Available online at www.sciencedirect.com
ScienceDirect
Hearing Research 229 (2007) 132–147
Hearing Research
www.elsevier.com/locate/heares

Research paper

Hearing speech sounds: Top-down influences on the interface between audition and speech perception

Matthew H. Davis [a,*], Ingrid S. Johnsrude [b]

[a] MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 7EF, UK
[b] Department of Psychology, Queen's University, Kingston, Ont., Canada

Received 6 September 2006; received in revised form 23 November 2006; accepted 3 January 2007
Available online 18 January 2007

Abstract

This paper focuses on the cognitive and neural mechanisms of speech perception: the rapid, and highly automatic processes by which complex time-varying speech signals are perceived as sequences of meaningful linguistic units. We will review four processes that contribute to the perception of speech: perceptual grouping, lexical segmentation, perceptual learning and categorical perception, in each case presenting perceptual evidence to support highly interactive processes with top-down information flow driving and constraining interpretations of spoken input. The cognitive and neural underpinnings of these interactive processes appear to depend on two distinct representations of heard speech: an auditory, echoic representation of incoming speech, and a motoric/somatotopic representation of speech as it would be produced. We review the neuroanatomical system supporting these two key properties of speech perception and discuss how this system incorporates interactive processes and two parallel echoic and somato-motoric representations, drawing on evidence from functional neuroimaging studies in humans and from comparative anatomical studies. We propose that top-down interactive mechanisms within auditory networks play an important role in explaining the perception of spoken language.

Keywords: Speech perception; Perceptual grouping; Lexical segmentation; Perceptual learning; Categorical perception; fMRI; Auditory cortex; Temporal lobe; Frontal lobe; Feedback

1. Introduction

You receive an unexpected call on your mobile phone. Despite the background noise on the line you immediately recognise your colleague's voice and can hear that she is excited about something. Catching her breath, she tells you that your joint grant application has been approved for funding and that you should meet to celebrate. In the space of a few seconds, this phone conversation has communicated a vital piece of information, conveyed the emotional significance of this news and provided physical information about the talker. While such exciting news is almost certainly not a daily occurrence, the cognitive and neural mechanisms that are at the heart of this scenario are so ubiquitous as to go largely unnoticed in our day-to-day life. We invariably focus on the information being communicated rather than the means by which it is conveyed, even in difficult listening situations.g.[1]

This paper will focus on the cognitive and neural mechanisms by which a complex time-varying acoustic signal is perceived as sequences of sounds that convey meaning; addressing precisely those stages of processing that occur so rapidly, automatically and effortlessly as to be beneath our notice. We suggest that a complete account of speech

* Corresponding author. Tel.: +44 1223 273 637; fax: +44 1223 359 062.
E-mail address: matt.davis@mrc-cbu.cam.ac.uk (M.H. Davis).

[1] We acknowledge that in much of our everyday experience, hearing speech is accompanied by seeing a talking face. However since the auditory modality provides the dominant input for speech perception, this paper will focus on mechanisms for perceiving heard speech. We consider visual information to be a source of valuable information qualitatively similar to the others that we explore, which all serve to tune and constrain interpretations of the speech input.

0378-5955/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.heares.2007.01.014

perception requires an understanding of both basic auditory and higher-level cognitive processes (see Plomp, 2001, for similar arguments). We will present evidence for an interactive processing system in which bottom-up and top-down processes combine to support speech perception. This interactive account provides mechanisms by which perceptual processing can rapidly change so as to optimally perceive and comprehend speech – including those important mobile-phone calls.

In the first section of the paper we will review behavioural evidence for interactive processes playing a critical role in speech perception. The background provided by these several decades of behavioural evidence must be accounted for by any neural account of speech perception and therefore constitutes the majority of the evidence presented here. Building on this behavioural evidence, the second section of the paper describes two types of representation that are integral to the implementation of an interactive account of speech perception. These multiple, parallel representations of the speech input make distinct contributions to the robustness of speech perception. In the third and final section of the paper we briefly review evidence from the anatomy of the auditory system that is consistent with this computational account, reviewing evidence both for interactive processes, and for multiple perceptual pathways.

2. Evidence for interactivity in speech perception

In this section, we will discuss four processes that contribute to speech perception: (1) perceptual grouping of speech sounds into a single coherent stream, (2) segmentation of speech into meaningful (lexical) units, (3) perceptual learning mechanisms by which distorted and degraded speech is perceived and comprehended, and (4) mechanisms for perceiving variable forms of speech in a categorical fashion. For each of these four cases we suggest that evidence supports highly interactive processes with top-down information flow often driving and constraining interpretation.

2.1. Perceptual grouping of speech

As shown in Fig. 1a, speech is a highly complex, rapidly changing acoustic signal. A number of very different acoustic elements (periodic sounds, aperiodic noise and silence) can be contained in a single spoken sequence. Yet, despite dramatic changes in spectral composition, and the lack of any obvious acoustic correlate of the somewhat regular rhythm that we hear in spoken sentences (Lehiste, 1977), we hear speech as a single coherent stream of sound, produced by a single source (the human voice). As Remez et al. (1994) have pointed out, many primitive auditory cues (such as dissimilarity of frequency, pitch or timbre, Bregman, 1990) would segregate a single spoken sentence into distinct acoustic elements corresponding to frication noises, isolated vowels, nasal formants, and so on. However, our knowledge of speech, along with other primitive cues (such as the harmonic structure of the vocal source) lead us to group these acoustically disparate sounds into one single stream.

The importance of experience in driving perceptual organization of speech is illustrated by considering the perception of clicks in spoken language. Clicks are plosive sounds created by the sudden release of a pocket of low-pressure air, trapped between the tongue and the roof of the mouth. Speakers of English and other languages in which clicks are non-linguistic typically assign these non-speech sounds to a different perceptual stream from the rest of speech. Thus, artificial clicks placed in sentences are difficult to locate in time with respect to the speech signal and are perceived as being displaced towards phrase boundaries (Fodor and Bever, 1965; Garrett et al., 1965). However, for speakers of sub-Saharan languages that include these sounds in their phonetic inventory, clicks are perceptually integrated into the stream of speech and tightly bound to the words in which they occur. This is a salient example of experience-driven or schema-governed grouping mechanisms (Bregman, 1990) that influence the perceptual organization of speech.

Experiments conducted by Liberman and colleagues illustrate the primacy of higher-level speech percepts in auditory perceptual organisation. They removed single formant transitions from synthetic CV syllables, so that the impoverished CV was ambiguous between a /da/ and a /ga/. They then presented the isolated formant transition either to the opposite ear (Mann and Liberman, 1983), or at an elevated amplitude in the same ear as the remaining portions of the speech signal (Whalen and Liberman, 1987). Listeners report hearing this isolated formant as a nonspeech chirp or whistle and yet simultaneously integrate this formant such that it influences the perception of the remaining fragments of speech in the same way as if it were perceptually-fused with the remaining speech. This duplex perception illustrates that humans organize the auditory world to yield speech whenever possible. This high-level schema-driven grouping mechanism 'trumps' the Gestalt principle of exclusive allocation in perception.

Another example of schema-based mechanisms dominating in the perception of speech comes from the 'migration' paradigm in which spoken materials are perceptually recombined between two concurrently presented sequences of syllables. Migrations frequently occur when the two sequences are presented in the same voice and in close spatial proximity (Cutting, 1975; Kolinsky and Morais, 1996). For instance, if the syllables "pay" and "lay" are presented simultaneously, one to each ear, listeners frequently report a fused syllable "play" (Cutting, 1975). Critically, migrations are not only affected by physical similarity between the two speech stimuli but also by higher-level, lexical properties. For instance, migrations are more common for pseudoword than word sequences (Mattys, 1997), and migrations that create illusory words (i.e. words that are not present in either of the two stimuli) occur more fre-

quently than those that create pseudowords (Kolinsky and Morais, 1992, reviewed in Kolinsky and Morais, 1996).

Similar influences of lexical knowledge occur in the 'verbal transformation effect' created by rapid and repeated presentations of short utterances at regular intervals (Warren and Gregory, 1958). For instance, hearing the word "spike" repeated regularly every 500 ms will, in time, evoke transformed speech percepts such as "spy", "spite", "bike", etc. Many of these verbal transformations result from spontaneous segregation of certain acoustic elements in speech into a separate stream (Pitt and Shoaf, 2002; Warren, 1968). For instance, participants experiencing the transformed percept "bike" for "spike" report hearing a burst of noise as a separate auditory stream and if provided with a speech editor can generate the same 'bike' percept by removing the frication noise from "spike" (Pitt and Shoaf, 2002). While changes in grouping can occur for non-speech sequences (Carlyon et al., 2001), research has shown that the frequency of verbal transformations depends on linguistic as well as physical properties of speech since real words like "spike" return to their veridical percept more frequently than nonwords like "spipe" (MacKay et al., 1993; Shoaf and Pitt, 2002). Thus evidence from migrations and verbal transformations indicate that language-specific knowledge of words influences the way in which speech sounds are grouped together in perception.

Another compelling example of top-down influences on perceptual grouping of speech comes from experiments exploring the intelligibility of sine-wave speech (Remez et al., 1981). As shown in Fig. 1b, sine-wave speech is created by synthesising sine-waves that track the formant frequencies of an utterance. Sine-wave

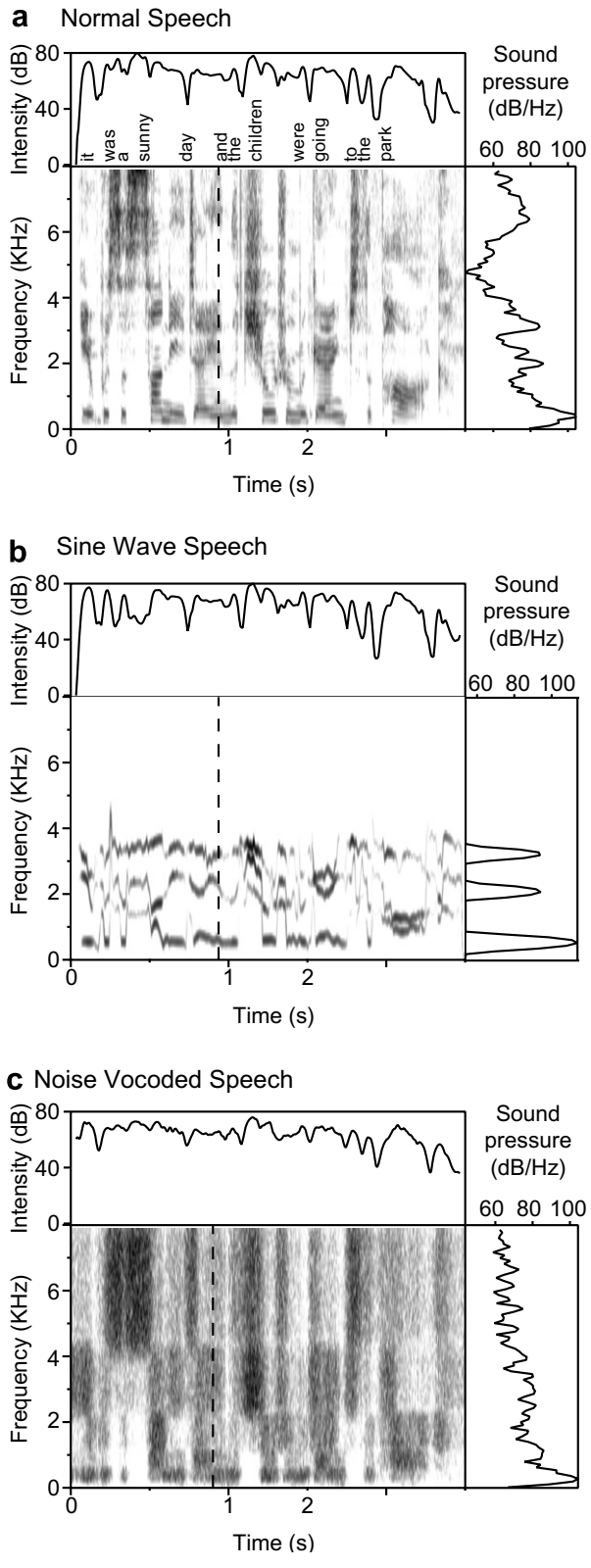

Fig. 1. (a) A broadband time-frequency spectrogram (middle panel) and intensity envelope (top panel) of the sentence "It was a sunny day and the children were going to the park". The approximate location of the onset of different words in the sentence is shown. The spectrogram shows the rapid alternations of periodic vowel sounds (containing both low frequency voicing and higher frequency formants), aperiodic noise (e.g. during the /s/ of "sunny"), and silence (during the closure of the /k/ in "park"). The left-hand panel shows the spectral profile, illustrating the formant frequencies present during the vowel-nasal transition of the word "day", marked with a vertical line in the spectrogram. (b) A sine-wave replica (cf. Remez et al., 1981) of the sentence "It was a sunny day and the children were going to the park", depicted as in (a). This stimulus was synthesised from the original recording using Praat software (version 4.4, <http://www.praat.org>) and a script written by Chris Darwin (http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts). As can be seen by comparing the spectrogram and spectral profile with (a), the centre frequencies of the three sine waves track the centre frequencies of the formants in the original sentence. For appropriately informed listeners, speech remains intelligible despite there being no trace of the original vocal source in the signal. Readers are encouraged to listen to the example sentence from (b), which can be heard online at http://www.mrc-cbu.cam.ac.uk/~matt.davis/sine-wave-speech/. (c) A noise-vocoded (cf. Shannon et al., 1995) version of the sentence "It was a sunny day and the children were going to the park", depicted as in (a). This stimulus was synthesised from the original recording using Praat software (as before) and a modified version of a script supplied by Chris Darwin (see Davis et al., 2005 for further details). NV speech contains modulated noise in six non-overlapping frequency bands that track the amplitude envelope of original speech (see (a) for comparison). Again, speech remains highly intelligible to trained listeners, despite all the spectral detail from the original sentence being replaced with noise. Examples can be heard online at http://www.mrc-cbu.cam.ac.uk/~matt.davis/vocode/.

speech lacks both harmonic structure and amplitude comodulation, two important cues for grouping speech formants into a single auditory stream. Naïve listeners report that sine-wave speech sounds like a number of independent whistles, reflecting the fact that these harmonically unrelated sinusoids do not perceptually cohere. However, a dramatic change in grouping, and hence perceptual experience, can be induced on the basis of top-down knowledge (Remez et al., 1981). If listeners are informed that what they are hearing is speech, or better still, told the identity of the original sentence, then these same sine-wave stimuli can evoke a clear and (somewhat) intelligible percept of a spoken sentence. Readers are encouraged to listen to example sine-wave speech stimuli which can be found online (http://www.mrc-cbu.cam.ac.uk/~mattd/sine-wave-speech/).

Effects of prior experience on perception have also been reported for other forms of distorted speech, such as noise-vocoded speech (Davis et al., 2005) and foreign-accented speech (Clarke and Garrett, 2004; Weill, 2003). The intelligibility of these sounds increases dramatically with experience – a form of perceptual learning that will be discussed in a later section. However, what is striking is that the perceived clarity of these forms of distorted speech can be almost immediately improved by providing information about the content of the sentence (much in the same way that written subtitles are sometimes provided to improve the clarity of poorly recorded interviews on television). Equally compelling results have been shown for speech presented in background noise. The perceived clarity of speech in noise is substantially improved by stimulus repetition (Jacoby et al., 1988), particularly if the same talker produces both first and subsequent presentations (Goldinger et al., 1999). Thus listeners estimate background noise to be quieter if they are familiar with the content of noise-masked spoken materials than if they are hearing novel materials, demonstrating an influence of top-down processes on more basic perceptual processes such as are involved in loudness estimation.

This section has presented a range of evidence suggesting that perceptual grouping of speech is driven not only by primitive grouping cues, such as similarity of pitch, timbre and timing, but also by powerful experience-driven mechanisms sensitive to high-level, linguistic, characteristics of speech such as lexicality, context and expectations. More primitive grouping cues also play a role (such as in the continuity illusion; Carlyon et al., 2002; Warren, 1970), however, the examples presented here show that these bottom-up cues can be overruled. Having discussing the means by which speech signals cohere into a single stream, we will next discuss the way in which the stream of speech is segmented by the listener into meaningful units such as morphemes, words and phrases. We propose that higher-level interpretation processes are the dominant influence on listeners' segmentation of connected speech.

2.2. Segmenting connected speech

In order to comprehend a sentence, listeners must recognise the individual words[2], access the meanings of these words, and combine them to compute the correct meaning. However, large linguistic elements such as words do not correspond to any discrete acoustic unit to be found in the speech signal: there are no simple, unambiguous analogues of the spaces, hyphens and other punctuation that appear between printed words. Furthermore, coarticulation produces dramatic changes to the pronunciation of words in different contexts. For instance, the final consonant of the word "stand" is pronounced differently in contexts such as "stand down", "stand back" or "stand close" (for discussion, see Gaskell and Marslen-Wilson, 1996). Nonetheless, when we listen to a spoken sentence our subjective experience is of hearing a sequence of discrete, invariant sounds that correspond to the meaning-carrying units stored in our mental lexicon. That this is an illusion created in the mind of the listeners is apparent if we consider the experience of hearing an unfamiliar foreign language. In the absence of comprehension, we hear a continuous stream of speech, without knowing where individual linguistic units begin and end. A substantial perceptual problem faced by the listener is therefore how to segment these relatively discrete linguistic units from a continuous stream of speech.[3]

A number of different segmentation cues have been distinguished in recent accounts of lexical segmentation (for reviews see Davis et al., 2002; Mattys et al., 2005). For instance, listeners can use acoustic cues that explicitly mark certain word boundaries (such as the acoustic differences that exist between pairs that differ in segmentation like "grey day" and "grade A" Nakatani and Dukes, 1977). Listeners also use statistical regularities in the sequence of speech sounds that tend to co-occur with word bound-

[2] There is a substantial debate in the literature on spoken language comprehension concerning whether the fundamental unit of lexical storage corresponds to a dictionary word, or whether complex words like "darkness" are broken down into smaller morphemic units (in this case, "dark" and "ness" – Marslen-Wilson et al., 1994b), and conversely whether even larger multiple-word units and phrases (e.g. "greasy spoon") are also stored (Harris, 1994). For ease of presentation this article assumes that single words are the primary stored unit while acknowledging that both larger and smaller units are often critical for correct interpretation.

[3] In the literature on speech perception, another, related segmentation problem – how the speech signal is divided into individual speech sounds or phonemes – has also been considered. It is often assumed that some categorically perceived element such as the phoneme is the primary unit of perceptual analysis of speech (Nearey, 2001). Yet, awareness of the phonemes in speech (e.g. breaking the word into "cat" into /k/, /æ/, /t/) is largely restricted to users of alphabetic written languages (Morais et al., 1979; Morais et al., 1986; Read et al., 1986) and it has been argued that phonemic awareness is a consequence of recognising individual words rather than a bottom-up process that precedes identification (Marslen-Wilson and Warren, 1994a). This paper will focus on the problem of segmenting larger, meaning-carrying units (such as words, morphemes or phrases) from spoken language, since comprehension necessarily requires accessing stored representations of these units.

aries, such as stressed syllables (Cutler and Carter, 1987; Cutler and Norris, 1988), or low-probability transitions between speech sounds (Cairns et al., 1997; McQueen, 1998). However, none of these cues appear to be sufficient (singly, or in combination) to identify all of the critical boundaries in connected speech (Brent and Cartwright, 1996; Christiansen et al., 1998). For this reason, accounts of speech perception incorporate mechanisms by which segmentation is achieved as a direct consequence of word recognition rather than as a necessary precursor to it. For instance, in TRACE (McClelland and Elman, 1986) and other competition-based models (e.g. Shortlist, Norris, 1994), inhibitory connections between similar sounding words ensure that the network settles into an state in which only mutually consistent segmentations of the input are activated. Similar mechanisms which achieve segmentation by identifying the most probable lexical interpretation of ongoing input are also incorporated into recurrent network (Davis, 2003) and symbolic AI models (Brent, 1997).

The presence of multiple, probabilistic and contextually dependent cues to segmentation in connected speech raises the problem of determining how listeners optimally combine potentially conflicting cues in order to identify ongoing input. Recent research on this topic provides evidence that the use of bottom-up, acoustic or statistical segmentation is largely confined to situations in which lexical information is absent, ambiguous or made unreliable by background noise or distortion (Davis et al., 2002; Mattys, 2004; Mattys et al., 2005). That is, multiple bottom-up segmentation cues are overlooked in favour of higher-level information sources when these are available (Mattys et al., 2005). This heuristic is reminiscent of the greater importance assigned to experience-driven, compared to primitive, grouping cues in determining the perceptual organisation of the speech stream. This observation similarly challenges any conventional assumption that bottom-up, stimulus-driven processes are the sole determinant of the segmentation and identification of words in speech. Instead, lexical segmentation arises as a consequence of the recognition of meaningful units, reflecting the ultimate goal of speech perception as being to perceive and understand entire spoken utterances rather than to identify individual units within each utterance (see Davis, 2003; Bybee and McClelland, 2005 for further discussion). We will next discuss influences of higher-level knowledge on lower-level processes in the context of perceptual learning. These top-down influences provide for robust speech perception in spite of degraded and variable speech input.

2.3. Perceptual learning of distorted speech

An important problem for accounts of speech perception and comprehension concerns how the recognition system tolerates variability in the heard form of speech. One example of this robustness has already been described: sine-wave (SW) speech (Fig. 1b) is surprisingly intelligible for appropriately informed listeners (Barker and Cooke, 1999; Remez et al., 1981). Robust comprehension of SW speech is particularly interesting given that another form of speech degradation; noise vocoding, which involves imposing the temporal envelope of speech onto a noise carrier, does not destroy intelligibility (Shannon et al., 1995, shown in Fig. 1c). There are few acoustic cues shared by both forms of intelligible speech: SW speech lacks broadband acoustic energy, but retains rapidly changing spectral cues for speech formants. NV speech consists entirely of slowly changing, broadband noises with little trace of speech formants. That both SW and NV speech remain intelligible is, we argue, due to the operation of perceptual learning mechanisms that enable listeners to comprehend forms of speech in which conventional speech cues are absent or degraded.

A number of demonstrations of perceptual learning of NV speech have been reported by our group (Davis et al., 2005; Hervais-Adelman et al., submitted for publication). Listeners presented with 6-band NV sentences were initially very poor at repeating back these sentences (reporting less than one word correctly from the first sentence), but report scores improved rapidly, such that for the last 10 sentences of a 30-sentence experiment, listeners reported around 50% of the words in each NV sentence.[4] Similar improvements have also been shown for listeners reporting isolated NV words (Hervais-Adelman et al., submitted for publication). The observation that learning generalises to words not previously heard as NV speech implies a sub-lexical locus for learning. We propose that listeners are retuning peripheral perceptual representations (e.g. acoustic-phonetic feature representations) that are shared among multiple lexical items.

In subsequent experiments, we (Davis et al., 2005) explored training conditions that significantly improved the efficacy of perceptual learning, enabling listeners to report over 75% of words correctly by the end of a 30-sentence experiment. Presenting each NV sentence twice significantly enhanced learning – particularly if the second presentation came after a clear (non-distorted) version of the same sentence (Clear then Distorted, or CD, feedback). NV speech is also heard more clearly when sentence content is known (cf. Jacoby et al., 1988; Goldinger et al., 1999, for speech in noise). Our results with NV speech show that this perceptual enhancement – or 'pop-out' effect – accompanies more rapid perceptual learning (Davis et al., 2005). Neither pop-out nor enhanced perceptual learning is observed if feedback on sentence content is provided after distorted speech (Distorted then Clear [DC]

[4] Although this finding is superficially similar to the poor comprehension reported for naïve listeners presented with SW sentences (Remez et al., 1981), the limited initial intelligibility of NV speech does not reflect a similar failure of perceptual grouping. NV speech is perceptually coherent, it is ordinarily heard as speech, and participants in the experiments of Davis et al. (2005) knew to expect speech and had previously understood a highly intelligible 30-band NV sentence.

feedback was equivalent to Distorted Only feedback). Furthermore, feedback need not be spoken: enhanced perceptual learning was also observed if feedback on sentence content was provided in written form (Written then Distorted feedback). It is therefore the high-level, linguistic content of the sentence and not clear presentation of its spoken form that is most important for enhancing perceptual learning. Effects of feedback manipulations have also been observed for perceptual learning of isolated NV words (Hervais-Adelman et al., submitted for publication): perceptual learning was again observed with CD feedback, but only marginal improvements of word report occurred with DC feedback.

Further evidence for higher-level influences on perceptual learning comes from studies in which the linguistic content of the training materials is manipulated. For NV sentences, Davis et al. (2005) showed that CD-feedback training with 20 NV sentences composed entirely of nonwords (e.g. "*cho tekine garund pid ga sumeun*"), produced no improvement of word report scores when listeners were subsequently tested on English NV sentences. Perceptual learning remained absent, even when written feedback was provided to support short-term memory representations of nonword sentences during training. The fact that listeners trained with English sentences learn, but those trained with nonword sentences do not, suggests that top-down support from lexical information is critical for retuning lower-level, prelexical representations. However, a similar experiment conducted with isolated words (Hervais-Adelman et al., submitted for publication) showed equivalent perceptual learning from NV word and nonword training blocks (both presented with CD feedback). These two findings can be reconciled by suggesting that feedback from non-lexical representations can support learning if a robust, phonological representation of the clear form of speech remains active when distorted speech is presented. However, for sentences, this phonological representation can be derived online since listeners can learn from single distorted presentations of NV English sentences, without external feedback (Experiment 1, Davis et al., 2005). We propose that ongoing prediction of upcoming words in combination with some form of internal, echoic memory is likely to be crucial for supporting perceptual retuning for sentences in the absence of external feedback. Hence, perceptual learning of NV speech from sentence stimuli is dependent on lexical information, even if learning from isolated nonwords can occur with appropriate external support. In general, we hypothesise that the presence or absence of external feedback may not be so crucial as the presence of some constraint on the interpretation of distorted speech that permits listeners to reinforce accurate perceptual hypotheses and make alterations that can correct inaccurate hypotheses. While many of the critical experiments remain to be done, it seems likely that other sources of external support such as visual speech (Thomas and Pilling, 2006), or semantic or pragmatic context, could also enhance perceptual learning.

Investigations of perceptual learning of other forms of artificially distorted speech (in particular time-compressed speech and sine-wave speech) also suggest that phonological and lexical representations are used to retune lower-level acoustic and phonetic processes in order to optimally perceive subsequent speech input (Barker and Cooke, 1999; Peelle and Wingfield, 2005; discussion in Davis et al., 2005). Results of experiments in which listeners perceive naturally occurring forms of speech variation, such as comprehending speech produced in an unfamiliar foreign or regional accent, suggest a similarly rapid, and effective form of perceptual learning (Clarke and Garrett, 2004; Maye et al., in press; Weill, 2003). These findings illustrate a role for perceptual learning mechanisms in training and tuning phonetic categories in naturally occurring speech. In the next section we will provide evidence that similar mechanisms are also involved in the categorical perception of speech.

2.4. Perceiving speech categorically

Categorical perception has long been seen as indicative of the interface between an analogue, continuously varying acoustic signal and the digital, symbolic properties of the linguistic objects conveyed by speech (see, Harnad, 1986). Classic demonstrations of categorical perception arise when listeners are presented with speech tokens that are acoustically intermediate between naturally produced syllables. For example, naturally produced speech with a 60 ms delay between a bilabial release burst and the onset of voicing (voice onset time, VOT), will be heard as a token of *pay*; without this delay the syllable *bay* is heard. Speech synthesised with an intermediate VOT value is perceived categorically as one or other syllable. Listeners report hearing artificial syllables with a VOT below 20 ms as *bay*, and syllables with a VOT above 40 ms as *pay*, with a discrete transition at some intermediate VOT value. Critically, discrimination between syllables with different VOT values (for instance, in an ABX task) is limited; two stimuli that fall on opposite sides of the category boundary are readily discriminated whereas two stimuli with an equivalent physical difference that fall on the same side of the category boundary are more difficult to discriminate. Such phenomena seem to reflect a very different mode of perception than is classically observed for judgements of (say) the pitch or loudness of stimuli, for which discrimination performance comfortably exceeds categorisation (MacMillan, 1986).

We have described categorical perception in the context of studies which assess the influence of a single acoustic cue (VOT) on the perception of specific phonetic contrasts (/b/ vs /p/). However, speech contrasts are not conveyed either by single acoustic cues or invariant combinations of acoustic cues in natural speech. For instance, in studies of the perception of fricative-stop syllables (such as *star*, *spar*, *scar*, etc) all of the acoustic differences measured in natural productions of these syllables can potentially inform perception of the stop consonant (Bailey and Summerfield,

1980). No single cue was necessary and many different cues were sufficient for correct perception. This result is difficult to reconcile with there being any stable, context-independent acoustic elements in speech which serve to cue perception. Instead, accumulating evidence would suggest that top-down processes are responsible for both generating and maintaining categorical perception in the face of the variability that is encountered in the speech input.

Evidence in support of these top-down processes has come from the observation that phonetic category boundaries are altered by the experimental and linguistic context in which stimuli are presented. Shifts in category boundaries occur if an unbalanced or restricted range of stimuli are used (such as only a subset of natural VOT values Brady and Darwin, 1978; Rosen, 1979). Similar (but more pronounced) shifts occur for boundaries between vowel categories (Sawusch and Nusbaum, 1979). Studies also reveal effects of apparent speech rate on the perception of speech segments (Miller, 1981; Miller and Liberman, 1979). Strong influences of lexical context on phonetic categorisation have also been observed. The widely studied "Ganong effect" (Ganong, 1980) refers to a shift in the category boundary depending on the lexical status of the context in which a phonetic segment is placed. For instance, a segment that is ambiguous between /g/ and /k/ may be perceived as /g/ in a context that forms the word *gift* or as /k/ in a context that forms *kiss* (Ganong, 1980; Pitt and Samuel, 1993). Although controversial, one common interpretation of the Ganong effect is that it reflects top-down influences of lexical information on lower-level perceptual processes (Elman and McClelland, 1988; Pitt and McQueen, 1998; Magnuson et al., 2003; McClelland et al., 2006).

Perhaps the most persuasive demonstrations of top-down influences on categorical perception have come from recent demonstrations of perceptual learning of phonetic category boundaries. Norris et al. (2003) showed long-lasting influences of exposure to Ganong-type stimuli with ambiguous, word-final segments. Two groups of participants heard words containing an ambiguous fricative midway between an /f/ and /s/, in contexts that either favoured an /f/ interpretation (at the offset of words like *cliff* or *beef*) or an /s/ interpretation (words like *kiss* or *peace*). After training, listeners changed their interpretation of an ambiguous fricative presented in isolation – favouring the segment that was lexically biased during training. Importantly, a similar shift in perception was not observed for participants that heard the same ambiguous fricative in contexts in which lexical information did not favour one or other interpretation (e.g. nonwords such as *driff* or *driss*). This form of perceptual learning persists for at least twelve hours after initial training (Eisner and McQueen, 2006), as long as stimuli that bias against newly learnt interpretations of ambiguous segments are not presented in the meantime (Kraljic and Samuel, 2005). Even previously sceptical authors have interpreted this finding as evidence for top-down, lexically guided retuning of lower-level perception (Norris et al., 2003).

The learning mechanism implied by these perceptual-learning studies has properties that are strongly reminiscent of the top-down process that was described for perceptual learning of noise-vocoded speech (Davis et al., 2005): both occur quickly (within a matter of minutes), require exposure to only a handful of distorted stimuli (as few as 20 items in some studies) and are dependent on lexical or other higher-level information in the input.[5] In one respect, however, the conclusions drawn from studies of noise-vocoded speech go beyond those of studies using phonetically ambiguous stimuli. Davis et al. (2005) demonstrated that feedback conditions that altered perception of a distorted sentence ('pop-out') also enhanced perceptual learning. A parsimonious interpretation is that a top-down feedback process, in which the bottom-up perception of distorted or degraded input is retuned on the basis of comparisons with higher-level linguistic representations, is responsible for both immediate shifts in phonetic categorisation (such as the Ganong effect), and for the longer-lasting changes in phonetic categorisation observed in perceptual-learning studies (Eisner and McQueen, 2006; Kraljic and Samuel, 2005; Norris et al., 2003).

Additional evidence that top-down mechanisms are responsible for categorical perception comes from experiments in which the time course of categorical effects is explored. Although categorical perception leads to poor discrimination in ABX tasks, sensitivity to within-category variation can be shown under certain circumstances. Participants performing a speeded 'same-different' task are faster to respond 'same' if two stimuli are both acoustically and phonetically identical than if they are acoustically different yet phonetically identical (Pisoni and Tash, 1974). Pisoni suggests that an initial acoustic stage of encoding precedes phonetic perception and can speed these 'same' responses if an acoustic match occurs. Further experiments by Howell (Howell, 1978; Howell and Darwin, 1977) show that this initial acoustic representation decays rapidly – an RT advantage for acoustically identical 'same' presentations was not observed for an ISI of greater than 500 ms. Further evidence that categorical perception builds up over time comes from studies that use speech-contingent eye movements (McMurray et al., 2002) or cross-modal priming (Andruski et al., 1994; Marslen-Wilson et al., 1996) to assess ongoing interpretations of speech stimuli that include non-prototypical segments. Initial perceptual processing is clearly influenced by the presence of within-category phonetic variation that does not lead to differences in later perceptual awareness. Similar results abound in other psycholinguistic domains – fine phonetic detail (that is, variation that occurs between stimuli that are all perceived as exemplars of a single phonological category; Hawkins, 2003) influences lexical segmentation (Davis

[5] We can distinguish between a rapid learning process that retunes existing phoneme categories and the much slower and more effortful learning that is required to learn a new phoneme category; for instance, in learning a foreign language (Logan et al., 1991; McCandliss et al., 2002).

et al., 2002; Salverda et al., 2003), perception of coarticulation (Marslen-Wilson and Warren, 1994a) and morphological parsing (Kemps et al., 2005), particularly if online measures of lexical activation such as cross-modal priming and speech-contingent eye tracking are used to assess listeners' interpretations. Yet at the same time, listeners' subjective experience of speech remains largely categorical; listeners are typically unaware of the input differences that drive these perceptual effects. Thus, categorical influences on perception gradually emerge over the time-course of recognition, and are accompanied by an apparent decline in the influence of representations of fine phonetic detail as originally demonstrated by Pisoni and Tash (1974).[6]

This body of work suggests that categorical perception of speech is generated by bottom-up and top-down processes working in concert. Initial bottom-up processes represent the full detail of the speech input, activating possible interpretations at multiple levels of representation. Recognition involves a multiple-constraint satisfaction process that selects the most appropriate interpretation of the current input from the activated set. Following recognition, top-down processes ensure that bottom-up stimulus-driven processes are retuned to generate 'correct' recognition with reduced competition from inappropriate interpretations. This retuning ensures that the perceptual system is optimally configured to efficiently comprehend subsequent, similar speech. These top-down processes also produce categorical perceptual awareness with only one 'winning' interpretation remaining active. These interactions between acoustic-phonetic and higher-level lexical and phonological processes allow for invariant recognition of varying forms of spoken input. Hence, categorical perception of speech, far from being achieved by bottom-up processes at an early stage of the perceptual system, is an emergent property that arises from complex interactions between higher-level interpretive and lower-level acoustic processes.

This combination of bottom-up and top-down information flow is reminiscent of that proposed in the TRACE model of speech perception (McClelland and Elman, 1986), and recent modifications of that model to incorporate Hebbian learning processes (Mirman et al., in press). Similar processes are also suggested for distributed connectionist models such as the Distributed Cohort Model (DCM, Gaskell and Marslen-Wilson, 1997) which incorporate back-propagation or other error-driven learning algorithms. In both cases, lower-level perceptual computations are altered on the basis of top-down influences. In the next section of the paper we consider what computational processes are required by this interactive account of speech perception.

[6] Although some recent studies have demonstrated long-term memory representations for non-phonetic aspects of previously heard words (e.g. Goldinger, 1996; Luce and Lyons, 1998), existing evidence suggests that long-term effects of fine phonetic detail arise primarily in tasks tapping explicit memory rather than word recognition (Luce and Lyons, 1998; Pallier et al., 2001) and that observed influences of fine phonetic detail decay rapidly (Goldinger, 1998). While it might be suggested that memory representations for voice identification retain certain forms of acoustic detail, it is likely that speaker recognition itself depends on highly abstract, non-acoustic representations (e.g. Remez et al., 1997; Sheffert et al., 2002).

3. Computational requirements for interactive processes in speech perception

We have reviewed four domains in which top-down processes appear to contribute to speech perception: in promoting perceptual grouping, in achieving lexical segmentation, in supporting perceptual learning of distorted speech, and in maintaining categorical perception of speech segments. In this section, we will address the computational implications of such interactions and suggest that: (1) top-down influences act on auditory, echoic representations of incoming speech, and (2) top-down influences (in part) arise from the interface between speech perception and speech production.

3.1. Auditory and echoic representations of speech

Because of the sequential nature of speech, higher-level interpretations cannot be immediately derived from the input. For instance, in order to recognise the word 'cabbage', information in the second syllable of the word must be perceived so as to rule out other words that start with the same syllable ('cabin', 'cabaret', etc, cf. Marslen-Wilson, 1984). Therefore, lexical influences on the perception of the initial phoneme of a word (as in speech migrations, or the Ganong effect) require some temporary storage or buffer to maintain ongoing spoken input until top-down information is available. Maintenance of auditory representations of incoming speech was also proposed in accounting for perceptual learning of distorted speech and categorical perception data. Evidence suggests that both higher-level interpretations and lower-level input representations must be simultaneously available in order to support learning (see for instance, the effects of feedback order reported by Davis et al., 2005). These findings therefore motivate an account in which a relatively unanalysed, auditory representation of the speech input is transiently maintained until words can be recognised, and top-down processes can arise.

The proposal that speech perception relies on transient storage of the incoming signal is consistent with an extensive literature on auditory echoic memory for verbal materials. In short-term memory studies, it is often reported that the last stimulus in a list of spoken materials is better remembered than preceding items in the list (Crowder and Morton, 1969). This auditory recency effect is not observed if the same materials are presented visually (Crowder and Morton, 1969; Watkins and Watkins, 1980). The auditory recency effect is also readily disrupted by the presentation of irrelevant sounds at the offset of the list, a suffix effect that decays over time, but is absent if distracting sounds are assigned to a different perceptual stream (Frankish, 1989; Morton et al., 1971). Such studies provide evidence

for a transient, rapidly fading auditory store ('echoic memory') that can support later recall of auditory materials. It has long been thought that echoic memory plays an important role in speech perception, both in supporting non-categorical comparisons among spoken materials (Crowder, 1983; cf. Pisoni and Tash, 1974), and in providing an auditory record that permits the processing of longer-lasting, supra-segmental structures in connected speech (Frankish, 1989). In line with this proposal, we suggest that ongoing maintenance of auditory information, at multiple levels of representation, plays an important role in permitting top-down information to influence perceptual processing of speech.

3.2. Speech perception is tied to speech production

Throughout this paper we have suggested that top-down support for lower-level perceptual processing comes from systems involved in deriving meaning from speech. In particular, we have reviewed evidence of a role for processes that recognise familiar words. However, the mental lexicon embodies many different kinds of information concerning the heard, spoken and written form of words, as well as associated meanings and syntactic functions. Although all of these representations probably play some role in constraining speech perception, a number of authors have argued that links between spoken input and motoric representations involved in speech production have a privileged status in speech perception (Liberman and Whalen, 2000; Liberman et al., 1967; Rizzolatti and Arbib, 1998). We would extend this somewhat, and add that such motor representations have strong somatosensory correlates (cf. Guenther et al., 2006); for this reason, we will refer to them together as 'somatomotor representations'. We will review a range of evidence indicating that somatomotor representations are implicated in speech perception. We suggest that speech perception may, in part, be driven by processing interactions between auditory and higher-level somatomotor representations of speech (see Poeppel et al., in press, for a similar view).

Links between perception and production develop in infancy through babbling; a process that infants use to tune their production of speech to match the speech sounds that they perceive in their linguistic environment (Doupe and Kuhl, 1999; Kello and Plaut, 2004; Werker and Tees, 1999). These links play a central role in the development and maintenance of categorical representations of speech. The developmental stage at which babbling is observed coincides with a decline in perceptual sensitivity to non-native phonetic contrasts, and increased robustness of categorical perception for native speech contrasts (for a review see Kuhl, 2004). Auditory and somatosensory feedback during speech production also plays an important role in controlling speech production in adults (Guenther et al., 2006; Houde and Jordan, 1998; Perkell et al., 1997; Purcell and Munhall, 2006). Phonetic alterations to auditory feedback lead speakers to systematically change their productions so that they perceive their own speech correctly (Houde and Jordan, 1998; Purcell and Munhall, 2006). This finding suggests that interactions between perception and production help ensure phonetic constancy in production. In adults, links between perception and production are also central to models of short-term memory, in which an articulatory loop provides for maintenance of verbal materials through overt or covert rehearsal processes (Baddeley, 1986). This longer-term maintenance of verbal stimuli (in contrast to the short-term, echoic processes described previously) operates equivalently on visual and auditory input, and necessarily implies the loss of fine phonetic detail, such that only categorical interpretations of speech are retained (cf. Hartley and Houghton, 1996).

Links between speech perception and production also promote parity of phonological representations between conversational partners. We have described perceptual mechanisms that allow listeners to adjust their speech perception system in order to optimally perceive speech with unfamiliar or mismatching phonetic-category boundaries (cf. Norris et al., 2003). In addition to these perceptual adjustments, research has documented a number of ways in which speakers spontaneously and involuntarily imitate heard speech (see Goldinger, 1998; Krauss and Pardo, 2006; Shockley et al., 2004). One salient form of adjustment occurs when we find ourselves (partly) adopting the accent of someone that we are talking with, although more subtle forms of adjustment are also observed (Shockley et al., 2004). Over the course of several exchanges, conversational partners tend to converge on maximally similar speech patterns (Krauss and Pardo, 2006), aligning their phonetic representations (see Garrod and Pickering, 2004; Pickering and Garrod, 2004, for a more general discussion of alignment). This process assists communication by ensuring that, in ongoing conversation; both speakers use a common, mutually intelligible phonetic code (i.e. promoting parity between speakers, Liberman and Whalen, 2000; Rizzolatti and Arbib, 1998). Parity between speakers is achieved through bidirectional interactions between speech perception and speech production mechanisms within each speaker. Hence, interactions between auditory, echoic processes involved in the perception of speech and somatomotor representations that are involved in speech production help ensure that heard speech is perceived categorically, and that produced speech successfully evokes a categorical percept in the mind of the listener.

4. Towards a neuroanatomical account of speech perception

This section will discuss the neural basis of the two central propositions that we make concerning speech perception: (1) that bidirectional, interactive connectivity allows higher-level constraints to influence ongoing speech perception and support the rapid retuning of perceptual processes, and (2) that parallel processing pathways support both an auditory-echoic record of incoming speech and the mapping of heard speech onto somatomotor

representations involved in speech production. In attempting to map speech perception onto neuroanatomical pathways for processing auditory information, we must necessarily contend with the fact that much of what is known about the anatomical and functional organization of the human auditory system can only be inferred from work in non-human primates. This is obviously problematic when considering speech perception, since large regions of cortex that are important for speech perception (particularly the superior temporal sulcus and middle temporal gyrus) have unknown homologies in those species in which anatomy and physiology have been well studied. Nonetheless, we believe that spoken language evolved out of a capacity, shared with other species, for vocal and gestural communication and therefore general principles of anatomical organization for auditory and gestural systems can be inferred from work in non-human primates and other species with auditory specializations (such as owls, bats, and cats). These studies provide a framework within which the findings of human functional imaging studies can be interpreted.

4.1. Anatomical organization consistent with top-down influences on speech perception

The traditional view of sensory-neural processing pathways is that they are largely unisensory and feedforward. Processing begins in a peripheral receptor array which is mapped topographically in the brain (Weinberg, 1997). Subsequent stages of processing are essentially integrative, involving computation of larger and more complex receptive fields until the objects of perception are achieved (Felleman and Essen, 1991; Winer, 2006). In the auditory domain, this classic view comes up short against the observations of descending projections – as massive and specific as the feedforward connections – at all levels (de la Mothe et al., 2006; Diamond et al., 1969; Petrides and Pandya, 2006). Descending projections carry information from higher processing centres to lower ones, potentially from higher-order cortical regions, through primary cortex, thalamus and brainstem, all the way back to the cochlea (Huffman and Henson, 1990; Winer, 2006; Xiao and Suga, 2002). These descending projections are also multiple and organized in parallel (Huffman and Henson, 1990; Winer, 2006), and many demonstrations of subcortical and primary auditory cortical plasticity have been attributed to the action of descending connections, in various species including primates (Fritz et al., 2005; Gao and Suga, 2000; Perrot et al., 2006; Xiao and Suga, 2002).

Descending connections from higher- to lower-level areas are such an important feature of auditory organization that such connections can exist without feedforward homologues (de la Mothe et al., 2006; Huffman and Henson, 1990; Winer, 2006). For instance, a recent anatomical study of marmoset auditory cortex (de la Mothe et al., 2006) used tracer injections to reveal the connectivity of primary auditory cortex, and demonstrated direct projections from parabelt to primary auditory cortex (A1). This is surprising, since feedforward connections between A1 and parabelt are exclusively indirect, via an obligatory relay in belt cortex (Hackett et al., 1998). The functional implication is that, whereas bottom-up processing proceeds in a sequential, cascaded fashion, with information flowing through the system one cortical stage at a time, top-down processes have privileged access and can influence lower-level processes directly. These descending connections likely play an important role in achieving rapid, task-related plasticity in auditory cortex (see Fritz et al., 2005), By extension, descending connections may play an important role in producing the rapid perceptual tuning responsible for the robust perception of highly variable forms of speech.

At present, however, there is only sparse evidence from human functional imaging to support equivalent, interactive processes in speech perception. The relatively poor temporal resolution of fMRI does not permit us to unambiguously distinguish bottom-up and top-down processes in speech perception, however, two sets of fMRI findings do provide some preliminary evidence consistent with activity in higher-order frontal regions modulating activity in lower-order temporal regions. First, presentation of distorted, yet still intelligible, speech leads to increased activity in frontal regions that is accompanied by an increased response in peri-auditory regions (Davis and Johnsrude, 2003; Giraud et al., 2004). That is, both frontal and peri-auditory regions show an elevated response to speech stimuli when listeners exert more effort to perceive these stimuli. This finding suggests top-down influences from frontal regions on peri-auditory responses since basic, acoustic properties of the stimuli were either very well matched (Davis and Johnsrude, 2003) or identical but with changes to listeners expectations (Giraud et al., 2004). A second source of evidence comes from an fMRI study (Davis et al., in preparation) in which speech and non-speech stimuli were presented to participants at varying levels of awareness (using controlled administrations of a sedative drug, propofol). Increased sedation produced a dramatic decline in prefrontal and premotor responses to speech. Reduced frontal activity was accompanied by a moderate decline in temporal lobe responses to speech, despite responses to non-speech noises being unaffected by sedation. These fMRI observations are consistent with the conclusion that changes in frontal lobe responses can modulate temporal lobe responses to speech through top-down influences on lower-level auditory processes.

We anticipate that electrophysiological techniques (EEG and MEG), which provide better temporal resolution, will offer a greater opportunity for observing top-down influences on processing in lower areas. The temporal resolution of both EEG and MEG is sufficient to permit classification of processes, based on joint consideration of their anatomical locus and the time they occur, as either feedforward (bottom-up) or feedback (top-down) processes. However the time frames reflective of 'early' and

'late' processing are a topic of some debate (e.g., Foxe & Simpson, 2002) even in the visual literature in which the onset of stimulation and the onset of information processing necessarily coincide. Since speech unfolds over time, it is difficult to align evoked neural responses to information-carrying acoustic events for anything other than very simple and frequently repeated speech stimuli. Resolving this alignment problem presents an obstacle to using timing information to distinguish bottom-up and top-down processing of speech and suggests that assessment of non-time locked induced EEG/MEG responses will be valuable (see Ahissar et al., 2001).

The extensive network of connections that we have documented among various levels in the auditory system (core, belt, parabelt and beyond), may support mechanisms by which higher-level interpretations of speech are tested against incoming auditory information. In visual object perception, it has been suggested that connections between primary and higher-order cortices implement a Bayesian inference process in which prior knowledge about the visual environment is combined with image features to determine the most probable interpretation of current input (Kersten and Yuille, 2003). We propose that similar mechanisms are involved in speech perception, selecting the most likely interpretation based on the combination of multiple sources of evidence, including phonological, lexical and syntactic information about speech input. These sources of information combine to constrain ongoing perception, and through the operation of learning mechanisms, support perceptual retuning that helps the speech system to adapt to novel and changing linguistic environments.

4.2. Multiple, parallel processes in speech perception

We have proposed that two parallel computational mechanisms are critical for an interactive account of speech perception: transient echoic representations of incoming speech, and tight coupling with neural mechanisms that support speech production. The role of these two systems does not map in any simple way onto single cognitive processes involved in speech perception (e.g. segmentation, categorisation, etc); rather interactions between these two processes are critical for successful perception. In this section we will review evidence that supports the idea that multiple parallel networks are critical for speech perception, and that one putative functional distinction among these multiple pathways is to segregate echoic and somatomotor representations of speech.

Processing of auditory information is highly parallel at multiple levels of the primate auditory system. Even in the earliest cortical receiving areas, multiple different representations of the input are available (Jones, 2003). The organization of the cortical auditory system is likewise parallel and cascaded, with hierarchical connections among auditory core, belt, and parabelt areas suggesting at least three discrete levels of processing (Hackett and Kaas, 2004; Kaas et al., 1999). A distributed, interconnected, set of fields in superior temporal gyrus and sulcus, in the inferior parietal lobule, and in prefrontal cortex, receive inputs from belt and parabelt; constituting a potential fourth stage of processing (Hackett and Kaas, 2004).

Within this network, recent accounts emphasise two main processing pathways that radiate out from primary auditory regions on the superior temporal plane (Hackett and Kaas, 2004; Hackett et al., 1999; also see Petrides and Pandya, 1988; Romanski et al., 1999a). The 'dual-stream' account is based in the observation that temporal, parietal and frontal connections of macaque auditory cortex are topographically organized. Anterior belt, and parabelt and associated anterior temporal-lobe regions inter-connect with anterior and ventral frontal cortical sites (the ventral auditory stream). In contrast, more posterior belt, parabelt and associated posterior temporal regions inter-connect with more posterior and dorsal frontal cortical sites (the dorsal auditory stream) (Hackett et al., 1999; Petrides and Pandya, 1988; Romanski et al., 1999a; Romanski et al., 1999b; Seltzer and Pandya, 1989). These two routes appear to converge in dorsolateral regions of frontal cortex in the macaque, and in the frontal eye fields and in area 46, known to be important for working memory.

Despite difficulties in determining the homologies of these auditory pathways in humans, functional connectivity analyses of fMRI responses to spoken sentences (Johnsrude et al., in preparation) and to spoken or written words during short-term memory tasks (Buchsbaum et al., 2005) have highlighted trial-by-trial correlations between activity in anatomically distant, but putatively connected frontal- and temporal-lobe regions. In particular, these studies show strong correlations between activity in anterior regions of the superior/middle temporal gyri and ventral, anterior frontal regions, and between posterior regions of the superior temporal gyrus and more posterior and dorsal frontal regions. Since a likely cause of correlated neural activity is underlying anatomical connections, these functional connectivity data in humans suggest anatomical connections similar to those that we have reviewed in non-human primates. These results therefore suggest apparent homologies between macaque and human processing pathways and provide support for accounts that propose functional specialisation of dorsal and ventral processing streams for speech perception (Hickok and Poeppel, 2004; Scott and Johnsrude, 2003).

The anterior auditory pathway in humans may be responsible for maintaining auditory-based representations of speech (Buchsbaum et al., 2005), and other sounds (cf. Price et al., 2005). We propose that more anterior regions of the temporal lobe construct progressively higher-level representations of the speech stream, spanning longer stretches of spoken input, with greater abstraction from auditory form representations. Processes such as lexical segmentation, prosodic and syntactic analysis that are necessary for sentence comprehension have a greater

requirement for long-term integration of information than does perception of isolated words or syllables (Mattys, 1997; Rosen, 1992). This requirement is therefore consistent with observations of additional anterior temporal activation for sentences, compared to isolated words or randomly ordered word lists (Friederici et al., 2000; Humphries et al., 2005, 2006; Indefrey and Cutler, 2004).

Functional imaging studies that have assessed responses to distorted utterances of varying intelligibility also report elevated responses in anterior temporal regions (Davis and Johnsrude, 2003; Narain et al., 2003; Scott et al., 2000). Those studies that contrast different forms of distorted speech suggest progressively greater abstraction from acoustic properties of speech at more anterior processing stages (Davis and Johnsrude, 2003; Johnsrude et al., in preparation). Within this hierarchically organised anterior network, successive processing stages achieve greater abstraction from the acoustic input, while maintaining multiple possible interpretations of incoming speech. In line with the account that we derived from behavioural data, we propose that interactions with the dorsal-stream network ensure that speech is perceived categorically, and these top-down influences are responsible for the decay of echoic representations of the acoustic-phonetic detail of speech.

A number of accounts have proposed that the posterior auditory pathway plays a critical role in integrating heard speech with systems involved in speech production (Hickok and Poeppel, 2004; Scott and Johnsrude, 2003). Evidence from other mammals suggests that somatosensory rather than motor correlates of vocal production converge with auditory information at peripheral levels in the central nervous system. Somatosensory inputs from systems controlling respiration and vocalization are observed in the mammalian cochlear nucleus (Shore and Zhou, 2006). Furthermore, caudomedial regions of belt cortex in primates receive feedforward somatosensory inputs and neurons in these regions are responsive to somatosensory stimuli (de la Mothe et al., 2006; Schroeder et al., 2003). However links to the motor system are also observed: core, belt, and parabelt regions all project into the dorsal caudate and putamen – components of the basal ganglia which are traditionally considered to serve a primarily motor function (Yeterian and Pandya, 1998). Finally, a complex set of links interconnects auditory belt and parabelt with superior temporal sulcus, multiple sites in the inferior parietal lobule (anterior supramarginal gyrus), frontal areas including BA 46, 8, 45, and 44, premotor and motor cortex (Petrides and Pandya, 2002; Rozzi et al., 2006; Seltzer and Pandya, 1991).

Physiological and neuroimaging data from humans also support rapid links between auditory perception and vocal production. Depth-electrode stimulation and electrophysiological recording in neurosurgical patients demonstrate activity in circuit involving primary auditory cortex, a posterolateral region of the superior temporal gyrus, inferior frontal gyrus (pars triangularis and opercularis) and orofacial motor cortex (Brugge et al., 2003; Greenlee et al., 2004; Howard et al., 2000). A host of recent neuroimaging and TMS studies also provide evidence that motor regions are active during speech perception (Fadiga et al., 2002; Pulvermuller et al., 2006; Uppenkamp et al., 2006; Watkins and Paus, 2004; Watkins et al., 2003; Wilson et al., 2004), with various authors proposing that coupling between auditory and motor activity is modulated by activity in posterior temporal and inferior frontal systems (Watkins and Paus, 2004; Wilson and Iacoboni, 2006; Wise et al., 2001).

We propose that somatomotor representations may be critically involved in transforming the echoic representations that encode the acoustic/phonetic detail of heard speech into categorical representations, suitable for articulation. Indeed fMRI studies demonstrate that stimulus items that evoke distinct phonological categories produce an additional response (compared to repeated or non-categorically perceived stimuli) in the posterior superior temporal gyrus, angular gyrus and supra-marginal gyrus (Callan et al., 2003; Golestani and Zatorre, 2004; Jacquemot et al., 2003; Wilson and Iacoboni, 2006). Similar areas also show activation for the contrast between naïve and trained responses to sine-wave stimuli (Dehaene-Lambertz et al., 2005; Mottonen et al., 2006). These results implicate posterior temporal and inferior parietal systems in categorical perception of speech. We note, however, that studies of categorical perception typically employ short, meaningless syllables which perhaps explains why less activity is reported in anterior temporal regions than in speech perception studies that use sentence-length stimuli (e.g. Davis and Johnsrude, 2003; Scott et al., 2000). These results therefore also suggest a functional distinction that we drew earlier between the initial use of fine phonetic detail to access the meaning of spoken words (involving auditory echoic representations within the anterior portions of the superior and middle temporal gyri), and later awareness of the sounds of speech, which is largely categorical in nature and arises through interactions between posterior temporal and inferior frontal regions.

It has been suggested that motor activity during speech perception reflects the activation of articulatory representations which permit the listener to derive the intended spoken gestures of the speaker (Hickok and Poeppel, 2004; Scott and Johnsrude, 2003); this proposal is reminiscent of ideas long associated with the motor theory of speech perception (Liberman and Whalen, 2000; Liberman et al., 1967). As we have described, modifications to speech representations that are jointly involved in speech perception and production provide for parity; the development of a shared code between speaker and listener, essential for successful speech communication (Liberman and Whalen, 2000; Rizzolatti and Arbib, 1998). While we agree with this claim, we would also suggest that the multiple parallel paths by which speech information can be processed indicate that higher-level linguistic interpretations can be computed from and tested against several different kinds of

representations (auditory-echoic, auditory-somatosensory, and auditory-motor) simultaneously. Speech perception likely proceeds by reconciling interpretations generated on multiple time scales, at multiple linguistic levels (including lexical, semantic and syntactic), and in multiple representational domains.

5. Concluding remarks

> *"Whereas elementary functions of a tissue can, by definition, have a precise localization in particular cell groups, there can of course be no question of the localization of complex functional systems in limited areas of the brain or of its cortex."* Luria (1976), p. 30.

In this paper we have proposed a multiple-pathway account of auditory processes that are critically important for a complex and uniquely human function – the comprehension of spoken language. As the quotation from Luria indicates, we must account for complex interactions among multiple brain areas in order to begin the task of providing a detailed neuroanatomical account of speech perception. Cognitive and behavioural explorations of speech perception converge with neuroscientific evidence in suggesting that interactions between higher-level linguistic knowledge and bottom-up perceptual processes are necessary for successful speech perception.

Acknowledgements

Preparation of this paper was supported by the UK Medical Research Council, and the Canada Research Chairs program. We thank Maggie Kemmner, Sarah Hawkins and two anonymous reviewers for comments on an earlier draft of the paper.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M.M., 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc. Natl. Acad. Sci. USA 98, 13367–13372.

Andruski, J.E., Blumstein, S.E., Burton, M., 1994. The effect of subphonetic differences on lexical access. Cognition 52, 163–187.

Baddeley, A.D., 1986. Working Memory. Clarendon Press, Oxford, Oxfordshire.

Bailey, P.J., Summerfield, Q., 1980. Information in speech: observations on the perception of [s]-stop clusters. J. Exp. Psychol. Hum. Percept. Perform. 6, 536–563.

Barker, J., Cooke, M., 1999. Is the sine-wave speech cocktail party worth attending? Speech Commun. 27, 159–174.

Brady, S.A., Darwin, C.J., 1978. Range effect in the perception of voicing. J. Acoust. Soc. Am. 63, 1556–1558.

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambridge, MA, USA.

Brent, M.R., 1997. Towards a unified model of lexical acquisition and lexical access. J. Psycholinguist. Res. 26, 363–375.

Brent, M.R., Cartwright, T.A., 1996. Distributional regularity and phonotactic constraints are useful for segmentation. Cognition 61, 93–125.

Brugge, J.F., Volkov, I.O., Garell, P.C., Reale, R.A., Howard, M.A., 2003. Functional connections between auditory cortex on Heschl's gyrus and on the lateral superior temporal gyrus in humans. J. Neurophysiol. 90, 3750–3763.

Buchsbaum, B.R., Olsen, R.K., Koch, P., Berman, K.F., 2005. Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. Neuron 48, 687–697.

Bybee, J., McClelland, J.L., 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. Linguist. Rev. 22, 381–410.

Cairns, P., Shillcock, R., Chater, N., Levy, J., 1997. Bootstrapping word boundaries: a bottom-up corpus based approach to speech segmentation. Cognitive Psychol. 33, 111–153.

Callan, D.E., Tajima, K., Callan, A.M., Kubo, R., Masaki, S., Akahane-Yamada, R., 2003. Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. Neuroimage 19, 113–124.

Carlyon, R.P., Cusack, R., Foxton, J.M., Robertson, I.H., 2001. Effects of attention and unilateral neglect on auditory stream segregation. J. Exp. Psychol. Hum. Percept. Perform. 27, 115–127.

Carlyon, R.P., Deeks, J., Norris, D., Butterfield, S., 2002. The continuity illusion and vowel identification. Acta Acust. Unit. Acust. 88, 408–415.

Christiansen, M.H., Allen, J., Seidenberg, M.S., 1998. Learning to segment speech using multiple cues: a connectionist model. Lang. Cognitive Process., 13.

Clarke, C.M., Garrett, M.F., 2004. Rapid adaptation to foreign-accented English. J. Acoust. Soc. Am. 116, 3647–3658.

Crowder, R.G., 1983. The purity of auditory memory. Philos. Trans. Royal Soc. Lond. B Biol. Sci. 302, 251–265.

Crowder, R.G., Morton, J., 1969. Precategorical acoustic storage. Percept. Psychophy. 5, 365–373.

Cutler, A., Carter, D.M., 1987. The predominance of strong initial syllables in the English vocabulary. Comput. Speech Lang. 2, 133–142.

Cutler, A., Norris, D., 1988. The role of strong syllables in segmentation for lexical access. J. Exp. Psychol. Hum. Percept. Perform. 14, 113–121.

Cutting, J.E., 1975. Aspects of phonological fusion. J. Exp. Psychol. Hum. Percept. Perform. 104, 105–120.

Davis, M.H., 2003. Connectionist modelling of lexical segmentation and vocabulary acquisition. In: Quinlan, P.T. (Ed.), Connectionist Models of Development. Psychology Press, Hove, UK.

Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J. Neurosci. 23, 3423–3431.

Davis, M.H., Marslen-Wilson, W.D., Gaskell, M.G., 2002. Leading up the lexical garden path: segmentation and ambiguity in spoken word recognition. J. Exp. Psychol. Hum. Percept. Perform. 28, 218–244.

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., McGettigan, C., 2005. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. J. Exp. Psychol. Gen. 134, 222–241.

Davis, M.H., Coleman, M.R., Absalom, A., Rodd, J.M., Johnsrude, I.S., Matta, B., Owen, A.M., Menon, D.K., in preparation. Dissociating Speech Perception and Comprehension at Reduced Levels of Awareness: An fMRI Study with Graded Propofol Sedation.

de la Mothe, L.A., Blumell, S., Kajikawa, Y., Hackett, T.A., 2006. Cortical connections of the auditory cortex in marmoset monkeys: core and medial belt regions. J. Comp. Neurol. 496, 27–71.

Dehaene-Lambertz, G., Pallier, C., Serniclaes, W., Sprenger-Charolles, L., Jobert, A., Dehaene, S., 2005. Neural correlates of switching from auditory to speech perception. Neuroimage 24, 21–33.

Diamond, I.T., Jones, E.G., Powell, T.P., 1969. The projection of the auditory cortex upon the diencephalon and brain stem in the cat. Brain Res. 15, 305–340.

Doupe, A.J., Kuhl, P.K., 1999. Birdsong and human speech: common themes and mechanisms. Annu. Rev. Neurosci. 22, 567–631.

Eisner, F., McQueen, J.M., 2006. Perceptual learning in speech: stability over time. J. Acoust. Soc. Am. 119, 1950–1953.

Elman, J.L., McClelland, J.L., 1988. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. J. Mem. Lang. 27, 143–165.

Fadiga, L., Craighero, L., Buccino, G., Rizzolatti, G., 2002. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur. J. Neurosci. 15, 399–402.

Felleman, D.J., Essen, D.C.V., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex 1, 1–47.

Fodor, J.A., Bever, T.G., 1965. The psychological reality of linguistic segments. J. Verb. Learn. Verb. Behav. 4, 414–420.

Foxe, J.J., Simpson, G.V., 2002. Flow of activation from V1 to frontal cortex in humans. A framework for defining "early" visual processing. Exp. Brain Res. 142, 139–150.

Frankish, C., 1989. Perceptual organization and precategorical acoustic storage. J. Exp. Psychol. Learn Mem. Cogn. 15, 469–479.

Friederici, A.D., Meyer, M., von Cramon, D.Y., 2000. Auditory language comprehension: an event-related fMRI study on the processing of syntactic and lexical information. Brain Lang. 74, 289–300.

Fritz, J.B., Elhilali, M., Shamma, S.A., 2005. Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks. J. Neurosci. 25, 7623–7635.

Ganong 3rd, W.F., 1980. Phonetic categorization in auditory word perception. J. Exp. Psychol. Hum. Percept. Perform. 6, 110–125.

Gao, E., Suga, N., 2000. Experience-dependent plasticity in the auditory cortex and the inferior colliculus of bats: role of the corticofugal system. Proc. Natl. Acad. Sci. USA 97, 8081–8086.

Garrett, M., Bever, T.G., Fodor, J.A., 1965. The active use of grammar in speech perception. Percept. Psychophys. 1, 30–32.

Garrod, S., Pickering, M.J., 2004. Why is conversation so easy? Trends Cogn. Sci. 8, 8–11.

Gaskell, M.G., Marslen-Wilson, W.D., 1996. Phonological variation and inference in lexical access. J. Exp. Psychol. Hum. Percept. Perform. 22, 144–158.

Gaskell, M.G., Marslen-Wilson, W.D., 1997. Integrating form and meaning: a distributed model of speech perception. Lang. Cognitive Process. 12, 613–656.

Giraud, A.L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M.O., Preibisch, C., Kleinschmidt, A., 2004. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cereb. Cortex 14, 247–255.

Goldinger, S.D., 1996. Words and voices: episodic traces in spoken word identification and recognition memory. J. Exp. Psychol. Learn. Mem. Cogn. 22, 1166–1183.

Goldinger, S.D., 1998. Echoes of echoes? An episodic theory of lexical access. Psychol. Rev. 105, 251–279.

Goldinger, S.D., Kleider, H.M., Shelley, E., 1999. The marriage of perception and memory: Creating two-way illusions with words and voices. Mem. Cogn. 27, 328–338.

Golestani, N., Zatorre, R.J., 2004. Learning new sounds of speech: reallocation of neural substrates. Neuroimage 21, 494–506.

Greenlee, J.D., Oya, H., Kawasaki, H., Volkov, I.O., Kaufman, O.P., Kovach, C., Howard, M.A., Brugge, J.F., 2004. A functional connection between inferior frontal gyrus and orofacial motor cortex in human. J. Neurophysiol. 92, 1153–1164.

Guenther, F.H., Ghosh, S.S., Tourville, J.A., 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. Brain Lang. 96, 280–301.

Hackett, T.A., Kaas, J.H., 2004. Auditory cortex in primates: functional subdivisions and processing streams. The Cognitive Neurosciences, third ed.

Hackett, T.A., Stepniewska, I., Kaas, J.H., 1998. Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. J. Comp. Neurol. 394, 475–495.

Hackett, T.A., Stepniewska, I., Kaas, J.H., 1999. Prefrontal connections of the parabelt auditory cortex in macaque monkeys. Brain Res. 817, 45–58.

Harnad, S., 1986. Categorical Perception: The Groundwork of Cognition. Cambridge University Press, Cambridge, UK.

Harris, C.L., 1994. Coarse coding and the lexicon. In: Fuchs, C., Victorri, B. (Eds.), Continuity in Linguistic Semantics. John Benjamins, Amsterdam, Holland.

Hartley, T., Houghton, G., 1996. A linguistically constrained model of short term memory for nonwords. J. Mem. Lang. 35, 1–31.

Hawkins, S., 2003. Roles and representations of systematic fine phonetic detail in speech understanding. J. Phonetics 31, 373–405.

Hervais-Adelman, A., Davis, M.H., Carlyon, R.P., Johnsrude, I.S., submitted for publication. Perceptual learning of noise vocoded words: Effects of feedback and lexicality.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.

Houde, J.F., Jordan, M.I., 1998. Sensorimotor adaptation in speech production. Science 279, 1213–1216.

Howard, M.A., Volkov, I.O., Mirsky, R., Garell, P.C., Noh, M.D., Granner, M., Damasio, H., Steinschneider, M., Reale, R.A., Hind, J.E., Brugge, J.F., 2000. Auditory cortex on the human posterior superior temporal gyrus. J. Comp. Neurol. 416, 79–92.

Howell, P., 1978. Syllabic and phonemic representations for short-term memory of speech stimuli. Percept. Psychophys. 24, 496–500.

Howell, P., Darwin, C.J., 1977. Some properties of auditory memory for rapid formant transitions. Mem. Cogn. 5, 700–708.

Huffman, R.F., Henson, O.W., 1990. The descending auditory pathway and acousticomotor systems: connections with the inferior colliculus. Brain Res. Brain Res. Rev. 15, 295–323.

Humphries, C., Love, T., Swinney, D., Hickok, G., 2005. Response of anterior temporal cortex to syntactic and prosodic manipulations during sentence processing. Hum. Brain Mapp. 26, 128–138.

Humphries, C., Binder, J.R., Medler, D.A., Liebenthal, E., 2006. Syntactic and semantic modulation of neural activity during auditory sentence comprehension. J. Cogn. Neurosci. 18, 665–679.

Indefrey, P., Cutler, A., 2004. Prelexical and lexical processing in listening. In: Gazzaniga, M.S. (Ed.), The Cognitive Neurosciences, third ed. MIT Press, Cambridge, MA, pp. 759–774.

Jacoby, L.L., Allan, L.G., Collins, J.C., Larwill, L.K., 1988. Memory influences subjective experience: noise judgements. J. Exp. Psychol. Learn. Mem. Cogn. 14, 240–247.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., Dupoux, E., 2003. Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. J. Neurosci. 23, 9541–9546.

Johnsrude, I.S., Davis, M.H., Horwitz, B., in preparation. Multiple processing pathways in speech comprehension.

Jones, E.G., 2003. Chemically defined parallel pathways in the monkey auditory system. Ann. NY Acad. Sci. 999, 218–233.

Kaas, J.H., Hackett, T.A., Tramo, M.J., 1999. Auditory processing in primate cerebral cortex. Curr. Opin. Neurobiol. 9, 164–170.

Kello, C.T., Plaut, D.C., 2004. A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters. J. Acoust. Soc. Am. 116, 2354–2364.

Kemps, R.J., Ernestus, M., Schreuder, R., Baayen, R.H., 2005. Prosodic cues for morphological complexity: the case of Dutch plural nouns. Mem. Cogn. 33, 430–446.

Kersten, D., Yuille, A., 2003. Bayesian models of object perception. Curr. Opin. Neurobiol. 13, 150–158.

Kolinsky, R., Morais, J., 1992. Représentations intermédiaires dans la reconnaissance de la parole: Apports de la technique de création de mots illusoires. In: Proceedings of the 19th journées d'étude sur la parole. SFA, ULB, Brussels, pp. 129–133.

Kolinsky, R., Morais, J., 1996. Migrations in speech recognition. Lang. Cogn. Process. 11, 611–619.

Kraljic, T., Samuel, A.G., 2005. Perceptual learning for speech: Is there a return to normal? Cogn. Psychol. 51, 141–178.

Krauss, R.M., Pardo, J.S., 2006. Speaker perception and social behavior: bridging social psychology and speech science. In: Van Lange, P.A.M. (Ed.), Bridging Social Psychology: Benefits of Transdisciplinary Approaches. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Kuhl, P.K., 2004. Early language acquisition: cracking the speech code. Nat. Rev. Neurosci. 5, 831–843.

Lehiste, I., 1977. Isochrony reconsidered. J. Phonetics 5, 253–263.

Liberman, A.M., Whalen, D.H., 2000. On the relation of speech to language. Trends in Cognitive Science 4, 187–196.
Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M., 1967. Perception of the speech code. Psychol. Rev. 74, 431–461.
Logan, J.S., Lively, S.E., Pisoni, D.B., 1991. Training Japanese listeners to identify English /r/ and /l/: a first report. J. Acoust. Soc. Am. 89, 874–886.
Luce, P.A., Lyons, E.A., 1998. Specificity of memory representations for spoken words. Mem. Cogn. 26, 708–715.
Luria, A.R., 1976. The working brain: An introduction to neuropsychology (Basil Haigh, Trans.). Penguin Books, London, UK.
MacKay, D.G., Wulf, G., Yin, C., Abrams, L., 1993. Relations between word perception and production: New theory and data on the verbal transformation effect. J. Mem. Lang. 32, 624–646.
MacMillan, N.A., 1986. Beyond the categorical/continuous distinction: a psychophysical approach to processing modes. In: Harnad, S. (Ed.), Categorical Perception. Cambridge University Press, Cambridge, UK.
Magnuson, J.S., McMurray, B., Tanenhaus, M.K., Aslin, R.N., 2003. Lexical effects on compensation for coarticulation: The ghost of Christmash past. Cogn. Sci. 27, 285–298.
Mann, V.A., Liberman, A.M., 1983. Some differences between phonetic and auditory modes of perception. Cognition 14, 211–235.
Marslen-Wilson, W., 1984. Function and processing in spoken word recognition: a tutorial review. In: Bouma, H., Bouwhuis, D.G. (Eds.), Attention and Performance X: Control of Language Processing. Erlbaum, Hillsdale NJ.
Marslen-Wilson, W., Warren, P., 1994a. Levels of perceptual representation and process in lexical access: words, phonemes, and features. Psychol. Rev. 101, 653–675.
Marslen-Wilson, W., Moss, H.E., van Halen, S., 1996. Perceptual distance and competition in lexical access. J. Exp. Psychol. Hum. Percept. Perform. 22, 1376–1392.
Marslen-Wilson, W., Tyler, L.K., Waksler, R., Older, L., 1994b. Morphology and meaning in the English mental lexicon. Psychol. Rev. 101, 3–33.
Mattys, S.L., 1997. The use of time during lexical processing and segmentation: A review. Psychon. Bull. Rev. 4, 310–329.
Mattys, S.L., 2004. Stress versus coarticulation: toward an integrated approach to explicit speech segmentation. J. Exp. Psychol. Hum. Percept. Perform. 30, 397–408.
Mattys, S.L., White, L., Melhorn, J.F., 2005. Integration of multiple speech segmentation cues: a hierarchical framework. J. Exp. Psychol. Gen. 134, 477–500.
Maye, J., Aslin, R.N., Tanenhaus, M.K., in press. The Weckud Wetch of the Wast: Lexical adaptation to a novel accent.
McCandliss, B.D., Fiez, J.A., Protopapas, A., Conway, M., McClelland, J.L., 2002. Success and failure in teaching the [r]-[l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. Cogn. Affect Behav. Neurosci. 2, 89–108.
McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cogn. Psychol. 18, 1–86.
McClelland, J.L., Mirman, D., Holt, L.L., 2006. Are there interactive processes in speech perception? Trends Cogn. Sci. 10, 363–369.
McMurray, B., Tanenhaus, M.K., Aslin, R.N., 2002. Gradient effects of within-category phonetic variation on lexical access. Cognition 86, B33–B42.
McQueen, J.M., 1998. Segmentation of continuous speech using phonotactics. J. Mem. Lang. 39, 21–46.
Miller, J.L., 1981. The effect of speaking rate on segmental distinctions: acoustic variation and perceptual compensation. In: Eimas, P.D., Miller, J.L. (Eds.), Perspectives on the Study of Speech. Erlbaum, Hillsdale, NJ.
Miller, J.L., Liberman, A.M., 1979. Some effects of later-occurring information on the perception of stop consonant and semivowel. Percept. Psychophys. 25, 457–465.
Mirman, D., McClelland, J.L., Holt, L.L., in press. An Interactive Hebbian Account of Lexically Guided Tuning of Speech Perception. Psychonomic Bulletin and Review. Available from: <http://www.psychonomic.org/PBR/>.
Morais, J., Cary, L., Alegria, J., Bertelson, P., 1979. Does awareness of speech as a sequence of phones arise spontaneously? Cognition 7, 323–331.
Morais, J., Bertelson, P., Cary, L., Alegria, J., 1986. Literacy training and speech segmentation. Cognition 24, 45–64.
Morton, J., Crowder, R.G., Prussin, H.A., 1971. Experiments with the stimulus suffix effect. J. Exp. Psychol. 91, 169–190.
Mottonen, R., Calvert, G.A., Jaaskelainen, I.P., Matthews, P.M., Thesen, T., Tuomainen, J., Sams, M., 2006. Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. Neuroimage 30, 563–569.
Nakatani, L.H., Dukes, K.D., 1977. Locus of segmental cues for word juncture. J. Acoust. Soc. Am. 62, 715–719.
Narain, C., Scott, S.K., Wise, R.J., Rosen, S., Leff, A., Iversen, S.D., Matthews, P.M., 2003. Defining a left-lateralized response specific to intelligible speech using fMRI. Cereb. Cortex 13, 1362–1368.
Nearey, T., 2001. Phoneme-like units and speech perception. Lang. Cogn. Process. 16, 673–681.
Norris, D., 1994. Shortlist: a connectionist model of continuous speech recognition. Cognition 52, 189–234.
Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cogn. Psychol. 47, 204–238.
Pallier, C., Colome, A., Sebastian-Galles, N., 2001. The influence of native-language phonology on lexical access: exemplar-Based versus abstract lexical entries. Psychol. Sci. 12, 445–449.
Peelle, J.E., Wingfield, A., 2005. Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. J. Exp. Psychol. Hum. Percept. Perform. 31, 1315–1330.
Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., Guiod, P., 1997. Speech motor control: acoustic goals, saturation effects, auditory feedback and internal models. Speech Commun. 22, 227–250.
Perrot, X., Ryvlin, P., Isnard, J., Guenot, M., Catenoix, H., Fischer, C., Mauguiere, F., Collet, L., 2006. Evidence for corticofugal modulation of peripheral auditory activity in humans. Cereb. Cortex 16, 941–948.
Petrides, M., Pandya, D.N., 1988. Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. J. Comp. Neurol. 273, 52–66.
Petrides, M., Pandya, D.N., 2002. Association pathways of the prefrontal cortex and functional observations. Prin. Frontal Lobe Funct., 31–50.
Petrides, M., Pandya, D.N., 2006. Efferent association pathways originating in the caudal prefrontal cortex in the macaque monkey. J. Comp. Neurol. 498, 227–251.
Pickering, M.J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behav. Brain Sci. 27, 169–190, discussion 190–226.
Pisoni, D.B., Tash, J., 1974. Reaction times to comparisons within and across phonetic categories. Percept. Psychophys. 15, 285–290.
Pitt, M., McQueen, J., 1998. Is compensation for coarticulation mediated by the lexicon? J. Mem. Lang. 39, 347–370.
Pitt, M.A., Samuel, A.G., 1993. An empirical and meta-analytic evaluation of the phoneme identification task. J. Exp. Psychol. Hum. Percept. Perform. 19, 699–725.
Pitt, M.A., Shoaf, L., 2002. Linking verbal transformations to their causes. J. Exp. Psychol. Hum. Percept. Perform. 28, 150–162.
Plomp, R., 2001. The Intelligent Ear. Lawrence Erlbaum Associates, Mahwah, NJ.
Poeppel, D., Idsardi, W.J., van Wassenhove, V., in press. Speech perception at the interface of neurobiology and linguistics. In: Proceedings of the Royal Society of London.
Price, C., Thierry, G., Griffiths, T., 2005. Speech-specific auditory processing: where is it? Trends Cogn. Sci. 9, 271–276.
Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. USA 103, 7865–7870.

Purcell, D.W., Munhall, K.G., 2006. Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. J. Acoust. Soc. Am. 120, 966–977.

Read, C., Zhang, Y.F., Nie, H.Y., Ding, B.Q., 1986. The ability to manipulate speech sounds depends on knowing alphabetic writing. Cognition 24, 31–44.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. J. Exp. Psychol. Hum. Percept. Perform. 23, 651–666.

Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D., 1981. Speech perception without traditional speech cues. Science 212, 947–949.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., Lang, J.M., 1994. On the perceptual organization of speech. Psychol. Rev. 101, 129–156.

Rizzolatti, G., Arbib, M.A., 1998. Language within our grasp. Trends Neurosci. 21, 188–194.

Romanski, L., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P., Rauschecker, J., 1999a. Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat. Neurosci. 2, 1131–1136.

Romanski, L.M., Bates, J.F., Goldman-Rakic, P.S., 1999b. Auditory belt and parabelt projections to the prefrontal cortex in the rhesus monkey. J. Comp. Neurol. 403, 141–157.

Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. Royal Soc. Lond. B Biol. Sci. 336, 367–373.

Rosen, S.M., 1979. Range and frequency effects in consonant categorization. J. Phonetics 7, 393–402.

Rozzi, S., Calzavara, R., Belmalih, A., Borra, E., Gregoriou, G.G., Matelli, M., Luppino, G., 2006. Cortical connections of the inferior parietal cortical convexity of the macaque monkey. Cereb. Cortex 16, 1389–1417.

Salverda, A.P., Dahan, D., McQueen, J.M., 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition 90, 51–89.

Sawusch, J.R., Nusbaum, H.C., 1979. Contextual effects in vowel perception I: anchor-induced contrast effects. Percept. Psychophys. 25, 292–302.

Schroeder, C.E., Smiley, J., Fu, K.G., McGinnis, T., O'Connell, M.N., Hackett, T.A., 2003. Anatomical mechanisms and functional implications of multisensory convergence in early cortical processing. Int. J. Psychophysiol. 50, 5–17.

Scott, S.K., Johnsrude, I.S., 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci. 26, 100–107.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J.S., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.

Seltzer, B., Pandya, D.N., 1989. Frontal lobe connections of the superior temporal sulcus in the rhesus monkey. J. Comp. Neurol. 281, 97–113.

Seltzer, B., Pandya, D.N., 1991. Post-rolandic cortical projections of the superior temporal sulcus in the rhesus monkey. J. Comp. Neurol. 312, 625–640.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.

Sheffert, S.M., Pisoni, D.B., Fellowes, J.M., Remez, R.E., 2002. Learning to recognize talkers from natural, sinewave, and reversed speech samples. J. Exp. Psychol. Hum. Percept. Perform. 28, 1447–1469.

Shoaf, L.C., Pitt, M.A., 2002. Does node stability underlie the verbal transformation effect? A test of node structure theory. Percept. Psychophys. 64, 795–803.

Shockley, K., Sabadini, L., Fowler, C.A., 2004. Imitation in shadowing words. Percept. Psychophys. 66, 422–429.

Shore, S.E., Zhou, J., 2006. Somatosensory influence on the cochlear nucleus and beyond. Hear. Res. 217, 90–99.

Thomas, S. M., Pilling, M., 2006. Auditory and audiovisual perceptual training using a simulation of a cochlear-implant system. Poster presented at the British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness. Cambridge, UK, September 2006.

Uppenkamp, S., Johnsrude, I.S., Norris, D., Marslen-Wilson, W., Patterson, R.D., 2006. Locating the initial stages of speech-sound processing in human temporal cortex. Neuroimage 31, 1284–1296.

Warren, R., 1968. Verbal transformation effect and auditory perceptual mechanisms. Psychol. Bull. 70, 261–270.

Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167, 392–393.

Warren, R.M., Gregory, R.L., 1958. An auditory analogue of the visual reversible figure. Am. J. Psychol. 71, 612–613.

Watkins, K., Paus, T., 2004. Modulation of motor excitability during speech perception: the role of Broca's area. J. Cogn. Neurosci. 16, 978–987.

Watkins, K.E., Strafella, A.P., Paus, T., 2003. Seeing and hearing speech excites the motor system involved in speech production. Neuropsychologia 41, 989–994.

Watkins, O.C., Watkins, M.J., 1980. The modality effect and echoic persistence. J. Exp. Psychol. Gen. 109, 251–278.

Weill, S.A., 2003. The Impact of Perceptual Dissimilarity on the Perception of Foreign Accented Speech. Ph.D,, Ohio State University.

Weinberg, R.J., 1997. Are topographic maps fundamental to sensory processing? Brain Res. Bull. 44, 113–116.

Werker, J.F., Tees, R.C., 1999. Influences on infant speech processing: toward a new synthesis. Ann. Rev. Psychol. 50, 509–535.

Whalen, D.H., Liberman, A.M., 1987. Speech perception takes precedence over nonspeech perception. Science 237, 169–171.

Wilson, S.M., Iacoboni, M., 2006. Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. Neuroimage 33, 316–325.

Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7, 701–702.

Winer, J.A., 2006. Decoding the auditory corticofugal systems. Hear. Res. 212, 1–8.

Wise, R.J., Scott, S.K., Blank, S.C., Mummery, C.J., Murphy, K., Warburton, E.A., 2001. Separate neural subsystems within 'Wernicke's area'. Brain 124, 83–95.

Xiao, Z., Suga, N., 2002. Modulation of cochlear hair cells by the auditory cortex in the mustached bat. Nat. Neurosci. 5, 57–63.

Yeterian, E.H., Pandya, D.N., 1998. Corticostriatal connections of the superior temporal region in rhesus monkeys. J. Comp. Neurol. 399, 384–402.