LANGUAGE AND COGNITIVE PROCESSES
2010, 25 (6), 808–839

Psychology Press
Taylor & Francis Group

Auditory sensitivity to formant ratios: Toward an account of vowel normalisation

Philip J. Monahan[1] and William J. Idsardi[2]

[1]*Basque Center on Cognition, Brain and Language, Donostia-San Sebastián, Spain,* [2]*Department of Linguistics, Neuroscience and Cognitive Science Program, University of Maryland, College Park, MD, USA*

A long-standing question in speech perception research is how listeners extract linguistic content from a highly variable acoustic input. In the domain of vowel perception, formant ratios, or the calculation of relative bark differences between vowel formants, have been a sporadically proposed solution. We propose a novel formant ratio algorithm in which the first (F1) and second (F2) formants are compared against the third formant (F3). Results from two magnetoencephalographic experiments are presented that suggest auditory cortex is sensitive to formant ratios. Our findings also demonstrate that the perceptual system shows heightened sensitivity to formant ratios for tokens located in more crowded regions of the vowel space. Additionally, we present statistical evidence that this algorithm eliminates speaker-dependent variation based on age and gender from vowel productions. We conclude that these results present an impetus to reconsider formant ratios as a legitimate mechanistic component in the solution to the problem of speaker normalisation.

Keywords: Formant ratios; MEG; Auditory cortex; Vowel normalisation; M100.

INTRODUCTION

The perceptual and biological computations responsible for mapping time-varying acoustic input onto linguistic representations are still poorly understood (Fitch, Miller, & Tallal, 1997; Hickok & Poeppel, 2007; Phillips,

Correspondence should be addressed to Philip J. Monahan, Basque Center on Cognition, Brain and Language (BCBL), Paseo Mikeletegi No. 69, 2nd Floor, 20009 Donostia-San Sebastián, Spain. E-mail: p.monahan@bcbl.eu

We would like to thank Jeffrey Walker for invaluable lab assistance and David Poeppel for useful suggestions in the preparation of this manuscript. This work was funded by NIH 7R01DC005660-07 to David Poeppel and William J. Idsardi.

http://www.psypress.com/lcp DOI: 10.1080/01690965.2010.490047

2001; Scott & Johnsrude, 2003; Sussman, 2000). The speech signal includes not only the content of an utterance but also cues that allow listeners to infer sociolinguistic and physical characteristics about the speaker (Ladefoged & Broadbent, 1957). These cues, however, also serve to introduce significant variation, obscuring any straightforward one-to-one mapping between acoustic features and phonetic or phonological representations. Some of the most compelling demonstrations of this variability in the speech signal have been presented in acoustic analyses of vowel distributions across different talkers (Peterson & Barney, 1952; Potter & Steinberg, 1950). Given their tractable nature and well-understood spectral properties, vowels have played a central role in understanding the mechanisms that underlie speaker variation and normalisation (Rosner & Pickering, 1994). The primary acoustic characteristic of spoken vowels is their formants: resonant frequencies of particular vocal tract configurations superimposed on the harmonic resonances of the glottal pulse source during production (Fant, 1960). Within speakers, the first (F1) and second (F2) formants are the principal determinants of vowel type—F1 varies as a function of vowel height and F2 varies as a function of vowel backness (the third formant (F3) primarily cues rhoticity; Broad & Wakita, 1977). F1 and F2 can secondarily be affected by other articulatory gestures, such as tongue-root position and lip posture (Stevens, 1998).

While the relative pattern of formants remains approximately constant across speakers for a given vowel type (Potter & Steinberg, 1950), the absolute formant frequencies for a given vowel token vary as a function of vocal-tract length (Huber, Stathopoulos, Curione, Ash, & Johnson, 1999). Using magnetic resonance imaging of the vocal tract, Fitch and Giedd (1999) demonstrate that vocal-tract length positively correlates with age, and within adults, gender. Despite these differences, however, listeners are quite good at recognising phonemes across a number of different speakers (Strange, Jenkins, & Johnson, 1983), can reliably categorise vowel tokens synthesised with varied vocal-tract lengths, both within and outside the normal range (Smith, Patterson, Turner, Kawahara, & Irnio, 2005) and, furthermore, can estimate speaker size from modulations of vocal-tract length with a minimal amount of speech input (Ives, Smith, & Patterson, 2005; Smith & Patterson, 2005). It seems, then, that auditory cortex segregates the incoming speech signal into information that allows listeners to recover both the vocal-tract size (formant scales) and vocal-tract shape (formant ratios) contemporaneously with one another (Irino & Patterson, 2002; Smith et al., 2005). In addition to this ability to cope with significant variation due to the physical characteristics of a talker, listeners can reliably identify the sociolinguistic background of a speaker with as little information as "hello", as has been documented in cases of housing discrimination (Purnell, Idsardi, & Baugh, 1999). Moreover, pre-linguistic infants ignore speaker-dependent acoustic variation and successfully

categorise vowels across different talkers (Kuhl, 1979, 1983). Thus, whatever normalisation procedures are available to listeners are deployed without significant linguistic experience or exposure to novel talkers, and consequently, accumulating a large amount of speaker-dependent information is unnecessary to adequately normalise across speakers. In order for a vowel normalisation algorithm to be plausible, it must, at least, normalise the vowel space and be computable by the auditory system during online speech perception.

Recent functional neuroimaging and electrophysiological work has identified different cortical networks subserving the processing separation of speaker-dependent (“who” is speaking) from speaker-invariant (“what” is being said) features in vowel perception (Bonte, Valente, & Formisano, 2009; Formisano, de Martino, Bonte, & Goebel, 2008). Specifically, Formisano et al. (2008) showed that the cortical networks responsible for distinguishing vowel categories independent of speaker were more bilaterally distributed in superior temporal cortex and involved the anterior–lateral portion of Heschl’s gyrus, planum temporale (mostly left lateralised), and extended areas of superior temporal sulcus (STS)/superior temporal gyrus (STG) bilaterally. In contrast, the networks underlying speaker identification independent of vowel category were far more right lateralised and included the lateral part of Heschl’s gyrus and three regions along the anterior–posterior axis of the right STS that were adjacent to areas in vowel discrimination. The conclusion that the neurobiology segregates the processing of speaker and vowel is consistent with recent perceptual learning (McQueen, Cutler, & Norris, 2006; Norris, McQueen, & Cutler, 2003) and neurophysiological work (see Obleser & Eisner, 2009) arguing that listeners construct abstract prelexical representations of phonological categories that are independent of particular speakers and that episodic traces (e.g., Bybee, 2001; Goldinger, 1996; Johnson, 1997, 2005; Pierrehumbert, 2002; Pisoni, 1997) cannot be the only representational schema employed in speech perception.

The primary goal of this paper is to revisit an idea that has received sporadic attention within the literature in attempting to solve the speaker normalisation problem: formant ratios or the calculation of relative log differences between formants in vowel perception (Lloyd, 1890; Miller, 1989; Peterson, 1951, 1961; Potter & Steinberg, 1950; Syrdal & Gopal, 1986; see Johnson, 2005 for criticisms). We pursue here a specific instantiation of formant ratios: namely, information in higher formants, specifically the third formant (F3), acts as the normalising factor (Deng & O’Shaughnessy, 2003; Peterson, 1951). A consequence of this proposal, then, is that the appropriate dimensions for the vowel space are the ratios F1/F3 and F2/F3 (or logarithmic-like transforms of these quantities, such as Mel (Stevens & Volkmann, 1940) or Bark (Zwicker, 1961) difference scores) as opposed to

the traditional F2 by F1 vowel space (Figure 1). In this article, we calculate the extent to which this particular hypothesis (F1/F3 by F2/F3) removes inter-speaker variation based on age and gender of the talkers from the Hillenbrand, Getty, Clark, and Wheeler (1995) corpus of American English vowels. Subsequently, we present data from two magnetoencephalography (MEG) experiments that suggest that auditory cortex is sensitive to modulations of the F1/F3 ratio. Additionally, our findings indicate that the perceptual system displays heightened sensitivity to formant ratios in more densely populated regions of the vowel space (in English, this would be for front and back vowels and not central vowels).

The third formant in perception and normalisation

The centre frequency of F3 appears to vary correlationally with a given speaker's fundamental frequency and remains fairly constant across vowels for that speaker (Deng & O'Shaughnessy, 2003; Potter & Steinberg, 1950). Given that F3 appears to be relatively stable across vowel tokens within a given speaker, but varies as a function of vocal-tract length inter-talker, and is present in different types of speech (e.g., whispered, nonphonated speech), a possible solution to vowel normalisation would be to take the ratio of the

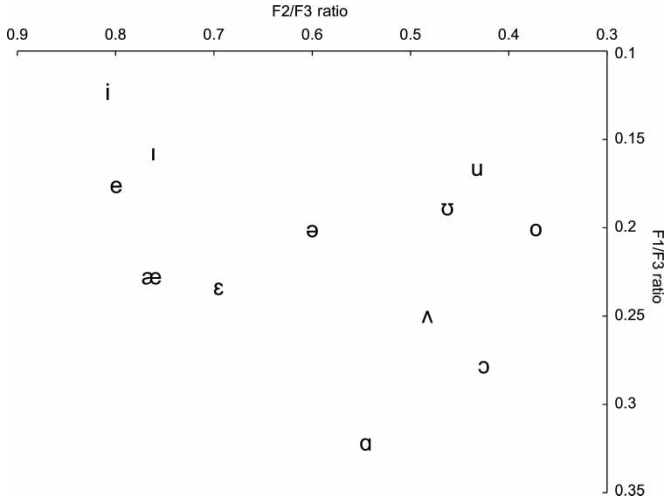

Figure 1. Vowel space normalised against F3. Note: Traditional vowel space plotted in the proposed normalised vowel space (F3 as the normalising factor). Formant values from which ratios were computed are from Hillenbrand et al. (1995) and averaged across age and gender per vowel category (except for /ə/ which was not collected in Hillenbrand et al. (1995); instead, those values were computed from the frequency values used in the experiments reported here (before F3 modulation)).

first and second formants against the third. Consequently, the vowel space would not be best represented as F2 plotted against F1 or the difference between F2 and F1 plotted against F1 (Ladefoged & Maddieson, 1996, p. 288), but rather as the ratio of F1 to F3 plotted against the ratio of F2 to F3. The third formant (F3) is useful in the identification and discrimination of a variety of speech contrasts (e.g., rhoticisation on vowels (Broad & Wakita, 1977), /l/–/r/ discrimination (Miyawaki et al., 1975), stop consonant place of articulation identification (Fox, Jacewicz, & Feth, 2008)), has been shown to affect the perception of vowels (Fujisaki & Kawashima, 1968; Nearey, 1989; Slawson, 1968), provides a good estimate of vocal-tract length in automatic speech recognition (Claes, Dologlou, ten Bosch, & van Compernolle, 1998), is useful in normalising whispered vowels (Halberstam & Raphael, 2004; although their data on the role of F3 in normalising phonated vowels were inconclusive), and it has also been shown that the higher formants are as important, if not more so, than pitch in normalising noise-excited vowels (Fujisaki & Kawashima, 1968). Given these results, we can be confident that listeners are able to use and exploit information contained within this frequency range.

Peterson (1951) converted vowel frequencies into Mel space and plotted F1/F3 against F2/F3 in the final two pages of his article. Taking vowel productions from one man, one woman, and one child, he showed that, impressionistically, these ratios remove much of the variation seen when F2 is plotted against F1. Unfortunately, little discussion or further results are provided, and it seems that this particular algorithm has not been pursued subsequently in the formant ratio literature. A similar solution was echoed in Deng and O'Shaughnessy (2003), where they write: "Since F3 and higher formants tend to be relatively constant for a given speaker, F3 and perhaps F4 provide a simple reference, which has been used in automatic recognizers, although there is little clear evidence from human perception experiments". One of half of the algorithm we propose (F2/F3) is present in previous formant ratio algorithms (Miller, 1989; Syrdal & Gopal, 1986). Therefore, in the experiments reported here, we concentrate on finding neurophysiological evidence for the more novel ratio, the F1/F3 ratio.

While the objective of this paper is to demonstrate human perceptual sensitivity to formant ratios, it is useful to assess how well our proposed algorithm eliminates variance due to speaker differences. The corpus data used to test our model are from Hillenbrand et al. (1995). In a replication of Peterson and Barney (1952), Hillenbrand et al. (1995) collected the productions of 12 American English vowels from 45 men, 48 women, and 46 children in an /hVd/ frame. In an acoustic analysis of the data, they identified a point centrally located in the steady-state portion of the vowel and measured the fundamental frequency (f0), as well as the centre frequencies for the first through fourth formants (F1–F4) for each token.

Prior to the analysis reported here, vowel tokens missing a value for one of their formants (F1–F4) in the corpus were eliminated from subsequent calculations (16.3% of the data).

To assess the amount of inter-speaker variation in the data as a function of speaker age and gender, we performed linear mixed effects modelling (Baayen, 2008) using the nlme package in R (R Development Core Team, 2006) with Subject as a random variable on the Hillenbrand et al. (1995) data comparing the effects of age (“adult”, “child”) and gender (“male”, “female”) on the raw frequency values for f0–F3 and subsequently on the transformed F1/F3 and F2/F3 ratios. Given previous results on differences in the fundamental frequency and formant frequencies of vowels across men, women, and children, we predict to find reliable main effects of gender and age for the untransformed data for all four measures (f0–F3).

For f0, we found a significant Age × Gender interaction [$F(1, 135) = 133.4$, $p < .0001$] and significant main effects for both age [$F(1, 135) = 255.0$, $p < .0001$] and gender [$F(1, 135) = 302.4$, $p < .0001$]. For F1, we found a significant Age × Gender interaction [$F(1, 135) = 19.8$, $p < .0001$], as well as significant main effects of both age [$F(1, 135) = 70.8$, $p < .001$] and gender [$F(1, 135) = 66.6$, $p < .0001$]. For F2, we again found a marginal Age × Gender interaction [$F(1, 135) = 3.4$, $p = .07$], and significant main effects of age [$F(1, 135) = 64.6$, $p < .0001$] and gender [$F(1, 135) = 45.3$, $p < .0001$]. Finally, for F3, we also find a significant Age × Gender interaction [$F(1, 135) = 24.0$, $p < .0001$] and significant main effects of both age [$F(1, 135) = 306.5$, $p < .0001$] and gender [$F(1, 135) = 211.1$, $p < .0001$]. These results confirm that the raw, untransformed values of formants across men, women, and children are highly variable, and effects of age and gender contribute greatly and interactively to this variability. Thus, a simple mapping between frequency information and speaker-independent representations is inadequate.

To provide an initial demonstration as to how well our proposed algorithm eliminates speaker variation, we ran the same model above on the transformed (F1/F3; F2/F3) corpus data (Figure 2). The significant main effects of age and gender, as well as the significant interactions of age and gender were completely eliminated. For F1/F3, there were no main effects of age [$F(1, 135) = 0.08$, $p = .78$] or gender [$F(1, 135) = 0.3$, $p = .57$] and no Age × Gender interaction [$F(1, 135) = 2.3$, $p = .12$]. For F2/F3, we also find no main effects of age [$F(1, 135) = 1.0$, $p = .32$] or gender [$F(1, 135) = 0.77$, $p = .38$] and no Age × Gender interaction [$F(1, 135) = 0.001$, $p = .97$].

The results of the linear mixed effects models on the transformed Hillenbrand et al. (1995) data demonstrate that our proposed algorithm, whereby F1 and F2 are ratioed against F3 successfully eliminates the variance due to effects of age and gender found in productions of vowel tokens across different speakers. The question we now focus on, and the primary aim of this paper, is to demonstrate that auditory cortex is sensitive

to one of the two dimensions of our proposed formant ratio algorithm, namely F1/F3. To do this, we present data from two MEG experiments on vowel perception. Our findings confirm that auditory cortex appears to be sensitive to modulations of the F1/F3 ratio (the more novel of the two computations in the proposed algorithm).

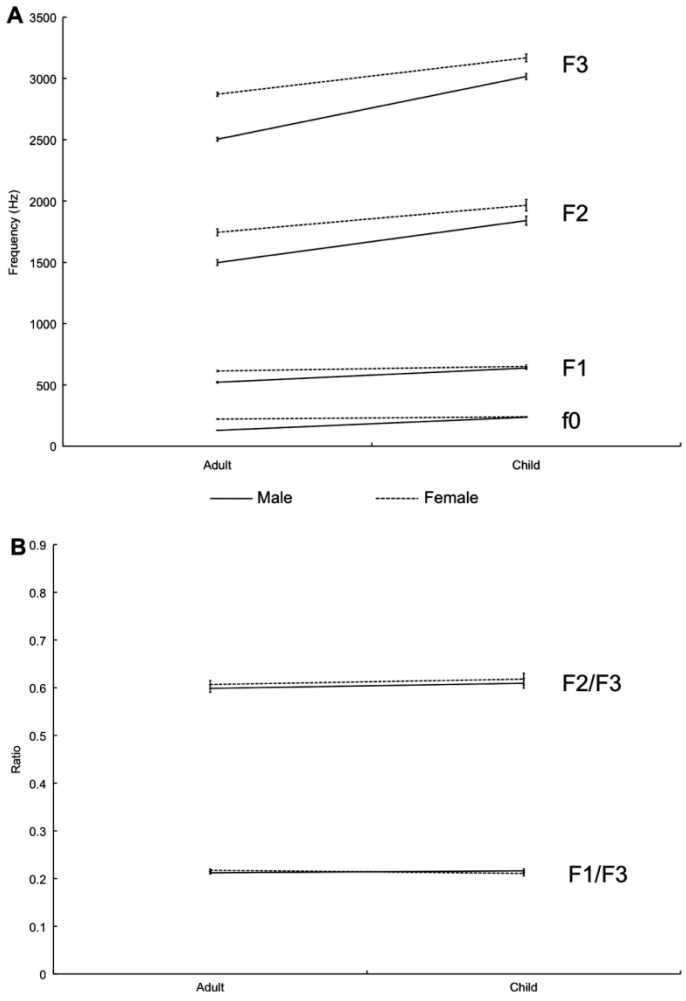

Figure 2. Comparison of mean formant values. Note: (A) Comparison of the untransformed formant values for fundamental frequency (f0), first formant (F1), second formant (F2), and third formant (F3) by age and gender of speaker. Mean formant values by group calculated from Hillenbrand et al. (1995). (B) Comparison of the transformed values for F1/F3 and F2/F3 by age and gender of speaker. Error bars represent one standard error of the mean.

The contribution of magnetoencephalography (MEG)

MEG is an electrophysiological recording technique that measures fluctuations in magnetic field strength caused by the electrical currents in neuronal signalling (Frye, Rezaie, & Papanicolaou, 2009; Hari, Levänen, & Raij, 2000; Lounasmaa, Hämäläinen, Hari, & Salmelin, 1996) and is particularly adept at recording potentials from auditory cortex (Roberts, Ferrari, Stufflebeam, & Poeppel, 2000). Combining its excellent temporal resolution (1 ms; and fair spatial resolution $\sim$ 2–5 cm) and aptitude for recording from auditory cortex, it provides a powerful tool in understanding how humans process speech in real time, whose temporal properties are both fast and fleeting. In the two experiments reported below, we exploit the response latency of an early, evoked neuromagnetic potential, the M100 (or N1m), which is the MEG equivalent of the N1 ERP component (Eulitz, Diesch, Pantev, Hampson, & Elbert, 1995; Virtanen, Ahveninen, Ilmoniemi, Näätänen, & Pekkonen, 1998). The electrical N1 in electroencephalography (EEG) is a negative-going potential comprising several subcomponents, with a primary subcomponent localising to primary auditory cortex (A1; Picton, Woods, Baribeau-Braun, & Healey, 1976). It is an exogenous response evoked by any auditory stimulus with a clear onset, and is found regardless of the task performed by participant, or his/her attentional state (Näätänen & Picton, 1987). Its MEG counterpart, the N1m or M100 (Figure 3), appears to be the magnetic equivalent of the primary subcomponent that localises to A1 in supratemporal auditory cortex (Eulitz et al., 1995; Hari, Aittoniemi, Järvinen, Katila, & Varpula, 1980; Virtanen et al., 1998), thereby making it a more focused dependent measure for use in understanding auditory processing (Roberts et al., 2000). The dependent measures of the evoked M100 (latency, amplitude) typically reflect spectral properties of the acoustic stimulus (frequency, loudness, fine structure of the waveform, etc.), as opposed to later evoked components (e.g., magnetic mismatch negativity [MMN]), and integrate only over the first 40 ms of the auditory stimulus (Gage, Roberts, & Hickok, 2006). Given its robustness and replicability, the M100 has been used extensively to study early auditory cortical processing, and we have a fair understanding of the types of stimulus dependent factors to which the M100 is sensitive (Roberts et al., 2000). Sinusoids closest to 1 KHz elicit the shortest evoked M100 response latency, while moving outward from 1 KHz in either direction (both lower and higher in frequency) elicit longer evoked latencies (Roberts & Poeppel, 1996). Relevant to the current work, the M100 response properties to vowels have been fairly well characterised. In particular, the M100 seems to be sensitive to F1 (Diesch, Eulitz, Hampson, & Ross, 1996; Govindarajan, Phillips, Poeppel, Roberts, & Marantz, 1998; Poeppel et al., 1997; Roberts et al., 2000; Roberts, Flagg, & Gage, 2004; Tiitinen, Mäkelä, Mäkinen, May, & Alku, 2005) independent of differences in fundamental frequency

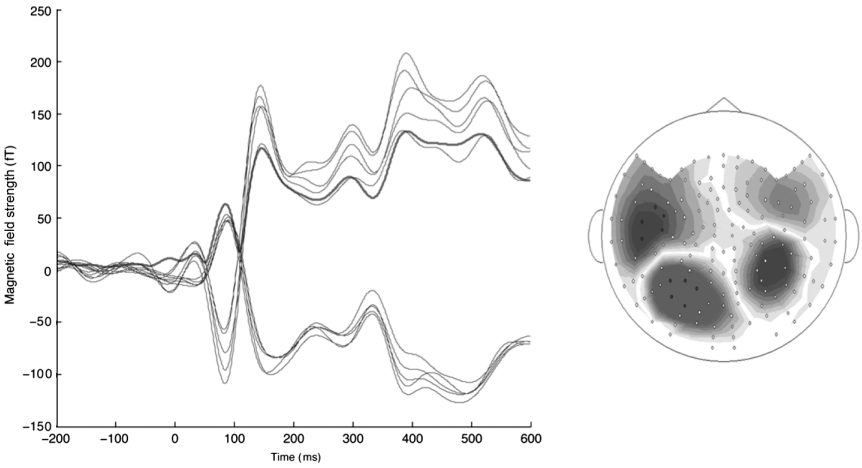

Figure 3. Evoked M100 temporal waveform and magnetic field contour. Note: Temporal waveform from 10 left hemisphere channels (RMS: thick black line superimposed) and the magnetic field distribution at peak latency of the M100 for a representative subject. For the magnetic field distributions, the left anterior and right posterior channels are measuring the ingoing magnetic field sinks and the left posterior and right anterior channels are measuring the outgoing magnetic field sources.

(Govindarajan et al., 1998; Poeppel et al., 1997). Diesch et al. (1996) compared the evoked latencies of the M100 to four different synthesised German vowels (/a/, /i/, /u/, and /æ/) and found that /a/ and /æ/, having higher F1 values, elicited reliably shorter latencies than /u/. Poeppel et al. (1997) synthesised three English vowels (/i/, /u/, and /a/) and also report a reliable difference in the evoked M100 latency between /a/ and /u/, with /a/ eliciting a shorter M100 latency and do not report a difference between /i/ and /u/. This finding was replicated in Govindarajan et al. (1998), who also found that both one and three formant synthesised tokens of /a/ elicit reliably faster M100 evoked latencies in English listeners than one and three formant synthesised tokens of /u/, respectively. Moreover, Tiitinen et al. (2005) replicated these findings in Finnish speakers, showing again that /a/ elicits faster M100 latencies than /u/ using semi-synthetic speech. The interpretation for the directionality of these effects, namely that /a/ elicits reliably shorter M100 evoked latencies than /u/, is that the spectral energy in F1 is driving the M100 response (Govindarajan et al., 1998; Poeppel et al., 1997; Roberts et al., 2000), and /a/ elicits shorter latencies because the F1 in /a/ ($\sim$ 700 Hz) is considerably closer to 1 KHz than the F1 in /u/ ($\sim$ 300 Hz), consistent with the sinusoid data (Roberts & Poeppel, 1996). This effect does not seem to be speech specific, however (Diesch et al., 1996; Govindarajan et al., 1998).

These findings have been confirmed and extended in more recent work. Roberts et al. (2004) showed that unlike responses to sinusoids, where the M100 response latency follows a smooth 1/frequency function (at least up to approximately 1000 Hz, above that point the latency again increases), the latency of the M100 to vowels (F1, in particular) seems to respect vowel category boundaries. They synthesised tokens of /a/ and /u/ and modulated F1 in 50 Hz increments between 250 and 750 Hz while keeping the values for F2 (1000 Hz) and F3 (2500 Hz) constant, albeit with broader than normal formant bandwidths. Instead of following the smooth 1/frequency function, M100 latencies clustered into three distinct bins, the lowest F1 values (250–350 Hz) elicited the longest latencies, the middle F1 values (400–600 Hz) elicited reliably shorter latencies and the high F1 values (650–750 Hz) elicited even shorter M100 latencies. The bin with the lowest F1 values also represents the natural range of F1 in /u/ and the bin with the highest F1 values represents the natural range of F1 in /a/ tokens. Roberts et al. (2004), therefore, concluded that the latency of the M100 is sensitive to information about the F1 frequency distributions of different vowel categories. In summary, the primary conclusion drawn from these results is that the M100 is sensitive to F1 in vowel perception, as vowel categories with a higher F1 (closer to 1000 Hz) consistently elicit shorter evoked latencies of the M100.

EXPERIMENT 1

The goal of the following experiments is to determine if the auditory system is sensitive to formant ratios, and in particular, if it is sensitive to the F1/F3 ratio. Given our hypothesis regarding the algorithm that is (at least partly) responsible for vowel normalisation, combined with previous MEG findings on vowel perception (Diesch et al., 1996; Govindarajan et al., 1998; Poeppel et al., 1997; Roberts et al., 2000, 2004; Tiitinen et al., 2005), we propose that the M100 is actually sensitive to the ratio of the first formant (F1) against the third (F3), instead of F1 alone. In order for us to test this representational and normalisation hypothesis with the M100, the M100 must be able to index more complex auditory operations performed on the input and not solely reflect surface properties of the stimulus. The results from Roberts et al. (2004) and work on inferential pitch perception that has shown that the M100 is modulated by a missing fundamental component (Fujioka et al., 2003; Monahan, de Souza, & Idsardi, 2008) demonstrate that the M100 can index more complex and abstract auditory operations that integrate information from across the acoustic spectrum.

In the first experiment, we presented participants with synthesised tokens of the mid-vowel categories /ɛ/ and /ə/, holding F1 (and F2)

constant while manipulating the value of F3 for each type. We modulated F3 both higher and lower by 4% in Mel space from the mean/standard F3 value (8% overall difference between the two tokens for a given vowel type). We predict that vowels with a lower F3 (larger F1/F3 ratio) should elicit faster M100 latencies than vowels with a higher F3 value (smaller F1/F3 ratio). This directional prediction is derived from recalculating the formant values of Poeppel et al. (1997). Converting their vowel tokens (for the male fundamental frequency) into F1/F3 Mel space, we find that the token of /a/ used in their experiment had a larger F1/F3 ratio than the token of /u/ and that these two tokens are 20% apart in this transformed space. Given that /a/ elicits an M100 latency than /u/ (Diesch et al., 1996; Govindarajan et al., 1998; Poeppel et al., 1997; Roberts et al., 2000, 2004; Tiitinen et al., 2005), we therefore predict that tokens with a larger F1/F3 ratio should elicit shorter M100 latencies than tokens with a smaller F1/F3 ratio.

METHODS

Materials

Vowel tokens were synthesised using HLSyn (Stevens & Bickley, 1991) with a sampling frequency of 11,025 Hz and an intensity of approximately 70 dB SPL (range: 69.2–71.2 dB SPL). Two tokens for each vowel type (mid-vowels /ɛ/ and /ə/) were synthesised, for a total of four tokens. A fundamental frequency of 150 Hz was used for all tokens. Using an f0 between typical male and female speaker values allowed for greater flexibility in possible F3 values. Moreover, a fundamental frequency of 150 Hz is not outside the possible range for either male or female speakers.

As previously mentioned, the values of F1 and F2 remained consistent across the tokens within each type. The Hertz (Hz) values were converted into Mel space, and we modulated the F3 value 4% higher and 4% lower in the transformed Mel space. Each token was 250 ms in duration with a 10 ms $\cos^2$ on- and off-ramp. The values for F1, F2, and F3, and their respective bandwidths are presented in Table 1 and a comparison of the linear predictive coding (LPC)-based spectral envelopes of the vowel tokens are presented in Figure 4. The F1, F2, and F3 values for [ə] are standard values (Stevens, 1998). The F1 and F2 values (and the F3 value for which we computed from) for [ɛ] are taken from a corpus of American English vowel formant frequencies (Hillenbrand et al., 1995) extracted from the steady-state portion of the vowel in [hVd] syllables. For present purposes, we used the average formant values for male speakers.

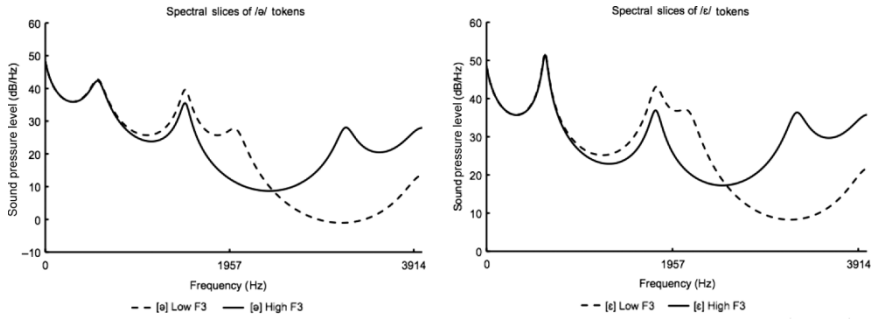

Figure 4. Experiment 1: spectral slices of vowel tokens. Note: LPC-based spectral envelopes of the vowel sounds used in Experiment 1. The solid line indicates the token with a higher F3 (smaller F1/F3 ratio) and the dashed line indicates tokens with a lower F3 (larger F1/F3) ratio. Spectral envelopes smoothed with six-pole LPC filter.

Participants

Thirteen monolingual English participants (five females; mean age: 20 yrs old) participated in the experiment. Two participants were excluded from statistical analysis: for one participant, the evoked waveform did not show a reliable M100, and the other participant showed exceptionally fast M100 responses (< 80 ms). Consequently, the data from 11 participants were analysed. Participants reported no hearing deficits. All participants provided written informed consent approved by the University of Maryland Institutional Review Board (IRB) and scored strongly right handed on the Edinburgh Handedness Survey (Oldfield, 1971). Each participant was compensated $10/hour. The typical session lasted approximately 1½–2 hours.

TABLE 1
Spectral characteristics of the four vowel tokens used in Experiment 1

		F1		*F2*		*F3*	
Vowel type	*F3 height*	*Centre frequency*	*Bandwidth*	*Centre frequency*	*Bandwidth*	*Centre frequency*	*Bandwidth*
/ə/	Low	500	80	1500	90	2040	150
/ə/	High	500	80	1500	90	3179	150
/ɛ/	Low	580	80	1712	90	2156	150
/ɛ/	High	580	80	1712	90	3247	150

Note: The centre frequency and bandwidth for each of the first three formants are provided in Hertz. The stimuli were synthesised using KLSyn (Stevens & Bickley, 1991), a user interface for the HLSyn speech synthesiser. The formant ratio calculations were performed in Mel frequency space and then converted back into Hertz for the speech synthesis.

Procedure

Participants lay supine in a dimly lit magnetically shielded room, as stimulus-evoked magnetic fields were passively recorded by a whole-head 157-channel axial-gradiometer MEG system (Kanazawa Institute of Technology, Kanazawa, Japan). The stimuli were delivered binaurally into the magnetically shielded room via Etymotic ER3A insert earphones that were calibrated and equalised to have a flat frequency response between 100 and 5000 Hz. Prior to the experiment, a hearing test was administered to the participants within the MEG system to ensure normal hearing and that the auditory stimuli were appropriately delivered by the earphones. Subsequently, a pretest localiser was performed. Participants were presented with roughly 100 tokens each of four pure sinusoids: 125, 250, 1000, and 4000 Hz. The neuromagnetic-evoked responses to the sinusoids were epoched and averaged online. The pretest was done to ensure good positioning of the participant's head within the system, as well as guaranteeing that he/she would show a reliable M100 response. The experiment began subsequent to the hearing test and pretest localiser.

For the experiment itself, participants listened to both vowel tokens and pure sinusoids. The four vowel tokens (/ə/: high F3, low F3; /ɛ/: high F3, low F3) were each presented 300 times in pseudo-randomised order (1200 vowel tokens in total), ensuring a good signal-to-noise ratio in the MEG signal. Sinusoids of 250 and 1000 Hz were pseudo-randomly presented 50 times each throughout the experiment. Participants were asked to listen passively to the vowel tokens and discriminate between the 250 and 1000 Hz sinusoids by pressing one of two labelled buttons depending on the sinusoid they heard. The inter-trial interval pseudo-randomly varied between 700 and 1300 ms.

Recording and analysis

Neuromagnetic signals were acquired in DC (no high pass filter) at a sampling frequency of 1 KHz. An online low pass filter of 200 Hz and a 60 Hz notch filter were applied during recording. Noise reduction was performed on the MEG data using a multi-shift PCA noise reduction algorithm (de Cheveigné & Simon, 2007). We extracted epochs of 800 ms intervals, including 200 ms of prestimulus baseline with the zero point set at stimulus presentation onset, from the continuous, noise-reduced data file. During the averaging process, any trials with artifacts exceeding 2.5 pT in amplitude during their epoch were removed from the analysis (6.2% of the total data). Off-line filtering (digital band pass filter with a hamming window, range: 0.03–30 Hz) and baseline correction (100 ms prior to onset of the vowel) were performed on the averaged data.

Evoked waveform analysis

Ten channels from each hemisphere that best correlated with the sink (ingoing magnetic field; five channels) and source (outgoing magnetic field; five channels) of the signal were selected for statistical analysis on a participant-by-participant basis. The same channels were used across the four conditions for the within subjects analysis. The peak latency and amplitude of the root mean square (RMS) of the evoked M100 component in the MEG temporal waveform for each hemisphere were carried forward for statistical analysis.

Equivalent current dipole (ECD) source location analysis

In addition to the latency and amplitude analyses of the RMS of the evoked M100, the equivalent current dipole (ECD) solution for the four distinct vowel tokens was calculated. First, we defined an orthogonal left-handed head-frame; x projected from the inion through to the nasion and z projected through the 10–20 Cz location. Thus, the lateral–medial dimension was defined by x coordinates, the anterior–posterior dimension was defined by y coordinates, and the superior–inferior dimension was defined by z coordinates. Then, a sphere, whose centre position and radius were calculated in headframe coordinates, was fit for each participant covering the entire surface of his/her digitised head shape. A single ECD model in a spherical volume conductor was used for source modelling analysis (Diesch & Luce, 1997; Sarvas, 1987) of the neuromagnetic data. For a source analysis of the data, sensors were selected from each hemisphere for each vowel token within each participant (mean number of channels per hemisphere = 25). The ECD was calculated based on a single point in time located during the final 30 ms of rise time to peak amplitude of the RMS waveform. The minimum goodness of fit (GoF) for inclusion in the analysis was 90% (mean GoF = 95.8%). We, thus, obtained the x, y, and z coordinates for each vowel token (/ɛ/ high F3, /ɛ/ low F3, etc.) in each hemisphere for each participant. From the 11 participants included in the evoked waveform analysis, one additional participant was excluded due to an inability to calculate a GoF > 90% (statistics on ECD source location, $n = 10$). Given that we do not have access to structural magnetic resonance images (MRIs) for each of our participants, we are unable to anatomically localise our findings; instead, our comparisons are based on the relative source location positions for each ECD fit for each vowel token.

Results

We conducted a linear mixed effects model on the M100 latencies and amplitudes for each vowel type with the factors hemisphere (left and right hemisphere) and F3 (high and low) with Subject as a random effect using the lme() package in R statistical software. For the M100 latencies to the

vowel type /ɛ/, only the main effect of F3 had an *F*-value greater than 1 (hemisphere and hemisphere × F3: $F < 1$). Given that neither hemisphere nor the interaction with hemisphere approached significance, we conducted a one-tailed (given our directional prediction: larger F1/F3 ratios should elicit shorter latencies) paired *t*-test comparing the latencies to the token of /ɛ/ with a high F3 to the token of /ɛ/ with a low F3. As predicted, the /ɛ/ token with the lower F3 (larger F1/F3 ratio) elicited a significantly shorter M100 latency than the /ɛ/ token with the higher F3 [smaller F1/F3 ratio; $t(21) = 3.05$; $p < .005$]. When the hemispheres are compared independently of one another, we also find the reliable differences in the predicted directions [one-tailed paired *t*-tests; Left Hemisphere (LH): $t(10) = 1.88$, $p < .05$; Right Hemisphere (RH): $t(10) = 2.35$, $p < .05$; Figure 5]. Comparing the tokens of /ə/ using the same model as above, we did not obtain reliable differences in the latency of the RMS of the M100 response (all Fs < 1). Analysing the amplitude of the RMS M100 waveform, we again modelled the data using a linear mixed effects model with Subject as a random effect and the factors hemisphere (left and right hemisphere) and F3 (high and low). We find a main effect of hemisphere for each vowel type [right hemisphere showing significantly larger amplitudes than the left hemisphere; /ɛ/: $F(1, 30) = 143.3$, $p < .0001$; /ə/: $F(1, 30) = 181.7$, $p < .0001$], but no main effects F3 or interactions of hemisphere × F3 (all Fs < 1). At this point in time, we do not have an explanation for why the right hemisphere shows reliably larger amplitudes than the left hemisphere.

To summarise, we found a reliable difference in the latency of the evoked M100 component in the predicted direction, vowel tokens with a larger

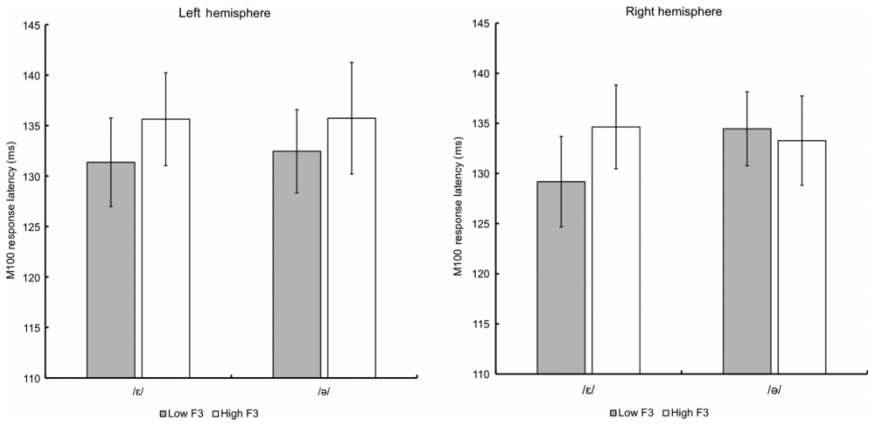

Figure 5. Experiment 1: M100 response latencies by vowel type. Note: Mean M100 response latencies across participants to the vowel tokens in Experiment 1. Gray bars refer to tokens with a Low F3 (large F1/F3 ratio) and white bars refer to tokens with a high F3 (small F1/F3 ratio). Error bars represent one standard error of mean.

F1/F3 ratio elicit a shorter M100 latency than tokens with a smaller F1/F3 ratio, which is consistent with our reinterpretation of the previous M100 results on F1 (Diesch et al., 1996; Govindarajan et al., 1998; Poeppel et al., 1997; Roberts et al., 2000, 2004; Tiitinen et al., 2005). We only find this effect for the front vowel /ɛ/, however, and not the central vowel /ə/. In the discussion, we speculate on some potential explanations for this pattern of results. In general, however, we take these results to suggest that the perceptual system is sensitive to formant ratios, and the F1/F3 ratio in particular.

M100 equivalent current dipole (ECD) source location

The primary aim of this study was to determine if the auditory perceptual system is sensitive to formant ratios in general, and moreover, if manipulations of the F1/F3 ratio would modulate the response latency of the RMS of the evoked auditory M100 component. In addition to the latency and amplitude analysis of the M100, we also conducted a source analysis on the data to determine if there were any localisation differences between the vowels. Obleser, Lahiri, and Eulitz (2004), using MEG, calculated the ECD source location for seven distinct German vowels. They found that front vowels tend to map onto a more anterior portion of auditory cortex, while back vowels map onto a more posterior region, retaining the front/back distinction of vowel categories on the anterior/posterior dimension of auditory cortex. It should be noted that Obleser et al. (2004) did not include mid-vowels in their experiment; however, given the directionality of their effects, we might expect to find reliable differences between the tokens of /ɛ/ and /ə/, with the front vowel /ɛ/ localising to more anterior regions than the mid-vowel /ə/. We performed a linear mixed effects model on each coordinate axis (i.e., x, y, and z) in each hemisphere independently with the factors vowel (/ə/ and /ɛ/) and F3 (high and low) and Subject as a random effect. In the left hemisphere, along the lateral–medial dimension, we find no main effects or Vowel × F3 interaction (all Fs < 1) and along the superior–inferior dimension, we again find no main effects (all Fs < 1), but we do find an interaction of Vowel × F3 [$F(1, 27) = 5.38$, $p < .05$]. Finally, along the inferior–posterior dimension, the dimension in which Obleser et al. (2004) found reliable differences in the ECD source location between front and back vowels, we also find no main effects of vowel or F3 and no interaction of Vowel × F3 (all Fs < 1).

In the right hemisphere, along the lateral–medial dimension, we again find no main effects and no interaction of Vowel × F3 (all ps $< .1$), and along the superior–inferior dimension, we also find no main effects and no interaction of Vowel × F3 (all Fs < 1.0). Finally, along the anterior–posterior dimension, we again find no main effects and no interaction of Vowel × F3 (all ps $< .2$). Finally, we performed a one-tailed sign test on the values in the anterior–posterior dimension across vowel type for each hemisphere to determine

whether the front vowels (the two tokens of /ɛ/) were located more anterior than the mid-vowels (the two tokens of /ə/). In the left hemisphere, we found no difference between the tokens with a high F3 ($S = 6$; $p = .5$) or between the tokens with a low F3 ($S = 4$; $p = .89$), and in the right hemisphere, we find no difference between vowel types for the tokens with a high F3 ($S = 4$; $p = .89$) or with a low F3 ($S = 4$; $p = .89$) along the anterior–posterior dimension.

Discussion

These findings suggest that auditory cortex (minimally, the neurobiological generators of the M100) is sensitive to formant ratios, and in particular, to modulations of the F1/F3 ratio. The latency difference was robust across participants for the /ɛ/ vowel and was in the predicted direction for nearly all subjects for the vowel type /ɛ/ (Sign Test: $S = 16$; $p < .05$). Any more concrete conclusions, however, should be taken cautiously, given that we did find such an effect for the mid-central vowel /ə/.

The immediate question is why we found an effect of F3 manipulation for /ɛ/ but not for /ə/. The lack of a result for /ə/ is not likely due to a lack of power in the experiment, given that 300 tokens of each vowel are more than sufficient to obtain a good signal-to-noise ratio. Moreover, the fact that we found an effect with /ɛ/ suggests that this asymmetry is due to some intrinsic properties of the vowels or their location in vowel space. It is this latter possibility that we explore in the second experiment. In particular, the asymmetry found in Experiment 1 could be a consequence of the location in vowel space that /ɛ/ and /ə/ occupy. The front mid-vowel /ɛ/ occupies a more crowded portion of the vowel space relative to that occupied by /ə/ (i.e., there are many more phonetic categories in close proximity to the distribution of /ɛ/ as opposed /ə/ in the vowel space), where categorisation might be more critical than the middle of the vowel space.

In Experiment 2, we test the hypothesis that the asymmetry found in Experiment 1 is due to the location in vowel space of each vowel. Consequently, we test two hypotheses. First, we aim to replicate the null effect with /ə/ that we found in Experiment 1. To accomplish this, we test the same /ə/ tokens with a different set of participants. Second, to test whether it is the demands for categorisation that drive the perceptual system's sensitivity to formant ratios in more crowded portions of the vowel space, we test two tokens of /o/ with the same manipulations we performed on /ɛ/ in the first experiment. The back vowel /o/, like /ɛ/, also occupies a more crowded portion of the vowel space than /ə/.

EXPERIMENT 2

The second experiment was nearly identical to Experiment 1; however, instead of testing tokens of /ɛ/, we tested synthesised tokens of /o/, a vowel

produced in the back of the vocal tract. The back mid-rounded vowel /o/, like /ɛ/, resides in a more crowded portion of the vowel space, at least when compared with the central vowel /ə/. Practically speaking, our hypothesis predicts that we should find M100 latency differences for vowels located in more crowded portions of the vowel space. Therefore, we should find effects for /o/ and replicate our null effects for /ə/. We speculate that the reason for this particular pattern is likely due to greater competition within the category space, which drives the perceptual system's heightened sensitivity to the formant ratios in these more densely populated regions.

METHODS

Materials

For the /ə/ stimuli, we used the same tokens used in Experiment 1. For the /o/ stimuli, we synthesised two new tokens using HLSyn (Stevens & Bickley, 1991) with a sampling frequency of 11,025 Hz and an average intensity level of 70 dB SPL (range: 69.5–71.2 dB SPL). The F1 and F2 values were taken from Hillenbrand et al. (1995). Again, we converted the Hz frequency values into Mel space. Using the F3 value (transformed into Mel space) from Hillenbrand et al. (1995) as the standard, we computed the new F3 values for our experimental tokens by moving 4% in either direction of the F1/F3 ratio space. Therefore, the overall distance in F1/F3 ratio space between the tokens was 8%. As before, we predict a M100 latency facilitation for the token with the lower F3 (the larger F1/F3 ratio). The F1, F2, and F3 values for the four tokens used in Experiment 2 are presented in Table 2 and a comparison of the LPC-based spectral envelopes of the vowel tokens are presented in Figure 6.

TABLE 2
Spectral characteristics of the vowel tokens used in Experiment 2

		F1		*F2*		*F3*	
Vowel type	*F3 height*	*Centre frequency*	*Bandwidth*	*Centre frequency*	*Bandwidth*	*Centre frequency*	*Bandwidth*
/ə/	Low	500	80	1500	90	2040	150
/ə/	High	500	80	1500	90	3179	150
/o/	Low	497	80	938	90	2011	150
/o/	High	497	80	938	90	3118	150

Note: The centre frequency and bandwidth for each of the first three formants are provided in Hertz. The stimuli were synthesised using KLSyn (Stevens & Bickley, 1991), a user interface for the HLSyn speech synthesiser. The formant ratio calculations were performed in Mel frequency space and then converted back into Hertz for the speech synthesis.

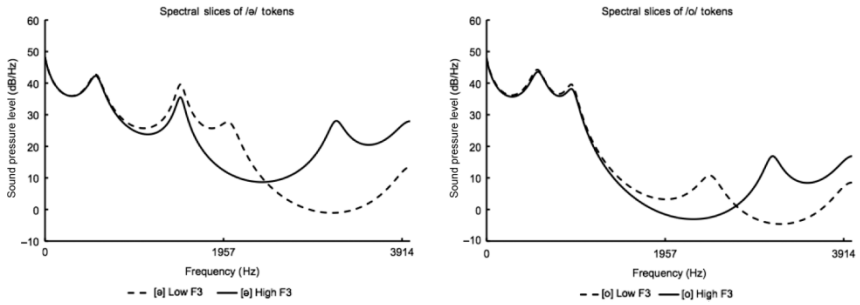

Figure 6. Experiment 2: spectral slices of vowel tokens. Note: LPC-based spectral envelopes of the vowel sounds used in Experiment 2. The solid line indicates the token with a high F3 (smaller F1/F3 ratio) and the dashed line indicates the token with a low F3 (larger F1/F3) ratio. Spectral envelopes smoothed with six-pole LPC filter.

Participants

Fifteen monolingual English participants (nine females; mean age: 20 yrs old) participated in the experiment. Six participants were excluded from analysis on various grounds: two participants were not included in the analysis, as there was no discernable M100 in the data; two participants were excluded from the analysis because the source distribution of the component did not match that of an M100; one participant was excluded because the peak latency of their M100 was over 200 ms; and finally, one participant was excluded due to hardware failure. Consequently, for the analysis, the data from nine participants (five females) were analysed. All participants had normal hearing. All participants provided written informed consent approved by the University of Maryland IRB and scored strongly right handed on the Edinburgh Handedness Survey (Oldfield, 1971). Each participant was compensated $10/hour. The typical session lasted approximately 1½–2 hours.

Procedure

The procedure was identical to Experiment 1.

Recording and analysis

The recording parameters and analysis procedures used in Experiment 2 were identical to those used in Experiment 1. All trials with artifacts above 2.5 pT in the noise-reduced data were eliminated from analysis (5.2% of the total data). The filtering and baseline correct parameters are identical to those used in Experiment 1.

Evoked waveform analysis

The methods for channel selection and calculation of the RMS of the evoked waveform carried forward for statistical analysis were identical to those used in Experiment 1.

Equivalent current dipole (ECD) source location analysis

The methods used for calculation of the ECD source locations in Experiment 2 are identical to those in Experiment 1. The mean number of channels per hemisphere across participants for each measurement was 28. The minimum GoF for inclusion in the analysis was 90% (mean GoF = 95%). From the nine participants included in the evoked waveform analysis, one additional participant was excluded due to an inability to calculate a GoF > 90% (statistics on ECD source location, $n = 8$). Again, given that we do not have access to structural MRIs for each of our participants, we are unable to anatomically localise our findings; instead, our comparisons are based on the relative source location positions for each ECD fit for each vowel token.

Results

Given the results from Experiment 1 and our hypothesis that the perceptual system displays a greater sensitivity to formant ratios for vowels located in more densely populated regions of the vowel space, we predict to find a reliable difference between the two tokens of /o/, with the token with a lower F3 (larger F1/F3) eliciting a shorter M100 latency, while we expect to replicate the null difference for the two tokens of /ə/ that we found in Experiment 1. We again conducted a linear mixed effects model on the M100 latencies and amplitudes for each vowel type with the factors hemisphere (left and right hemisphere) and F3 (high and low) with Subject as a random effect using the lme() package in R statistical software. For the M100 latencies to the vowel type /o/, we find main effects of F3 [$F(1, 24) = 15.36$, $p < .001$] and hemisphere [$F(1, 24) = 4.95$, $p < .05$] but no hemisphere × F3 interaction ($F < 1$). The response latencies of the RMS of the M100 temporal waveform were approximately 6 ms shorter than those in the left hemisphere [paired two-tailed t-test: $t(17) = 2.93$, $p < .01$]. We did not have a hypothesis regarding hemispheric differences in latencies, a finding not reported in other similar experiments (e.g., Diesch et al., 1996; Tiitinen, Sivonen, Alku, Virtanen, & Näätänen, 1999), and consequently, we are hesitant to draw any significant conclusions based on this result. The main effect of F3 motivates our planned comparison: a comparison of the latencies to the token of /o/ with a high F3 to the token of /o/ with a low F3 using a paired one-tailed t-test. Our findings demonstrate that the /o/ token with the lower F3 (larger

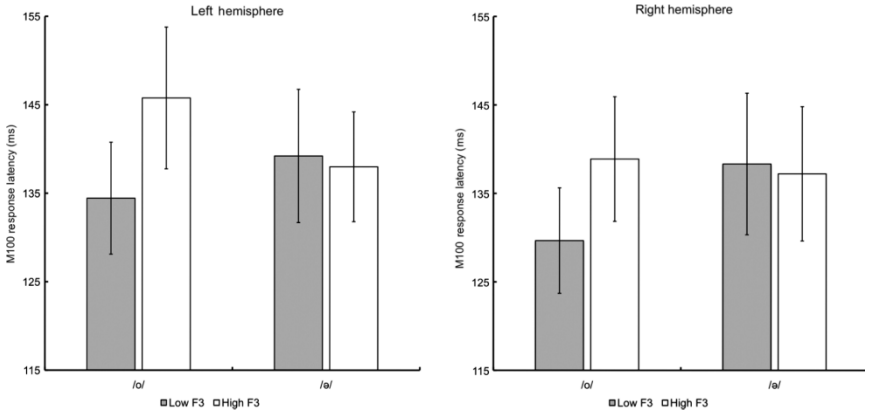

Figure 7. Experiment 2: M100 response latencies by vowel type. Note: Mean M100 response latencies across participants to the vowel tokens in Experiment 2. Gray bars refer to tokens with a Low F3 (large F1/F3 ratio) and white refer to tokens with a high F3 (small F1/F3 ratio). Error bars represent one standard error of mean.

F1/F3 ratio) elicited a reliably faster M100 latency than the /o/ token with the higher F3 [smaller F1/F3 ratio; $t(17) = 3.81$; $p < .001$]. Subsequently, we compared the hemispheres independently of one another, and we again find reliable differences in the predicted directions [one-tailed paired t-tests; LH: $t(8) = 2.93$, $p < .01$; RH: $t(8) = 2.33$, $p < .05$; Figure 7]. In order to assess whether we were able to replicate the null finding for the tokens of /ə/ from Experiment 1, we used the same model as above. As expected, consistent with the hypothesis that there is an influence on location in vowel space with sensitivity to formant ratios, we did not find reliable differences in the latency of the RMS of the M100 response between the tokens of /ə/ (all Fs < 1). Next, we turn to an analysis of the amplitudes of the RMS of the M100 temporal waveform. We modelled the data using a linear mixed effects model with Subject as a random effect and the factors hemisphere (left and right hemisphere) and F3 (high and low) on each vowel type separately. We found no main effects or interactions for either vowel type (i.e., /o/ and /ə/; all Fs $<$ 1). The findings from Experiment 2 confirm that the auditory perceptual system (at least the neurobiological generators of the M100) is sensitive to formant ratios, and in particular at least the F1/F3 ratio, and that the perceptual system shows greater sensitivity to formant ratios in regions of the vowel space that are more densely populated.

M100 equivalent current dipole (ECD) source localisation

To assess whether the vowels presented to participants in Experiment 2 elicited differences in their source localisation as well as the latency of the

evoked M100, we calculated the ECD solution for the four distinct vowel tokens on an intra-subject and intra-hemispheric basis. Identical to the statistical analysis performed for the ECD source data in Experiment 1, we performed a linear mixed effects model on each coordinate axis (i.e., *x*, *y*, and *z*) in each hemisphere independently with the factors vowel (/ə/ and /o/) and F3 (high and low) and Subject as a random effect. We first report our findings from the left hemisphere. Along the lateral–medial dimension, we find no main effects or interaction (all ps $< .15$). Moreover, along the superior–inferior dimension, we again find no main effects or interaction (all ps $< .1$). Finally, along the inferior–posterior dimension, we again find no main effects and no interaction (all ps $< .1$). In the right hemisphere, along the lateral–medial dimension, we also find no main effects and no interaction (all Fs < 1), and along the superior–inferior dimension, we again find no main effects and no interaction (all ps $< .15$). Finally, along the anterior–posterior dimension, there are no main effects and no interactions (all ps $< .25$). To determine if there are directional differences in the location of the ECD between the vowels along the anterior–posterior dimension, we performed a Sign Test on the different vowels types within each hemisphere. In the left hemisphere, we find a strong directional difference for the tokens with a high F3 ($S = 1$; $p < .05$), in the opposite direction, with the ECD source location of the token of /o/ with a high F3 localising to a more anterior position along the anterior–posterior dimension than the token of /ə/ with a high F3. We find no difference between the tokens with a low F3 ($S = 5$; $p = .86$). In the right hemisphere, we neither found effect between /ə/ and /o/ with low F3s ($S = 6$; $p = .14$), nor did we find an effect between /ə/ and /o/ with high F3s ($S = 4$; $p = .64$).

Discussion

The motivation for Experiment 2 was to determine if the density of speech sound categories in perceptual space affects the sensitivity of the perceptual system to formant ratios. Recall that in Experiment 1, we found a significant M100 latency difference for the /ɛ/ token with a larger F1/F3 ratio but not for the /ə/ token with a larger F1/F3 ratio. If an adequate explanation for the findings in Experiment 1 is that the sensitivity of our perceptual system to the F1/F3 ratio is a function of how dense the space is, and consequently, how much more competitive categorisation is, then we also predict to find a significant difference for tokens of /o/ that vary on the F1/F3 ratio. As predicted, the token of /o/ with a larger F1/F3 ratio elicited a shorter M100 latency than the token of /o/ with a smaller F1/F3 ratio. And equally important, we replicated the null effect for /ə/. This reaffirms our findings from Experiment 1 that the auditory system is sensitive to F1/F3 ratios, lending further support to using ratios in normalisation algorithms.

And moreover, it demonstrates that formant ratios are psychologically plausible computations that can be exploited in the course of speaker normalisation.

An alternative explanation of the results we report is that the M100 response latency is sensitive to the entire spectrum, and therefore is also sensitive to modulations of F3 or perhaps even to differences in the power spectral density (PSD) of the vowel tokens (see Roberts et al., 2000 for results that suggest the M100 is sensitive to PSD). However, given that the differences in F3 between the tokens for each category (/ɛ/: $\Delta = 1091$ Hz; /ə/: $\Delta = 1139$ Hz; /o/: $\Delta = 1107$ Hz) are roughly equivalent in raw Hz space and moreover, the differences between tokens within each category are equivalent in Mel space (8% difference in the Mel space), this alternative does not adequately account for the M100 latency findings. Additionally, differences in the central moment of the PSD of the tokens (PSD; /ɛ/: $\Delta = 11$ Hz; /ə/: $\Delta = 14$ Hz; /o/: $\Delta = 10$ Hz) cannot account for the differences either, as /o/ has a smaller difference than /ə/ and yet, we found a reliable difference in the M100 response latency for /o/ and not for /ə/. While we accept that the overall PSD may contribute considerably to the response (differences in formant ratios lead to differences in power spectral densities), our results suggest that this alternative is insufficient as the sole property responsible for our findings.

GENERAL DISCUSSION

Understanding how listeners normalise the highly variable speech signal across different talkers has been a long-standing problem in speech perception research (see Johnson, 2005 for an overview of the various approaches to speaker normalisation). Within the domain of vowel perception, a variety of different proposals have been offered to account for how listeners cope with this variation (Adank, Smits, & van Hout, 2004; Disner, 1980; Irino & Patterson, 2002; Miller, 1989; Nearey, 1989; Rosner & Pickering, 1994; Strange, 1989; Zahorian & Jagharghi, 1993). Here, we revisited an idea that has been sporadically proposed in the literature: listeners are sensitive to the relative differences between formants (formant ratios) and not their absolute values (Lloyd, 1890; Miller, 1989; Peterson, 1951, 1961; Peterson & Barney, 1952; Syrdal & Gopal, 1986). We put forward a (relatively) novel formant ratio algorithm in which the first (F1) and second (F2) formants are ratioed against the third formant (F3). Higher formants, such as F3, may act as an adequate normalising factor (Deng & O'Shaughnessy, 2003) and have been, at least impressionistically, judged to eliminate speaker-dependent variation (Peterson, 1951), the sort of variation that exists in vowel productions between men, women, and children. Previous work on formant ratios has employed

algorithms that require large corpora to adequately eliminate speaker variation (e.g., Miller, 1989). One of the advantages to the algorithm we propose here is that it appears to be an efficient computation for online speaker normalisation that can be performed with little exposure to a given speaker, which is consistent with what we know about dialect identification (Purnell et al., 1999), the perceptual abilities of infants (Kuhl, 1979, 1983), and listeners' abilities to make speaker size estimates (Ives et al., 2005; Smith et al., 2005).

In this paper, we investigated whether the perceptual system is sensitive to the F1/F3 ratio (the less novel of the two ratios; F2/F3 has appeared in previous ratio algorithms (Miller, 1989; Syrdal & Gopal, 1986)). We reported data from two MEG experiments that demonstrate that the neurobiological generators of the M100, an early, auditory evoked neuromagnetic component is sensitive to modulation of the F1/F3 ratio. The M100 had been previously reported to show sensitivity to the frequency of F1 in vowel perception (Diesch et al., 1996; Govindarajan et al., 1998; Poeppel et al., 1997; Roberts et al., 2000, 2004; Tiitinen et al., 2005). Given our hypothesis regarding the algorithm involved in vowel normalisation and the consequent representational nature of the vowel space (F1/F3 by F2/F3), we reinterpreted the previous MEG findings to conclude that the M100 is actually sensitive to the F1/F3 ratio and not F1 alone. The frequency of the third formant (F3) was not typically modulated in the previous MEG experiments, only F1. Therefore, we hypothesised that if we varied the value of the F3, and consequently, the F1/F3 ratio, we should be able to modulate the latency of the M100 in a predicted direction if the neurobiological generators of the M100 are sensitive to the F1/F3 ratio.

Our findings suggests the perceptual system can calculate formant ratios (or something substantially equivalent), lending further support to the notion that this is a plausible normalisation algorithm, and moreover, that the M100 is sensitive to the F1/F3 ratio and not F1 alone. Furthermore, we calculated the statistical effectiveness of this algorithm in eliminating variance that is a function of the age and gender of a speaker on a large corpus of productions of American English vowels (Hillenbrand et al., 1995). While the statistical analysis was perfunctory in many respects (e.g., we did not calculate how well the vowel space categorises or how well particular tokens are classified as is normally done, which was beyond the scope of this paper), the calculations demonstrate that speaker-dependent variation, when we compare vowel utterances across different talkers, as a function of age and gender, was eliminated.

While we found that auditory cortex is sensitive to modulations of the F1/F3 ratio, the pattern of effects suggests a more nuanced conclusion. In the first experiment, we found a reliable difference in the predicted direction only for the front mid-vowel /ɛ/, but no difference between in the

response latency of the M100 for the two tokens of /ə/. As a result of this asymmetric result and the direction of the pattern, we hypothesised that the perceptual system displays heightened sensitivity to modulations of the F1/F3 ratio only when mapping acoustic information into more crowded regions of the vowel space. Experiment 2 was designed to test this hypothesis. As predicted, we found a reliable difference in the latency of the M100 between the two tokens of /o/ in the predicted direction and we replicated the null effect for /ə/, demonstrating that the sensitivity of auditory system is not equal across the vowel space. To place these findings within a theoretical framework, in English, the front and back portions of the vowel space are more densely populated and therefore, categorisation can be though of as being "more competitive". In other words, the acoustic distribution of a vowel can afford to be more diffuse in central portions of the vowel space where no other categories exist, as compared to more densely populated regions of the space, where more different vowel categories are located. This provides an intuitive explanation for why we might find a greater sensitivity of the neurobiological generators of the M100 to vowels located in the front and back of the vowel space as compared with vowels located in the centre of the space.

As a point about the M100 component itself, we can be confident that the M100 is sensitive not only to F1, but that higher regions of the frequency space also play a role in modulating its latency. In particular, we conclude that the response latency of the M100 indexes more abstract computations that have been performed on the stimulus and in fact reflect complex representational schemas in auditory cortex. This conclusion is consistent with other work done on the relation between the M100 and F1 (Roberts et al., 2004) and findings that demonstrate that the M100 is sensitive to differences in the inferred pitch of complex tone stimuli that are missing a fundamental component (Fujioka et al., 2003; Monahan et al., 2008).

The question of how the brain computes formant ratios is a tractable one; and one that we believe is a point where biology and psycholinguistics can fruitfully combine to provide a fairly complete account of a perceptual linguistic phenomenon. Since Delattre, Liberman, Cooper, and Gerstman (1952) presented participants with synthetic one- and two-formant vowels and showed that listeners were able to reliably judge vowel category based on this information alone, the working hypothesis within the field is that listeners extract formant information from vowel tokens (i.e., the "formant extraction" principle). One of the possible ways in which the brain encodes formant information is via rate encoding at various characteristic frequencies (CF) of auditory nerve fibres (Sachs & Young, 1979; Young & Sachs, 1979). For example, Sachs and Young (1979) recorded the rate response properties of populations of neurons

with different CF in auditory nerve fibres in anaesthetised cats to the steady-state synthetic vowels /ɪ/, /ɛ/, and /a/. For stimulus presentations below 70 dB SPL, they report increases in the normalised rate of auditory nerve fibres whose CF matches the formant peaks of the synthetic vowels; for the vowels /ɪ/ and /ɛ/, they show a clear separation between the peaks of discharge rates of nerve fibres with CF corresponding to F1 and F2 of the synthesised vowel (the separation of discharge rates of the CF between F2 and F3 was not as clear—it should be noted, however, that the distance between F2 and F3 in the spectral envelope of their synthetic vowels was not particularly large for these two vowel types). For the vowel /a/, which has the most closely spaced F1 and F2 of all English vowels (and thus, the most distanced F2 and F3 of the vowels they tested), they report what appears to be a separation in the peak discharge rates at CF corresponding to F1, F2, and F3 separately at the lower presentation levels. At higher presentation levels (>70 dB SPL), the distinct peaks appear to rate saturate. Provided these results, one could conclude that the auditory representation of vowel spectra is in terms of place (CF) and rate encoding. Given the inability to find reliable distinct peaks in the discharge rate at sound intensity levels greater than 70 dB SPL, however, in a follow-up paper, Young and Sachs (1979) replaced normalised rate and instead measured the temporal response patterns (defined as the amount of synchronisation between the peak of a harmonic in the Fourier transform of the vowel and the discharge rate at that particular harmonic) of the nerve fibres, again at different CF along the auditory nerve. They find a better representation of the vowel spectra (including separation of F1, F2, and F3) throughout the range of sound levels used in the experiment. A combination of rate, place, and temporal coding provides an interpretable representation of vowels in auditory nerve fibres. Delgutte and Kang (1984) report similar findings using two-formant steady-state synthesised vowels, whereby the CF of the auditory nerve fibres closest to the spectral peaks of the auditory stimulus dominated the responses; subsequently, they delineate the tonotopically arranged fibres into five distinct CF regions centred around the largest spectral peaks in the vowel stimuli. Much of the work on vowel perception, however, has demonstrated the need for a transformation of the vowel space (i.e., a simple F2 by F1 coordinate system is an inadequate representation of the vowel space; see the formant ratio literature cited above and Rosner & Pickering, 1994 for an overview).

Ohl and Scheich (1997) measured cortical patterns in response to vowels using 2-Fluro-2-Deoxy-d-[^{14}C(U)]Glucose (FDG) Autoradiography in euthanised gerbils. They found a vertical stripe of activity caused by vowel excitation along consecutive horizontal slices of A1 in auditory cortex. Boundaries of activity along the dorsal–ventral direction appeared to

correlate with the distance between the first two formants; that is, the vowel /i/, which has a larger F2–F1 distance showed stripes that extended further dorsally than vowels with a smaller F2–F1 distance (e.g., /o/). Single formant vowels produced maximal vertical excitation across the cortical slices, suggesting neuronal inhibition in the calculation of relative differences between formants. These results provide support for the notion that auditory cortex calculates the relative differences between formant peaks as a means towards vowel perception, a result consistent with the central intuition of formant ratios as a mechanism in solving speaker normalisation. Recent electrophysiological work is also consistent with this result. For example, Diesch and Luce (1997), using MEG, showed that the properties of the N1m/M100 (latency, ECD moment, and ECD location) differed between responses to composite (two-formant vowel tokens) and the linear sum of its components, suggesting that the component spectral properties of a vowel (e.g., formants, harmonics, etc.) interact. Additionally, also using MEG, Mäkelä, Alku, and Tiitinen (2003) showed that vowels having equal F2–F1 differences elicited equally strong N1m/M100 responses—again, suggesting that the calculation of differences between formants is a plausible algorithm employed by auditory cortex. It seems then, that the neurophysiological evidence supports the extraction of formant peaks in the auditory nerve and, with additional evidence from human electrophysiology, sensitivity to relative differences between formant peaks can be found in cortical responses. Whether these calculations occur prior to cortex remains to be seen. While the particular algorithm proposed in this paper has not been tested using such methods, the general notion that listeners are calculating relative differences between spectral peaks gains considerable evidence from this work—and provides support for the idea that work of this sort provides a tractable bridge between linguistics/psychology and neuroscience.

CONCLUSION

The goal of this paper was to test whether auditory cortex, and in particular, the neurobiological generators of the M100 located in auditory cortex, were sensitive to formant ratios. In a pair of experiments using MEG, we found that the latency of the M100 is modulated by the F1/F3 formant ratio. These results also suggest, however, that the auditory system shows differential sensitivity to formant ratios depending upon where in vowel space the vowel categories are located. In particular, we found significant M100 latency differences to modulations of the F1/F3 ratio for tokens of the vowel categories /ɛ/ and /o/ but not /ə/. While we are hesitant to conclude that this is the algorithm wholly responsible for successfully eliminating variance based

on inter-speaker variation in vowel perception, we suggest that the exploitation of higher formants in vowel normalisation, in particular F3, could provide valuable insight into furthering our understanding of the perceptual and neurobiological mechanisms underlying speaker normalisation.

Manuscript received 11 June 2009
Revised manuscript received 26 April 2010

REFERENCES

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, *116*, 3099–3107.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bonte, M., Valente, G., & Formisano, E. (2009). Dynamic and task-dependent encoding of speech and voice by phase reorganization of cortical oscillations. *Journal of Neuroscience*, *29*, 1699–1706.

Broad, D. J., & Wakita, H. (1977). Piecewise-planar representation of vowel formant frequencies. *Journal of the Acoustical Society of America*, *62*, 1467–1473.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Claes, T., Dologlou, I., ten Bosch, L., & van Compernolle, D. (1998). A novel feature transformation for vocal tract length normalization in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, *6*, 549–557.

de Cheveigné, A., & Simon, J. Z. (2007). Denoising based on time-shift PCA. *Journal of Neuroscience Methods*, *165*, 297–305.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word*, *8*, 195–210.

Delgutte, B., & Kang, N. Y. S. (1984). Speech coding in the auditory nerve: I. Vowel-like sounds. *Journal of the Acoustical Society of America*, *75*, 866–878.

Deng, L., & O'Shaughnessy, D. (2003). *Speech processing: A dynamic and optimization-oriented approach*. New York: Marcel Dekker.

Diesch, E., Eulitz, C., Hampson, S., & Ross, B. (1996). The neurotopography of vowels as mirrored by evoked magnetic field measurements. *Brain and Language*, *53*, 143–168.

Diesch, E., & Luce, T. (1997). Magnetic fields elicited by tones and vowel formants reveal tonotopy and nonlinear summation of cortical activation. *Psychophysiology*, *34*, 501–510.

Disner, S. F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, *67*, 253–261.

Eulitz, C., Diesch, E., Pantev, C., Hampson, S., & Elbert, T. (1995). Magnetic and electric brain activity evoked by the processing of tone and vowel stimuli. *Journal of Neuroscience*, *15*, 2748–2755.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fitch, R. H., Miller, S., & Tallal, P. (1997). Neurobiology of speech perception. *Annual Review of Neuroscience*, *20*, 331–353.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, *106*, 1511–1522.

Formisano, E., de Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "What"? Brain-based decoding of human voice and speech. *Science*, *322*, 970–973.

Fox, R. A., Jacewicz, E., & Feth, L. L. (2008). Spectral integration of dynamic cues in the perception of syllable initial stops. *Phonetica*, *65*, 19–44.

Frye, R. E., Rezaie, R., & Papanicolaou, A. C. (2009). Functional neuroimaging of language using magnetoencephalography. *Physics of Life Reviews*, *6*, 1–10.

Fujioka, T., Ross, B., Okamoto, H., Takeshima, Y., Kakigi, R., & Pantev, C. (2003). Tonotopic representation of missing fundamental complex sounds in the human auditory cortex. *European Journal of Neuroscience*, *18*, 432–440.

Fujisaki, H., & Kawashima, T. (1968). The role of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics, AU-16*, 73–77.

Gage, N., Roberts, T. P. L., & Hickok, G. (2006). Temporal resolution properties of human auditory cortex: Reflections in the neuromagnetic auditory evoked m100 component. *Brain Research*, *1069*, 166–171.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 1166–1183.

Govindarajan, K. K., Phillips, C., Poeppel, D., Roberts, T. P. L., & Marantz, A. (1998). Latency of MEG M100 response indexes first formant frequency. *Journal of the Acoustical Society of America*, *103*, 2982–2983.

Halberstam, B., & Raphael, L. J. (2004). Vowel normalization: The role of fundamental frequency and upper formants. *Journal of Phonetics*, *32*, 423–434.

Hari, R., Aittoniemi, K., Järvinen, M. L., Katila, T., & Varpula, T. (1980). Auditory evoked transient and sustained magnetic fields of the human brain: Localization of neural generators. *Experimental Brain Research*, *40*, 237–240.

Hari, R., Levänen, S., & Raij, T. (2000). Timing of human cortical functions during cognition. *Trends in Cognitive Sciences*, *4*, 455–462.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*, 393–402.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999). Formants of children, women, and men: The effects of vocal intensity variation. *Journal of the Acoustical Society of America*, *106*, 1532–1542.

Irino, T., & Patterson, R. D. (2002). Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-mellin transform. *Speech Communication*, *36*, 181–203.

Ives, D. T., Smith, D. R. R., & Patterson, R. D. (2005). Discrimination of speaker size from syllable phrases. *Journal of the Acoustical Society of America*, *118*, 3816–3822.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Oxford: Blackwell.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, *66*, 1668–1679.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior & Development*, *6*, 263–285.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*, 98–104.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford: Blackwell.

Lloyd, R. J. (1890). Speech sounds: Their nature and causation. *Phonetische Studien*, *3*, 251–278.

Lounasmaa, O. V., Hämäläinen, M., Hari, R., & Salmelin, R. (1996). Information processing in the human brain: Magnetoencephalographic approach. *Proceedings of the National Academy of Sciences*, *93*, 8809–8815.

Mäkelä, A. M., Alku, P., & Tiitinen, H. (2003). The auditory N1m reveals the left-hemispheric representation of vowel identity in humans. *Neurocsience Letters*, *353*, 111–114.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*, 1113–1126.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, *85*, 2114–2134.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*, 331–340.

Monahan, P. J., de Souza, K., & Idsardi, W. J. (2008). Neuromagnetic evidence for early auditory restoration of fundamental pitch. *PLoS ONE, 3*, e2900.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, *24*, 375–425.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*, 2088–2113.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Obleser, J., & Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, *13*, 14–19.

Obleser, J., Lahiri, A., & Eulitz, C. (2004). Magnetic brain response mirrors extraction of phonological features from spoken vowels. *Journal of Cognitive Neuroscience*, *16*, 31–39.

Ohl, F. W., & Scheich, H. (1997). Orderly cortical representation of vowels based on formant interaction. *Proceedings of the National Academy of Sciences*, *94*, 9440–9444.

Oldfield, R. C. (1971). Assessment and analysis of handedness: Edinburgh inventory. *Neuropsychologia*, *9*, 97–113.

Peterson, G. E. (1951). The phonetic value of vowels. *Language*, *27*, 541–553.

Peterson, G. E. (1961). Parameters of vowel quality. *Journal of Speech and Hearing Research*, *4*, 10–29.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

Phillips, C. (2001). Levels of representation in the electrophysiology of speech perception. *Cognitive Science*, *25*, 711–731.

Picton, W., Woods, D. L., Baribeau-Braun, J., & Healey, T. M. (1976). Evoked potential audiometry. *Journal of Otolaryngology*, *6*, 90–119.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 101–139). Berlin: Mouton de Gruyter.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–31). San Diego, CA: Academic Press.

Poeppel, D., Phillips, C., Yellin, E., Rowley, H. A., Roberts, T. P. L., & Marantz, A. (1997). Processing of vowels in supratemporal auditory cortex. *Neuroscience Letters*, *221*, 145–148.

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America*, *22*, 807–820.

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, *18*, 10–30.

R Development Core Team (2006). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Roberts, T. P. L., Ferrari, P., Stufflebeam, S. M., & Poeppel, D. (2000). Latency of the auditory evoked neuromagnetic field components: Stimulus dependence and insights toward perception. *Journal of Clinical Neurophysiology*, *17*, 114–129.

Roberts, T. P. L., Flagg, E. J., & Gage, N. M. (2004). Vowel categorization induces departure of M100 latency from acoustic prediction. *Neuroreport*, *15*, 1679–1682.

Roberts, T. P. L., & Poeppel, D. (1996). Latency of auditory evoked m100 as a function of tone frequency. *Neuroreport*, *7*, 1138–1140.

Rosner, B. S., & Pickering, J. B. (1994). *Vowel perception and production*. Oxford: Oxford University Press.

Sachs, M. B., & Young, E. D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. *Journal of the Acoustical Society of America*, *66*, 470–479.

Sarvas, J. (1987). Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. *Physics in Medicine and Biology*, *32*, 11–22.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*, 100–107.

Slawson, A. W. (1968). Vowel quality and musical timbre as functions of spectrum envelopes and fundamental frequency. *Journal of the Acoustical Society of America*, *43*, 87–101.

Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *Journal of the Acoustical Society of America*, *118*, 3177–3186.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., & Irnio, T. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America*, *117*, 305–318.

Stevens, K. N. (1998). *Acoustic phonetics*. Cambridge, MA: MIT Press.

Stevens, K. N., & Bickley, C. (1991). Constraints among parameters simplify control of klatt formant synthesizer. *Journal of Phonetics*, *19*, 161–174.

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*, 329–353.

Strange, W. (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America*, *85*, 2081–2087.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, *74*, 695–705.

Sussman, H. M. (2000). Phonemic representation: A twenty-first century challenge. *Brain and Language*, *71*, 237–240.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, *79*, 1086–1100.

Tiitinen, H., Mäkelä, A. M., Mäkinen, V., May, P. J., & Alku, P. (2005). Disentangling the effects of phonation and articulation: Hemispheric asymmetries in the auditory N1m response of the human brain. *BMC Neuroscience*, *6*, 62.

Tiitinen, H., Sivonen, P., Alku, P., Virtanen, J., & Näätänen, R. (1999). Electromagnetic recordings reveal latency differences in speech and tone processing in humans. *Cognitive Brain Research*, *8*, 355–363.

Virtanen, J., Ahveninen, J., Ilmoniemi, R. J., Näätänen, R., & Pekkonen, E. (1998). Replicability of MEG and EEG measures of the auditory N1/N1m-response. *Electroencephalography and Clinical Neurophysiology*, *108*, 291–298.

Young, E. D., & Sachs, M. B. (1979). Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *Journal of the Acoustical Society of America*, *66*, 1381–1403.

Zahorian, S. A., & Jagharghi, J. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, *94*, 1966–1982.

Zwicker, E. (1961). Subdivision of the audible frequency rage into critical bands. *The Journal of the Acoustical Society of America*, *33*, 248.