PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2007) **363**, 917–921
doi:10.1098/rstb.2007.2195
Published online 7 September 2007

Introduction. The perception of speech: from sound to meaning

Brian C. J. Moore[1,*], Lorraine K. Tyler[1] and William Marslen-Wilson[2]

[1]*Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK*
[2]*MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK*

Spoken language communication is arguably the most important activity that distinguishes humans from non-human species. This paper provides an overview of the review papers that make up this theme issue on the processes underlying speech communication. The volume includes contributions from researchers who specialize in a wide range of topics within the general area of speech perception and language processing. It also includes contributions from key researchers in neuroanatomy and functional neuro-imaging, in an effort to cut across traditional disciplinary boundaries and foster cross-disciplinary interactions in this important and rapidly developing area of the biological and cognitive sciences.

Keywords: speech perception; speech production; psycholinguistics; audition; phonetics; functional imaging

1. INTRODUCTION

Spoken language communication is arguably the most important activity that distinguishes humans from non-human species. While many animal species communicate and exchange information using sound, humans are unique in the complexity of the information that can be conveyed using speech, and in the range of ideas, thoughts and emotions that can be expressed.

Despite the importance of speech communication for the entire structure of human society, there are many aspects of the speech communication process that are not fully understood. Research on speech and language is typically carried out by different groups of scientists working on separate aspects of the underlying functional and neural systems. Research from an auditory perspective focuses on the acoustical properties of speech sounds, the representation of speech sounds in the auditory system and how that representation is used to extract phonetic information. Research from psycholinguistic perspectives studies the processes by which representations of meaning are extracted from the acoustic–phonetic sequence, and how these are linked to the construction of higher-level linguistic interpretation in terms of sentences and discourse. However, there has been relatively little interaction between speech researchers from these two groups.

In addition, there has been a dramatic expansion in recent years of research into the neural bases of auditory and linguistic functions. Developments in the neuroanatomy and neurophysiology of the auditory system of non-human primates provide the basis for mapping out the basic organization of the structures and pathways that support the processing of auditory information in the primate brain. Complementary developments in neuroimaging techniques for visualizing the activity of the intact brain are allowing scientists to probe the dynamic spatio-temporal patterns of neural activity that underlie the representation and processing of speech and language in the human brain.

Despite this ferment of activity across a variety of fields, there has been relatively little interaction between researchers working on these various topics, and perhaps a lack of recognition that they are all participating in the same overall scientific process of understanding how the motor gestures of a speaker are transformed to sounds and how those sounds are mapped onto meaning in the comprehension of spoken language. This volume addresses these issues. It includes contributions from researchers who specialize in a wide range of topics within the general area of speech perception and language processing. It also includes contributions from key researchers in neuroanatomy and functional neuro-imaging, in an effort to cut across traditional disciplinary boundaries and foster cross-disciplinary interactions in this important and rapidly developing area of the biological and cognitive sciences.

2. OVERVIEW OF THE SPECIAL ISSUE

The paper by Young (2008) describes the representation of speech sounds in the auditory nerve and at higher levels in the central nervous system, focusing especially on vowel sounds. The experimental data are derived mainly from animal models (especially the cat), so some caution is needed in interpreting the results in terms of the human auditory system. However, it seems probable that at least the early stages of auditory processing, as measured in the auditory nerve, are similar across all mammals. A key feature of the representation of sounds is that it is tonotopic; speech signals are decomposed into sinusoidal frequency

* Author for correspondence (bcjm@cam.ac.uk).

One contribution of 13 to a Theme Issue 'The perception of speech: from sound to meaning'.

components or groups of components and different frequency components are represented in different populations of neurons. In other words, the short-term spectrum of the sound is represented in the relative amount of neural activity in neurons that are tuned to different frequencies. This tonotopic organization is preserved throughout the auditory system, although at higher levels in the auditory system there may be multiple 'maps'. Another critical feature of the representation is nonlinear suppression, whereby strong neural activity in one group of neurons (all 'tuned' to similar frequencies) suppresses activity in neurons tuned to adjacent frequencies. This suppression is essential for maintaining the representation of the spectral content of sounds over a wide range of sound levels. Spectral features may also be represented in the detailed timing of the neural activity (phase locking), although the role of this 'temporal fine structure' is still controversial. The representation of speech sounds in central auditory neurons is more robust than at the periphery to changes in stimulus intensity and it also becomes more transient. Furthermore, Young argues that it is probable that the form of the representation at the auditory cortex is fundamentally different from the representation at lower levels, in that stimulus features other than the distribution of energy across frequency are analysed.

The paper by Moore (2008) reviews basic aspects of auditory processing that play a role in the perception of speech. Here, the data are mainly derived from perceptual experiments using human listeners. The frequency selectivity of the auditory system refers to the ability to resolve the sinusoidal components in complex sounds, and is closely related to the tonotopic representation described by Young. Moore describes how frequency selectivity can be quantified using masking experiments. The 'auditory filters' inferred from the results can be used to calculate the internal representation of the spectrum of speech sounds in the peripheral auditory system. This representation is called the excitation pattern. The perception of timbre and distinctions in quality between vowels are related to both static and dynamic aspects of the spectra of sounds, as represented in the excitation pattern. The pitch of speech sounds is related to their fundamental frequency, which is in turn related to the rate of vibration of the vocal folds. Moore describes the mechanisms by which the auditory system extracts the pitch of speech sounds and the role that pitch patterns play in speech perception, especially the perception of intonation.

Although some speech sounds, such as vowels, can be characterized in terms of their long-term spectral properties, speech perception in general depends strongly on the dynamic nature of speech sounds, and the way that they change over time. Moore describes the limits of the ability of the auditory system to follow rapid changes, and describes how temporal resolution can be modelled using the concept of a sliding temporal integrator. The combined effects of limited frequency selectivity and limited temporal resolution can be modelled by calculation of the spectro-temporal excitation pattern, which gives good insight into the representation of speech sounds in the auditory system. Moore argues that, for speech presented in quiet, the resolution of the auditory system in frequency and time usually markedly exceeds the resolution necessary for the identification or discrimination of speech sounds, which partly accounts for the robust nature of speech perception. However, people with impaired hearing have reduced frequency selectivity and can hear comfortably over a smaller than normal range of sound levels. For such people, speech perception is often much less robust than for normally hearing people.

The paper by Diehl (2008) considers further the robust nature of speech perception. For people with normal hearing, speech can be understood even under conditions when there is considerable background noise or reverberation, or when the speech is distorted in a variety of ways. Diehl considers how the acoustical and auditory properties of vowels and consonants help to ensure intelligibility. The properties of speech sounds can be understood by considering the sounds as resulting from a source of sound energy, such as vibration of the vocal folds or turbulence produced by forcing air through a narrow constriction, followed by a filter (the vocal tract) which modifies the spectrum of the source. Diehl describes this 'source–filter' theory, and demonstrates how it can account for the relationship between vocal-tract properties and formant patterns. He points out that certain types of speech sounds (e.g. the resonance patterns or 'formant' frequencies of specific vowel sounds) occur commonly in the languages of the world, while others occur much more rarely. He presents two theories that have been proposed to account for the structure of these 'preferred sound inventories': quantal theory and dispersion theory.

Quantal theory (Stevens 1989) is based on the fact that nonlinearities exist in the mapping between articulatory (i.e. vocal-tract) configurations of talkers and acoustic outputs. For certain regions of articulatory 'space', perturbations in the articulatory parameters result in small changes in the acoustic output, whereas in other regions perturbations of similar size yield large acoustic changes. Given these regions of acoustic stability and instability, quantal theory is based on the idea that preferred sound categories are selected to occupy the stable regions and to be separated by unstable regions. Dispersion theory (Liljencrants & Lindblom 1972), like quantal theory, is based on the idea that speech sound inventories are structured to maintain perceptual distinctiveness. However, in dispersion theory, distinctiveness is viewed as a global property of an entire inventory of sound categories. A vowel or consonant inventory is said to be maximally distinctive if the sounds are maximally dispersed (i.e. separated from each other) in the available 'phonetic space'. Diehl discusses the strengths and limitations of each theory, and proposes that certain aspects of the two theories can be unified in a principled way so as to achieve reasonably accurate predictions of the properties of preferred sound inventories.

The paper by Kuhl *et al.* (2008) describes the development of language during the early years of life, and the mechanisms that appear to underlie that development. Infants' speech perception skills show two types of changes towards the end of the first year of

life. First, the ability to perceive phonetic distinctions in a non-native language declines. Second, skills at making phonetic distinctions in the child's own language improve. The paper presents new data showing that both native and non-native phonetic perception skills of infants predict their later language ability, but in opposite directions. Better *native*-language skill at seven months predicts faster language advancement, whereas better *non-native*-language skill predicts slower advancement. Kuhl *et al.* suggest that native-language phonetic performance is indicative of commitment of neural circuitry to the native language, while non-native phonetic performance reveals uncommitted neural circuitry. This paper describes a revised version of a model previously proposed by Kuhl and co-workers, the native language magnet model.

The paper by Campbell (2008) emphasizes the fact that speech perception is multimodal; what we perceive as speech is influenced by what we see on the face of the talker as well as by what is received at the two ears. This is illustrated by the McGurk effect (McGurk & MacDonald 1976), which is produced when a video recording of one utterance is combined with an audio recording of another utterance. What is heard is influenced by what is seen. For example, an acoustic 'mama' paired with a video 'tata' is heard as 'nana'. The influence of vision on speech perception is also illustrated by the fact that, in noisy situations, speech can be understood much better when the face of the talker is visible than when it is invisible (Erber 1974).

Campbell proposes that there are two main ways or 'modes' in which visual information may influence speech perception. The first is a complementary mode, whereby vision provides information more efficiently than hearing for some under-specified parts of the speech stream. For example, the acoustic cues signalling the distinction between 'ba' and 'ga' may be relatively weak and easily masked by background sounds, whereas, visually, these two sounds are very distinct. The second is a correlated mode, whereby vision partially duplicates auditory information about dynamic articulatory patterning.

Campbell reviews evidence suggesting that these two modes are not reflected in discrete cortical processing systems, but that they reflect somewhat differentiated access to two major streams for the processing of natural language—a 'what' and a 'how' stream. The 'what' stream makes particular use of the inferior occipito-temporal regions of the cortex and of the ventral visual processing stream which can specify image details effectively. It can therefore serve as a useful route for complementary visual information to be processed. A major projection of this stream is to association areas in middle and superior temporal cortex. In contrast to this, the 'how' stream for the analysis of auditory speech may be readily accessed by natural visible speech, which is characterized by dynamic features that correspond with those available acoustically. Processing that requires sequential segmental analysis (e.g. identifying syllables or words individually or in lists) will differentially engage this posterior stream. It is in this stream that the correlational structure of seen and heard speech is best reflected. The visual input to these analyses arises primarily in the lateral temporo-occipital regions that track visual movement.

Although the great majority of studies of speech perception have been conducted using speech sounds presented in quiet with little reverberation, speech communication in everyday life often takes place in the presence of background sounds and reverberation. The issues raised by this are considered in the paper by Darwin (2008). He points out that irrelevant background sounds can cause severe problems for computer-based speech recognition algorithms and for people with hearing impairment, but that people with normal hearing are remarkably little affected. A variety of perceptual problems are created by the presence of background sounds. These include: complete or partial masking of some parts of the target speech; the need to decide which parts of the sound 'belong to' each sound source; and the recognition of speech sounds based on partial information. Darwin examines the effectiveness of the cues, which can be used to separate target speech from a background of other sounds (including competing speech), focusing particularly on the role of fundamental frequency, onset asynchronies and binaural cues. At present, human listeners perform far better than any computer system in separating mixtures of sounds. A fuller understanding of how humans do this would have important practical applications.

The paper by Patterson & Johnsrude (2008) places the study of auditory processing, as applied to speech, squarely in a neuro-biological and neuro-imaging context. Cross-species studies—especially in the macaque—provide a well-developed neuroanatomical and neurophysiological account of the primate auditory processing system. This leads to concrete hypotheses both about the detailed functional architecture of the human system, with subcortical auditory processing systems feeding into primary auditory cortex, and about the local and the global connectivity of these areas with other regions of the brain. In this general framework, Patterson & Johnsrude go on to consider some of the basic functional challenges that speech variation presents to the listener, and how these challenges are met in the primate auditory system. Two major sources of variation are differences in pitch and vocal-tract length, which mean, for example, that the same vowel (in terms of its linguistic label) spoken by a child or an adult will vary markedly in its acoustic properties.

Patterson & Johnsrude present an innovative account of how adaptive mechanisms, operating before speech analysis can take place, may provide information sufficient to allow the system to normalize for pitch and vocal-tract variation. This account combines a computational model of auditory processing with psychophysically constrained neuro-imaging investigations of the spatial locations in auditory processing areas (in and around Heschl's gyrus) that are particularly sensitive to the relevant acoustic and phonetic contrasts. An important role is played here by magnetoencephalography (MEG), where high temporal resolution is accompanied by significantly improved spatial resolution, relative to electroencephalography (EEG). Recent studies using MEG are beginning to tease out the spatio-temporal details of

the cortical processing events underlying the extraction and perception of pitch information.

In a final section Patterson & Johnsrude address the central question of how the cortical system moves from general auditory processing to potentially voice- and speech-specific processing activities. Research into this question is still in its early stages, but the evidence suggests that the transformation from an auditory signal to speech is localizable, but is not straightforwardly hierarchical. The emergence of a vowel percept, for example, from the building blocks provided by sub-processes concerned with glottal pulse rate, vocal-tract length and so forth, seems to be distributed across several neural loci, situated around but not directly in core auditory cortex, and with possible links further afield to structures in premotor and motor cortex involved in speech production. This suggests that motor theories of speech perception (Liberman *et al.* 1967; Liberman & Mattingly 1985) may be due for a revival.

The next paper, by Zatorre & Gandour (2008), is highly complementary, with its focus on neural specializations for speech and pitch, and provides a balanced and informative account of possible hemispheric differences in these domains. They argue against standard approaches to this issue, which have led to a polarized debate asking whether speech processing is underpinned either by encapsulated, specialized domain-specific mechanisms or whether it piggybacks on general-purpose neural mechanisms for processing sound which are sensitive to the acoustic features that are present in speech. Zatorre & Gandour propose a more integrated approach, arguing that the brain's response to low-level acoustic features is modulated by linguistic factors, affecting the specificity of hemispheric function.

They outline the considerable evidence that has now accumulated for hemispheric differences in sensitivity to both the spectral and temporal properties of auditory inputs, but go on to argue that these differences can be modulated by the linguistic status of the input. In addition to neuro-imaging data on English, they also discuss extensive data from tonal languages, where Gandour and colleagues have been the pioneers in applying neuro-imaging techniques to the evaluation of neural contrasts in how pitch is processed as a function of its linguistic role. A clear outcome of these studies, in languages like Mandarin and Thai, is that when linguistically relevant pitch patterns carried by tones cue linguistic differences, activity tends to be left-lateralized, but when they do not, then activity is right-lateralized. Zatorre & Gandour conclude by arguing for an approach to speech processing that recognizes the complexity of hemispheric interactions between general sensory-motor and cognitive processes, modulated by the specific processing demands of different linguistic environments.

The paper by Poeppel *et al.* (2008), although it covers some of the same ground as the two preceding papers, addresses the neurobiology of speech in the brain from very different and more 'external' theoretical perspectives. Poeppel and colleagues revive—and significantly rework—the classic methodological framework put forward by David Marr in the 1980s for the analysis of complex neuro-cognitive systems, and they give linguistic theory equal status with neurobiology and auditory neuroscience in placing fundamental constraints on the realization of speech in the brain. Their key assumption is that speech perception is about the construction of abstract phonological representations, structured in such a way that they can interface with lexical representations as characterized in current linguistic theory. In their Marrist framework, this requirement is related to three levels of scientific description. The highest, computational, level refers to the commitment to a representational theory in terms of phonological distinctive features. The two lower levels—the implementational and the algorithmic—describe how the system is organized to generate a linguistically relevant output specified in these terms.

The implementational level centres around the notion of multi-time resolution processing (also considered by Zatorre & Gandour), where speech signals are simultaneously processed on a short (25–80 ms) time scale, and on a longer time scale of roughly syllabic length (approx. 200 ms), and where there are hemispheric asymmetries associated with these two temporal domains. The output of these processes is the input to an analysis-by-synthesis process—specified at the algorithmic level—that interacts with lexical hypotheses and a partial feature matrix to generate a contextually acceptable lexical outcome. The analysis-by-synthesis approach—linked to current developments in Bayesian methodology and to the notion of a 'forward model'—is well suited to these authors' proposal of a 'phonological primal sketch' at the segmental level. Preliminary, broad-brush hypotheses about feature content can be tested and elaborated relative to stored knowledge about possible lexical analyses.

The next contribution, by Tyler & Marslen-Wilson (2008), moves away from the specifics of auditory speech processing to focus on higher levels of the neural language system, combining cognitive accounts of language function with neuro-imaging studies of healthy subjects and patients who have specific language deficits. This research complements standard subtractive analyses of the fMRI data with connectivity analyses in order to better understand the relationship between frontal and temporal regions in the processing of different aspects of language function. These studies develop a general contrast between a core set of morphological and syntactic linguistic functions, likely to be combinatorial in nature, and requiring an intact left hemisphere perisylvian language network, with more general processes of semantic and pragmatic interpretation whose neural substrate is more distributed and more bilateral in nature.

The first part of the paper focuses on the processing of regularly inflected forms in English, as a prominent example of a linguistic process likely to involve the decomposition of a complex linguistic form (such as the past tense *jumped*) into its morphemic components (the stem *jump* and the grammatical affix *-ed*). A growing body of neuropsychological and neuro-imaging evidence points to a decompositional morphemic substrate for lexical processing, which requires an intact fronto-temporal network linking left posterior temporal lobe regions with left inferior frontal cortex (classical Broca's area). The second part of the paper, focusing on syntactic

and semantic processing—and their disruption following left hemisphere damage caused by stroke—confirms the critical dependency of syntactic (but not semantic) processes on a left fronto-temporal network that partially overlaps with the network revealed for morphological processes. Tyler & Marslen-Wilson interpret this overlap as indicating that different linguistic processes are not carried out in neural regions that are functionally specialized. Instead, each language function requires the co-activation in time of multiple regions within the fronto-temporal-parietal system, providing a different perspective on structure–function relationships in human language processing.

Also considering the global structure of the speech comprehension process, Hagoort's paper (2008) discusses the speed with which spoken language is processed, focusing on research using EEG, a time-sensitive methodology for probing the moment-by-moment processing of language. As is now well-established, spoken word recognition is a remarkably rapid process whereby multiple word candidates are activated on the basis of the sensory input and word recognition occurs when one candidate emerges as having the best fit. This produces a system in which words are identified well before their offset, through a process of activation, selection and integration with the prior context. Hagoort describes EEG studies which confirm the earliness of word identification and the structure of the system which underpins lexical processing. He complements EEG studies on single-word processing with experiments showing how sentence and discourse contexts modulate the processing of individual words. These experiments show that context speeds up lexical selection, and add to previous findings by relating different aspects of processing to different event-related potential components. He describes EEG data that help to develop models of language processing in which the processing of individual words is immediately affected by the discourse and real-world context.

The paper by Tanenhaus & Brown-Schmidt (2008) continues with the theme of the facilitatory effects of higher-level context on lexical processing, but does so in the framework of the 'visual world paradigm', in order to generate a more naturalistic environment in which to study language comprehension. These studies, using eye movement monitoring techniques, show that that multiple sources of linguistic and visual information are used to constrain the real-time analysis of spoken language processing. In the second part of the paper, the authors describe studies which also focus on language use in naturalistic contexts, but here the emphasis is upon natural conversation, on the assumption that language use is typically an interactive process whereby speakers and listeners share common communicative goals. These types of naturalistic context may generate different models of language use compared with those based on more impoverished contexts. Subjects in these studies engage in a referential communicative task while gaze and speech are monitored. The results show that subjects closely coordinate referential domains as the conversation develops. The wider implication of this work is that behavioural context, including attention and intention, affects even basic perceptual processes involved in language processing.

May 2007

REFERENCES

Campbell, R. 2008 The processing of audiovisual speech: empirical and neural bases. *Phil. Trans. R. Soc. B* **363**, 1001–1010. (doi:10.1098/rstb/2007.2155)

Darwin, C. J. 2008 Listening to speech in the presence of other sounds. *Phil. Trans. R. Soc. B* **363**, 1011–1021. (doi:10.1098/rstb/2007.2156)

Diehl, R. L. 2008 Acoustic and auditory phonetics: the adaptive design of speech sound systems. *Phil. Trans. R. Soc. B* **363**, 965–978. (doi:10.1098/rstb/2007.2153)

Erber, N. P. 1974 Auditory–visual perception of speech: a survey. In *Visual and audio-visual perception of speech* (eds H. Birk Nielsen & E. Kampp), pp. 12–30. Stockholm, Sweden: Almquist & Wiksell.

Hagoort, P. 2008 The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Phil. Trans. R. Soc. B* **363**, 1055–1069. (doi:10.1098/rstb/2007.2159)

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M. & Nelson, T. 2008 Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Phil. Trans. R. Soc. B* **363**, 979–1000. (doi:10.1098/rstb/2007.2154)

Liberman, A. M. & Mattingly, I. G. 1985 The motor theory of speech perception revised. *Cognition* **21**, 1–36. (doi:10.1016/0010-0277(85)90021-6)

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. 1967 Perception of the speech code. *Psychol. Rev.* **74**, 431–461. (doi:10.1037/h0020279)

Liljencrants, J. & Lindblom, B. 1972 Numerical simulation of vowel quality systems: the role of perceptual contrast. *Language* **48**, 839–862. (doi:10.2307/411991)

McGurk, H. & MacDonald, J. 1976 Hearing lips and seeing voices. *Nature* **264**, 746–748. (doi:10.1038/264746a0)

Moore, B. C. J. 2008 Basic auditory processes involved in the analysis of speech sounds. *Phil. Trans. R. Soc. B* **363**, 947–963. (doi:10.1098/rstb/2007.2152)

Patterson, R. D. & Johnsrude, I. S. 2008 Functional imaging of the auditory processing applied to speech sounds. *Phil. Trans. R. Soc. B* **363**, 1023–1035. (doi:10.1098/rstb/2007.2157)

Poeppel, D., Idsardi, W. J. & van Wassenhove, V. 2008 Speech perception at the interface of neurobiology and linguistics. *Phil. Trans. R. Soc. B* **363**, 1071–1086. (doi:10.1098/rstb/2007.2160)

Stevens, K. N. 1989 On the quantal nature of speech. *J. Phon.* **17**, 3–45.

Tanenhaus, M. K. & Brown-Schmidt, S. 2008 Language processing in the natural world. *Phil. Trans. R. Soc. B* **363**, 1105–1122. (doi:10.1098/rstb/2007.2162)

Tyler, L. K. & Marslen-Wilson, W. 2008 Fronto-temporal brain systems supporting spoken language comprehension. *Phil. Trans. R. Soc. B* **363**, 1037–1054. (doi:10.1098/rstb/2007.2158)

Young, E. D. 2008 Neural representation of spectral and temporal information in speech. *Phil. Trans. R. Soc. B* **363**, 923–945. (doi:10.1098/rstb/2007.2151)

Zatorre, R. J. & Gandour, J. T. 2008 Neural specializations for speech and pitch: moving beyond the dichotomies. *Phil. Trans. R. Soc. B* **363**, 1087–1104. (doi:10.1098/rstb/2007.2161)