Brain Mechanisms of Speech Perception: A Preliminary Report*

Colin Phillips[†]*, Alec Marantz, Martha McGinnis*[†]*, David Pesetsky,*
Ken Wexler, Elron Yellin: Massachusetts Institute of Technology
David Poeppel, Tim Roberts, Howard Rowley: University of California,
San Francisco

1. Introduction

This paper is a progress report on our work on speech perception using brain imaging techniques. We review our goals, methods, and two experimental paradigms we have used to study classic categorical perception effects, like the one illustrated in Figure 1. This graph illustrates one subject's identification of stimuli from a synthesized /dæ/–/tæ/ continuum.

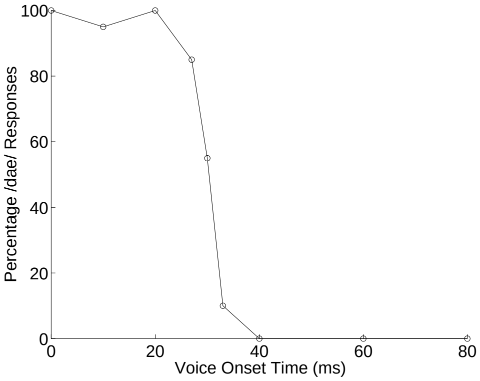

Figure 1
Categorical Perception of a /dæ/–/tæ/ Continuum

* [†]Authors of this paper. We would like to acknowledge the generous financial support for this work of a number of organizations: the School of Social Sciences and Humanities at MIT, NEC Corporation, NSF SGER grant #SBR-9423242, the McDonnell-Pew Center for Cognitive Neuroscience, and the NSF-funded RTG Program 'Language: Acquisition and Computation' (#DIR9113607). McGinnis was supported by a graduate fellowship from the Social Sciences and Humanities Research Council of Canada (#752-93-2393), and Poeppel was supported by a graduate fellowship from the McDonnell-Pew Center for Cognitive Neuroscience. We would also like to thank Al Galaburda for neuroanatomical advice; Susanne Honma, without whose skill and patience we would probably still have collected no data; Noel Massey for crucial assistance in preparation of synthetic speech stimuli; and Sam Williamson for his invaluable help at the start of this project and his continuing advice.

MIT Working Papers in Linguistics 26, 000–000
Papers on Language Processing and Acquisition. © 1995 Colin Phillips, Alec Marantz, Martha McGinnis et al.

The paper is aimed for a general linguistic audience rather than for experts in brain imaging or the physiology of speech perception, and we therefore assume no technical background in either of these areas. Specialists should feel free to skip sections, and may wish to refer to our technical papers for a more complete description of our procedures and results.

1.1 Goals

It is sometimes tempting to think that our task as linguists would suddenly become much easier if only we could open up the head and look at what's going on inside when people are speaking and understanding language. Instead of reasoning backwards to what's going on inside, from speakers' intuitions about acceptability and interpretation or from measurements of reaction time, we could just look and see. In reality, looking for language inside the head tends not to be all that revealing, because we have so little idea of what we're looking for. The structures and processes we search for in the brain don't look much like the structures and processes we talk about when studying language. Worse still, it's not that easy to look inside the head. This paper is a brief report on our progress in trying to get around these problems, using non-invasive techniques for measuring and localizing brain activity. Our goal is to give an idea of the kinds of linguistic questions that can be usefully addressed at present by studying brain activity, and the tools we are using to pursue these questions.

The technique we use to look inside the brain is known as magnetic source imaging (MSI). The main component of the technique is magnetoencephalography (MEG), a method that uses passive measurements of minute magnetic fields outside the head to make inferences about the timing and location of electrical activity in the brain. MEG offers an unusually good combination of temporal and spatial resolution, and data from MEG recordings can be combined with standard clinical magnetic resonance imaging (MRI) scans to identify the brain structures responsible for the activity. We give an overview of how MEG works, and what it is good for, in section 2.

1.2 A well-constrained problem

The series of studies we describe in sections 3 and 4 focuses on the neural basis of classic categorical perception effects in phonetics. The behavioral phenomenon is quite simple. Consider a series of synthesized syllables which form a continuum from /dæ/ to /tæ/, with the only difference among the syllables consisting in the timing lag between the initial noise burst and the onset of voicing (Voice Onset Time: VOT). It was found in the 1950s that when speakers of English are asked to identify members of this continuum, a syllable with a VOT below 30 ms or so will be perceived as voiced /dæ/, while one with a VOT above 30 ms will be heard as voiceless /tæ/. Despite the gradual change in VOT along this acoustic continuum, there is a sharp perceptual boundary between the categories of voiced and voiceless sounds (see figure 1). Also, when speakers are presented with pairs of syllables from the continuum, and asked whether they are acoustically the same or different, accurate discrimination is far more likely when the two syllables fall on either side of the 30 ms boundary than when both syllables are drawn from the same category.

The ability to distinguish phonetic categories is one of the most basic linguistic capacities humans possess. This ability makes it possible to interpret a stream of acoustic information as a series of speech sounds, which can then be assembled into words and phrases and associated with a meaning. This in itself is sufficient to make it very worthwhile to study categorical perception, but there is an important additional reason why this is a particularly promising area for brain-imaging studies of language. Existing work in a number of fields makes it possible for us to formulate specific hypotheses about the patterns of brain activity that we might expect to find.

The attraction of this domain is the fact that so much is already known in closely related areas. The acoustic stimuli are easily described and there is an enormous literature on behavioral studies of phonetic perception. We can also draw on a growing literature on both behavioral and neuroscientific studies of phonetic discrimination in animals, as well as on some useful studies of evoked responses to acoustic and phonetic stimuli using electroencephalograms (EEG) and MEG. These existing studies make it possible to formulate hypotheses about the kinds of activity we expect speech stimuli to elicit; they also help to constrain our interpretation of what we find. Almost none of these constraints would be available if we were studying syntactic phenomena, for example. It is not particularly difficult to find differences in brain waves, but it is much more difficult to interpret these differences when there are almost no constraints on the range of possible interpretations.

We studied the phenomenon of categorical perception in two sets of experiments, using brain imaging techniques to study neural activity in auditory cortex. We wanted to know first whether representations of phonetic categories are even accessible to auditory cortex, and secondly, if so, how these categories are represented in auditory cortex.

The experiments discussed in sections 3 and 4 address the role of auditory cortex in the perception of phonetic categories, specifically the feature [±voice], but the two sets of studies approach the question in different ways. The experiments discussed in section 3 extend a well-studied experimental paradigm to show that auditory cortex has access to phonetic representations. These studies tell us nothing about the question of *how* phonetic categories are represented in auditory cortex. To address this further question, the studies in section 4 test a proposal about the particular neural responses which lead to categorical perception of a VOT continuum.

2. Magnetic Source Imaging

In this section we outline the basics of the technique we use to measure brain activity. It is not crucial to understanding the point of the experiments described in sections 3 and 4, so the reader may wish to skip ahead to section 3, but it is useful for understanding the current possibilities and limitations of brain recording techniques, and why we chose the paradigms we did.

2.1 The basic idea, and the basic problems

Just as the flow of electrical current in a wire generates a magnetic field, so does electrical activity in neurons within the brain. In both cases, the direction of the magnetic field relative to the electrical current can be determined by the simple right-hand rule: imagine that an electric current is flowing along your right thumb, towards your thumbnail. The magnetic field that is generated by this current goes around the thumb in the direction of the fingers. The field strength can be calculated using the inverse square law, which states that the strength of a magnetic field decreases in inverse proportion to the square of the distance from the source. The simple relationship between current flow and magnetic fields forms the basis of the brain imaging technique we use, which is known as magnetoencephalography (MEG). Given measurements of a magnetic field in a number of locations outside the head, it should be possible to figure out the strength and position of the current source that generated the field. What is more, since the magnetic fields change in complete synchrony with changes in current flow, it should be possible to track changes in brain activity millisecond by millisecond.

When these ideas are applied in practice, of course, many complications arise. In the remainder of this section we discuss how some of the main difficulties have been overcome in implementing the ideas behind magnetic source imaging, and the constraints that these difficulties place on experimentation. The principal difficulties encountered are the following:

i. *Signal to noise ratio.* The strongest evoked brain responses generate fields with strengths on the order of a hundred femtoTesla (1 fT = 10^{-15} T), many orders of magnitude weaker than the magnetic noise in a normal working environment, and 100 billion times weaker than the earth's steady magnetic field. Measurements must therefore be extremely sensitive to tiny magnetic fields, and yet insensitive to much larger fields.

ii. *Irrelevant activity.* Most brain activity that we are likely to be interested in as linguists is drowned out by other activity in the brain, which for our purposes is just noise. So even with extremely sensitive field measurements, it's difficult to detect interesting activity.

iii. *Unconstrained inverse problem.* Despite the lawful relation between electric current and magnetic fields, the relation is not one-to-one. Any given pattern of magnetic fields measured at the scalp could have been generated by one of an infinite number of current-source configurations. These possibilities wouldn't pose too much of a problem if the brain were a serial processor, doing only one thing at a time, in one place at a time, but many current sources in the brain can be active at the same time. Therefore, even if we've solved the problems in (i) and (ii), the solution to the inverse problem is badly underdetermined.

iv. *Unmeasurable fields.* The problem in (iii) is exacerbated by the fact that certain configurations of current flow which are known to be widespread in the brain have the property that the magnetic fields they generate partially cancel one another out or they are shielded by the scalp and therefore have no measurable effect outside the head. These configurations involve both particular kinds of cells and cells in particular orientations relative to the scalp.

v. *Lack of structural information.* Even if we are successful in determining the current source that generates a pattern of magnetic fields at the scalp, the localization is to an arbitrary point in space rather than to an area of the brain. MEG doesn't tell us which brain structures are responsible for the activity we're measuring.

In the next few pages, we explain how each of these problems is solved, and also point out the practical constraints they place on experimentation.

2.2 Measuring minute fields

Any instrument for measuring magnetic fields generated by the brain has to solve two problems. First, it needs to be sensitive enough to measure magnetic fields that are vanishingly weak. Secondly, it needs to be able to distinguish these tiny fields from the many stronger fields in the environment.

The first problem is solved by the technology of superconducting quantum interference devices (SQUIDS). The heart of the MEG sensor is a coil cooled to -269˚ Celsius in a bath of liquid helium. At this temperature the coil is a superconductor, and very small changes in magnetic field induce changes in the current flow through the coil. The SQUID electronics amplify these changes.

The problem of distinguishing weak nearby sources from strong distant sources is solved in two ways. First, the subject and the MEG sensors are enclosed in a magnetically shielded room. Second, the inverse square law can be used to advantage. As noted above, the strength of a magnetic field falls off in inverse proportion to the square of the distance from the source. The source of the earth's steady field is many thousands of miles away from the MEG sensors. The source of a magnetic field generated by activity in the brain, however, is only a few centimeters away. If we measure the magnetic field at a pair of coils 2 centimeters apart, the difference in the strength of the earth's steady field between the two coils is negligible, even though the field itself is very strong. On the other hand, two coils placed 2 cm apart can register a considerable difference in the strength of the magnetic field generated by a source in the brain just 4 or 5 centimeters away.[1]

To exploit this difference between distant and nearby sources, MEG sensors often contain first-order gradiometers, consisting of a pair of superconducting coils wound in opposite directions, with one slightly closer to the scalp than the other. Current induced in the two coils by fields with distant sources cancels out (because the coils have opposite directions), causing no net current flow. Current induced by nearby cortical magnetic fields does not cancel out, and therefore has a measurable net effect on the current flow in the sensor.

2.3 Identifying activity of interest

Sorting out brain activity of interest from the immense amount of background noise in the brain at any moment in time is a problem for magnetic measurements as well as for any other technique of measuring brain activity. The straightforward solution is to average over responses to a large number of identical stimulus presentations.

The number of stimulus presentations needed to obtain a useful signal-to-noise ratio depends on which kinds of responses are of interest. For the middle-latency (50–200 ms) auditory evoked responses we focus on here, about 100 presentations are needed for each stimulus. Since subjects must remain awake and motionless during an experiment, this requirement obviously constrains the number and length of feasible stimuli. As an illustration, in the experiments we describe in section 3 we were interested in the responses to two 300 ms stimuli. These had to occur 100 times each, and the design of the paradigm required that they occur only one-eighth of the time. Even when stimuli are presented at a rate of one per second, this led to runs of 30 minutes each, which approaches subjects' tolerance level.

2.4 Inverse problem

The main advantage of MEG over EEG—and what justifies the massive extra cost of MEG—is the improved opportunity it offers for accurate localization of neural activity. Unlike the electrical fields measured by EEG, the magnetic fields measured by MEG are not strongly distorted by brain tissue and the skull. By measuring magnetic fields at a number of different points outside the head, it is therefore possible to infer the location of the current source inside the head which is generating the field. The system we use in our experiments (Magnes, BTi, San Diego) consists of 37 sensors arranged on the surface of a sphere with a diameter of 14.4 cm. Figures 2a and 2b show averaged and low-pass filtered waveforms from the evoked field measured at each of the sensors, and a corresponding contour map of the magnetic field at one instant in time.

[1] The reason for this is that the retio of $(n+2)^2$ to n^2 decreases, asymptotically approaching 1 as n increases.

Since the field extrema of the dipolar field are captured within the array, the location of its source can be estimated with a high level of confidence.

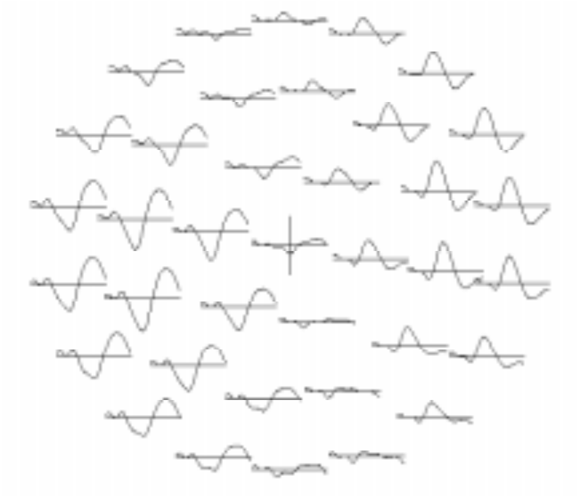

Figure 2a
Sensor Array Display of M100 Response
(200 ms interval, 120 ms marked on middle sensor)

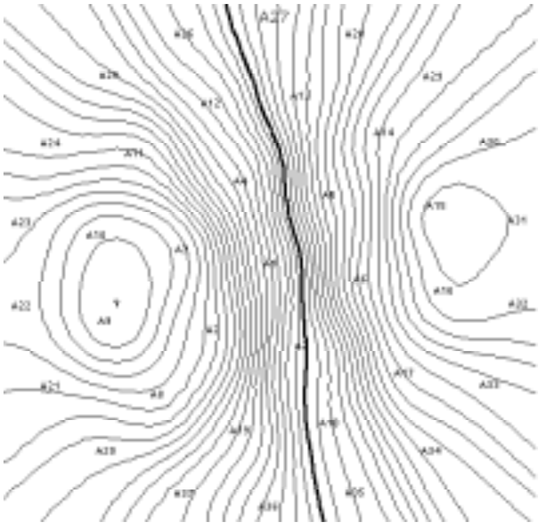

Figure 2b
Contour Map derived from Sensor Array in Figure 2a at 120 ms

Reasoning backwards from the scalp distribution of magnetic fields to the current that generated them, however, is far from simple. For any pattern of magnetic fields at the scalp there are, in principle, infinitely many source configurations. The task of the localization algorithm is to choose the source configuration most likely to be correct. Most of the widely used localization algorithms simplify this task by constraining the localization algorithm to look for the *single* current source, or dipole, that best fits the distribution of magnetic fields on the scalp.

Although a single-dipole model clearly cannot model multiple simultaneous processes in the brain, it appears to be remarkably successful in some domains. Single-dipole models of the auditory evoked responses we describe below regularly achieve a correlation of 0.99 or better between the model's predictions and the observed magnetic fields.

So long as we have access only to a single-dipole model, we are largely constrained to studying brain responses that can be described to a reasonable degree of accuracy with a single dipole model. Several groups are working on developing localization algorithms that allow more complex models, but relaxing the constraints on the model massively increases the computational complexity of the inverse problem.

2.5 What kinds of activity are detected?

A good deal of electrical activity in the brain simply isn't detectable using MEG. First, the current generated by cells in certain configurations fails to have any effect on the magnetic field measured at the scalp. Second, in order to generate an evoked field that is strong enough to be detected and modelled using existing techniques, hundreds of thousands of cells must be acting together. There is sure to be a large amount of interesting neural activity that does not meet this criterion.

The net magnetic field generated by a given cell depends very much on the arrangement of a cell's dendrites. If electrical current is flowing concurrently in dendrites pointing in all directions from a cell's body, then the net magnetic field associated with that cell will be effectively zero, since the fields generated by the individual dendrites cancel one another out. Unfortunately, this factor alone eliminates an enormous amount of brain activity from study. However, it does make life easier in some ways. It is thought that much of the activity detected by MEG originates in pyramidal cells in layer III of cortex, cells whose dendritic patterns are not self-cancelling. It is also known that these cells are generally oriented perpendicular to the surface of cortex. Thus, it is possible to figure out which areas of cortex are most likely to generate a strong MEG signal, using the following reasoning.[2]

i. The surface of cortex is convoluted into numerous folds, which can be separated into concave fissures, or sulci, and convex ridges, or gyri.

[2] For a more detailed, but highly accessible, discussion of what kinds of activity are believed to be detectable using MEG, see Lewine & Orrison (1995).

Pyramidal cells generate net *current flow* in an orientation perpendicular to this surface.

ii. Given the right hand rule, the *magnetic field* generated by pyramidal cells is oriented tangential to the cortical surface.

iii. It is independently known that fields oriented tangentially to the scalp are effectively shielded by the skull, and thus unmeasurable outside it. Only fields with a radial component relative to the scalp are detectable outside the head.

iv. Therefore, the electrical activity due to pyramidal cells inside sulci will be relatively easy to detect, since it generates a magnetic field with a large radial component, but activity due to cells on gyri will not be at all easy to detect, since most of a gyrus is parallel to the scalp.

The difficulty of detecting activity with a gyral source is a major problem for anybody interested in studying parts of cortex known to be located on gyri. Fortunately for our purposes, however, much of auditory cortex is known to be located on the upward-facing surface of the deep sylvian fissure, so activity in auditory cortex is relatively easy to measure.

2.6 Structural information

Assuming that a plausible dipole model has been computed, all we know from MEG are the coordinates of the current source in space. We do not know what area in the subject's brain is represented by those coordinates. To localize the current source in the brain, we combine MEG technology with information derived from structural MRI scans. By using reference points common to both MRI and MEG data, MEG localization data can be superimposed on structural MRI images. An experienced radiologist or neurologist can then interpret the results. This technique, known as magnetic source imaging (MSI), has proved especially valuable in clinical practice. For example, MSI data have been used to identify the location of epileptic foci in the brain, and so determine the minimal surgery needed to reduce epileptic seizures. Since the structure and function of auditory cortex are little understood, the MSI technique is of limited use in cognitive research on auditory processes, such as understanding speech. Correlations with MRI scans do, however, act as an additional test for the plausibility of MEG localization data. To take a simple example, since the fields detected by MEG are probably generated by sources in cortical grey matter, any localization which points to other areas may be treated with some skepticism.

To summarize this section, then, while MEG is a remarkably powerful tool, as with other functional neuroimaging methods it still leaves us far from being able to "open up the head and look inside." At present there are still extremely tight constraints on the kinds of brain activity that can usefully be studied using MEG. We now turn to a description of some of our experiments, which are an attempt to use MEG technology to address questions that linguists might be interested in.

3. Phonetic contrasts in auditory cortex: Evidence from MMF

This section describes a series of experiments which we believe lead to the conclusion that auditory cortex accesses representations of phonetic categories, in particular the categorial representations of /dæ/ and /tæ/. We show this using a novel variant of the widely studied “oddball” paradigm.

The oddball paradigm evokes a characteristic brain response to an infrequently presented stimulus when it is appears among a series of frequently presented stimuli. This characteristic response is known as a “mismatch” response. Previous research using the oddball paradigm has shown that a mismatch response is elicited by a wide range of acoustic contrasts involving properties such as pitch, length and amplitude. These contrasts can be at or below the threshold of conscious discrimination. It has also been shown that the generator of the mismatch response localizes to an area of auditory cortex close to the generator of the auditory M100 response.[3]

A number of studies have addressed the question of whether a mismatch response may be elicited by phonetic contrasts, but with differing conclusions. The largest obstacle to answering this question is the fact that phonetic contrasts are almost always also acoustic contrasts.[4] In this section we first review some of these existing studies and show how their design made it difficult to differentiate responses to acoustic and phonetic properties of the stimuli. We then go on to describe a series of our own experiments in which acoustic and phonetic contrasts are much easier to distinguish.

3.1 The oddball paradigm

In the oddball paradigm, a subject is presented with a series of identical acoustic stimuli (“standards”), among which contrasting stimuli (“oddballs” or “deviants”) occur with a frequency of 20% or less. An example of the design of a typical oddball study, contrasting tones at 1000Hz and 1100Hz, is shown in figure 3.

[3] The M100 is an automatic response elicited by auditory stimuli, with a latency of typically 100–130 ms. The “M” of M100 stands for “magnetic.” It is the magnetic correlate of the negative wave N100 measured using EEG. The characteristic labeling of evoked waves as either positive (e.g., P300) or negative (e.g., N200) in the EEG literature is not followed in the MEG literature, presumably because of the emphasis in MEG work on measuring both positive and negative field extrema of any response in order to localize its source.

[4] See below for discussion of a rare situation in which a phonetic contrast is not also an acoustic contrast.

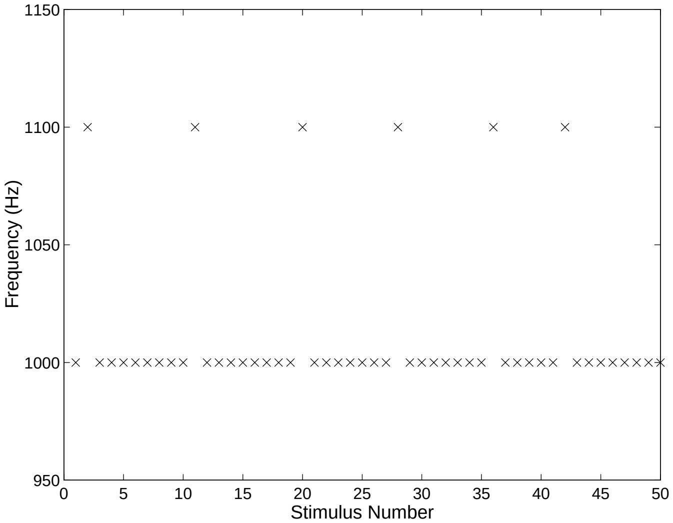

Figure 3
Design of oddball study using 2 stimuli.

The oddball stimulus has been found to elicit a certain response, occurring 180–250 ms after stimulus onset, and localizable to auditory cortex (Näätänen et al. 1978; Näätänen 1992; Hari et al. 1984; Sams et al. 1985). This response is known in the EEG literature as the mismatch negativity (MMN), or in the MEG literature as the magnetic mismatch field (MMF). The MMN/MMF is typically derived by subtracting the averaged response to standard stimuli from the averaged response to oddball stimuli.[5] The extra response component to the oddball stimulus which emerges from the subtraction of standard from deviant (figure 4) has been shown to be due to an independent response in auditory cortex.

[5] Ideally, a study using the oddball paradigm should come in two parts, as follows. In the first part of the experiment, one of the stimuli, A, is the standard, while B is the oddball. In the second part, B is the standard and A is the oddball. This makes it possible to compare responses to A as standard against A as deviant, and B as standard against B as deviant. In practice, however, it is common to compare A as standard against B as deviant. This kind of comparison obviously introduces possible confounds due to the fact that A and B are physically different stimuli.

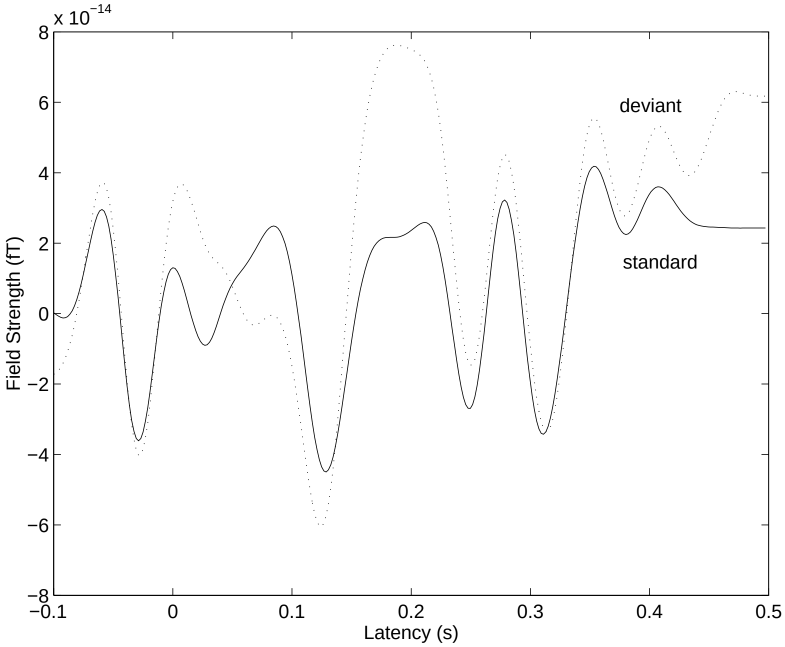

Figure 4
Averaged waveforms to standards and deviants at one sensor position

3.2 Previous studies of phonetic contrasts and the mismatch response

As mentioned already, a number of different kinds of acoustic contrast have been shown to elicit mismatch responses. These include contrasts in frequency, duration, amplitude, and interstimulus interval, to name just a few. A natural question which arises is whether phonetic contrasts also elicit a mismatch response. This turns out to be a rather difficult question to ask, since phonetic contrasts are also acoustic contrasts. For any mismatch response elicited using a phonetic contrast, we have to determine whether the response could be explained solely on the basis of the acoustic contrast between the stimuli. A number of studies have tried to circumvent this difficulty in various ways.

3.2.1 Dimensional additivity

One strategy for separating effects of phonetic versus acoustic contrasts is to take advantage of the fact that a fixed acoustic contrast may be phonetically relevant in some contexts and not in others. For example, the difference between a 20 ms VOT stop and a 35 ms VOT stop is phonetically relevant for most speakers of English, but the difference between a 20 ms VOT stop and a 5 ms VOT stop is not phonetically relevant for the same speakers of English, although the acoustic contrast is the same in both pairs of syllables.

It is known from mismatch studies using contrasts between pure tones that larger acoustic differences along a single dimension yield stronger mismatch responses (Aaltonen et al. 1987). It has also been shown that when deviants contrast with standards on two acoustic dimensions (e.g. length and pitch) a larger mismatch response is elicited than when standards and deviants contrast on just one of those dimensions. In other words, the strength of the mismatch response is dimensionally additive (or at least the dimensions are positively interactive). A number of studies have taken as a premise the assumption that phonetic contrasts between standards and deviants should be treated as just another dimension, and reasoned from this that if the generator of the mismatch response is sensitive to phonetic contrasts then a fixed acoustic contrast should elicit a stronger mismatch field when it also spans a phonetic boundary (Sams et al. 1990; Sharma et al. 1993; Maiste et al. 1995).

For example, Sharma et al. (1993) used two pairs of stimuli along a synthetic /da/–/ga/ continuum in their study. The continuum varied the starting frequencies of the second and third formants. In both conditions they used as the standard stimulus a syllable categorized as /da/ with a 2537Hz F3 onset. In one condition the deviant had a 175Hz higher F3 onset, and was categorized as /da/, and in the second condition the deviant had a 175Hz lower F3 onset, and was categorized as /ga/. Therefore, the acoustic difference between standards and deviants was identical in both conditions, but in the second condition there was also a phonetic contrast between the standard and the deviant.

As it turned out, none of these studies found any increase in the mismatch response when the acoustic contrast was augmented by a phonetic contrast. Sharma et al. (1993) concluded from this that the generator of the mismatch reponse is not sensitive to phonetic contrasts, and accesses a purely acoustic representation of stimuli. However, as Maiste et al. (1995) point out, there is another possible reason why dimensional additivity was not found in these studies: it could be that the generator of the mismatch response is sensitive to phonetic contrasts, but that the effect of this may be much smaller than the mismatch response generated by acoustic contrasts. Since there is always a constant acoustic contrast between standards and deviants, the response to this contrast may obscure the additional response to the phonetic contrast.

3.2.2 Variation among standards

A study by Aulanko et al. (1993) took a rather different approach to separating acoustic and phonetic contrasts in the oddball paradigm, one which we build on in our studies. Aulanko et al. assumed that if two stimuli are acoustically distinct and perceptibly different, they can only be treated as a group if they share some phonetic property. Their study contrasted syllables which were perceived as /bæ/ and /gæ/, but used a number of different tokens of each category, varying in fundamental frequency. This meant that instances of the phonetic categories /b/ and /g/ were being heard in the requisite proportions to elicit a mismatch response, although no single acoustic stimulus was frequent.

When the waveforms were averaged over the acoustically different tokens of each phonetic category, a mismatch effect appeared. Aulanko et al. argued that their result could only be produced if subjects were grouping together stimuli of each phonetic category, and the generator of the mismatch response was able to access these groupings.

Aulanko et al. concluded that the effect was purely phonetic, since the stimuli grouped together were acoustically different. However, their result could be explained without reference to phonetic contrasts. Within each phonetic category stimuli did differ in F0 frequency, but the different stimuli also shared the characteristic formant transitions distinguishing a labial /bæ/ from a velar onset /gæ/. Therefore the stimuli may have been identified by an acoustic rather than a phonetic categorization: if subjects attended just to frequencies above 200Hz in the stimuli, i.e., to the formants but not to the fundamental, then they would have been hearing just two stimuli, distinguished by a fixed acoustic difference.

Although we are not convinced by the interpretation that Aulanko et al. gave to their results, their design represents a crucial step towards differentiating acoustic and phonetic contrasts, and we build on this in our studies.

3.3 Phonetics without acoustics: Experiment 1

3.3.1 Design

In our studies we follow the method of Aulanko et al. (1993), in that we use a number of acoustically different stimuli from two phonetic categories that we contrast. Our design differs from the one used by Aulanko et al. in that the acoustic parameter that varies within categories is the same acoustic parameter that distinguishes between the two categories.

We contrasted four acoustically different tokens from each of the two phonetic categories /dæ/ and /tæ/. These tokens differed along the VOT dimension, which also served to distinguish the categories themselves and therefore the differences within and across category were identical. Moreover, there was no fixed acoustic property that could be used to distinguish tokens of one category from tokens of the other category. Everything that stimuli from one category had in common was common to *all* stimuli in the experiment.

We presented subjects with a randomized series of the eight acoustic stimuli, proportioned so that the two phonetic categories occurred in a 7:1 ratio. 87% of the stimuli were randomly selected from the 4 tokens of the “standard” category, and 13% of the stimuli were randomly selected from the 4 tokens of the “deviant” category. Figure 5 shows the kind of series of stimuli that would be used for a hypothetical subject with a perceptual boundary of 28 ms VOT for /dæ/ versus /tæ/. In the first half of the experiment /dæ/ is the standard and /tæ/ is the deviant; in the second half of the experiment /tæ/ becomes the standard.

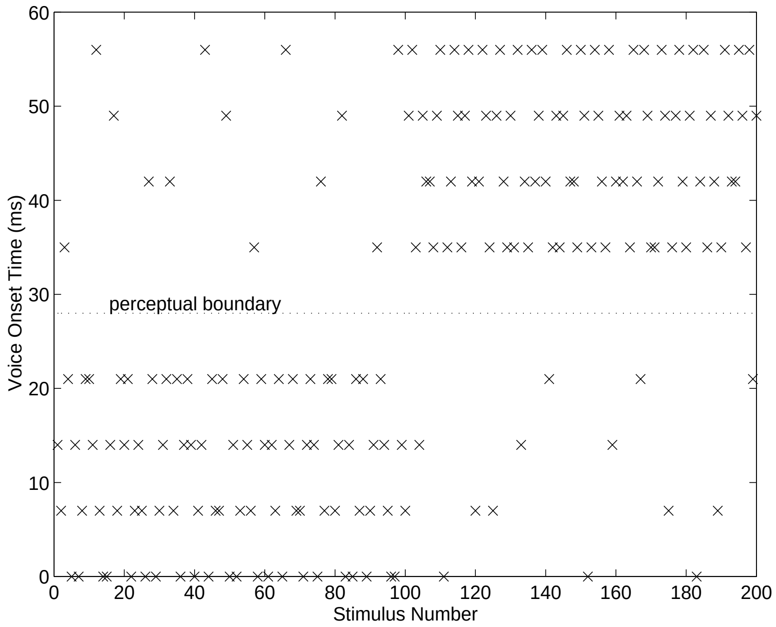

Figure 5
Design of Phonetic Mismatch Experiment

This design makes it easier to distinguish phonetic and acoustic contributions to the mismatch reponse. If the stimuli are being labelled phonetically, the two stimuli /dæ/ and /tæ/ occur in a ratio sufficient to elicit a mismatch response. However, if the stimuli are being labelled acoustically, eight distinct stimuli occur, each with no more than 22% probability. Moreover, the difference between consecutive stimuli could be as much as 21 ms VOT for stimuli of the same phonetic category, and as little as 14 ms VOT for stimuli from different categories. Therefore, if a mismatch response is elicited, it must be due to the perceptual grouping of acoustically different stimuli, and we can be fairly confident that the generator of the mismatch response is able to access phonetic representations.

3.3.2 Procedure

We tested 9 normal, right-handed native speakers of English (4 male, 5 female). Eight appropriate stimuli from a /dæ/–/tæ/ continuum, synthesized using a Klatt speech synthesizer (Klatt 1980) were selected for each subject, based on the results of a behavioral pretest (forced choice identification) which measured subjects' individual perceptual boundaries. 4 of the 8 stimuli used in the MEG experiment were reliably classified as /dæ/ in the pretest (90%+ accuracy), and the remaining 4 stimuli were reliably classified as /tæ/ in the pretest. The 4 different stimuli from each category were acoustically evenly spaced, with increments of 6–8 ms VOT between each stimulus. The acoustic difference between the longest VOT /dæ/ and the shortest VOT /tæ/ was double the within-category increment size. For example, for a typical subject with a 28 ms VOT perceptual boundary we used stimuli with VOTs 0 ms, 7 ms, 14 ms, 21 ms, 35 ms, 42 ms, 49 ms, 56 ms. One consequence of this is that the acoustic difference between categories (e.g., 35ms – 21ms = 14ms) is smaller than the largest within-category acoustic difference (e.g., 56ms – 35ms = 21ms).

The first half of the experiment had standard /dæ/ and oddball /tæ/ (see figure 5 above); in the second half, /tæ/ was standard and /dæ/ was the oddball. The standard:deviant ratio was 7:1. Stimuli were presented in pseudo-random order, with no consecutive oddballs. Each stimulus lasted 300 ms, and the inter-stimulus interval varied randomly between 700 and 850 ms. Each run consisted of 1600 epochs, with 700 instances of each category as standard and 100 instances of each category as an oddball. Subjects were lying on their side on a bed inside a magnetically shielded room; they were instructed to close their eyes and to remain awake and not to attend to the stimuli.[6] Runs lasted around 30 minutes, and included 4 breaks of a few seconds which allowed subjects to adjust their position if necessary.

We measured the extracranial magnetic field using a 37-channel first-order gradiometer (Magnes, BTi, San Diego, CA). The dewar containing the sensor array was positioned above the left ear, so as to record from left auditory cortex. Stimuli were presented to the right ear through an eartube. The appropriate sensor position was determined by playing a series of 500 Hz tones, which are known to generate an M100 response in auditory cortex. The sensor array was adjusted until it was centered directly above the source of the M100 response to these tones.

Responses were recorded with a sampling rate of 500Hz and a bandwidth of 200Hz. Epochs of 600 ms were recorded, lasting from 100 ms pre-stimulus onset to 500 ms after stimulus onset. Recording epochs were first low-pass filtered from 1–20Hz, and then averaged according to event type, where each phonetic category as a standard and as an oddball constituted a different event type.

In order to reduce the variance of the grand average waveforms, the averaging process had two stages. For each of the 4 event types, stimuli were first grouped into 10 subaverages, each derived from 10% of the filtered epochs of that event type. Next, each set of 10 subaverages was combined to form a grand average waveform. Since a good deal of irrelevant brain activity is

[6] It is common practice in experiments employing the mismatch paradigm to distract subjects from the auditory stimuli using a visual stimulus: subjects either read a self-chosen book or watch a silent movie. This was not possible in the present study for practical reasons. However, the absence of a distractor task is unlikely to have affected our results. The visual distractor was introduced into EEG studies of MMN in order to remove the auditory N200 response, which tends to obscure the MMN response, but which is suppressed when the subject is not attending to the auditory stimulus. M200 and MMF do not overlap to the same extent in MEG recordings.

cancelled out in the subaverages, the variance of the grand averages is much smaller than it would be if individual recording epochs were averaged in one stage.

The difference between grand-average waveforms for responses to standards and deviants was tested using *t*-tests, one for each recording sample (i.e., every 2 ms over a 600 ms time window) and for each of the 37 recording sensors.

3.3.3 Results

We compared responses to standards and oddballs from each of the perceptual phonetic categories, /dæ/ and /tæ/. Across subjects, responses to standards and oddballs were extremely similar until a latency of 180–240 ms, when larger differences between responses to standards and oddballs appeared. Raw difference waves are shown for a representative subject in figure 6a; figure 6b displays the same difference waves as a function of standard deviations of the grand-average responses to the deviants. Figure 6b shows that differences between standards and deviants first reach significance at around 200 ms latency.[7]

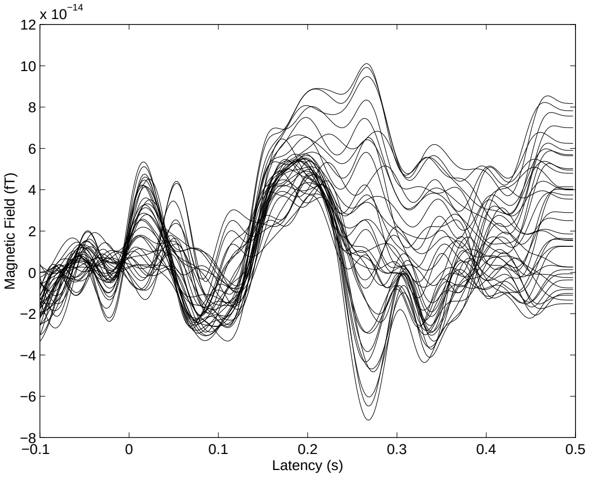

Figure 6a
Subject AZ, difference wave in /dæ/ mismatch study

[7] It is generally suspect practice to draw conclusions from the results of 12480 *t*-tests for each comparisons and for each subject. However, a number of factors make us believe that it is warranted to interpret these significance levels here. First, the greatest difference between standards and deviants falls in the time period that we predict, based on existing studies using the mismatch paradigm. Second, the area of greatest difference is consistent across subjects. Third, when *t*-tests yield significance over a number of successive time samples, the normal risks of multiple *t*-tests are lessened (Guthrie and Buchwald 1991).

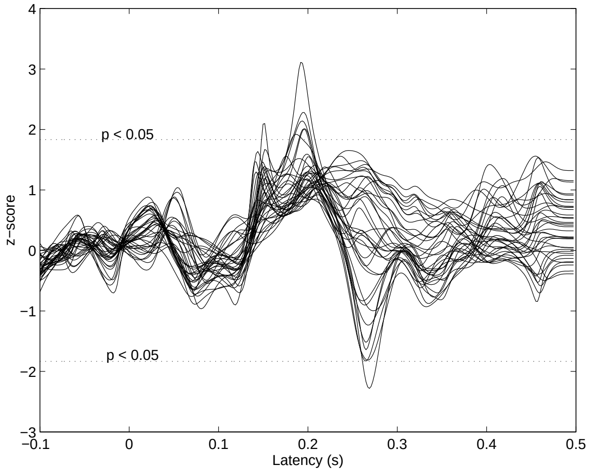

Figure 6b
Subject AZ, normalized difference waves in /dæ/ mismatch study

Figures 7a and 7b show that both the temporal and spatial distribution of the difference waves correspond to what is reported in the literature for acoustic oddball responses. We therefore conclude that our phonetic oddball paradigm is eliciting a classic magnetic mismatch field (MMF).

Figure 7a shows the distribution across the 37 sensor positions of the difference waves in figures 6a and 6b. The vertical lines mark the peak of the difference waves at a latency of 190ms. Comparison of the difference waves in figure 7a with the M100 waves in figure 2a shows that the difference waves have a very similar distribution to the M100 waves, although the peaks of the difference waves are slightly more anterior to the peaks of the M100 waves. This suggests that the source of the difference waves is located slightly anterior to the generator of the M100 response, which is known to be located in auditory cortex.

Since the difference waves in figure 7a have a clearly dipolar distribution, it is feasible to ask what source in the brain is responsible for generating the difference waves. Figure 7b shows an overlay of the localization of the difference magnetic field in figure 7a on slices of the subject’s structural MRI scan. As can be seen from the axial slice (upper right: image displayed as if looking from below, eyes at the top) and the coronal slice (lower left: looking from the front), the difference field localizes to a source on the lower surface of the deep sylvian fissure in the left hemisphere. This is supratemporal auditory cortex, which is where acoustic oddball responses are known to be generated.

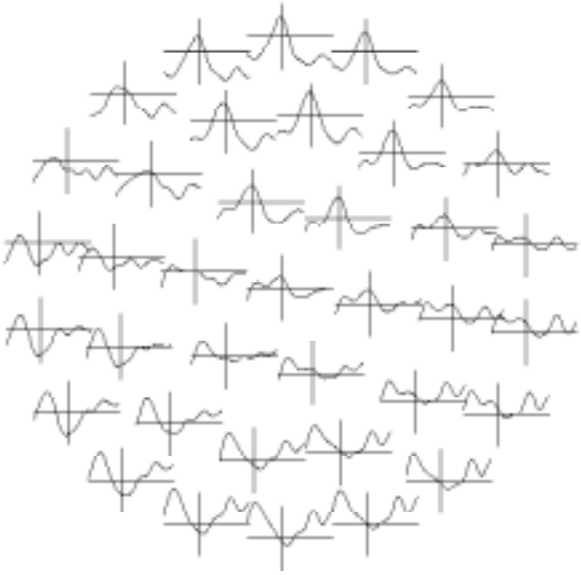

Figure 7a
Subject AZ, Sensor Array Display of /dæ/ mismatch
(latency 190 ms; 200ms time interval displayed)

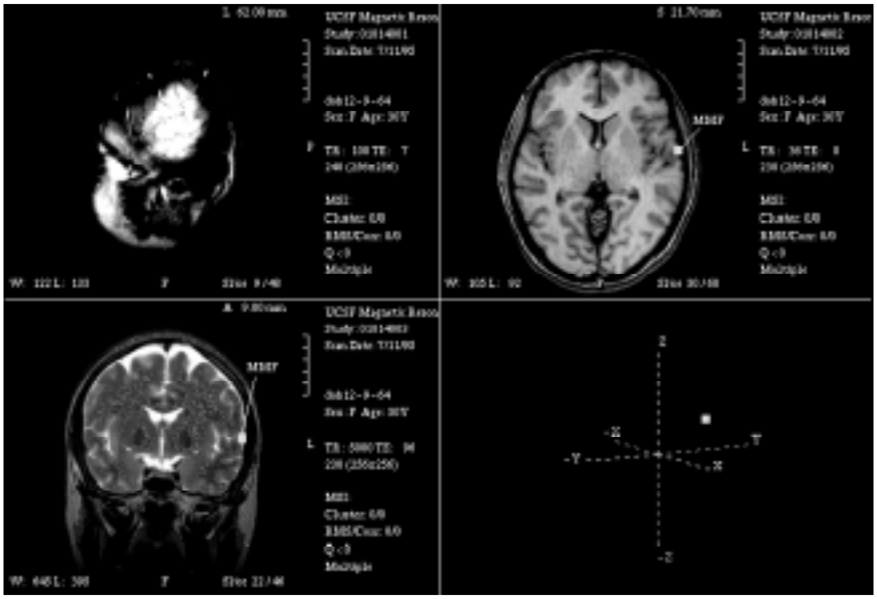

Figure 7b
Subject AZ, MRI Overlay of /dæ/ mismatch field, 190ms latency

Given that our design contained no acoustic oddballs, only phonetic oddballs, we conclude that the generator of the MMF in auditory cortex accesses phonetic category information about /dæ/ and /tæ/.

3.4 Is this really phonetics?

We have argued that the MMF elicited by experiment 1 could only be the result of accessing phonetic categories. However, a skeptic might object that although no individual acoustic stimulus occurs frequently in experiment 1, stimuli might be grouped into two ad hoc acoustic categories, each containing stimuli which are acoustically similar in some respect. If these ad hoc categories can be constructed based on acoustic groupings, then the result of experiment 1 could be explained without making any reference to phonetic categories. We designed two control experiments to test whether the MMF observed in experiment 1 was an effect of phonetic or acoustic grouping of stimuli. One control tested directly for the possibility of acoustic grouping by using a set of stimuli which were acoustically analogous, but not phonetically analogous, to those used in experiment 1. The other control used speakers of two different languages and a phonetic contrast that is used in only one of the languages. This control shows that when using the design in experiment 1, a given set of stimuli elicits an MMF only if the grouping of stimuli is phonetically salient for the speaker being tested.

3.4.1 Experiment 2: Longer VOTs

3.4.1.1 Design

The design of the first control was identical to that of the main experiment, except that the VOT of all stimuli were increased by 15–20 ms, so that they showed the same acoustic clustering as in experiment 1, but the acoustic groupings no longer corresponded to phonetic groupings. For example, for a subject with a 28 ms perceptual boundary on the /dæ/–/tæ/ continuum, we used the following stimuli in experiment 2: 15, 22, 29, 35, 50, 56, 63, 70 ms. Now only two of the stimuli were from the /dæ/ category. All other features of the design of experiment 1 were held constant. “Low VOT” stimuli and “high VOT” stimuli were still presented in a 7:1 ratio. However, the perceptual categories /dæ/ and /tæ/, which in the first half of experiment 1 occurred with probabilities of 88% and 12% respectively, occurred with probabilities 44% and 56% in the first half of experiment 2. This is because the perceptual boundary now falls between stimuli 2 and 3 from the “low VOT” group, as figure 8 shows.

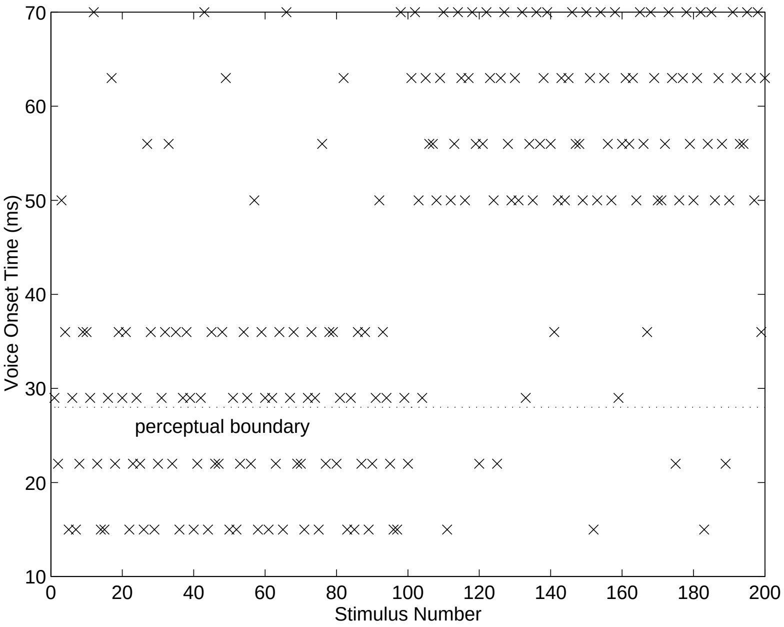

Figure 8
Design of Mismatch Control Experiment

In the second half of experiment 2 standards and deviants were reversed, just as in experiment 1. When the standards and deviants are reversed in this experiment, a phonetic oddball appears, because now all of the high VOT standards and half of the low VOT deviants are phonetic /tæ/s, and only the remaining deviants, which make up only 6% of the stimuli, are phonetic /dæ/s.

We predicted that if the MMF in the main experiment was due to acoustic grouping of stimuli, then the same MMF would appear in the control experiment, both in the difference between high VOT standards and deviants, and in the difference between low VOT standards and deviants. If, on the other hand, we were correct in claiming that the MMF observed in experiment 1 was due to phonetic grouping of stimuli, then the MMF should not be found in the comparison of high VOT standards and deviants in experiment 2. A mismatch response might be expected in the comparison between low VOT standards and deviants, due to the effect of the extremely infrequent /dæ/ stimuli when low VOTs are deviants.

Six of the subjects from experiment 1 participated in this experiment (3 men, 3 women). Stimulus presentation, data acquisition and data processing were identical to experiment 1.

3.4.1.2 Results

We compared the grand-average responses to the standards with the grand-average responses to the deviants for each of the two acoustic groups “low VOT” and “high VOT.”

No mismatch response was observed in the difference between high VOT standards and deviants. For some subjects a mismatch response was observed in the difference between low VOT standards and deviants. Figures 9a and 9b show the relevant contrasts for a representative subject. Figure 9a shows a large difference wave around a latency of 200 ms for the contrast between /tæ/ standards and deviants, while figure 9b shows no large difference wave for the contrast between high-VOT standards and deviants.[8] As before, the magnitude of the difference wave is displayed in multiples of the standard deviation of the grand-average waveform for the deviants.

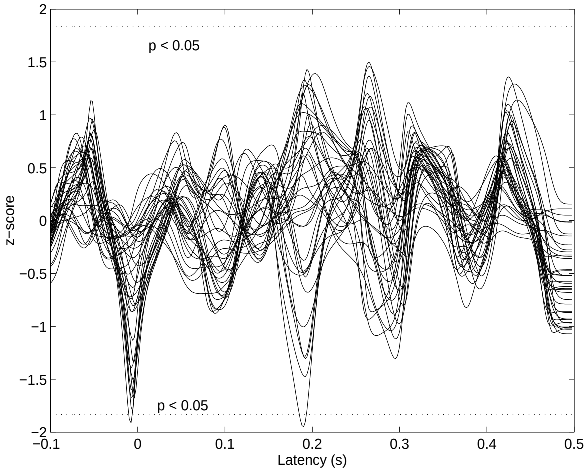

Figure 9a
Subject LS, normalized difference wave, /tæ/ mismatch

[8] The extremely large normalized difference wave in figure 9a around -10 ms latency is the result of a “spill-over” of the response to the preceding stimulus. The raw difference waves at this point are not particularly large, but because they have such a small variance the difference converts to large z-scores.

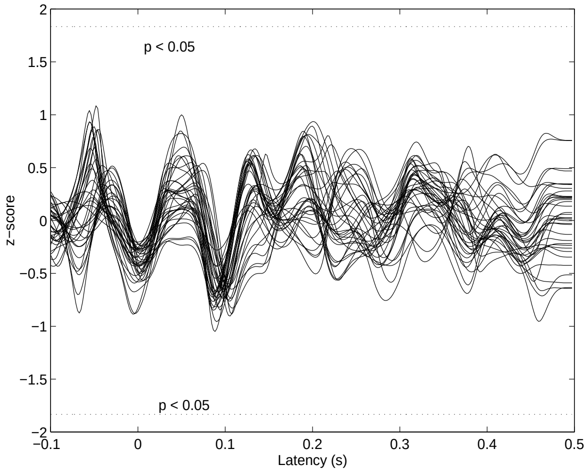

Figure 9b
Subject LS, normalized difference wave, high VOT mismatch control

Subjects varied with respect to whether a significant mismatch response was observed in the difference between low VOT standards and deviants. This is not surprising, given the fact that the 4 low VOT stimuli consist of 2 instances of /dæ/ and 2 instances of /tæ/. This means that low VOT deviants are a combination of phonetic standards and phonetic deviants, a factor which could easily obscure any phonetic mismatch response in the averaged waveforms.

Therefore, the results of experiment 2 lend further support to our interpretation of experiment 1: the generator of the mismatch response in auditory cortex must have access to representations of phonetic categories, since the observed MMF is not the result of ad hoc acoustic categories.

3.4.2 Experiments 3 and 4: Cross-linguistic differences

The second control is work in progress, carried out in collaboration with Morihiro Sugishita and colleagues at the University of Tokyo Medical School. The design of this control is identical to that of our main experiment, except that in this control we are using a synthesized /la/–/ra/ continuum.[9]

As is well known, the distinction between /l/ and /r/ is not phonetically relevant in Japanese, and consequently native speakers of Japanese typically do not show categorical perception of a /la/–/ra/ continuum, unlike speakers of English. Figure 10 shows this contrast in the results of a behavioral pretest for a

[9] Our /la/–/ra/ continuum consisted of 13 syllables synthesized using a Klatt synthesizer. To increase the naturalness of the stimuli, we varied a total of 3 acoustic parameters: (i) F2/F3 onset frequency, (ii) duration of F1 transition, (iii) duration of F2 bandwidth transition. Full details of stimulus parameters are available on request from the authors.

typical Japanese subject. American subjects show identification functions similar to figure 1 above.

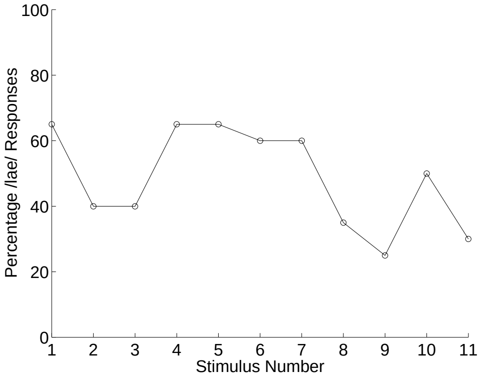

Figure 10
/r/–/l/ identification, subject MK (Japanese)

This cross-linguistic difference in the salience of the contrast between /l/ and /r/ makes it possible to test our conclusions from experiment 1 in another way. If the mismatch response elicited in experiment 1 for the /dæ/–/tæ/ contrast is due to phonetic representations, then we should expect to see a similar response to an identically presented /la/–/ra/ contrast for English speakers (experiment 3), but not for Japanese speakers (experiment 4).

Preliminary results from both experiment 3 (English speakers) and experiment 4 (Japanese speakers) confirm our predictions: a mismatch response is elicited in English subjects, for whom the distinction between /l/ and /r/ is phonetically salient, and not in Japanese subjects, for whom the contrast is not phonetically salient. More complete results of these studies will be presented elsewhere.[10]

3.5 Discussion

Our series of experiments using the oddball paradigm lead to the conclusion that the generator of the mismatch response, which is located in auditory cortex, is able to access phonetic representations. This finding is interesting simply because it implicates a specific part of the brain in a specific kind of cognitive information processing. Such specific conclusions are generally hard to draw. The result is also interesting because of the connection it draws between phonetic category representations and auditory cortex, which is generally

[10] Experiment 1 (/dæ/–/tæ/ contrast) is also being conducted with the Japanese subjects, in order to establish that these subjects do show a phonetic mismatch response to constrasts which are phonetically salient in Japanese.

assumed to be responsible only for “early” stages in the processing of auditory stimuli.

However, there are some fairly clear limits on the interpretation of the results of our oddball studies. First, since so little is known about auditory processing in specific areas of the human brain, apart from in the auditory nerve and other very early subcortical relays, it is hard to evaluate the significance or plausibility of the claim that auditory cortex accesses phonetic categories. Second, we are being necessarily vague when we say that auditory cortex “accesses” phonetic representations: our results do not allow us to conclude that auditory cortex actually contains representations of phonetic categories, only that the generator of the mismatch response can access these representations, wherever they might be. It is entirely possible that phonetic categories are not represented in auditory cortex, but that descending auditory pathways are able to relay information about these representations to the MMF generator. Third, even if auditory cortex is implicated in the processing of phonetic categories, the experiments in this section tell us nothing about *how* auditory cortex might be implicated in processing these categories. Although we have little enlightenment to offer on the first and second of these worries at this point, the experiments discussed in section 4 speak directly to the issue of how acoustic and phonetic information is processed in auditory cortex.

4. Continuous and categorical in stop perception

Our second set of studies again focuses on processing a VOT continuum, but in a different way from the studies described in section 3. In section 3 we asked whether the mismatch response was sensitive to the categories which are the endpoint of the mapping from acoustics to phonetics. The studies described in this section investigate whether an earlier evoked response reflects an intermediate stage in the acoustics-phonetics mapping. We argue that certain properties of the M100 response reflect just such an intermediate stage in the mapping of an acoustic VOT continuum onto [±voice] phonetic categories. We also relate our results to results from other MEG studies and from single-cell recordings from animals. Together these results begin to suggest an account of *why* it is that voice onset time is used distinctively in human languages.

4.1 Categorical perception as information loss

4.1.1 Where does information loss occur?

As noted above, classic findings about the perception of VOT continua show that identification of phonetic categories follows a step function, with reliable identification up to the perceptual VOT boundary, where there is a sudden shift from one category to the other (see figure 1 above). Pairs of stimuli are much easier to discriminate when they fall on opposite sides of the perceptual boundary, and stimuli falling on the same side of the perceptual boundary are usually perceived as identical. These findings have been interpreted as evidence for *loss of information* about the relative timing of the initial noise burst and voicing onset, which would otherwise serve to distinguish among stimuli within a given phonetic category. The only information retained is an indication that the VOT is either above or below the boundary value.

Given the difficulty that subjects generally experience in distinguishing stimuli which differ in VOT but are drawn from the same phonetic category, the information loss account of categorical perception has to be correct at some level or other. However, it is far from clear at what level of auditory processing we should say that information loss takes place, and this question has been the topic of a lively controversy in the literature on speech perception. At one extreme it could be that detailed timing information about VOT is lost very early in auditory processing, as a consequence of an inadequacy of general-purpose auditory processing mechanisms. At the other extreme, it could be the case that extremely detailed timing information about VOT is preserved right through until "higher level" auditory processing, but that within-category timing differences are inaccessible to conscious inspection because of their phonetic irrelevance. These extremes, together with a number of intermediate positions, have been extensively discussed in the literature, generally under the heading of the special systems versus general mechanisms debate.

In what follows, we discuss a number of previous studies which point to the conclusion that at least some of the timing information about VOT—specifically, VOTs in the [+voice] range—may be lost at an early stage in auditory processing, due to general limitations of the auditory system. We then go on to show results from some of our MEG studies which support the same conclusion.

4.1.2 Animal studies

Some animals have demonstrated categorical perception of a /dæ/–/tæ/ continuum in behavioral tasks (Kuhl & Miller 1975, 1978; Kuhl & Padden 1982). These findings have given rise to a series of studies of the neural basis of this ability using single-unit recordings from chinchillas (Sinex & McDonald 1988, 1989) and macaques (Steinschneider et al. 1994, 1995). These studies suggest that there is asymmetric loss of timing information about VOT early in the auditory pathway, at or before primary auditory cortex. Specifically, timing distinctions in the [+voice] range are lost, such that all low VOTs are treated alike, but in the [-voice] range timing distinctions are preserved, and differences in VOT are represented. We refer to this situation as *partial information loss*.

Recording from individual fibers in the auditory nerve of anaesthetized chinchillas, Sinex and McDonald (1988, 1989) measured responses to two series of synthetic syllables, heard by humans as /ga/–/ka/ or /da/–/ta/. In each series, VOTs ranged from 0–80 ms in steps of 10 ms. Sinex and McDonald found that in neurons tuned to frequencies near the first formant of the stimuli there was an increase in average discharge rate and synchronized discharge rate[11] either at 20 ms or at voicing onset—whichever of the two came *later*. In other words, for syllables with VOTs of 20 ms or less, the discharge-rate increases always occurred at 20 ms, regardless of whether VOT occurred at 0, 10 or 20 ms. For syllables with a VOT of 30 ms or longer, the discharge-rate increases occurred at the same latency as the VOT. Figure 11 illustrates the relationship between the timing of the voicing onset and the increase in discharge rate.

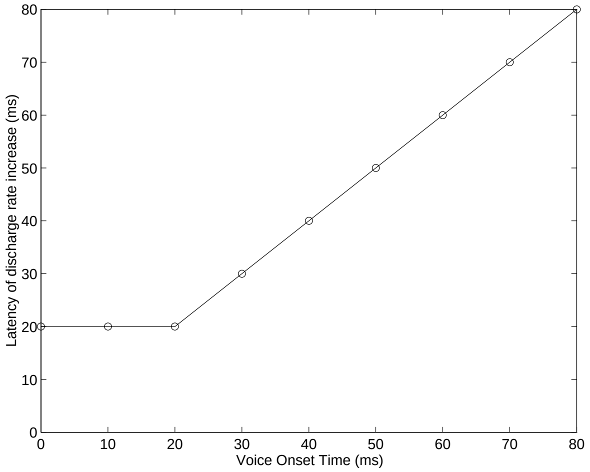

Figure 11
Latency of discharge rate changes as function of VOT
(adapted from Sinex et al. 1989)

As is clear from figure 11, the chinchilla auditory nerve fibers Sinex and McDonald were recording from showed partial information loss about VOT.

Along similar lines, Steinschneider et al. (1994, 1995) recorded responses in layer IV of primary auditory cortex in awake macaque monkeys to stimuli varying along the same VOT continuum as the one we used in our experiments. For all stimuli the multiple unit activity (MUA)[12] at this site showed a sudden transitory increase in response to the noise burst at the stimulus onset. This response was identified as a standard “on” response, i.e., a response triggered specifically by the onset of a given kind of stimulation. Interestingly,

[11] Average discharge rate (ADR) is a measure of the firing frequency of a neuron, measured in spikes per second. Typical ADR values in Sinex and McDonald’s studies were around 1000 spikes/s. Synchronized discharge rate (SDR) is the result of a discrete Fourier transform of series of neural spikes, and is measured in spikes per second *at a given frequency*. SDR is a measure which makes it possible to see important changes in the timing patterns of a neuron’s firings, even when the neuron’s average firing rate is not changing.

[12] Multiple Unit Activity (MUA) is a measure of the net action potential activity generated by a group of neurons surrounding a single recording contact.

when VOT exceeded 30 ms an identical additional MUA spike was observed at the same recording site at a latency proportional to VOT. In other words, this is a second “on” response, elicited by the onset of voicing. Two properties of this second “on” response are of particular importance for our pursposes. First, the latency of the second MUA spike varied in tandem with variations in VOT, indicating preservation of timing information about VOT. Second, there was no second MUA spike at all for VOTs below 30 ms. These findings resemble those of Sinex and McDonald (1989) in that they show partial loss of information about VOT.

The findings of Sinex and McDonald with chinchillas and Steinschneider et al. with monkeys are important in a number of respects. Since the “boundary” in these neural recordings occurred at 20–30 ms VOT, it is hard to avoid trying to use these findings to accont for the earlier demonstrations of categorical perception in these animals in behavioral tasks. As a first approximation, it seems that the timing of the boundary between responses to low VOT and high VOT stimuli is related to the refractory period of “on” detectors at different points in the auditory pathway,[13] as Steinschneider et al. (1994, 1995) suggest.

If animal categorical perception of VOT continua is related to so specific a mechanism as the refractory period of certain cells, a natural question to ask is, could the same mechanism account for some aspects of categorical perception of the same stimuli in humans? Most of the remainder of this paper addresses this question.

4.1.3 Representation of time in human MEG studies

A couple of studies of human auditory perception using MEG have shown results similar to the animal studies in section 4.1.2 in certain respects. One study demonstrates the existence of multiple “on” responses to speech stimuli. Another study shows a pattern of partial loss of information about timing differences, but in response to non-speech stimuli.

It is well known that the onset of an acoustic stimulus elicits a strong response in auditory cortex with a latency of roughly 100 ms. This is the so-called M100 response. It is one of the strongest components of the auditory evoked response, and it could reasonably be described as an “on” response. A number of MEG studies have demonstrated that the M100 wave is not only elicited by the transition from silence to sound; it can also be elicited by certain kinds of major change in an acoustic stimulus. This means a single stimulus could in principle elicit more than one M100 response, if it contained appropriate major acoustic changes.

[13] Many cells or groups of cells have the property that after being triggered to produce their characteristic response, there follows a period in which they are no longer sensitive to their normal trigger. This period is known as the refractory period.

Distinct M100 waves have been elicited by onset of aspiration and onset of voicing in syllables like /ha/ and /sa/ in Finnish and Japanese (Hayashi 1993; Kaukoranta et al. 1987; Mäkelä et al. 1988). In these stimuli the onset of voicing comes 100 ms or more after the onset of aspiration, and the second M100 elicited in the MEG studies is correspondingly found 100 ms after the first M100. The 100 ms lag before voicing onset is longer than the VOT in a typical [-voice] stop, and these stimuli lack the initial noise burst which accompanies the onset of aspiration in stop consonants, so we cannot be sure to find distinct “on” responses to the noise burst and to the onset of voicing in voiceless stops. However, these studies do increase the plausibility that the beginning of stop consonants might elicit double “on” responses.

A study by Joliot et al. (1994) looks at another kind of “on” response in MEG recordings, and finds evidence for asymmetric loss of information about the relative timing of two events. Joliot et al. presented subjects with pairs of short clicks separated by intervals ranging from 3 to 30 ms. They argued that the first click always “resets” the cycle of 40Hz spontaneous oscillatory brain activity.[14] If the second click occurs with a latency of less than 12–15 ms, the 40Hz activity is not reset, but if the second click occurs 15 ms or longer after the first click, it is treated like a new event and the 40Hz activity is reset.

Although Joliot et al. used non-speech stimuli and the critical latency for the differentiation of clicks by the brain is somewhat shorter at 12–15 ms than the critical latency of on the VOT continuum, their study is extremely relevant to our interests here. It is the closest human correlate to the results of the animal studies in section 4.1.2 that we are aware of, because it shows the same pattern of asymmetric information loss.

4.2 Predictions

The existing studies have established two main points. First, early responses in the auditory pathway of chinchillas and macaques demonstrate asymmetric loss of VOT information for /dæ/ and /tæ/. Similar asymmetric information loss was found in humans by Joliot et al. (1994) for short and long separations between clicks. Secondly, separate M100 responses to the noise burst and the onset of aspiration occur for very long VOT (/ha/–/sa/) in humans. Our experiments tested the relevance of these findings to categorical perception in humans.

The voicing onset transition, from the broad-band noise of aspiration to the periodic waveforms of the vowel, has elicited a second M100 response in previous MEG studies. This same transition marks the onset of voicing in syllables along a /dæ/–/tæ/ continuum. However, in our VOT continuum, as in normal speech, the lag between aspiration and voicing onset is generally much shorter than 100 ms. We ask whether these two acoustic events elicit separate

[14] A number of studies have shown that this 40Hz oscillatory activity is reset by sensory stimulation—see references cited in Joliot et al. 1994.

M100 responses in normal syllables, as in the longer ones of previous studies. We also investigate whether within-category distinctions are preserved for long and short VOT stimuli alike in humans. According to the animal findings, within-category distinctions are lost in the [+voice] range.

From the conclusions of previous studies, we predict that properties of the M100 response will show partial loss of information about VOT. Neural responses to syllables with low VOT will be identical, while responses to syllables with high VOT will vary with VOT. For example, if a given response property decreases as VOT increases, we expect to see the first pattern in figure 12 below, rather than the second or third.

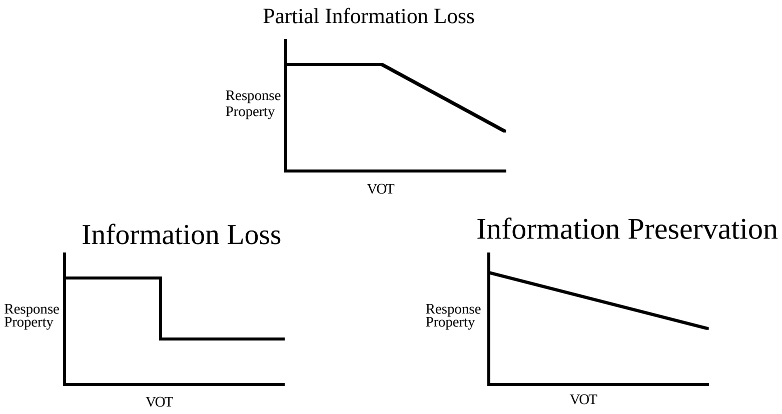

Figure 12
Possible patterns of preservation or loss of VOT information

4.3 Experimental procedure

We pretested 6 normal right-handed subjects with syllables chosen from the synthesized VOT continuum, to establish the individual perceptual boundary for each subject. We selected nine values along the VOT continuum to be presented during MEG recording. For each subject, four of the stimuli were clearly voiced syllables (/dæ/), four were clearly unvoiced syllables (/tæ/), and one fell at the perceptual boundary. The boundary value selected was a syllable categorized as belonging to each of the two phonetic categories half of the time. For example, for a subject with a perceptual boundary of 28 ms we used the following VOT values: 0 ms, 10 ms, 20 ms, 25 ms, 28 ms, 31 ms, 40 ms, 60 ms, 80 ms. During MEG data acquisition, the 9 stimuli were presented 100 times each in pseudo-random order. Subjects performed a binary forced-choice judgment of category membership by pressing one of two buttons. Their responses and reaction times were recorded.

MEG data acquisition was identical to the procedure described for experiment 1 in section 3, except that the sampling rate was 2048 Hz, with a passband of 800 Hz and a high-pass cut-off of 1.0 Hz. The responses to the pseudo-randomly presented stimuli were selectively averaged according to VOT and low-pass-filtered at 0–80Hz for further analysis.

4.4 Results

For all subjects, each of the nine stimuli elicited a strong M100 response to the initial noise burst. This response localized reliably to auditory association cortex. However, we focus here on differences in the timing and strength of the responses to stimuli of differing VOTs, rather than to localization differences among stimuli. A series of properties of the M100 response component show the partial loss of information about VOT that was observed in the animal studies.

4.4.1 Wave separation

All stimuli yielded an M100 response at a latency of about 120 ms. In the /tæ/ range, a second M100 emerged and became increasingly distant from the first with increasing VOT. Figure 13 illustrates this with averaged, filtered waveforms measured at one sensor for different VOTs for one representative subject. In the /tæ/ range the second response affected the strength and latency of the single observed peak until it appeared as a separate peak. Notice that the peak amplitude of the M100 is essentially identical for all of the stimuli in the /dæ/ range, but that the peak amplitude decreases as VOT increases in the /tæ/ range.

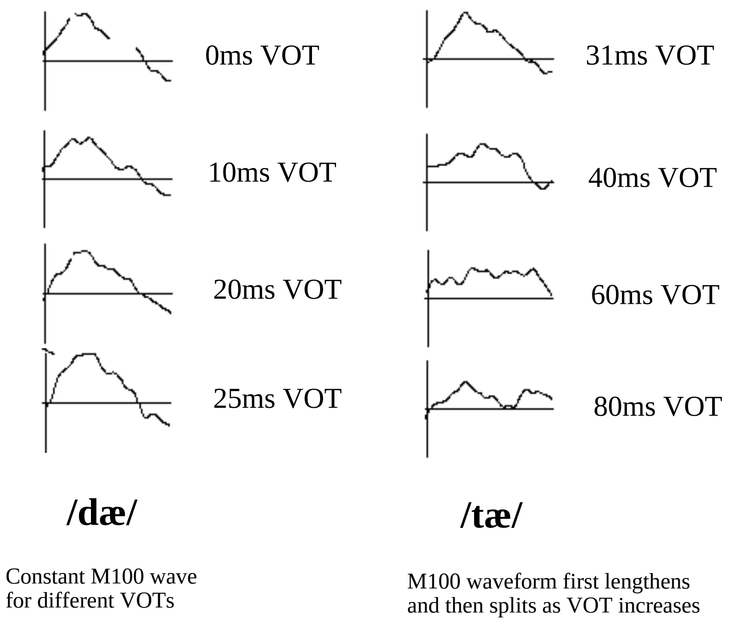

Figure 13
Distinct M100 waves for /tæ/, but not for /dæ/

This difference in responses to stimuli with different VOTs parallels the results of the animal and human studies discussed above, both because it shows separate responses to the initial noise burst and to the onset of voicing, and because there appears to be an asymmetric loss of timing information about VOT in these responses: stimuli perceived as /dæ/ show indistinguishable responses, whereas stimuli perceived as /tæ/ show easily distinguishable responses.[15]

Given the amplitude differences between the single M100 wave elicited by /dæ/s and the first of the two M100s elicited by high VOT /tæ/s, it is likely that the single M100 wave observed for the /dæ/ stimuli is actually an aggregate of two separate and very close M100 responses. The closeness of the two responses in time makes it impossible to distinguish between them in the magnetic brain recordings. However, even though the two responses are not separable in the /dæ/ range, if their relative timing is varying with VOT then they should combine in different ways at different VOTs. We do not see this. Therefore, the crucial property of the two responses for our purposes is that they are separated by a time interval which is proportional to VOT in the high VOT

[15] We can rule out the possibility that the difference between responses to low VOT and high VOT stimuli is an artifact of our post-recording low-pass filtering of waveforms. Since the low-pass cutoff we used in this experiment was 80Hz, it ought to be possible to see differences among responses to stimuli in the 10ms–25ms VOT range, if they indeed existed.

/tæ/ region of the continuum, but not in the low VOT /dæ/ region of the continuum.

In a separate study, not discussed here, subjects listened passively to a series of syllables which included one /tæ/ stimulus with a VOT of 100 ms. Responses to this stimulus showed distinct responses to stimulus onset and voicing onset more clearly than the 30 ms–80 ms /tæ/s used in the present experiment.

4.4.2 Response latency

The second finding also involves differences in response latencies due to changes in VOT, and also involves a contrast between stimuli in the [+voice] range and the [-voice] range of the VOT continuum.

For each of the 6 subjects in the study we identified the M100 response as the strongest derived magnetic field between 100 ms and 200 ms post stimulus onset in response to each of the 9 stimuli in the VOT continuum. The derived magnetic field strength (root mean squared: RMS) is a property of the single-dipole localization model computed from the measurements at each of the 37 sensor locations.

Figure 14 shows the latency of M100 as a function of VOT for all 6 subjects. In the /dæ/ range, there is close similarity in the latency of the M100 RMS peak both within and across subjects, with M100 peaks generally occurring between 115 ms and 135 ms post stimulus onset. On the other hand, the latencies of M100 responses in the /tæ/ range show enormous variability, both within and among subjects.

Subjects show very little variation in M100 latency for VOTs in the [+voice] range, whereas, for many subjects, M100 latencies increase in the [-voice] range. Notice, however, that the increase in [-voice] M100 latencies does not vary randomly. The largest latency increases are consistently found between 28 and 40 ms, just beyond the boundary VOT. At higher VOTs, M100 latencies increase more gradually and then reconverge on the short latency found in the [+voice] range.[16]

[16] The one subject whose M100 latency does not return to a low value at 80 ms VOT was tested in another experiment in which one of the stimuli was a 100ms VOT /tæ/. This stimulus clearly elicited two M100s, with the first having a latency similar to low VOT /dæ/s. This means that all subjects are alike in that M100 latencies return to the latency elicited by a /dæ/ when VOT is extremely long.

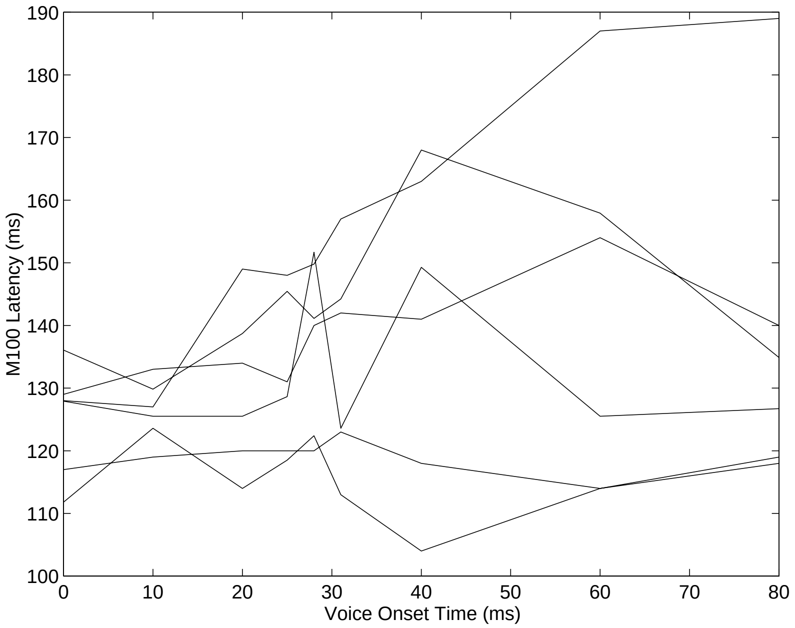

Figure 14
Latency of M100 Response, 6 Subjects

This pattern is consistent with the notion that /tæ/ elicits two M100 responses, and that the latency of the second of these M100s tracks VOT. When the two responses are close together in time, as shown in figure 13, both responses affect the latency of a single observed peak. At higher VOTs the separate M100 responses will begin to be discernible in the magnetic recordings. Exactly where this separation becomes visible will vary from subject to subject, depending on the the profile of the individual M100 waves for each subject. These factors can account both for the variation in latencies seen within the /tæ/ range, and for the return to short-latency M100s at extremely long VOTs.

4.4.3 Field strength

Our third finding involved changes in the strength of the M100 response as a function of voice onset time. For each subject and each of the 9 stimuli we selected the strongest RMS value between 100 ms and 150 ms post stimulus onset, and computed the mean RMS value in a time window around that peak. The window extended 25 ms before and after the RMS peak, or until the goodness-of-fit of the single-dipole model dropped below .95, whichever window was smaller.[17] In order to compensate for the enormous variation across

[17] It was extremely rare for goodness-of-fit measures to drop below .95 in the 50 ms period around the M100 RMS peak, so almost all of the means were based on RMS values across the entire 50ms period.

subjects in the absolute strength of the M100 RMS peaks, values were normalized prior to across-subject comparison, by dividing the mean RMS for each VOT by the mean RMS for all VOTs.

The strength (RMS) of the estimated current dipole decreased with increasing VOT, but in the /tæ/ range only. Mean RMS values were roughly constant for VOTs in the /dæ/ range, but decreased linearly for VOTs in the /tæ/ range. Figure 15a shows normalized mean RMS values around the M100 peak for 4 subjects.[18] Figure 15b shows the pooled values for these subjects.

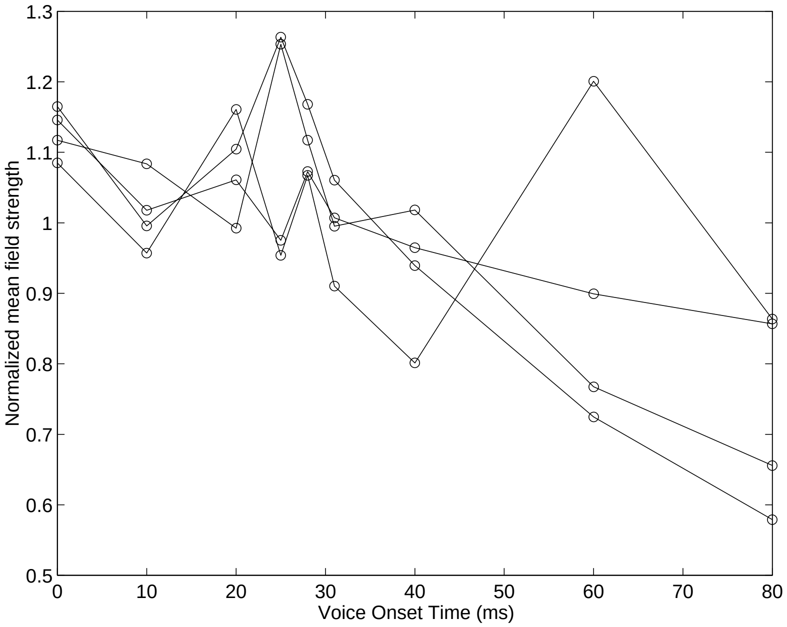

Figure 15a
Normalized mean field strength around M100

[18] The one subject whose M100 latencies continued to increase at VOTs of 60–80ms was excluded from this measure; another subject was not included because the relevant RMS data was not available at the time of writing.

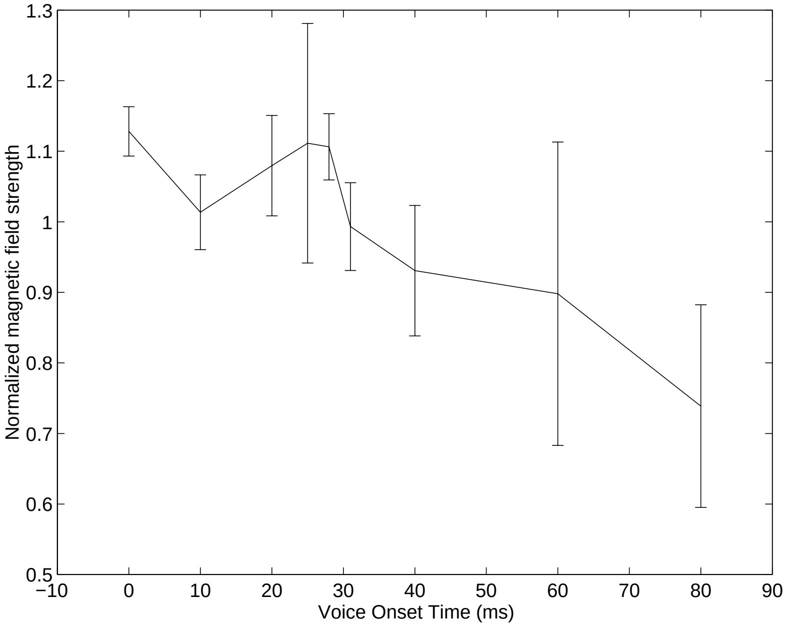

Figure 15b
Normalized mean field strength around M100, mean of 4 subjects

The regression line for the 5 stimuli in the [+voice] range does not have a slope which differs significantly from zero: *F*(1,19) = 0.036, *p* = 0.85. On the other hand, the regression line for the 5 stimuli in the [-voice] range has a negative slope which is highly significant: *F*(1,19) = 16.64, *p* < 0.001.

This finding is also explained by the asymmetric information loss model. The latency of the second response increases with VOT. As its latency increases, its contribution to the RMS of the observed M100 peak decreases.

4.4 Discussion

We have shown here that different properties of the M100 responses to the VOT continuum show the pattern characteristic of partial information loss, consistent with the notion that there is a response to the onset of voicing which varies with VOT in the [-voice] range, but not in the [+voice] range. Therefore, the automatic evoked M100 response appears to encode properties of an intermediate stage in the mapping of continuous acoustic information onto discrete phonetic categories. Since the M100 is a fairly early response and is known to be generated in auditory association cortex, we may infer that at least one important stage of the acoustics-to-phonetics mapping takes place at or before auditory cortex in humans.

As pointed out in section 4.1 above, the finding of partial information loss about VOT in animal direct recording studies had a natural account in terms of the refractory period of specific groups of neurons. Given that the results of the MEG experiment discussed in this section show a similar pattern of partial information loss, we must ask whether it is feasible to try to extend the account of the VOT perceptual boundary suggested for animals to humans. This is an attractive possibility, because it would go some way towards explaining a commonly exploited phonetic feature in terms of an independently existing physiological mechanism.

If we adopt the view that loss of information about VOT is related to the refractory period of relatively low-level auditory neurons, we are forced to reconsider the status of a body of existing evidence that has been used to argue for the view that loss of information about VOT is a rather high-level attentional process. A number of behavioral studies have shown evidence that speakers can distinguish among different exemplars of the same phonetic category, even if they are unable to reliably discriminate between them. This evidence typically consists of demonstrations that subjects are able to differentiate between "good" and "poor" exemplars of a given phonetic category, a process which requires the ability to access subphonetic contrasts (Samuel 1982; Kuhl 1991; see Miller 1994 for a recent review).[19]

Conflicting evidence for information preservation and information loss may be reconciled by assuming that both kinds of evidence are correct, even for the same categories, and that information about speech stimuli is processed in multiple parallel channels, some of which preserve subphonetic distinctions, while others do not. This possibility is not easy to verify, but it is supported by the evidence for within-category information preservation. For example, Samuel (1982) compared subjects' prototypes for /gæ/ with their identification boundary along a synthetic /gæ/–/kæ/ continuum, and found that there was no correlation between a subject's boundary and his prototype. This suggests that the evidence for information loss and information preservation draw on very different representations.

5. Conclusion

The main conclusion from these studies is that auditory cortex appears to be implicated in the perception of at least some phonetic contrasts. We have been encouraged by the specificity of the claims that we have been able to make about the relationship between brain activity and behavior. However, the reader will have noticed at a number of points that our findings are still consistent with a number of somewhat different accounts of the relatively straightforward mapping of VOT onto phonetic categories. Some of these questions can be

[19] A particularly interesting feature of Kuhl's (1991) study is that she found evidence for within-category discrimination (using the "perceptual magnet effect") in human adults and children, but not in monkeys. This finding raises the possibility that the loss of within-category information may be a very different process in humans and monkeys.

feasibly addressed given the current state of theory and technology; other questions are still hard to even formulate.

We would like to stress at this point, as we did at the beginning, the usefulness of being able to build on relevant findings in related fields in designing our experiments and hypotheses and interpreting our results. Psychoacoustics provides a wealth of such findings. In the absence of clearly statable hypotheses it is easy for brain-imaging studies to become no more than extremely expensive variants on grammaticality judgements or reaction-time data. It is for this reason that we believe that speech perception is such a valuable area for learning about how the brain represents language.

References

Aaltonen, O., P. Niemi, T. Nyrke & M. Tuhkanen (1987). Event-related brain potentials and the perception of a phonetic continuum. *Biological Psychology* 24, 197-207.

Aulanko, R., R. Hari, O.V. Lounasmaa, R. Näätänen & M. Sams (1993). Phonetic invariance in the human auditory cortex. *NeuroReport* 4, 1356-1358.

Guthrie, D. & J.S. Buchwald (1991). Significance Testing of Difference Potentials. *Psychophysiology* 28, 240-244.

Hari, R., M. Hämäläinen, R. Ilmoniemi, E. Kaukoranta, K. Reinikainen, J. Salminen, K. Alho, R. Näätänen & M. Sams (1984). Responses of the primary auditory cortex to pitch changes in a sequence of tone pips: Neuromagnetic recordings in man. *Neuroscience Letters* 50, 127-132.

Hayashi, M (1993). Auditory Neuromagnetic Fields Evoked by Spectral Transition of Syllables. *Journal of Robotics and Mechatronics* 5, 409-412.

Joliot, M., U. Ribary & R. Llinas (1994). Human oscillatory brain activity near 40 Hz coexists with cognitive temporal binding. *Proceedings of the National Academy of Science* USA 91, 11748-11751.

Kaukoranta, E., R. Hari & O.V. Lounasmaa (1987). Responses of the human auditory cortex to vowel onset after fricative consonants. *Experimental Brain Research* 69, 19-23.

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustic Society of America* 67, 971-995.

Kuhl, P.K. (1991). Human adults and infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50, 93-107.

Kuhl, P.K. & J.D. Miller (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science* 190, 69-72.

——— (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustic Society of America* 63, 905-917.

Kuhl, P.K. & D.M. Padden (1982). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Perceptual Psychophysiology* 32, 542-550.

Lewine, J.D. & W.W. Orrison (1995). Magnetoencephalography and Magnetic Source Imaging. In: *Functional Brain Imaging 1995,* Mosby-Year Book Inc., 369-417.

Maiste, A.C., A.S. Wiens, M.J. Hunt, M. Scherg & T.W. Picton (1995). Event-Related Potentials and the Categorical Perception of Speech Sounds. *Ear & Hearing* 16, 68-90.

Mäkelä, J.P., R. Hari, & L. Leinonen (1988). Magnetic responses of the human auditory cortex to noise/square wave transitions. *Electroencephalography and clinical neurophysiology* 69: 423-430.

Miller, J. (1994). On the internal structure of phonetic categories: A progress report. *Cognition* 50, 271-285.

Näätänen, R. (1992). *Attention and Brain Function*. Hillsdale, NJ: Erlbaum.

Näätänen, R., A.W.K. Gaillard & S. Mäntysalo (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica* 42, 313-329.

Sams, M., M. Hämäläinen, A. Antervo, E. Kaukoranta, K. Reinikainen & R. Hari (1985). Cerebral neuromagnetic responses evoked by short auditory stimuli. *Electroencephalography and clinical neurophysiology* 61, 254-266.

Sams, M., R. Aulanko, O. Aaltonen & R. Näätänen (1990). Event-related potentials to infrequent changes in synthesized phonetic stimuli. *Journal of Cognitive Neuroscience* 2, 344-357.

Samuel, A. (1982). Phonetic Prototypes. *Perception and Psychophysics* 31, 307-314.

Sharma, A., N. Kraus, T. McGee, T. Carrell & T. Nicol (1993). Acoustic vs. phonetic representation of speech stimuli as reflected by the mismatch negativity event-related potential. *Electroencephalography and clinical neurophysiology* 88, 64-71.

Sinex, D.G. & L.P. McDonald (1988). Average discharge rate representation of voice onset time in the chinchilla auditory nerve. *Journal of the Acoustic Society of America* 83, 1817-1827.

——— (1989). Synchronized discharge rate representation of voice onset time in the chinchilla auditory nerve. *Journal of the Acoustic Society of America* 85, 1995-2004

Steinschneider, M., C.E. Schroeder, J.C. Arezzo, & H.G. Vaughan (1994). Speech-Evoked Activity in Primary Auditory Cortex: Effects of Voice Onset Time. *Electroencephalography and clinical neurophysiology* 92, 30-43.

——— (1995). Physiologic Correlates of the Voice Onset Time Boundary in Primary Auditory Cortex (A1) of the Awake Monkey: Temporal Response Patterns. *Brain and Language* 48, 326-340.

Phillips *cphill@mit.edu*
Marantz *marantz@mit.edu*
McGinnis *marthajo@mit.edu*
Pesetsky *pesetsk@mit.edu*
Wexler *wexler@psyche.mit.edu*
Yellin *elronx@mit.edu*

Dept of Linguistics & Philosophy
20D-219, MIT
Cambridge, MA 02139

Poeppel
david_poeppel@rad-mac1.ucsf.edu
Roberts *tplr@basil.ucsf.edu*
Rowley
howard_rowley@rad-mac1.ucsf.edu

MSI Laboratory
S-362, UCSF
513 Parnassus Avenue
San Francisco, CA 94143