, Kello and Plaut (2004) focused specifically on the forward mapping from articulation to acoustics, by training a network on the mapping between actual articulatory recordings (EMA, laryngography, electropalatography) and their associated acoustic output. A key insight of both these papers is the idea that the representations that underlie

speech processing are best viewed as neither purely acoustic nor purely articulatory in nature, but rather are shaped by the covariance structure of the articulation-audition interface.

Methods

Model Architecture

The simulations described in this paper employed a 4-layer PDP network trained with continuous recurrent backpropagation (Pearlmutter, 1995). The input layer contained 46 units, divided into two separate filter banks and two task units. The first 22 units represented the presence of acoustic energy in 1 Bark bands, corresponding to the region of the spectrum from 0-8 kHz, spaced according to auditory acuity as a function of frequency. The second bank of 22 units had the same frequency spacing, but their activation corresponded to the presence or absence of periodicity in each frequency range. The task units indicated the delay at which the model was required to produce the response; these units were employed in the delayed repetition test described below. The input layer projected to a hidden layer of 150 units, which itself was recurrently connected to a second hidden layer containing 150 units as well. This second hidden layer projected recurrently to the output layer. All simulations used a learning rate of 0.01, and a Euclidean distance metric was used for scoring model performance.

The output layer contained 21 units, which represented a schematic articulatory mapping based on that reported in Keidel, Zevin, Kluender and Seidenberg (2003). Each word was coded as a series of articulatory targets, in line with the work of Browman and Goldstein (1992). Individual bits in each segment vector corresponded to constructs such as place of articulation (POA), constriction degree, tongue tip/body position and velar lowering.

Stimulus Design The input to the model consisted of schematic acoustic representations of CVC stimuli. To create these representations, we first recorded a native Southern British English speaker (SRW) producing tokens of 16 English consonants (six stops, eight fricatives, and two nasals) in onset position before 11 different vowels. These recordings were then analyzed to determine values for key acoustic parameters known to affect phonetic identification, such as burst spectrum, direction and magnitude of formant transitions, vowel formant values, and others described below. These values were then employed to create schematic time-varying acoustic input for the network to classify.

The perception of stop consonants is perhaps the best studied field in speech research, as they embody the interaction of multiple probabilistic cues perhaps better than any other type of segment. In our stimuli, voiced stops in onset position were represented as three separate events: 1) a burst portion corresponding to the release of pressure built up behind the constriction; 2) formant transitions resulting from the movement of the primary articulator from the closure into the vowel, and 3) the steady state vowel itself. Values for the bursts followed the characterization found in the work of Blumstein and Stevens (1979) and Repp and Lin (1989). Schematically, labial bursts were flat and diffuse, alveolar bursts were diffuse and rising, and velar bursts were compact. To prevent the model from tracking idiosyncratic features of production by our single speaker, the F1 transition was always calculated as the trajectory from a start value of 200 Hz to the steady-state F1 of the following vowel. The onset frequencies for F2 and F3 transitions were measured for each vowel context, and these served as the basis for the model's input representations. Values for the first three formants of each vowel were calculated as the averages of F1, F2 and F3 across the recorded consonant contexts.

Fricatives were represented as a steady-state frication period followed by formant transitions into the syllable nucleus. Values for the frication spectra were taken from spectrographic measurements of the model speaker. The formant transition values from the stop measurements were also used in some fricative contexts; thus the alveolar and post-alveolar fricatives received the transitions from the alveolar stops and the labiodental fricatives received the transitions of bilabial stops. For the interdental fricatives, transition values from the model speaker were measured and used to create the stimuli. Nasals were modeled by an initial murmur followed by transitions appropriate for the POA taken from the stop productions described above. Generally, coda consonants were represented as the time-reversed versions of their onset counterparts. However, in the case of voiceless stops, the distinction was produced by shortening the vocalic portion of the word, as this is the predominant cue to coda stop voicing in English.

The measured values described above were then converted into time-varying acoustic input vectors to the model (hereafter referred to as the 'acoustic matrix', reflecting the time x frequency nature of the stimuli). The matrix for each CVC word was 36 x 44, with 36 20 ms time steps and 44 frequency coefficients (22 for energy in a filter and 22 for presence or absence of periodicity). Stimuli were first vowel-centered, so that similar formant transitions (e.g., those in /d/ and /z/) would overlap in time. Next, measured acoustic values were inserted into the proper filters, according to a linearly weighted split of energy between the units representing the closest Bark values on either side of the given frequency. Formant transitions were created by linear interpolation between onset frequencies for transitions and steady-state values for the relevant vowel over a 60 ms time window, equivalent to 3 events along the time axis of the acoustic matrix.

To simulate effects of speaker variability, 15 versions of each token were created by adding noise to the vowel formants, then calculating transitions with respect to the new values (e.g., /da/ was specified with a 400 Hz falling F2 transition, so regardless of the formant values chosen in the noise procedure, F2 would fall 400 Hz into the steady state

F2 value). Additionally, Gaussian noise with an SD of 0.1 was added to the activation of the input units.

In order to validate the similarity of our acoustic representations to the actual speech signal, we trained the network on all possible combinations of the 16 consonants in both onset and coda position with the 11 vowels in the syllable nucleus (a total of 2816 stimuli; 10% of these stimuli were withheld to test generalization performance). At the conclusion of training, we tested the model on the trained stimuli in varying levels of noise, for comparison with the perceptual confusion data presented in Miller and Nicely (1955).

In this classic paper, the authors present identification data from four human listeners labeling thousands of syllables of the form /Ca/, where C ranged over the six stops, eight fricatives, and two onset nasals (/m/ and /n/) of the English sound system. The experiment was carried out under a number of different noise levels and bandpass filter widths, illustrating the gradual breakdown of the boundaries between phonetic categories as listening conditions deteriorate. The comparison of the model's misidentifications of words in noise with Miller and Nicely's (MN) data allows for an independent validation of the match between our acoustic representations and the statistical structure of the actual speech signal. That is, while any mapping is in principle learnable by a PDP network, it is by no means given that the breakdown in function induced by the addition of noise will follow that exhibited by human listeners.

Results

All results presented represent the average of 10 runs of the model. After 1.5M trials the models reached asymptote, identifying an average of 99% of the input patterns correctly based on a Euclidean distance criterion. Generalization performance was somewhat poorer, with an average of 93% of items named correctly.

Clearly, any model of speech perception must account for the foundational finding of MT: categorical perception. In the terminology of Liberman, Harris, Hoffman, & Griffith (1957), categorical perception occurs when identification predicts discrimination: if two stimuli from the same speaker are both identified as /ba/ then the listener will not be able to tell them apart. However, if the same amount of acoustic difference straddles a category boundary (such as that between /ba/ and /da/), then the two stimuli will be discriminable. To simulate this in the model, we interpolated the initial formant values used to create /ba/ and /da/ stimuli and generated a 10-step series between these stimuli. Importantly, the model had only been exposed to the endpoint stimuli in training; thus, the identification of the intermediate stimuli represents true generalization. Figure 1 shows the results of this test: as in human subjects, discriminability peaks sharply at the category boundary, and intermediate stimuli are identified as the closest endpoint in auditory space. Discrimination performance was calculated as the normalization of the Euclidean distance between hidden unit representations generated by three consecutive stimuli in the series.

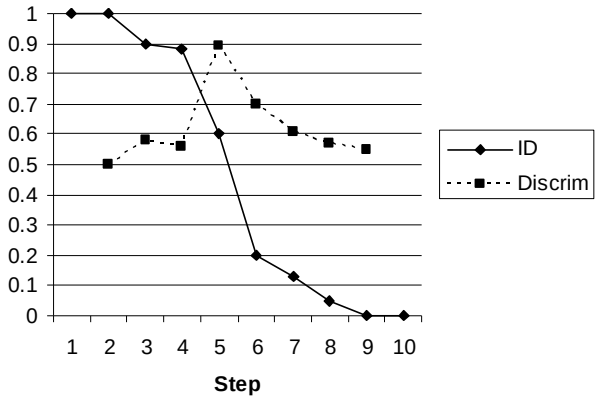

Figure 1. Identification and discrimination of a /ba/-/da/ series. Y-axis represents percent /ba/ identification for ID curve and correct discrimination of tokens (see text).

Figure 2 shows the correlation between the errors of human listeners perceiving speech at a signal-to-noise ratio (SNR) of -6 dB, and the average of 10 models' identifications of the training stimuli presented in Gaussian input noise with an SD of 0.35. This value was chosen to produce the same proportion of errors as the MN subjects independent of the errors' distribution. Because most of the cells of the confusion matrix are empty we only entered cells with error rates greater than .05 for one of the groups. The results of the analysis demonstrate a good fit to the human data: $r(27) = .54$, $\boldsymbol{p} < .05$. Similar results were found for the other noise levels tested by Miller and Nicely; in all cases the correlations between human and model data were significant.

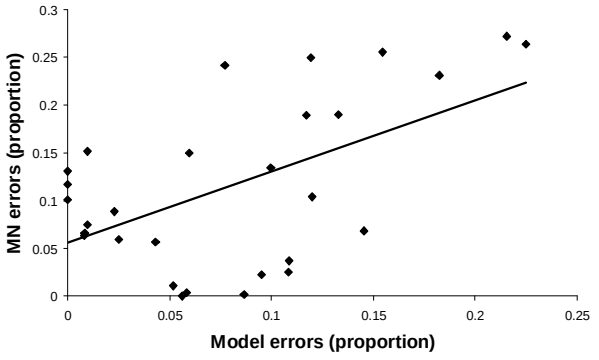

Figure 2. Correlation of model identification performance in noise and human participant data from Miller and Nicely

As a further test of the match between human and model behavior, we introduced high levels of noise to the second hidden layer, and tested the model's ability to repeat words (i.e., items from the model's training set) and nonwords (items from the generalization set). Performance on this task was evaluated in the same manner as above, viz. Euclidean

distance. Jefferies, Crisp and Lambon Ralph (2006) demonstrated that patients with phonological impairments following cerebrovascular accident showed an interaction between lexicality and delay, such that nonword repetition was significantly more impaired than word repetition at longer delays. As can be seen in Figure 3, the model demonstrates a very similar lexicality effect as unit noise increases, with lexical items much more resilient to increasing noise.

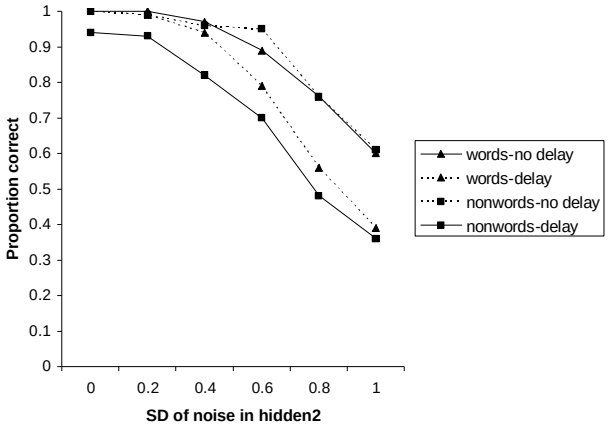

Figure 3. Model performance as a function of lexicality and naming delay.

Discussion

The results from identification in noise demonstrate that our acoustic representations accurately reflect the structure of the speech signal: error patterns in the simulations closely matched those observed in human listeners. At a general level, place errors were much more common than errors in voicing, and perception of nasal consonants was very robust even at high noise levels. At a more fine-grained level of detail, the models captured the high degree of confusability for interdental and labiodental fricatives, both voiced and voiceless. Additionally, the model captured an important dissociation in misidentification of stop consonants in noise. Specifically, for voiced stops it is /d/ and /g/ which are most likely to be confused, since listeners rely on the direction of F2 and F3 transitions, which are rather similar for these two segments. On this basis, then, one might expect a similar pattern for the voiceless stops, with a high degree of confusion between /t/ and /k/. However, it is actually /p/ and /k/ that are more confusable in this case, since the formants in voiceless stops are excited by low-energy aspiration noise and thus listeners appear to focus more on the release burst in identifying these sounds. In the case of /t/, the release burst is a very strong cue to a coronal POA, as it is high-energy and high-frequency. For the other two stops, however, the bursts are not quite as robust and thus it is the labial and the velar that are more commonly mistaken for one another when perceived in noise.

Dorsal stream speech processing is a highly complex behavior, requiring analysis of information at multiple levels of specificity. At the acoustic level, listeners are exquisitely sensitive to small variations in the signal (e.g., VOT changes on the order of 10s of milliseconds), yet are able to decode highly degraded signals in which most or all of these fine-grained cues have been destroyed. At the same time, speech perception is also strongly shaped by higher-level linguistic influences. While the many cues to phonetic identity are largely independent of meaning, listeners' online processing of the signal is demonstrably affected by their knowledge of the lexicon.

Further, the speech processing apparatus must fulfill two main functions. On the one hand, we listen to understand, and thus speech perception must provide a mapping from the acoustic signal to semantic representations—the function of the ventral stream in Hickok and Poeppel's model. However, both during linguistic development and in the adult state, we must also process speech signals in such a way as to be able to produce articulations similar to those which gave rise to the input. An important question, then, is at what point in the processing pathway do the cortical representations diverge?

Unsurprisingly, the answers that have been offered to this question reflect certain theoretical commitments that lead different researchers to different conclusions. For instance, proponents of MT and direct realism propose that this is simply a distinction without a difference, as the key tenet of MT is that it is recovery of the underlying articulations that allows lexical access. In this sense, then, we get the sound to meaning mapping 'for free': since the brain has inbuilt structure that allows us to know what the speaker intended to do with their vocal tract, the acquisition of a lexicon is to a first approximation simply a matter of rote memorization.

While MT and its descendants possess a great deal of intuitive appeal, they face a number of issues at both the implementational and theoretical levels. For any given speech waveform, there are an infinite number of vocal tract configurations that could have given rise to the signal—a relation known as the 'inverse problem'. The work presented here does not directly address this issue, as the preprocessing of the acoustic signal and the provision of veridical articulatory targets renders the mapping learnable. Nonetheless, the work spawned by MT has greatly increased our understanding of the speech processing apparatus, and the model presented here represents an attempt to reify many of the insight that this work has provided.

While we believe that this model represents an important first step toward a mechanistic implementation of recent theory in the study of speech processing, there obviously remain a number of issues to address. Chief among these is the addition of lexical/semantic knowledge, which would permit the exploration of the many interactions between bottom-up and top-down interaction in speech perception. Further, while the added noise to the input pattern prevents the model from learning entirely speaker-specific information, the results from this word do not directly

address the rather vexatious question of speaker variability. In this, however, we are not alone; even the most advanced commercial automatic speech recognition systems have not yet achieved the ideal of large-vocabulary speaker-independent identification.

In future work, we intend to employ the multi-layer architecture implemented here in order to investigate the internal representations that arise in layers whose input and output is impacted more by perception or production. For instance, the first hidden layer in our model receives direct input from the acoustic layer, while the second hidden layer receives a transformed version of this input which it must use to drive the articulatory output. It is likely that the internal representations within these layers (investigated with multivariate tools such as multidimensional scaling) reflect the different processing demands associated with these tasks.

Acknowledgements

This work was supported by a grant from the Gatsby Foundation (GAT2831).

References

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., et al. (2000). Human temporal lobe activation by speech and nonspeech sounds. ***Cerebral Cortex, 10***(5), 512-528.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic Invariance in Speech Production - Evidence from Measurements of the Spectral Characteristics of Stop Consonants. ***Journal of the Acoustical Society of America, 66***(4), 1001-1017.

Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology - an Overview. ***Phonetica, 49***(3-4), 155-180.

Geschwind, N. (1970). Organization of Language and Brain. ***Science, 170***(3961), 940-&.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. ***Nat Rev Neurosci, 8***(5), 393-402.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. ***Science, 279***(5354), 1213-1216.

Jefferies, E., Crisp, J., & Ralph, M. A. L. (2006). The impact of phonological or semantic impairment on delayed auditory repetition: Evidence from stroke aphasia and semantic dementia. ***Aphasiology, 20***(9-11), 963-992.

Keidel, J. L., Zevin, J. D., Kluender, K. R., & Seidenberg, M. S. (2003). Modeling the role of native language knowledge in perceiving nonnative speech contrasts. ***Proceedings of the 15th International Congress of Phonetic Sciences***, 2221-2224.

Kello, C. T., & Plaut, D. C. (2004). A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. ***Journal of the Acoustical Society of America, 116***(4), 2354-2364.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. ***Journal of Experimental Psychology, 54***(5), 358-368.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. ***Cognit Psychol, 18***(1), 1-86.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. ***Journal of the Acoustical Society of America, 27***, 338-352.

Pearlmutter, B. A. (1995). Gradient calculation for dynamic recurrent neural networks: a survey. ***IEEE Transactions on Neural Networks, 6***(5), 1212-1228.

Plaut, D. C., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), ***The Emergence of Language***. Mahwah, NJ: Erlbaum.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech-Perception without Traditional Speech Cues. ***Science, 212***(4497), 947-950.

Repp, B. H., & Lin, H. B. (1989). Acoustic Properties and Perception of Stop Consonant Release Transients. ***Journal of the Acoustical Society of America, 85***(1), 379-396.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. ***Nature, 398***(6730), 760.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. ***Science, 270***(5234), 303-304.