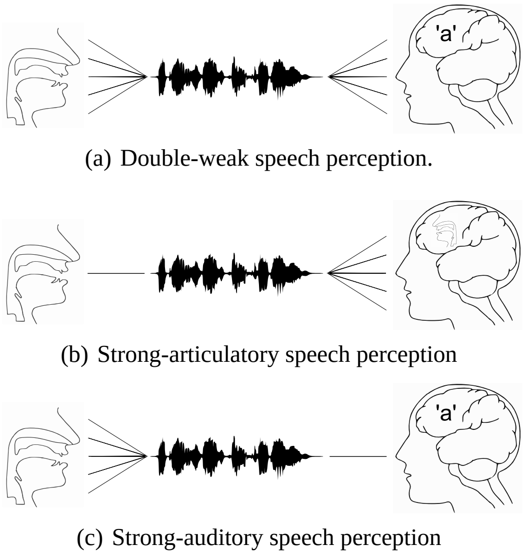
Speech Perception as Non-symbolic Pattern Recognition

S. F. Worgan and R. I. Damper
Information: Signals, Images, Systems (ISIS) Research Group,
School of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ, UK
sw205r@ecs.soton.ac.uk

Abstract

Despite ongoing research, the human ability of speech perception remains a mystery. Current phonetic theory is divided by two points of contention: the relationship from production to signal to audition and the object of perception/cognition. Here we discuss the role of current phonetic theory within this debate and propose our own hypothesis. We argue that human speech is enabled through loosely constrained articulation and audition coupled with the cognitive process of direct realism (DR). We also contend that disembodied pattern recognition is sufficient for the perception of phonetic tokens, as grounding can be maintained through the properties of real speech. However, to maintain this at the semantic level we feel that robotic embodiment will be necessary.

Although related to motor theory (MT), DR differs in a number of important ways. Significantly, speech perception is not held to be 'special' ... "and there is no more reason to propose a role for the speech motor system in speech perception than to propose an analogous role for the viewer's locomotor system in the visual perception of walking" (Fowler, 1996, p. 1731). Instead of forming cognitive representations of the external world (either gestural or acoustic), our senses cause the direct perception of the gesture through the acoustic signal.



(a) Double-weak speech perception.

(b) Strong-articulatory speech perception

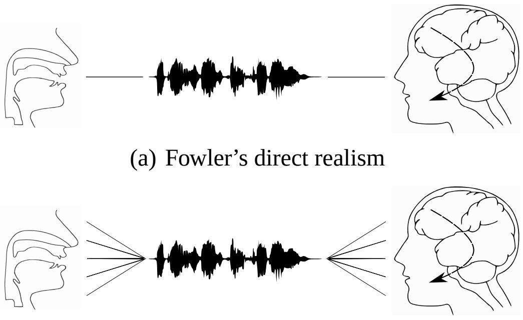
(c) Strong-auditory speech perception

Figure 1: Conflicting phonetic theories use evidence of strong constraints on articulation or audition to argue for different symbolic systems of perception.

DR faces various criticisms, arising through its association with MT, as they are often treated as one and the same, e.g., Sussman (1989); Ohala (1996). Other criticisms are more specific. What is the force enabling auditory distinctiveness if we only perceive the gesture? Surely we would be driven to maintain *articulatory* distinctiveness? Fowler argues that the acoustic signal still conveys information about the gesture, which accordingly must be sufficiently distinct. But it does not follow that a distinct signal is evidence for a symbolic auditory representation. Another objection is that those who can't speak can still perceive speech. Motor theorists believe that an "innate vocal-tract synthesizer" (Liberman and Mattingly, 1985) can overcome this objection. While Fowler reemphasises that the direct perception of speech derives from a general theory of perception, this "inability to reproduce heard gestures does not imply that they did not perceive gestures (any more that the typical person's inability to perform a triple axel implies that he or she cannot see them)" (p. 1738). DR does not have to imply a motor theory of speech perception. It only needs to agree with MT in the trivial sense—we obviously 'perceive' the vocal tract as it is the source of the speech signal. Where DR can provide insight is in determining the object of speech perception.

Using Nearey's (1997) framework, we can classify conflicting theories of perception into strong-auditory, strong-articulatory, double-strong and double-weak (see Figure 1). Strong-auditory theories include Stevens's (2002) well-known quantal theory. By contrast, strong-articulatory theories include MT and Fowler's direct realism. Double-weak theory defines a middle course, loosening constraints on both production and perception. However, many would consider it to be an auditory rather than articulatory theory.

Such disagreements arise because Nearey's classification only considers the means of production, the signal and perception of speech, whereas the current major source of disagreement is the form of the cognitive tokens. Auditory theories hold that these smallest tokens are resolved as



(a) Fowler's direct realism

(b) Proposed double-weak direct realism

Figure 2: A comparison of Fowler's direct realism and double-weak direct realism. The phonetic evidence suggests a double-weak approach, while our own work proposes a direct realist cognitive theory.

idealised symbolic phonetic tokens, whereas MT holds that the ultimate forms of perception are gestural tokens. Considered in these terms we can see that DR and MT (lumped together in Nearey's framework) are clearly different, as DR considers the perception of speech to be direct "unmediated by processes of hypothesis testing or inference making and unmediated by mental representations" (Fowler, 1996, p. 1731)—articulatory or acoustic. Freed from the need to lump all gesturalist theories into the strong-articulatory camp, we can see that DR is in fact a double-strong gesturalist theory (as opposed to motor theories strong-articulatory gesturalist approach). As clearly stated by Fowler: "phonological gestures are the public actions of the vocal tract that cause structure in acoustic speech signals. By hypothesis, they will be found to cause specifiers or invariants in the acoustic signal" (p. 1731).

We believe that speech is directly perceived; what is perceived (in the trivial sense) is the vocal tract. Although this appears to agree with Fowler, our theory differs in important respects. We question Fowler's naïve realism assertion that invariant "specifying acoustic properties is what allows perception of the phonological properties to be direct" (p. 1731). We feel that this plays into the hands of a number of arguments against the philosophy of DR. Rather we, like Nearey, are "genuinely impressed by the quality of the research by both auditorists and the gesturalists that is critical of the other position" (p. 3242). Given this we take a double-weak standpoint to the production and auditory perception of the speech signal. However, we do not believe that this double-weak approach necessarily precludes DR. As Figure 2(b) shows, in this new framework we can conceive of loosely-constrained articulation and perception coupled with the direct perception of speech, leading to a new double-weak direct realism. Clearly, there needs to be a de-coupling between the constraints on speech and the cognitive objects of perception.

To support this assertion, we have constructed a computational model that is able to acquire the phonetic structure of real speech using the details of this hypothesis. An artificial agent, equipped with a biologically plausible auditory system and vocal tract, is able to reproduce a range of phonemes after being exposed to real speech. Both its auditory and articulatory functions are loosely constrained (in accordance with double-weak theory) and at no time does it establish symbolic phonetic tokens with its cognitive abilities. Rather, complex auditory cues are used to enable the agent to reproduce the perceived phonemes. We can infer from this reproduction that the agent is capable of the direct perception of speech through pattern recognition. Why has this separation between the constraints present within the articulatory gesture and auditory signal not taken place before? Perhaps because evidence for a highly constrained vocal tract has been assumed to be evidence for abstract gestures as the objects of perception. Accordingly, a highly-constrained acoustic signal has been assumed to be evidence for abstract phonetic tokens. We argue that this is not necessarily the case.

Direct realism supposes that speech is perceived directly, in the absence of any idealised abstract tokens—either phonetic or articulatory. To test this hypothesis, our agents have been embodied in a real-speech environment avoiding the current symbolic phonetic systems which force a (potentially-ungrounded) symbolic solution. To develop our theory from the phonetic to the syntactic level, and to avoid a reversion to ungrounded symbolism, we will need to ground the evolved phonemes in real speech and the evolved syntax in the real world. Thus, future work will develop robotic agents to test further our notions of DR within language. Ultimately, DR has lead us to believe that the continued modelling of language will require embodiment through the use of robotics.

References

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99(3):1730–1741.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, 101(6):3241–3254.

Ohala, J. (1996). Speech perception is perceiving sounds not tongues. *Journal of the Acoustical Society of America*, 99(3):1718–1725.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, 111(4):1872–1891.

Sussman, H. (1989). Neural coding of relation invariance in speech: Human language analogs to the barn owl. *Psychological Review*, 96(4):631–642.