*Auditory Signal Processing: Physiology, Psychoacoustics, and Models.*
*Pressnitzer, D., de Cheveigné, A., McAdams, S.,and Collet, L. (Eds). Springer Verlag, 2004.*

# From sound to meaning: Hierarchical processing in speech comprehension

Ingrid Johnsrude, Matt Davis, and Alexis Hervais-Adelman

Medical Research Council Cognition and Brain Sciences Unit,
{ingrid.johnsrude,matt.davis,alexis.hervais-adelman}@mrc-cbu.cam.ac.uk

## 1 Introduction

Cognitive models of spoken language comprehension postulate several processing stages as sound is mapped onto meaning (e.g. Gaskell and Marslen-Wilson 1997; McClelland and Elman 1986). Some of these stages, operating on sound information, may map onto the sequential, hierarchical organization observed for general auditory processing in macaques (see Kaas and Hackett 2000; Rauschecker 1998 for reviews), whereas higher-level processing stages operate upon more abstract representations of linguistic, rather than acoustic, information. However, the degree to which higher-level linguistic processes can be distinguished from less-specialized auditory and sound-form-based processes remains unclear (Remez, Rubin, Berns, Pardo, and Lang 1994; Scott, Blank, Rosen, and Wise 2000; Whalen and Liberman 1987).

In this study, we alter (distort) the specific surface properties of speech in three different ways, and use a correlational design to relate brain activity to intelligibility, using functional magnetic resonance imaging (fMRI). We operationalize "intelligibility" as the amount of a sentence that is understood: an aggregate measure of the multiple, hierarchically organized, processes involved in comprehension. Within areas that correlate with intelligibility, we can differentiate regions that are sensitive to the type of distortion used (form-dependent), and thus probably involved in acoustic analysis; and those that are insensitive to distortion type (form-independent); these areas may be involved in higher-level, linguistic processes.

## 2 Methods

Methods are described in detail in Davis and Johnsrude (2003), which presents data from 12 of the 27 listeners discussed here.

## 2.1 Stimuli

Stimuli were 190 declarative English sentences 5 to 17 words (1.7 to 4.3 seconds) long, digitized at a sampling rate of 22.1Khz. Three forms of distortion were applied to these sentences using Praat software [www.praat.org]. All three forms of distortion preserved the duration, amplitude and average spectral composition of the original sentences but markedly altered the acoustic form.

**Segmented speech** was created by dividing the speech waveform into short chunks at fixed intervals and replacing even-numbered chunks of speech with a signal-correlated noise version of the original speech (Bashford, Warren, and Brown 1996). Signal-correlated noise is a waveform with the same spectral profile and amplitude envelope as the original speech but consists entirely of noise, and is totally unintelligible (Schroeder, 1968). The duration of clear speech was fixed at 200 ms and 500, 200, or 100 ms sections of speech were replaced by signal-correlated noise.

**Noise-vocoded speech** (Shannon, Zeng, Kamath, Wygonski, and Ekelid 1995) was created by dividing the speech signal between 50 and 8000 Hz into 4, 7 or 15 band-pass filtered frequency bands. Sentences were re-synthesised by replacing information in each frequency band with amplitude-modulated, bandpass noise.

**Speech in noise** was generated by adding a continuous white noise background to sentences at three signal-to-noise ratios (-1, -4, or -6 dB). The overall amplitude of each speech-in-noise stimulus was reduced to match the amplitude of the original sentence.

**Signal-correlated noise** (SCN) was generated as a totally unintelligible baseline stimulus using the same algorithm as for segmented speech, but without periods of clear speech. (Schroeder, 1968).

## 2.2 Pilot study

In order to ensure that a continuum of intelligibility was obtained for each form of distortion, 18 native English speakers heard single stimulus sentences over closed-ear headphones (BeyerDynamic DT770) played from the soundcard of a Dell laptop PC. Participants were required to either type as many words as they could understand or to rate intelligibility (on a nine-point scale) immediately after each item. Sentences were pseudorandomly assigned to a type and level of distortion. Word-report performance (calculated as the proportion of words per sentence that were reported correctly) and rated intelligibility were averaged over 5 items per condition per subject: these were reliably correlated ($r=.99$, $p<.001$). A total of six levels of intelligibility were tested for each form of distortion. We selected three levels of each form of distortion: a low-intelligibility condition (approximately 20% of words reported correctly); a medium-intelligibility condition (65% words correct); and a high-intelligibility condition (90% correct).

## 2.3 Subjects

Twenty-seven right-handed volunteers aged between 18 and 42 were scanned in two experiments. All subjects were native speakers of English, without any history of neurological illness, head injury, or hearing impairment. The study was approved by the Addenbrooke's Local Research Ethics Committee and written informed consent was obtained from all subjects.

## 2.4 Scanning procedure

Stimuli were presented diotically using a high-fidelity auditory stimulus-delivery system incorporating electrostatic headphones inserted into sound-attenuating ear defenders (Palmer, Bullock, and Chambers 1998). To further attenuate scanner noise, participants wore insert earplugs (www.aearo.com), rated to attenuate by approximately 30 dB. Twelve subjects were asked to rate the intelligibility of each item using a four-alternative button press with their right hand, after presentation of each sentence (Davis and Johnsrude 2003). The remaining 15 listened to the stimuli without performing a task. The ratings gathered from the first 12 listeners correlated very highly with the word-report scores from the pilot study ($r=.98$, $p<.001$), indicating that the fMRI sound-delivery system did not degrade stimulus quality.

We acquired imaging data using a Bruker Medspec (Ettlingen, Germany) 3-Tesla MR system. Echo-planar whole-brain image volumes (228 in total; resolution 2 x 2 x 4 mm) were acquired using a sparse imaging technique, in which stimuli are presented in the silent period between successive scans, minimizing acoustic interference (Edmister, Talavage, Ledden, and Weisskoff 1999; Hall, Haggard, Akeroyd, Palmer, Summerfield, Elliott, Gurney, and Bowtell 1999).

Each trial comprised a stimulus item followed by a tone pip and a single EPI volume. Stimuli were pseudorandomly drawn from the 11 experimental conditions (low- medium- and high- intelligibility conditions for each of three forms of distortion, plus signal-correlated noise and clear speech). There were 19 trials of each stimulus type and an additional 19 silent trials.

## 2.5 Data analysis

Data processing and analysis was accomplished using Statistical Parametric Mapping (SPM99, www.fil.ion.ucl.ac.uk/spm) Pre-processing steps included within-subject realignment, spatial normalization and spatial smoothing using a Gaussian kernel of 12 mm, suitable for random-effects analysis (Xiong, Rao, Jerabek, Zamarripa, Woldorff, Lancaster, and Fox 2000).

We first wished to identify areas within subjects in which activation correlated with intelligibility (as indexed by word-report scores from the pilot study; see Davis and Johnsrude 2003). Within these intelligibility-sensitive areas, we then wished to differentiate between areas of form dependence (activation that was sensitive to the acoustic form of the stimulus) and areas of form independence (areas that responded equivalently to the different forms of distortion). In addition, we identify areas involved in a preliminary cortical stage of auditory processing as those exhibiting elevated response to signal-correlated noise over silence, without a

correlation with intelligibility. Some spatial segregation among the three response types might indicate a hierarchy of processing within auditory cortices as stimulus characteristics become more complex, such as has been observed in the macaque (Rauschecker, 1998).

Single-subject analyses in the two sets of subjects (12 subjects with task, 15 without) were followed by a random-effects analysis on all 27 in which Task was included as a factor. The significance threshold was set at $p < .05$, corrected for comparisons across the whole brain.

## 3 Results

The effect of Task was not significant in any of the analyses presented here; and so data from all 27 subjects are combined.

Comparison of SCN and silence across subjects yielded activation bilaterally in Heschl's gyrus and surrounding areas, consistent with recruitment of core and belt auditory cortex (even with areas sensitive to intelligibility excluded; Fig. 1a).

BOLD signal was positively correlated with word-report score in voxels along the length of the superior and middle temporal gyri in the left hemisphere, extending outwards from auditory cortex towards the temporal pole and the temporoparietal junction. Similar, less extensive, activation was observed in the right superior and middle temporal gyri. A portion of left inferior frontal gyrus also showed a positive correlation with intelligibility, as did the body of the left hippocampal complex.

These activation foci can be divided into those showing sensitivity to acoustic form (form dependence) and those that are insensitive to the acoustic properties of sound. The intelligibility-responsive region was masked by all six possible contrasts between pairs of the three distortion types. A form-dependent response (in which at least one of the 6 contrasts was significant) was observed in the superior temporal gyrus, bilaterally (Fig 1b). Intelligibility-responsive areas in which none of these contrasts reach significance at $p<.00851$ were considered to be form-independent: this response pattern was observed in the anterior middle temporal gyrus bilaterally, and in the left posterior superior temporal sulcus, left inferior frontal gyrus, left hippocampus and left precuneus (Fig. 1c).

## 4 Discussion

Our observation of intelligibility-sensitive regions in the lateral temporal lobe replicates and extends the findings of previous functional imaging studies (Binder, Frost, Hammeke, Bellgowan, Springer, Kaufman, and Possing 2000; Scott et al., 2000; Vouloumanos, Kiehl, Werker, and Liddle 2001). By using multiple, acoustically different, distortions and a correlational design, we were able to overcome an important methodological limitation of earlier studies; namely that differences in intelligibility were confounded with specific acoustic differences between intelligible and unintelligible stimuli.

The results point clearly to anatomical segregation consistent with hierarchical processing of speech. Sound (compared to silence) produced activation in the probable location of primary auditory cortex (e.g., Rademacher, Morosan, Schormann, Schleicher, Werner, Freund and Zilles 2001). Importantly, activation here did not correlate reliably with intelligibility; instead, the bilateral temporal-lobe region in which activation correlated with intelligibility is adjacent to this initial processing area. The form-dependent portion of this intelligibility-sensitive region may include some core auditory cortex and probably includes both auditory belt and parabelt areas (and beyond), and so probably subserves more than one processing stage (e.g. Kaas and Hackett, 2000; Rauschecker, 1998), although our data cannot speak to further functional segregation.

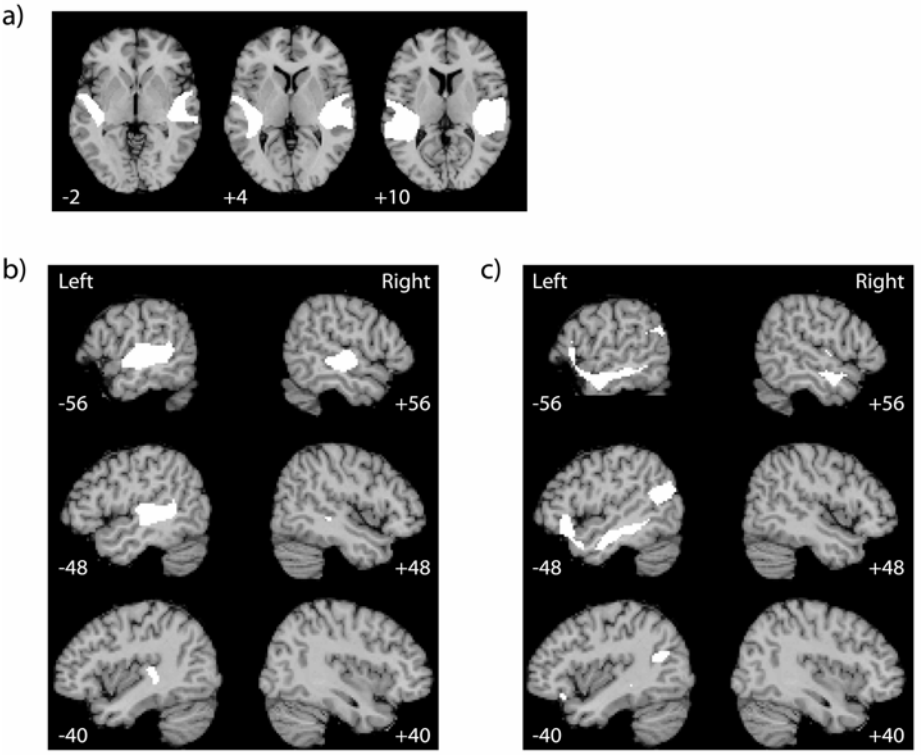


**Fig. 1.** Activations are shown superimposed on a canonical structural MR image from a single individual, and thresholded at $p < 0.001$, uncorrected for multiple comparisons. **(a)** Axial sections depicting areas in which activation was observed for signal-correlated noise relative to rest (excluding areas exhibiting a correlation with intelligibility) **(b)** Areas that correlate significantly with intelligibility, and show a significant difference in activation level across distortion types, indicating sensitivity to acoustic form. **(c)** Areas that correlate significantly with intelligibility, and do not differ among distortion types, indicating a lack of sensitivity to acoustic form.

Surrounding this periauditory form-dependent region anteriorly, posteriorly, and inferolaterally, we observed areas in which activation correlated significantly with intelligibility but was insensitive to acoustic differences among types of distortion. These areas may include some parabelt but are largely in what is probably polymodal cortex. We conclude that these form-independent areas are involved in processing speech at more abstract, non-acoustic levels of representation. The hierarchical structure that we infer from these results is consistent with cognitive accounts of spoken language comprehension (Gaskell and Marslen-Wilson 1997; McClelland and Elman 1986) in which lexical and semantic processes are driven by the output of lower-level acoustic and phonetic processes.

We also observed a form-independent, intelligibility-related response in left posterior superior temporal gyrus and left angular gyrus. These activations may be indicative of other, parallel streams of processing, extending posteriorly from auditory and form-dependent regions (Hickok and Poeppel 2000; Scott and Johnsrude 2003). Anatomical support for connections between auditory and inferior frontal cortex comes from studies of macaques (Hackett, Stepniewska and Kaas 1999; Romanski, Bates, and Goldman-Rakic 1999). Although the functional significance of these streams has yet to be firmly established, they may play a role in linking the perception and production of speech (Scott and Johnsrude 2003). In support of this account, a number of recent cognitive models have proposed separate processing pathways involved in phonological versus lexical processing of speech (e.g. Gaskell and Marslen-Wilson 1997).

## Acknowledgments

We thank the staff of the Wolfson Brain Imaging Centre, University of Cambridge for their help with data acquisition and Matthew Brett and Ian Nimmo-Smith for advice on image processing and statistical analysis. This work was supported by the Medical Research Council of the UK.

## References

Bashford, J.A.J., Warren, R.M., Brown, C.A. (1996) Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration. Percept. Psychophys. 58, 342-350.

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

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

Edmister, W.B., Talavage, T.M., Ledden, P.J., and Weisskoff, R.M. (1999) Improved auditory cortex imaging using clustered volume acquisitions. Hum. Brain Map. 7, 89-97.

Gaskell, M.G., and Marslen-Wilson, W.D. (1997) Integrating form and meaning: a distributed model of speech perception. Lang. Cog. Processes 12, 613-656.

Hackett, T., Stepniewska, I., and Kaas, J. (1999) Prefrontal connections of the parabelt auditory cortex in macaque monkeys. Brain Res. 817, 45-58.

Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., Gurney, E.M., and Bowtell, R.W. (1999) "Sparse" temporal sampling in auditory fMRI. Hum. Brain Map. 7, 213-223.

Hickok, G., and Poeppel, D. (2000) Towards a functional neuroanatomy of speech perception. Trends Cog. Sci. 4, 131-138.

Kaas, J., and Hackett, T. (2000) Subdivisions of auditory cortex and processing streams in primates. Proc. Natl. Acad. Sci. U. S. A. 97, 11793-11799.

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

Palmer, A.R., Bullock, D.C., and Chambers, J.D. (1998) A high-output, high-quality sound system for use in auditory fMRI. NeuroImage 7, S359.

Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werner, C., Freund, H.J., and Zilles, K. (2001) Probabilistic mapping and volume measurement of human primary auditory cortex. NeuroImage 13,669-683.

Rauschecker, J.P. (1998) Cortical processing of complex sounds. Curr. Opin. Neurobiol. 8, 516-521.

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

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

Schroeder, M. R., (1968) Reference signal for signal quality studies. J. Acoust. Soc. Am. 44, 1735-1736.

Scott, S.K., and Johnsrude, I.S. (2003) The neuroanatomical and functional organziation of speech perception. Trends Neurosci. 26, 100-107.

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

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

Vouloumanos, A., Kiehl, K.A., Werker, J.F., and Liddle, P.F. (2001) Detection of sounds in the auditory stream: Event-related fMRI evidence for differential activation to speech and nonspeech. J. Cog. Neurosci. 13, 994-1005.

Whalen, D.H., and Liberman, A.M. (1987) Speech perception takes precedence over nonspeech perception. Science 237, 169-171.

Xiong, J., Rao, S., Jerabek, P., Zamarripa, F., Woldorff, M., Lancaster, J., and Fox, P.T. (2000) Intersubject variability in cortical activations during a complex language task. NeuroImage 12, 326-339.

## Comments

### Comment by Duifhuis:

The activated areas you present are quite large and smooth, which might be due, at least partly, to the averaging across subjects, in addition to the smoothing window that you use. It would be interesting to see the individual data of at least two subjects per group in order for the reader to get an impression of the individual effects.

**Reply:**

Our studies were designed so that we would be able to generalize the results from the subjects we tested to the population at large. To that end, we collected a moderate amount of data from many subjects, and then used a random-effects analysis (Friston, Holmes, and Worsley 1999). Our study was not designed to gather enough data on each subject to make within-subjects analyses very meaningful. We recognize that this approach is different to the usual psychophysical approach of characterising the response of a few well-studied listeners, but believe it to be valuable, particularly in the early stages of a research program.

The large smooth areas of activation we observe may well overestimate the area of physiologically active tissue, since in imaging studies activation intensity (or significance) is confounded with activation extent. This would be true of the results of any neuroimaging study (PET or fMRI), even in single subjects. For this reason, inferences regarding the absolute size of activations are not easy to make. The conclusions that we draw in our paper do not rely on inferences about the absolute size of activations.

Friston, K.J., Holmes, A.P., and Worsley, K.J. (1999). How many subjects constitute a study? NeuroImage 10, 1–5.