Available online at www.sciencedirect.com
SciVerse ScienceDirect
www.elsevier.com/locate/brainres

Research Report

Attention modulates the use of spectral attributes in vowel discrimination: Behavioral and event-related potential evidence

J. Tuomainen[a,c,*], J. Savela[b], J. Obleser[d], O. Aaltonen[c,e]

[a]Department of Speech, Hearing and Phonetic Sciences, University College London, UK
[b]Department of Information Science, University of Turku, Finland
[c]Centre for Cognitive Neuroscience, University of Turku, Finland
[d]Max Planck Institute of Human Cognitive and Brain Sciences, Leipzig, Germany
[e]Institute of Behavioural Sciences, University of Helsinki, Finland

ARTICLE INFO

Article history:
Accepted 31 October 2012
Available online 19 November 2012

Keywords:
Vowel discrimination
Spectral moments
Formants
Attention
Event-related potentials

ABSTRACT

Speech contains a variety of acoustic cues to auditory and phonetic contrasts that are exploited by the listener in decoding the acoustic signal. In three experiments, we tried to elucidate whether listeners rely on formant peak frequencies or whole spectrum attributes in vowel discrimination. We created two vowel continua in which the acoustic distance in formant frequencies was constant but the continua differed in spectral moments (i.e., the whole spectrum modeled as a probability density function). In Experiment 1, we measured reaction times and response accuracy while listeners performed a go/no-go discrimination task. The results indicated that the performance of the listeners was based on the spectral moments (especially the first and second moments), and not on formant peaks. Behavioral results in Experiment 2 showed that, when the stimuli were presented in noise eliminating differences in spectral moments between the two continua, listeners employed formant peak frequencies. In Experiment 3, using the same listeners and stimuli as in Experiment 1, we measured an automatic brain potential, the mismatch negativity (MMN), when listeners did not attend to the auditory stimuli. Results showed that the MMN reflects sensitivity only to the formant structure of the vowels. We suggest that the auditory cortex automatically and pre-attentively encodes formant peak frequencies, whereas attention can be deployed for processing additional spectral information, such as spectral moments, to enhance vowel discrimination.

1. Introduction

Vowels and consonants can be considered as the main building blocks of spoken words. Their role in speech and language processing is, however, in many respects different (for a review, see Nespor et al., 2003). For example, vowels and consonants follow a distinct developmental course in speech perception with vowels emerging earlier than consonants

*Corresponding author at: University College London, Department of Speech, Hearing and Phonetic Sciences, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom. Fax: +44 20 7679 4238.
E-mail address: j.tuomainen@ucl.ac.uk (J. Tuomainen).

0006-8993/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.brainres.2012.10.067

both in perception (Bertoncini et al., 1988) and production (Kuhl et al., 2008). In contrast, consonants seem to be more important in the acquisition of vocabulary later in infancy (Nazzi and New, 2007). At the other extreme, processing of vowels and consonants can also be dissociated by brain damage in the adult brain (Caramazza et al. 2000). In addition, in the majority of languages vowels form the nucleus of the syllable and accordingly provide the major impetus for speech rhythm in terms of alternation of stressed and unstressed syllables (e.g., Cutler et al., 1997), and finally, speaker normalization seems to occur mostly by calibrating the perceptual mechanism of the listener by using the formant values of vowels produced by the speaker (Ladefoged and Broadbent, 1957).

Vowels have a relatively stable temporal and spectral structure. At least in clear speech, vowels as compared to consonants comprise steady-state parts of high energy (Stevens, 2002) in which the resonance properties of the vocal tract (formants) change relatively little, a property which lends itself suitable for the recognition of the identity of the vowel (Delattre et al., 1952). In contrast, consonants typically involve several acoustic cues that correlate with the identity of the consonants (Pickett, 1999). A well-known example is formant transitions, i.e., rapid changes in the resonance frequencies of the vocal tract that are critical cues to the place of articulation of adjacent consonants. Thus, given their less dynamic nature, vowels seem to provide a relatively stable frame of reference for many operations in speech perception, which has led some researchers to consider vowels as "islands of reliability" (Diehlet al., 1987). Nevertheless, there is still dispute over the distance metrics that determine the position of individual vowels in the perceptual space. In the present study, we asked the participants to discriminate vowels using a simple go/no-go task. We then examined the effects of two acoustic metrics on discrimination performance. We focused (a) on the distance of vowels as defined by the two lowest formant peak frequencies in Euclidean space, and (b) on the distance defined by using spectral moments. If the metric correlates with perception, the prediction is that for vowel pairs that are highly discriminable, the metric must indicate large distances, while those vowel pairs that are hard to discriminate the metric should indicate small distances in the vowel space (e.g., Bladon and Lindblom, 1981).

Another aim of the current study is to investigate whether attention plays a role in the availability of auditory and/or phonetic representations. Recent research suggest that, for example, new non-native speech contrasts may first be learned by selectively attending to auditory attributes that do not underlie native phonetic contrasts (Francis and Nusbaum, 2002). Furthermore, some context effects in speech perception once thought to be specific to phonetic interactions seem to involve abstract auditory representations separate from formant peaks e.g., (Holt, 2005). In accordance with this, neurophysiological evidence indicates that attention affects the functioning of the auditory cortex by reducing frequency-specificity (Petkov et al., 2004) suggesting that instead of detailed frequency information, spectral attributes with reduced spectral detail may be available to the listener through attentional modulation. At present, however, it is unclear what the role of attention is in the availability of different perceptual distance metrics in vowel discrimination. To summarize, the current paper focuses on the utilization of two distance metrics – formant peak frequencies and spectral moments – in vowel discrimination, and reports on three experiments in which the availability of these parameters as a function of attention were studied using both behavioral and electrophysiological methods.

1.1. *The metrics of vowel identification and discrimination–behavioral studies*

All (non-rhotacized) oral vowels can be identified on the basis of the two lowest formant frequencies (Assmann and Summerfield, 1989; Delattre et al., 1952; Klatt, 1982b; Rosner and Pickering, 1994), and the variation in the two lowest formant frequencies explains vowel identification better than other acoustic components of the spectrum (Carlson and Granström, 1979; Pickett, 1957; Pols et al., 1969). Furthermore, formant peaks provide a robust landmark even in adverse listening conditions such as those in the presence of simultaneous speakers and other background noise (for a recent review, see Assmann and Summerfield, 2004). Finally, the results from a series of studies conducted by Klatt (1979, 1982a, 1982b) suggest that formant peak frequencies best predict *phonetic distance* between vowels, and only moderately large changes in relative formant amplitude (i.e., spectral slope) induce a similar change in phonetic perception as changes in formant frequency (see also Aaltonen, 1985; Klatt, 1982a) also noted that amplitude and formant changes do not cancel each other out and that, qualitatively, manipulation of formant frequency tends to induce a change to a different vowel but formant amplitude changes prompt the perception of phonetic change towards nasality. Another important observation is that the type of the listener's *task* seems to be crucial in that other spectral attributes based on more global spectral aspects (such as spectral slope) may well be used in tasks in which *psychophysical* or *general auditory distance* is judged (Carlson and Granström, 1979). Despite the well-established role of formants as acoustic correlates of vowel identity, other perceptual distance metrics have been proposed focusing on the shape of the whole spectrum. In this study, we will consider one of them in some detail, namely, a quantification of the whole spectral shape as a probability density function yielding four spectral moments (mean, standard deviation, skewness and kurtosis).

Bladon and Lindblom (1981) argued that models of vowel perception that concentrate exclusively on the formant peaks may ignore important information about perceptually relevant auditory characteristics of vowels. Second, formants are elusive because they may fuse or interact with anti-formants. Furthermore, high fundamental frequencies produce problems for formant tracking because of sparse distribution of harmonics at the assumed locations of formants (e.g., De Cheveigné and Kawahara, 1999). Third, other spectral attributes than individual formant peaks are employed in vowel perception such as local fusion of (higher) formants (Assmann, 1991; Beddor and Hawkins, 1990; Chistovich, 1985) or whole spectral shape (Ito et al., 2001; Zahorian and Jagharghi, 1993). Bladon and Lindblom, 1981) proposed that

vowel perception is established by assessing the whole spectral energy distribution of the spectrum. In several experiments, they tested predictions derived from their model, and the results from vowel quality judgment tasks, in which participants estimated the similarity of pairs of two- and four-formant vowels, indicated that the performance could be predicted relatively well by a model in which each vowel was represented as a single spectral shape.

A further example of the use of the shape of the whole spectrum in vowel identification is the study by Ito et al. (2001) who set up experiments to investigate the relative importance of formant peaks and whole spectrum attributes of Japanese vowels. The results showed that formant amplitude and the amplitude ratio between the high and low spectral components (even in the presence of formant peak information) are equally important and sometimes even more important for the perception of the vowel identity. Taken together, all these findings suggest that listeners are capable of using other cues than local spectral peaks in vowel identification, and vowel timbre is not completely determined by formant peaks.

In the whole spectrum approach, the spectrum of a speech sound can be conceptualized as a probability density function, and four moments (i.e., center of gravity or mean, standard deviation, skewness and kurtosis) can be computed from the spectrum. These moments have been shown to play a role in the perception of fricatives (Forrest et al., 1988; Jongman et al., 2000; Tomiak, 1990) and laterals (Sawusch and Gagnon, 1995). For example, Sawusch and Gagnon (1995) studied the perception of sine wave speech (SWS) analogs of/la/ and/ra/ stimuli in which the formant tracks were replaced by three time varying sine waves (T1, T2, T3) by training the participants to classify the stimuli into two categories. The important aspect of these experiments was that the participants were under the impression that the stimuli were *non-speech signals*. Sawusch and Gagnon suggested that two distance metrics, one based on sine wave peak differences using formulae from Miller (1989) and another based on spectral moments could account for the results of labeling experiments. To evaluate quantitatively the sufficiency of these two perceptual cues, Sawusch and Gagnon fitted their data against a classification model by Nosofsky (1989) in which the Euclidean distance between the stimulus and the memory representation in a multidimensional space was compared. The results indicated a better fit for spectral moments than peak differences, and also that the first three spectral moments predicted listeners' performance well in the nonspeech tasks. Accordingly, they proposed a serial model of speech perception in which speech and nonspeech categories are both based on a rich abstract auditory code.

Besides being identified, vowels need to be discriminated from each other. The ability to discriminate between vowel tokens is a crucial prerequisite for acquiring the native language vowel repertoire in infancy (e.g., Werker and Tees (1984)) and for second language learners to acquire phonetic categories that do not exist in their native language (e.g., Flege et al. (1999). Some researchers have also suggested that the ability to discriminate native language vowel contrasts correlates with production accuracy (Perkell et al., 2004). Although it is well-established that formant peak frequencies are important for the identification of the identity or category of the vowel, it seems evident that when the task of the participant is changed from identification to discrimination then those properties reflecting category identity may not be sufficient as listeners are able to discriminate within-category tokens (Fry et al., 1962). The difference in task demands can be exemplified by focusing on the phenomenon of categorical perception (Harnad, 1987). The categorical perception effect refers to a relationship between perception of category identity and discrimination sensitivity which in an ideal situation is derived from the identification function (Liberman et al., 1957). Most effectively, stop consonants seem to form a speech category that displays a pattern of classical categorical perception. In contrast, identification and discrimination of vowels has been suggested to show a less straightforward relationship compared to consonants. In some instances vowels are not perceived categorically, and within-category discrimination scores well above chance are obtained (Fry et al., 1962; Fujisaki and Kawashima, 1969; Fujisaki and Kawashima, 1970; Pisoni, 1973). However, given "ideal" experimental conditions, such as short stimulus duration, long inter-stimulus interval in stimulus presentation and a specific discrimination task such as the classical ABX task – all of which tax the *auditory* short-term memory – vowels also show the typical characteristics of categorical perception (Gerrits and Schouten, 2004; Pisoni, 1973; Schouten and van Hessen, 1992).

Thus, even though vowel identification in several instances is based on formant peak picking mechanisms, certain experimental conditions (such as using auditory distance or quality judgment task) may direct the listener to employ other spectral attributes in vowel identification. Furthermore, in vowel discrimination experiments listeners are often asked to make judgments about differences of within-category stimuli which may also prompt the listener to use all available spectral information.

1.2. *Electrophysiological studies of vowel perception*

A complement to traditional behavioral vowel discrimination experiments are measurements of event-related potentials (ERPs) evoked by electric currents in the brain. Recording of ERPs offers a method with millisecond accuracy to investigate perceptual and other cognitive processes as they happen in the brain. Moreover, ongoing activity of the brain can be studied even when the participant is not attending to the stimuli of specific interest. This provides a suitable method to tease out the effects of attention and conscious processing from automatic and pre-attentive processing of the stimuli. The ERP component mostly used to examine the pre-attentive discrimination processes in audition is the mismatch negativity (MMN) (Näätänen et al., 1978). The MMN peaks fronto-centrally in the time window between 100 ms and 300 ms from stimulus onset depending on the stimulus type, presentation rate, and also on whether stimulus difference is defined in terms of simple physical parameters or more abstract representation (Näätänen and Alho, 1997; Näätänen and Winkler, 1999; Näätänen et al., 2005). The MMN is elicited by any discriminable change in some

repetitive aspect of on-going auditory stimulation irrespective of the direction of the participant's attention or task. When simple auditory stimuli are used, the latency and amplitude of the MMN depend directly from the acoustic distance between the standard and the deviant stimuli (Sams et al., 1985; for a detailed review, see, e.g., Näätänen et al., 2001). A common interpretation of the MMN is that it indexes an automatic change–detection process in which a discrepancy is discovered between the sensory-memory representation formed by the regular aspects of the repeating ("standard") auditory events and an infrequent ("deviant") auditory event (Näätänen, 2001; see Jääskeläinen et al. (2004) for an alternative view).

Early reports on the MMN using phonetic stimuli such as vowels or CV syllables found no speech-specific effects which lead to the conclusion that the MMN reflects a comparison process that is affected by the acoustic or other physical parameters such as frequency, intensity, and duration of the stimulus, inter-stimulus interval or the probability of occurrence of the deviant stimulus (e.g., Aaltonen et al., 1987). Nevertheless, recent studies have shown that phonetic and language-specific factors may affect the MMN amplitudes and/or latencies (Aaltonen et al., 1997; Dehaene-Lambertz et al., 2000; Eulitz and Lahiri, 2004; Ikeda et al., 2002; Näätänen et al., 1997; Phillips et al., 2000; Shestakova et al., 2002; Winkler et al., 1999a, 1999b. For example, Näätänen et al. (1997) presented vowels [e ø ɤ o] to Estonian and Finnish participants. Vowel [e] was always the standard, and the other three vowels were used as deviants with increasing acoustic distance in terms of the second formant frequency. The critical deviant vowel was the Estonian /ɤ/ that is prototypical for Estonian participants but not for Finnish participants (the vowel is located close to the category boundary between /ø/ and /o/ in Finnish). The results indicated decreased MMN amplitude for Finnish participants suggesting, according to Näätänen et al. (1997), the availability of less neural resources for processing that particular vowel compared to the Estonian participants. The main argument by Näätänen et al. (1997) rests on the fact that acoustic distance cannot explain the ERP pattern as the familiar Finnish vowel /ø/ which was acoustically closer to the standard vowel /e/, produced a larger MMN for Finnish participants than the unfamiliar Estonian vowel /ɤ/, which was acoustically further away from the standard vowel. The results supported the conclusions that first, two separate auditory and phonetic memory-based comparison processes operate in parallel, and second, that the language environment alters the sensitivity to phonetic contrasts already at the pre-attentive level of processing (see also Kazanina et al. (2006) for a report on language-specific effects on MMN between Russian and Korean speakers, which expands the results by Näätänen et al. (1997) in showing that besides phonetic variation also higher level speech processes related to phonotactic regularities modulate the MMN amplitude).

Similar findings to Näätänen et al. (1997) were obtained by Winkler et al. (1999a) who recorded larger MMN amplitudes when the vowel pair spanned a vowel category boundary (Hungarian /e/ and /ɛ/) compared to a condition in which the vowel pair was selected within a category (Finnish /e/). Another aspect related to the findings by Näätänen et al. (1997) is the prototypicality of the stimulus that seems to affect the MMN amplitudes. Ikeda et al. (2002) provided evidence that if the standard was a non-prototypical category-boundary stimulus, its trace acted as a poor adaptor for discrimination, and therefore the discrimination was more difficult, a result that was also reflected in the amplitude of the MMN response (see, however, Savela et al. (2003) for a different result). Näätänen (2001) refined the idea that pre-attentive discrimination utilizes two kinds of information, phonetic and general auditory suggesting that phonetic information increases the distinctiveness between the categories and inhibits discrimination within a category. This model resembles the dual-code model of speech perception first proposed by Fujisaki and Kawashima (1969; 1970) (see also Schouten and van Hessen (1992) and van Hessen, Schouten, 1992, who provide behavioral evidence that the dual-code mechanism is more suited to consonant perception, and a trace-context theory by Macmillan et al. (1988) models vowel perception more accurately).

Finally, in some instances the *latency* of the MMN may also reflect language-specific processes. Eulitz and Lahiri (2004) found that when a *dorsal* (i.e., back) vowel deviant interrupted a stream of *coronal* (i.e., front) vowel context, a later and less pronounced MMN response was elicited compared to the reverse case of a dorsal vowel stream being interrupted by a coronal vowel. The authors argued for a different hierarchical and status of the phonological place feature *coronal* in the mental lexicon, but most interestingly, the latency of the MMN was as indicative as the amplitude of the MMN response.

Recordings of the brain's magnetic activity seem to support the dual code view (Diesch et al., 1996; Diesch and Luce, 2000; Mäkelä et al. 2003; Obleser et al., 2003; 2004; Shestakova et al., 2004). For example, Diesch and Luce (2000) studied the neuromagnetic fields evoked by single and two-formant vowels with varying F1 values but a constant high F2 value while participants were actively listening to the stimuli. The results showed that both single and two-formant vowels elicited an N100m (a magnetic counterpart of the electric N1 response) for which the sources for higher frequency formants were more anterior than for lower frequency formants. The finding is somewhat surprising because the anterior–posterior axis is orthogonal to the mediolateral tonotopic axis. Diesch and Luce (2000) suggested that the result may reflect the activity outside the primary auditory cortex and the strength of the source could be a function of the variation of formant frequency. Alternatively, the results may be due to differences in the sharpness-of-tuning and inhibitory response-area asymmetry in the isofrequency axis of the AI (see also, Versnel and Shamma, 1998). Both alternatives point to the conclusion that the spectral structure of the vowels may be blurred at this level of processing and suggests that spectral envelop information might be abstracted from the detailed spectral composition, which in turn can be regarded as a prerequisite for perception of invariant phonetic objects.

Evidence for abstract representation of vowels in terms of F1–F2 plane was obtained by Mäkelä et al. (2003) who investigated the cortical correlates of vowel perception by using whole-head magnetoencephalography (MEG) and measuring the N1m response. They suggested that the reason

why only few studies have obtained clear-cut results supporting the idea that vowel space in the auditory cortex is tonotopically organized (Diesch et al., 1996; Diesch and Luce, 2000; Obleser et al., 2003) could be the simultaneous and uncontrolled variation in the F2–F1 differences of the vowels and in their locations in the F1–F2 space. This could blur the neuromagnetic (and in principle also the electrophysiological) responses evoked by the formant structure due to inhibitory neuronal processes reflecting the F2–F1 differences (possibly in the isofrequency axis). Accordingly, their stimuli, [a o u], were selected so that their F2–F1 difference was fixed to a mean of 350 Hz and all the stimuli were back vowels with minimal variation in the F2 frequency. Another variable was the acoustic distance in the Euclidean space, which was 460 Hz for [a] and [u], and 230 Hz for [a] and [o]. The results revealed an N1m response in both hemispheres at 120 ms that was equally strong for all vowels reflecting the equal F2–F1 difference. Furthermore, the differences in the acoustic distance between vowels was reflected in the growing distance of left-hemisphere dipole sources providing direct evidence for orderly (left-hemisphere) representation of vowels in the auditory cortex.

1.3. *Present study*

The present study had three major objectives. First, in Experiment 1, we investigated whether participants relied on formant peak frequencies or whole spectrum representation when they discriminated synthetic Finnish vowels. To this end, we created two vowel continua ([æ–e] and [æ–ϕ]), which had an equal acoustic distance between vowels on both continua as measured by formant peak frequencies in Euclidean space but differed in their spectral moments yielding a larger difference on the [æ–ϕ] continuum.

Second, Experiment 2 was designed to show that listeners have access to formant peak information especially in adverse conditions. The same vowel stimuli as in Experiment 1 were employed, but we also added white noise to the stimuli to eliminate the differences between stimuli in terms of spectral moments, so that listeners could only detect differences between vowels by focusing on the formant structure of the stimuli.

The third goal of the current study relates to the effects of attention on the usage of perceptual metrics (Experiment 3). By recording the MMN component of the ERPs when the participants did not attend to the same vowels as in Experiment 1, we attempted to investigate what kind of perceptual metric (based on formant peaks or spectral moments) is utilized in pre-attentive vowel discrimination as characterized by the MMN activity. The results should further highlight the role of attention in the perception of vowels.

2. Results

2.1. *Experiment 1*

The purpose of Experiment 1 was to investigate whether besides well-established formant peak frequencies listeners might be sensitive to other spectral attributes such as a whole spectrum representation in vowel discrimination. This was done by measuring vowel discrimination performance in a go/no-go task in which participants pressed a response button as fast as possible whenever they heard a deviant vowel stimulus (target) in a train of standard (repeating) vowel. Two vowel continua (Finnish [æ–e] and [æ–ϕ]) were employed which both had the same Euclidean distance between the non-target and three target stimuli, as measured by the distance of the two lowest formant frequencies. The rationale was that if discrimination were based on formant peaks, there should be no difference in reaction times or accuracy between the two continua. In contrast, different distances based on spectral moments (especially, on centre of gravity, CoG) were assigned along the two continua. More specifically, the distance between the stimuli on the [æ–ϕ] continuum was larger than on the [æ–e] continuum. Accordingly, faster and more accurate performance was predicted for the [æ–ϕ] continuum.

2.1.1. *Results and discussion*

The results are presented in Table 1. Statistical analyses were performed separately for reaction times and error rates using similar models and procedures. RTs longer or shorter than three standard deviations were excluded from the analysis comprising only 0.4% of the responses.

A repeated measures ANOVA with participants as random factor and reaction time as the dependent variable was conducted using the vowel continuum type ([æ–ϕ] vs. [æ–e]) and Acoustic distance (short, medium, long) as within-participants factors. The results showed a significant main effect of continuum type ($F(1,13)=21.441$, $p<0.001$, $\eta_p^2=0.623$). In general, the participants responded faster to the [æ–ϕ] continuum ($M=436$ ms, $SEM=14.5$) compared to the [æ–e] continuum ($M=474$ ms, $SEM=15.7$). The main effect of Acoustic distance was also significant ($F(2,26)=55.292$, $p<0.001$, $\eta_p^2=0.810$). This was due to faster reaction times in both the medium distance ($M=432$ ms, $SEM=14.5$) and long distance ($M=428$ ms, $SEM=15.3$) compared to the short distance condition ($M=505$ ms, $SEM=16.0$), both p's<0.001).

Table 1 – Mean reaction times (in ms) and miss rate (%) for the three targets on the /æ–e/ and /æ –ϕ/ vowel continua in Experiment 1. Standard deviations are indicated in parentheses.

	/æ–e/ continuum			/æ–ϕ/ continuum		
	Short distance	Medium distance	Long distance	Short distance	Medium distance	Long distance
RT (ms)	528 (63)	456 (63)	439 (65)	483 (65)	408 (53)	417 (54)
Miss rate (%)	13.2 (9.7)	1.1 (2.1)	0	9.3 (13.0)	0.7 (1.8)	0

The difference between medium and long distance conditions was not significant.

The overall error rate was 2.5% and the false alarm (FA) rate was 0.96%.

The repeated measures ANOVA on the averaged miss rates indicated that the Acoustic distance had a significant effect on the miss rate ($F(2,26)=19.277$, $p<0.001$, $\eta_p^2=0.597$) due to decreasing miss rate as a function of increasing acoustic distance (all comparisons significant at $p<0.019$). The main effect of continuum type ($F(1,13)=2.035$, $p=0.117$, $\eta_p^2=0.177$) or the interaction term ($F<1$) were not significant.

Taken together, these results suggest that discrimination of the two vowel continua is based on the whole spectral attributes such as the spectral moments. One should note that the differences in the spectral moments between the two continua did not affect the identification functions, which were practically identical in terms of overall shape, boundary location and the steepness of the slope. This finding also provides further evidence that vowel identification and discrimination can be accomplished using different spectral attributes.

2.2. Experiment 2

On the basis of the results of the discrimination task in Experiment 1, we conclude that listeners are capable of using differences in the whole spectral shape. A curious and repeated finding in vowel perception research is that listeners are capable of hearing differences between vowels that are drawn from the same category (Fry et al., 1962). This, however, does not by itself prove that discrimination is achieved by using formant peak information and not, for example, other information such as whole spectral attributes. This is a reasonable concern, since in most of the studies on vowel discrimination other spectral cues than formant peaks have usually not been controlled for and this is especially true of studies which have focused on formant frequency difference limens (DL) using stimuli with multiple formants (Hawks, 1994; Kewley-Port and Zheng, 1999; Nord and Svantelius, 1979); in this kind of stimuli, a change in the frequency of one formant alone will inevitably also affect the overall spectrum of the vowel. Furthermore, some of the results indicating asymmetric DLs may well be due to the availability of additional spectral information (Hawks, 1994; Nord and Svantelius, 1979).

One way of showing that formant peak information is sufficient to discriminate the stimuli used in Experiment 1 (and also for detecting within-category differences in front vowels) is to remove information about spectral moments is by adding white noise to the stimuli. This would provide the listener only with information about the formant peaks. Added noise effectively flattens the spectrum, and the remaining major spectral landmarks are the peaks of the two or three lowest formants (Assmann and Summerfield, 2004). The rationale is the following: When (i) noise is added and only formant peaks provide information about the acoustic distance of the targets, and (ii) if the performance pattern is still a function of the acoustic distance (e.g., faster and more accurate responses to more distant targets), and no performance differences between the two vowel continua are found, then the listeners have computed formant peak frequencies and used that information in detecting the target vowels.

2.2.1. Results and discussion

The results are presented in Table 2. Inspection of both the reaction times and accuracy suggest only an effect of target distance but no differences between continua. These observations were confirmed by repeated measures ANOVAs separately for RTs and errors with two within-participant factors, continuum type ([æ –e], [æ–ϕ]) and Acoustic distance between the standard and deviant stimuli (short, medium, long).

Using RTs as the dependent variable only the main effect of Acoustic distance was significant ($F(2,22)=25.995$, $p<0.001$, $\eta_p^2=0.703$). The shortest acoustic distances were perceived significantly more slowly (M=580 ms) than the longer ones (Medium: 516 ms, $p<0.001$; long: 498 ms, $p<0.001$, both Bonferroni corrected for multiple comparisons). The main effect of continuum type or the interaction term were not significant.

The overall error rate was 8.0% and the false alarm rate 1.23%. The miss rate for the deviants was 16.8%. A repeated measures ANOVA showed a statistically significant main effect of Acoustic distance ($F(2,22)=73.322$, $p<0.001$, $\eta_p^2=0.870$), which was due to decreasing miss rates as a function of increasing acoustic distance.

The results clearly indicate that there was no difference in the speed or accuracy between the two continua, strongly suggesting that vowel discrimination in noise employs formant peaks when available, and listeners can, as expected, use formant peak information to detect target vowels.

Taken together, the results from Experiments 1 and 2 show that listeners are capable of directing their attention to the perceptual information that is most distinctive and useful in a particular task. In terms of vowel discrimination, these two sources seem to be formant peaks and whole spectrum attributes such as spectral moments. One important issue concerns the effect of attention on the availability of these

Table 2 – Mean reaction times (in ms) and miss rates (%) for the three targets in /æ–e/ and /æ–ϕ/ vowel continua in Experiment 2, signal-to-noise ratio=0 dB. Standard deviations are indicated in parentheses.

	/æ–e/ continuum			/æ–ϕ/ continuum		
	Short distance	Medium distance	Long distance	Short distance	Medium distance	Long distance
RT (ms)	579 (62)	520 (61)	499 (58)	580 (89)	512 (97)	496 (101)
Miss rate (%)	42.8 (21.9)	6.7 (6.9)	0.6 (1.9)	44.4 (26.1)	5.6 (17.6)	1.1 (3.8)

resources to the listener. The final experiment addresses this issue by recording brain event-related potentials (ERPs) focusing on the mismatch negativity, MMN) response, which is an index of automatic discrimination of auditory information and can be recorded reliably even if the participants do not attend to the auditory stimuli (Näätänen, 2001). This approach may give important insight on what kind of information is extracted automatically from the acoustic signal and what other resources may be available for the listener only when attention is directed to the stimuli to exploit all potentially useful information for a specific task.

2.3. *Experiment 3*

The ERP recordings were used in order to compare the attentive and pre-attentive discrimination with the same stimuli as in Experiment 1. Straightforwardly, two alternative hypotheses can be put forth. First, if the automatic and pre-attentive discrimination process is sensitive only to formant peak frequencies, then we should not find any differences in the MMN pattern between vowel continua. This is based on the earlier literature which suggests that formant peaks are automatically extracted in the auditory pathway and represented in the primary auditory cortex as other spectral properties of sound (Mäkelä et al., 2003; Obleser et al., 2003; Ohl and Scheich, 1997; Shestakova et al., 2004). Accordingly, we should find a dissociation between the MMN and behavioral results in Experiment 1, in which faster and more accurate responses were recorded on the [æ–ϕ] continuum. The second hypothesis is that, if the MMN is also sensitive to other kind of spectral information such as the whole spectral shape, then the deviant stimuli on the [æ–ϕ] continuum should elicit larger amplitudes and/or shorter latencies than those on the [æ –e] continuum.

2.3.1. *Results and discussion*

The ERP waveforms to the standard and the deviant stimuli, the difference wave (obtained by subtracting the responses to the standard stimuli from those to the deviant stimuli), and scalp topographies of difference waves (150–210 ms time-frame) are presented in Fig. 1. The waveforms show typical ERP responses to simple speech stimuli, which are presented at fairly short ISIs (which usually results in a negative going "tail" due to the fact that the response has not yet returned to

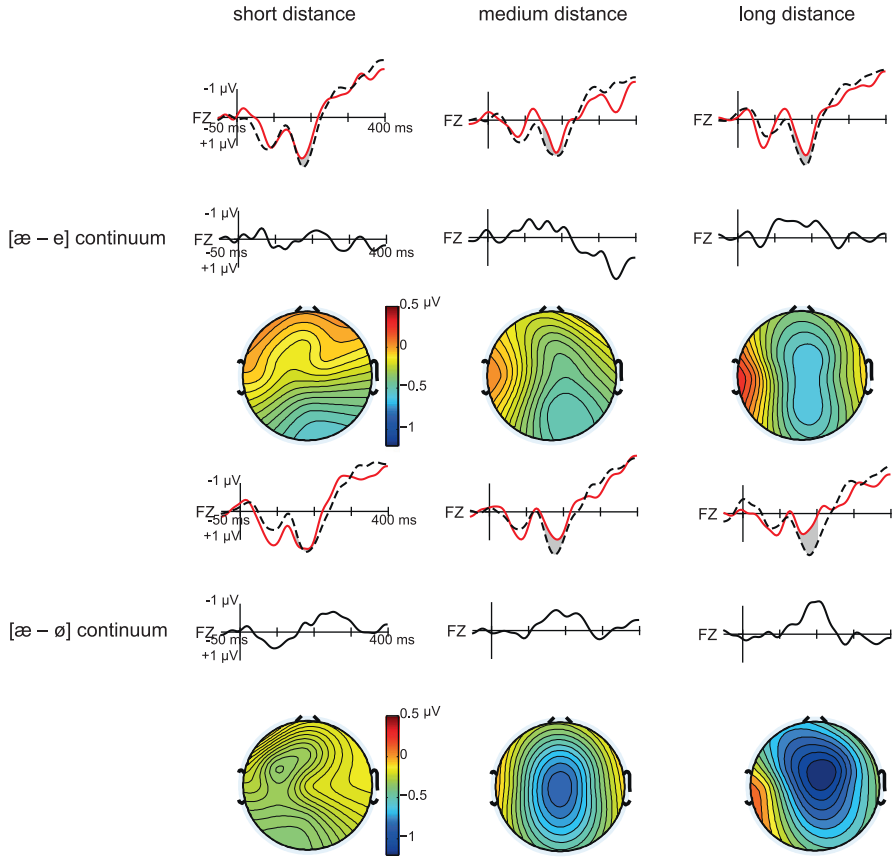

Fig. 1 – The grand average ERP waveforms for the three different stimulus conditions for the /æ–e/ continuum (upper panel of the figure) and /æ–ϕ/ continuum (lower half of the figure). On both panels, the first row depicts the waveforms for the standard (thin black line) and deviant stimulus (thick red line). The gray shading denotes the time window 150–210 ms which was used to quantify the MMN response. The second row depicts the MMN response as a difference wave, which was obtained by subtracting the ERPs to the standard stimuli from the ERPs to the deviant stimuli separately for each condition. The third row depicts the scalp topography of the MMN response (difference wave) quantified as the mean amplitude between 150 ms and 210 ms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3 – The peak MMN amplitudes (μV) and latencies (ms) in Experiment 3. All amplitudes differed significantly from 0 μV baseline (p's <0.001). Standard deviations are indicated in the parentheses.

	/æ–e/ continuum			/æ–ϕ/ continuum		
	Short distance	Medium distance	Long distance	Short distance	Medium distance	Long distance
Amplitude (μV)	−0.928 (0.87)	−1.218 (0.79)	−1.584 (1.03)	−1.363 (0.68)	−1.521 (1.26)	−1.797 (1.33)
Latency (ms)	220 (37)	202 (49)	192 (35)	233 (26)	199 (43)	206 (37)

the baseline). Inspection of the scalp maps indicate a central or fronto-central voltage distribution which is typical to MMN, especially in the conditions in which the acoustic distance is larger.

The individual MMN peak amplitudes and their latencies for each condition measured at Fz are presented in Table 3. A one-sample t-test showed that all amplitudes differed significantly from zero (in all condition $p<0.01$, two-tailed, Bonferroni corrected for multiple comparisons). The individual MMN peak amplitudes and their latencies for each condition were subjected to separate 2×3 repeated measures ANOVAs with Continuum type ([æ–e], [æ–ϕ]) and Acoustic distance between the deviant and standard stimuli (short, medium, long) as within-participant factors. The latency of the peak MMN amplitude was significantly affected by the acoustic distance between the standard and deviant stimuli ($F(2,26)=5.598$, $p=0.010$, $\eta_p^2=0.301$). The deviant with the shortest acoustic distance from the standard stimulus had a significantly longer mean latency ($M=227$ ms, SEM=5.2) compared to the other deviants (medium: $M=200$ ms, SEM=8.9, $p=0.023$; long: $M=199$ ms, SEM=8.4, $p=0.009$, Bonferroni corrected for multiple comparisons). The main effect of continuum type ($F(1,13)=2.075$, $p=0.173$) or the interaction term ($F<1$) showed no statistically significant effects.

Neither acoustic distance ($F(2,26)=1.975$, $p=0.159$) nor continuum type ($F(1,13)=1.884$, $p=0.193$) had any statistically significant effects on MMN amplitude, and the interaction term was also not significant ($F<1$).

The purpose of the experiment was to investigate how withdrawing attention from the auditory stimuli affects the processing of vowel stimuli. Specifically, the issue was whether formant peak frequencies or a whole spectrum attributes were extracted automatically in vowel discrimination. We hypothesized that if the auditory cortex automatically represents formant frequencies there would be no differences in MMN responses to vowel targets between vowel continua as the Euclidean distances were similar in terms of formants. However, if spectral moments (such as the mean of the spectrum, center of gravity) are computed automatically by the auditory cortex even if attention was focused on the visual modality then one should see larger amplitudes and shorter latencies on the [æ–ϕ] continuum because of the larger distance in whole spectrum attributes between the standard and the targets.

The results indicated no differences between vowel continua either in the MMN amplitude or latency. The main finding was that the MMN latencies only reflected the Euclidean acoustic distances between the standard and deviant similarly on both continua. Thus, the results suggest that the MMN response to vowel stimuli is best interpreted as an index of formant extraction possibly in Euclidean space (Diesch and Luce, 2000; Mäkelä et al., 2003). The reason why spectral moments were not automatically available may be directly related to the effects of attention on the functioning of the sensory cortices, and will be discussed in more detail in the next section.

3. General discussion

The goal of the present study was twofold: To investigate (i) what kind of stimulus information (formants or spectral moments) is used by the listener in vowel discrimination, and (ii) what the effect of attention (pre-attentive and attentive discrimination) is on the accessibility of these representations. The results showed, first, that when the stimuli were presented in good listening conditions, the participants were able to discriminate the vowels by focusing their attention on the global attributes of the vowel spectra rather than the local energy maxima of the two lowest vowel formants (Experiment 1). Second, when the stimulus quality was reduced by adding background noise, the primacy of the whole spectrum attributes disappeared and the participants employed formant peaks in the discrimination tasks (Experiment 2). Finally, in Experiment 3, records of the brain's event-related potentials (ERPs) showed that when the participants did not attend to the stimuli and both whole spectrum attributes and formant frequency information were accessible, the mismatch negativity (MMN) component of the ERPs was only sensitive to formant peak frequencies. Taken together, the results indicate that the participants are able to use the available auditory information differently in vowel discrimination depending on the task, the quality of the stimuli, and focus of attention. When the participants are asked to attend to the stimuli and the stimuli are of good quality, they use all the information available in the entire auditory spectrum. Furthermore, when the stimuli are of poor quality, then the participants rely on information conveyed by formants only. Similarly, when the participants are asked to ignore the stimuli, the brain automatically extracts information about the formant peaks, and seems not to be sensitive to other spectral factors suggesting that the memory traces for vowels could be presented on a two-dimensional plane based on the two lowest formants.

In general, our findings support the long prevailing view of the primacy of formants in vowel perception, which has most clearly been established for vowel identification (e.g., Assman

and Summerfield, 1989; Delattre et al., 1952; Klatt, 1982b; Rosner and Pickering, 1994). Furthermore, there is a large bulk of research which supports the view that formants are also important in vowel discrimination (reviewed, e.g., by Rosner and Pickering (1994)). Our results extend these findings in that they suggest that formants are also the primary objects of vowel perception when the task requires discrimination of vowel irrespective of whether attention is directed or not to the auditory stimuli. These results are well supported by behavioral and neurophysiological evidence, which all point to the fact that the human auditory system is adapted to perceiving energy concentrations produced by the vocal tract presumably for communicative purposes. Our results also indicate that the human auditory system provides different spectral representations that (together with temporal cues, e.g., Rosen, 1992; Shannon et al., 1995; Zeng et al., 2005) underlie the ability to recognition speech.

It should be noted that, for example, if a discrimination task is easy, such as when the acoustic distance of the targets from non-targets is large, the properties important in categorization prevail. In Experiment 1, target detection may have been based solely on formant peaks when the acoustic distance of the target from the standard stimulus was large, which was reflected by the leveling off of the response speed so that there were no significant differences in the continua between the most distant targets even though the whole spectrum distance indicated an advantage for [æ-ϕ] continuum. The participants, however, were under the time pressure to respond as fast as possible which may have led them to actively search for all available information that might help in target detection.

The primacy of formants in vowel perception may stem of the nature of human speech perception that needs robust landmarks in feedback processes of speech perception. The facts supporting the idea that formant peaks are primary properties used in vowel perception are, first, that formant peaks are by far the most informative parameter across different listening conditions, such that, for example, formants are resistant to noise (Assmann and Summerfield, 2004). Second, formants are very robustly presented throughout the auditory pathway suggesting that at least the non-human auditory system seems to be fully adapted to processing formant peak information (Delgutte, 1984; Delgutte and Kiang, 1984a; Eggermont, 2001; Sachs and Young, 1979; Shamma, 2001), even in background noise (Delgutte and Kiang, 1984b). Furthermore, vowel identification can easily be carried out based on formant frequencies without any other auditory cues (Delattre et al., 1952; Hillenbrand et al., 2006). Finally, Kiefte and Kluender (2005) provide evidence that whole spectrum attributes seem to work well only in the perception of monophthong vowels but not with diphthongs which provide dynamic information. Thus, those instances suitable for extracting whole spectrum information from continuous speech may not be plentiful. One should bear in mind, however, that current results do not favor a model of vowel perception in which a formant peak picking mechanism would be mainly responsible for extracting relevant spectral information. Instead, as already suggested by Klatt (1982a); (see also Hillenbrand et al., 2006), a more plausible model would be a spectral shape pattern matching process, which is more sensitive to local energy concentrations but pays less attention to relative intensities of formant peaks or to spectral valleys between formants. Third, distributional variance is an information carrying parameter (although distinct from informativeness of the cue) so that more variance implies higher importance of the parameter (Holt and Lotto, 2006). This again is reflected in the proficiency of the auditory system to carry out processing that is sensitive to stimulus dimensions that are varying (for a non-speech example, see Lutfi, 1993). Even though research is scarce on how much variance is present in whole spectrum attributes, it may be safe to assume that the variance represented by formant peak frequencies across difference speakers may well outweigh that of whole spectrum attributes. All these three components point to the supremacy of formant peaks in vowel perception.

Finally, one should consider the effect of the task on the availability of different representations, which we feel is the most important factor to explain current results. As pointed out in the introduction, it is well known that discrimination of vowel contrast exceeds that of identification. The category boundary effect is a good example. To put it simply, task specification predicts the performance of the listener in a discrimination task (Gerrits and Schouten, 2004; Schouten et al., 2003; Schouten and van Hessen, 1992). Furthermore, of particular interest is the study by Guenther et al. (1999) who with different stimulus sets of non-speech stimuli showed that participants use different types of stimulus information in identification and discrimination tasks. In discrimination, the listeners seemed to utilize richer stimulus representations, whereas identification appeared to be based on less sophisticated representations. In a follow-up study, Guenther et al. (2004) showed that the brain activation patterns reflected task differences such that discrimination training was followed by an increase in the overall area of brain activation during a discrimination task. In contrast, categorization training decreased the activated brain areas specifically in the left supra-temporal gyrus. Thus, the brain would shift neural resources away from those areas where discrimination of small differences between stimuli is not behaviorally feasible (category center) to the areas where accurate discrimination is more important (category boundary). The present results, however, suggest another interpretation; in a discrimination task, listeners try to maximize task performance by utilizing all available information sources such as formant frequencies and whole spectral attributes, which recruit larger brain areas than a task requiring extraction of information that is relevant for category identity. This kind of task could be accomplished using a simple pattern recognition mechanism (Nearey, 1997).

To conclude, attentive discrimination versus pre-attentive discrimination requires different perceptual strategies. The MMN reflects the discrimination of spectral elements that are represented in a memory trace used in pre-attentive discrimination whereas attentive discrimination shows the use of richer vowel representations. The problems determining the information used in vowel discrimination and identification may reflect the differences in how attention is focused on stimulus properties depending on the demands of the current task of the listener. The formants and spectral moments are used in parallel in attentive discrimination depending on the experimental task and stimulus quality whereas the pre-

attentive discrimination seems to be based on formant peak frequencies. It is obvious that in most natural conditions speech is heard in different levels of background noise, and formants are the prime candidates to survive in these conditions. However, the current results emphasize the dynamic nature of human auditory perception in that participants can exploit different types of information as a function of the task demands when they are attending to the stimuli.

4. Experimental Procedures

4.1. *Experiment 1*

4.1.1. *Stimulus selection*

In order to select the stimuli for the behavioral discrimination tasks and for the MMN recordings, 10 native speakers of Finnish were asked to identify the stimuli on two vowel continua. All participants were informed about the nature of the experiment and they all provided a verbal consent. Both continua consisted of 11 vowel stimuli synthesized by HLSyn Klatt synthesizer (Sensimetrics, Inc.) using a sampling frequency of 11025 Hz. On the [æ–e] continuum, the frequency of the first formant (F1) varied from 655 Hz to 484 Hz, and the second-formant frequency (F2) from 1756 Hz to 1933 Hz. On the [æ–ϕ] continuum the frequency of the F1 was varied similarly to the [æ–e] continuum from 655 Hz to 484 Hz but the F2 decreased from 1756 Hz to 1592 Hz, assigning a critical distinctive role to F2. The step size was set to 15 mels (Stevens and Volkman, 1940) in the Euclidean vowel space. The Hertz values were converted to a Mel scale using the formula $m=1127\ \ln(1+f/700)$. The higher formant frequencies were kept constant (F3=2474 Hz, F4=3500 Hz, and F5=4490 Hz). The fundamental frequency (f0) rose from 100 Hz to 120 Hz until 125 ms and declined during the rest of the stimulus to 80 Hz. The duration of the stimuli was 385 ms, and a linear ramp of 50 ms was used to smooth the onset and offsets of the stimuli. The four spectral moments were computed by Praat software (Boersma and Weenink, 2001). The spectrum was treated as a probability density function in which the center of gravity is the first moment (the mean frequency of the sound in the whole frequency domain in which the frequencies were weighted by their amplitudes in power spectrum), the second moment is the standard deviation, the third moment is skewness of the spectrum and the fourth moment is the kurtosis of the spectrum. A larger change in the spectral moments in the [æ–ϕ] continuum compared to [æ–ϕ] continuum is expected as F1 and F2 frequencies move parallel to each other on the [æ–ϕ] continuum (both frequencies get lower) whereas on the [æ–e] continuum the formant frequencies move in the opposite directions (the frequency of F1 lowers and that of F2 rises) which flattens the overall spectrum, and to some extent cancels out the movement in the center of gravity.

Identification performance was tested separately for both continua. Each stimulus was presented in a random order 10 times and the participants' task was to indicate whether they heard [æ] or [ϕ] (for the [æ–ϕ] continuum), or [æ] or [e] (for the [æ–e] continuum) by pressing one of the two buttons indicated by Finnish orthographic symbols <ä> and <ö> or <ä> and <e>, respectively. If they were not sure, they were encouraged to guess. The presentation of the stimuli was self-paced. The results were analyzed by fitting a probit function (Finney, 1971) to each participant's identification data on both continua. The 50% point of the fitted labeling curve indexing the location of the category boundary and the slope (or gradient) of the probit function characterizing the consistency of the identification performance were extracted. The results showed that the category boundary on the [æ–ϕ] continuum was located at 6.4 (s.d. 0.55) on the stimulus axis and the identification slope was–2.254 (s.d. 0.76), and the corresponding category boundary on the [æ–e] continuum was located at 6.6 (s.d. 0.65) and the identification slope was—3.04 (s.d. 0.70). An alpha level of 0.05 was used for all statistical testing. Paired samples *t*-tests showed no significant differences between continua for either of the parameters (boundary, $t(9)=-0.562$, $p=0.587$; slope, $t(9)=-1{,}925$, $p=0.0864$).

For both the behavioral discrimination and ERP experiments the following stimuli were selected on the basis of the results of the identification experiment. The reference stimulus (or the 'Standard') was always the second stimulus of the respective continua. Three target stimuli (or 'Deviants') were selected so that the first deviant (stimulus#4) was a within-category stimulus, the second deviant was located at the category boundary (stimulus#6) and the third deviant was a member of the adjacent category (stimulus#8; see Fig. 2). Since the category boundary on both continua was identical, so-called phoneme-boundary effects (such as a decrease in the number of errors or faster reaction times) were expected to be similar on both continua and thus could not be used as a potential alternative explanation for possible differences in the discrimination performance between continua.

4.1.2. *Discrimination experiment*

Participants. Fifteen native speakers of Finnish who were students at the University of Turku (mean age 23.6 years, range 20–25 years, all females) participated in the

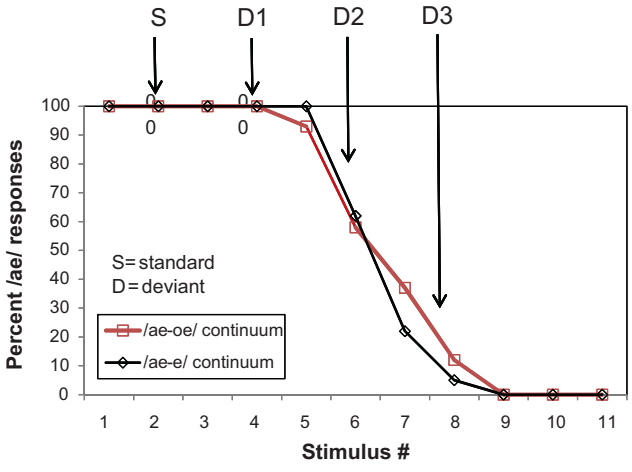

Fig. 2 – The identification functions for the /æ–e/ (filled squares) and /æ–ϕ/ continuum (open diamonds). The stimuli chosen for the discrimination tasks are indicated by arrows (the standard stimulus, S; the three target (or deviant) stimuli: D1=within-category stimulus, D2=category boundary stimulus, D3=between-category stimulus).

experiment. None reported hearing abnormalities. Two of them were left-handed. All participants were informed about the nature of the experiment and they all provided a verbal consent.

Stimuli and procedure. The stimuli were chosen on the basis of the identification test administered to a separate group of Finnish participants (see above Section 4.1.1). The measured differences of formant peaks between the standard and the three deviants were similar in the Euclidean space on both continua but different on the basis of all spectral moments (see Table 4, middle panel). Inspection of the table indicates that in all spectral moments, especially in the center of gravity, standard deviation and kurtosis, the differences are larger in the [æ–ϕ] continuum, suggesting that the mean of the spectrum varies more and the spectrum is more peaked than that of the [æ–e] continuum.

The four stimuli of both continua were played back to the participants in a pseudorandom order in separate blocks consisting of 275 ($p=0.82$) standard stimuli and 20 ($p=0.06$) of each of the three types of deviants. The inter-stimulus interval (ISI) was 400 ms. Order of the blocks was counterbalanced across the participants. Participants were asked to press a button as fast and accurately as possible when they heard a vowel deviating from the stream of the standard stimuli (yielding the task equivalent to a fixed standard go/no-go task). A practice block of 15 stimuli was administered to the participants to familiarize them with the task. The experimental session took about 20 min.

4.2. *Experiment 2*

Participants. Twelve students of the University of Turku (mean age 24.3 years, range 20–40 years, 6 females, one left-handed) participated in the experiment. They were all native speakers of Finnish with no reported hearing abnormalities. All participants were informed about the nature of the experiment and they all provided a verbal consent.

Stimuli and procedure. The stimuli used in Experiment 1 were embedded in white noise with same RMS amplitude as the stimuli (+0 Signal-to-Noise ratio) (see Fig. 3 for samples of the spectra of the stimuli presented in quiet or embedded in noise, and Table 4, bottom part, for the values of the formant peaks and spectral moments). The measured center of gravity differences for the [æ–e] continuum were -1 Hz, 47 Hz and 73 Hz, and for the [æ–ϕ] continuum 31 Hz, 106 Hz, and 157 Hz yielding all targets (especially those close to the standard stimuli) very difficult to discriminate on the basis of spectral moments (note that DLs for different spectral moments are not available, and those for formant frequencies are not applicable here). Furthermore, as expected the white noise evens out the amplitude differences yielding a spectrum that is flat with practically identical dispersion and shape of the spectrum in the two continua; however, the first two formant peaks are readily available for discrimination. Two blocks of stimuli were presented in the same oddball paradigm as in Experiment 1, each consisting of 206 ($p=0.82$) presentations of the standard stimulus and 15 ($p=0.06$) presentations of each of the three deviants. The inter-stimulus interval (ISI) was 400 ms. Participants were asked to push a button as fast and accurately as possible when they heard a sound deviating from the stream of the standard stimuli. The session took about 20 min.

4.3. *Experiment 3*

Participants. The same Finnish participants were used as in Experiment 1.

Stimuli and procedure. The stimuli of Experiment 1 were used for the MMN recordings to allow for a direct comparison of

Table 4 – The measured frequencies of the first (F1) and the second (F2) formant and the four spectral moments of the stimuli (1st=center of gravity (CoG), 2nd=standard deviation (SD), 3rd=skewness, 4th=kurtosis) on /æ–e/ and /æ–ϕ/ continua used in the discrimination tasks in Experiments 1 and 3 are presented on the upper part of the table. Standard denotes the non-target stimulus and deviants 1, 2, and 3 denote the target stimuli with increasing acoustic distance from the standard. The middle part of the table shows the distance in Euclidean space between the standards and deviants for formant peaks and the difference of the moments for the same stimuli. The bottom part shows the differences for the four spectral moments in the noise condition. Formant frequencies and the Euclidean distances in formant space are in mels. All other acoustic measurements are in Hz.

	/æ–e/ continuum				/æ–ϕ/ continuum			
	Standard	Deviant 1	Deviant 2	Deviant 3	Standard	Deviant 1	Deviant 2	Deviant 3
F1	735	703	677	642	731	703	673	646
F2	1778	1808	1842	1886	1744	1715	1671	1646
CoG	798	753	707	673	773	667	613	551
SD	616	617	613	624	587	531	474	435
Skewness	1.74	1.86	2.02	2.07	1.86	2.22	2.62	2.95
Kurtosis	3.11	3.40	3.89	3.89	3.85	6.00	8.86	11.64
Formant peaks	–	32	64	97	–	35	67	104
CoG	–	45	91	125	–	106	160	222
SD	–	−1	3	−8	–	56	113	152
Skewness	–	−0.12	−0.28	-0.33	–	−0.36	−0.76	−1.09
Kurtosis	–	−0.29	−0.78	-0.78	–	−2.15	−5.01	−7.79
CoG	–	−1	47	73	–	31	106	157
SD	–	−16	−20	−31	–	−22	−3	−3
Skewness	–	0	0	−0.01	–	0	−0.11	−0.11
Kurtosis	–	−0.06	0.01	0.01	–	0	0.013	0.013

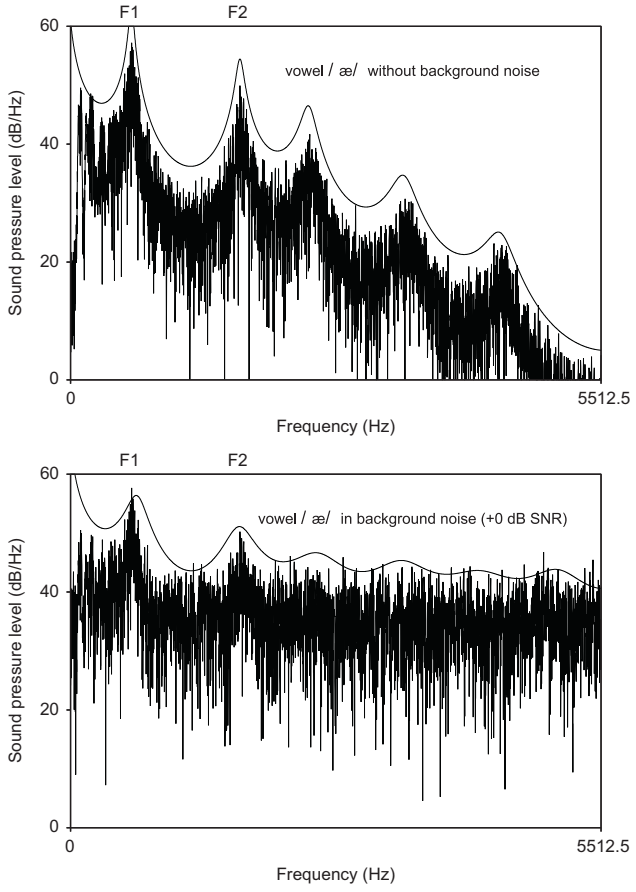

Fig. 3 – Spectra of the vowel /æ/ in quiet (upper panel) and in the background of white noise (+0 dB signal-to-noise ratio). The addition of white noise removes the spectral tilt drastically reducing the differences in the spectral moments. All formants except the first (F1) and the second (F2) are also effectively masked.

behavioral and the ERP results. All participants performed the behavioral task first followed by the ERP recording. In order to increase the signal-to-noise ratio of the ERP recordings, there were 120 of each of the deviant stimuli and 1440 standard stimuli in each of the blocks thus yielding the probability of the occurrence of the stimuli the same as in the behavioral discrimination test (i.e., $p=0.82$ for the standard stimulus, stimulus #2 from the continua, cf. Fig. 1 and $p=0.06$ for each of the deviant stimuli). Stimuli were presented in a pseudorandom order in the oddball paradigm (inter-stimulus interval 400 ms) by using the Neurostim program (Neuroscan Inc.). Participants watched a silent movie during the session and were instructed to ignore the vowel sounds, which were binaurally presented through earphones at a comfortable sound-pressure level (about 70 dB SPL). The recording session took about 1 hour.

Continuous EEG was recorded using a Braintronics 32-channel EEG amplifier connected to the NeuroScan EEG data acquisition and the stimulus presentation computer. The signal was sampled at 200 Hz, amplified (band pass 0.5–70 Hz with a 50 Hz notch filter), and stored on a hard disk of a personal computer. Ag/AgCl cup electrodes were used. Three electrodes (Fz, Cz, and Pz) were placed on the standard locations according to the 10–20-system. Six lateral electrodes (3 on the left and 3 on the right) were positioned at non-standard locations equidistantly placed on the coronal line connecting the mastoids through Fz; these locations were labeled as L1, R1, L2, R2, Lm (left mastoid), and Rm (right mastoid). Eye movements were recorded with two electrodes, one placed at the outer canthus of the right eye, and the other at Fpz. The reference electrode was placed on the tip of the nose. Electrode impedance was kept below 5 kΩ at all electrode sites. The EEG was analyzed offline by first extracting 450 ms epochs for ERP averaging, (including a 50 ms pre-stimulus baseline interval), after which the ERP epochs were digitally filtered by a 1.6–30 Hz band pass FIR filter. Epochs containing artifacts exceeding +/−70 μV caused by eye movements or other extracerebral sources were removed. The ERP responses to the standard stimulus preceding each deviant were subtracted from that to the deviant for each stimulus block. The MMN peak amplitudes and latencies were measured from the difference waveforms in the time window between 150 ms and 210 ms at Fz separately for each participant. This decision was based on visual inspection of the grand-average waveforms, which also suggested no effects of laterality or anterior–posterior effects between conditions.

Acknowledgments

We thank Emma Brint and Outi Tuomainen for constructive comments on an earlier version of the manuscript.

REFERENCES

Aaltonen, O., 1985. Effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. J. Phon. 13, 1–9.

Aaltonen, O., Eerola, O., Hellstrom, A., Uusipaikka, E., Lang, A.H., 1997. Perceptual magnet effect in the light of behavioral and psychophysiological data. J. Acoust. Soc. Am. 101, 1090–1105.

Aaltonen, O., Niemi, P., Nyrke, T., Tuhkanen, M., 1987. Event-related brain potentials and the perception of a phonetic continuum. Biol. Psychol. 24, 197–207.

Assmann, P.F., 1991. Perception of back vowels: centre of gravity hypothesis. Q. J. Exp. Psychol. 43A (3), 423–448.

Assmann, P.F., Summerfield, A.Q., 2004. Perception of speech in adverse conditions. In: Greenberg, S., Ainsworth, W.A., Popper, A., Fay, R. (Eds.), Speech Processing in the Auditory System. Springer Verlag, New York, pp. 231–308.

Assmann, P.F., Summerfield, A.Q., 1989. Modeling the perception of concurrent vowels: vowels with the same fundamental frequency. J. Acoust. Soc. Am. 85, 327–338.

Beddor, P.S., Hawkins, S., 1990. Influence of spectral prominence on perceived vowel quality. J. Acoust. Soc. Am. 87, 2684–2704.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P., Kennedy, L., Mehler, J., 1988. An investigation of young infants' perceptual representations of speech sounds. J. Exp. Psychol.: Gen. 117 (1), 21–33.

Bladon, R.A.W., Lindblom, B., 1981. Modeling the judgment of vowel quality differences. J. Acoust. Soc. Am. 69, 1414–1422.

Boersma, P. & Weenink, D. (2001). Praat: doing phonetics by computer (Version 4.0) [Computer software].

Caramazza, A., Chialant, D., Capasso, R., Miceli, G., 2000. Separate processing of consonants and vowels. Nature 403, 428–430.
Carlson, R., Granström, B., 1979. Model predictions of vowel dissimilarity. TMH-QPSR 3–4, 84–104.
Chistovich, L.A., 1985. Central auditory processing of peripheral vowel spectra. J. Acoust. Soc. Am. 77, 789–805.
Cutler, A., Dahan, D., van Donselaar., 1997. Prosody in the comprehension of spoken language: a literature review. Language and Speech 40, 141–201.
de Cheveigné, A., Kawahara, H., 1999. Missing-data model of vowel identification. J. Acoust. Soc. Am. 105, 3497–3508.
Dehaene-Lambertz, G., Dupoux, E., Gout, A., 2000. Electrophysiological correlates of phonological processing: a cross-linguistic study. J. Cogn. Neurosci. 12, 635–647.
Delattre, P.C., Liberman, A.M., Cooper, F.S., Gerstman, L.J., 1952. 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. (1984). Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds. J. Acoust. Soc. Am., 75, 879-886.
Delgutte, B., Kiang, N.Y.S., 1984a. Speech coding in the auditory nerve: I. Vowel-like sounds. J. Acoust. Soc. Am. 75, 866–878.
Delgutte, B., Kiang, N.Y.S., 1984b. Speech coding in the auditory nerve: V. Vowels in background noise. J. Acoust. Soc. Am. 75, 908–918.
Diehl, R.L., Kluender, K.R., Foss, D.J., Parker, E.M., Gernsbacher, M.A., 1987. Vowels as islands of reliability. J. Mem. Lang. 26, 564–573.
Diesch, E., Eulitz, C., Hampson, S., Ross, B., 1996. The neurotopography of vowels as mirrored by evoked magnetic field measurements. Brain Lang. 53, 143–168.
Diesch, E., Luce, T., 2000. Topographic and temporal indices of vowel spectral envelope extraction in the human auditory cortex. J. Cogn. Neurosci. 12, 878–893.
Eggermont, J.J., 2001. Between sound and perception: reviewing the search for a neural code. Hear. Res. 157, 1–42.
Eulitz, C., Lahiri, A., 2004. Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. J. Cogn. Neurosci. 16, 577–583.
Finney, D.J., 1971. Probit analysis. Cambridge University Press, Cambridge.
Flege, J.E., MacKay, I.R., Meador, D., 1999. Native Italian speakers' perception and perception of English vowels. J. Acoust. Soc. Am. 106, 2973–2987.
Forrest, K., Weismer, G., Milenkovic, P., Dougall, R.N., 1988. Statistical analysis of word-initial voiceless obstruents: preliminary data. J. Acoust. Soc. Am. 84, 115–123.
Francis, A.L., Nusbaum, H.C., 2002. Selective attention and the acquisition of new phonetic categories. J. Exp. Psychol.: Hum. Percept. Perform. 28, 349–366.
Fry, D.B., Abramson, A.S., Eimas, P.D., Liberman, A.M., 1962. Identification and discrimination of synthetic vowels. Lang. Speech 5, 171–189.
Fujisaki, H. & Kawashima, T. (1969). On the Modes and Mechanisms of Speech Perception. Annual Report of Engineering Research Institute, Faculty of Engineering, University of Tokyo, 28, 67–73.
Fujisaki, H. & Kawashima, T. (1970). Some Experiments on Speech Perception and a Model for Perceptual Mechanism. Annual Report of Engineering Research Institute, Faculty of Engineering, University of Tokyo, 67–73.
Gerrits, E., Schouten, M.E.H., 2004. Categorical perception depends on the discrimination task. Percept. Psychophys. 66, 363–376.
Guenther, F.H., Husain, F.T., Cohen, M.A., Shinn-Cunningham, B.G., 1999. Effects of categorization and discrimination training on auditory perceptual space. J. Acoust. Soc. Am. 106, 2900–2912.
Guenther, F.H., Nieto-Castanon, A., Ghosh, S.S., Tourville, J.A., 2004. Representation of sound categories in auditory cortical maps. J. Speech Lang. Hear. Res. 47, 46–57.
Harnad, S., 1987. Categorical perception. Cambridge University Press, Cambridge.
Hawks, J.W., 1994. Difference limens for formant patterns of vowel sounds. J. Acoust. Soc. Am. 95, 1074–1084.
Hillenbrand, J.M., Houde, R.A., Gayvert, R.T., 2006. Speech perception based on spectral peaks versus spectral shape. J. Acoust. Soc. Am. 119, 4041–4054.
Holt, L.L., 2005. Temporally nonadjacent nonlinguistic sounds affect speech categorization. Psychol. Sci. 16, 305–312.
Holt, L.L., Lotto, A.J., 2006. Cue weighting in auditory categorization: implications for first and second language acquisition. J. Acoust. Soc. Am. 119, 3059–3071.
Ikeda, K., Hayashi, A., Hashimoto, S., Otomo, K., Kanno, A., 2002. Asymmetrical mismatch negativity in humans as determined by phonetic but not physical difference. Neurosci. Lett. 321, 133–136.
Ito, M., Tsuchida, J., Yano, M., 2001. On the effectiveness of whole spectral shape for vowel perception. J. Acoust. Soc. Am. 110, 1141–1149.
Jongman, A., Wayland, R., Wong, S., 2000. Acoustic characteristics of English fricatives. J. Acoust. Soc. Am. 108, 1252–1263.
Jääskeläinen, I., Ahveninen, J., Bonmassar, G., Dale, A., Ilmoniemi, R., Levänen, Lin, F.H., et al., 2004. Human posterior auditory cortex gates novel sounds to consciousness. PNA S 101 (17), 6809–6814.
Kazanina, N., Phillips, C., Idsardi, W., 2006. The influence of meaning on the perception of speech sounds. PNAS 103 (30), 11381–11386.
Kewley-Port, D., Zheng, Y., 1999. Vowel formant discrimination: towards more ordinary listening conditions. J. Acoust. Soc. Am. 106, 2945–2958.
Kiefte, M., Kluender, K.R., 2005. The relative importance of spectral tilt in monophthongs and diphthongs. J. Acoust. Soc. Am. 117, 1395–1404.
Klatt, D.H., 1979. Perceptual comparisons among a set of vowels similar to/æ/: some differences between psychophysical distance and phonetic distance. J. Acoust. Soc. Am. 66, S86.
Klatt, D.H., 1982a. Prediction of perceived phonetic distance from critical band spectra: a first step. Proc. IEEE Int. Congr. Acoust., 1278–1281 Speech, Signal Processing.
Klatt, D.H., 1982b. Speech processing strategies based on auditory models. In: Granström, B., Carlson, R. (Eds.), Representation of Speech in the Peripheral Auditory System. Elsevier, Amsterdam, pp. 181–191.
Kuhl, P., Conboy, B., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., Nelson, T., 2008. Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). Philos. Trans. R. Soc. B 363, 979–1000.
Ladefoged, P., Broadbent, D.E., 1957. Information conveyed by vowels. Journal of the Acoustical Society of America 29, 326–332.
Liberman, A.S., Harris, K.S., Hoffman, H.S., Griffith, B.C., 1957. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54, 358–368.
Lutfi, R.A., 1993. A model of auditory pattern analysis based on component-relative-entropy. J. Acoust. Soc. Am. 94, 748–758.
Macmillan, N.A., Goldberg, R.F., Braida, L.D., 1988. Resolution for speech sounds: Basic sensitivity and context memory on vowel and consonant continua. J. Acoust. Soc. Am. 84, 1262–1280.
Mäkelä, A.M., Alku, P., Tiitinen, H., 2003. The auditory N1m reveals the left-hemispheric representation of vowel identity in humans. Neuroscience Letters 353, 111–114.
Miller, J.D., 1989. Auditory-perceptual interpretation of the vowel. J. Acoust. Soc. Am. 85, 2114–2134.
Näätänen, R., 2001. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). Psychophysiology 38, 1–21.

Näätänen, R., Alho, K., 1997. Mismatch negativity (MMN) - the measure for central sound representation accuracy. Audiology Neuro-Otology 2, 341–353.
Näätänen, R., Gaillard, A.W., Mäntysalo, S., 1978. Early selective-attention effect on evoked potential reinterpreted. Acta Psychologica 42, 313–329.
Näätänen, R., Jacobsen, T., Winkler, I., 2005. Memory-based or afferent processes in mismatch negativity (MMN): A review of the evidence. Psychophysiology 42, 25–32.
Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., Winkler, I., 2001. 'Primitive intelligence' in the auditory cortex. Trends in Neuroscience 24, 283–288.
Näätänen, R., Winkler, I., 1999. The concept of auditory stimulus representation in cognitive neuroscience. Psychological Bulletin 125, 826–859.
Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al., 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. Nature 385, 432–434.
Nazzi, T., New, B., 2007. Beyond stop consonants: Consonantal specificity in early lexical acquisition. Cognitive Development 22, 271–279.
Nearey, T.M., 1997. Speech perception as pattern recognition. J. Acoust. Soc. Am. 101, 3241–3254.
Nespor, M., Pena, M., Mehler, J., 2003. On the different roles of vowels and consonants in speech processing and language acquisition. Lingue e Linguaggio 2, 221–247.
Nord, L., Svantelius, E., 1979. Analysis and prediction of difference limen data for formant frequencies. Perilus 1, 24–37.
Nosofsky, R.M., 1989. Further tests of an exemplar-similarity approach to relating identification and categorization. Perception & Psychophysics 45, 279–290.
Obleser, J., Elbert, T.E., Lahiri, A., Eulitz, C., 2003. Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies. Brain Research Cognitive Brain Research 15, 207–213.
Obleser, J., Lahiri, A., Eulitz, C., 2004. Magnetic brain response mirrors extraction of phonological features from spoken vowels. J. Cogn. Neurosci. 16, 31–39.
Ohl, F.W., Scheich, H., 1997. Orderly cortical representation of vowels based on formant interaction. PNAS 94, 9440–9444.
Petkov, C.I., Kang, X., Alho, K., Bertrand, O., Yund, E.W., Woods, D.L., 2004. Attentional modulation of human auditory cortex. Nature Neuroscience 7, 658–663.
Perkell, J., Guenther, F., Lane, H., Matthies, M., Stockmann, E., Tiede, M., Zandipour, M., 2004. The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. Journal of the Acoustical Society of America 116 (4), 2338–2344.
Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., et al., 2000. Auditory cortex accesses phonological categories: an MEG mismatch study. J. Cogn. Neurosci. 12, 1038–1055.
Pickett, J.M., 1957. Perception of vowels heard in noises of various spectra. J. Acoust. Soc. Am. 29, 613–620.
Pickett, J.M. (1999). The acoustics of speech communication: Fundamentals, speech perception theory, and technology. Needham Heights, MA: Allyn and Bacon.
Pisoni, D.B., 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Perception & Psychophysics 13, 253–260.
Pols, L.C.W., van der Kamp, L.J.T., Plomp, R., 1969. Perceptual and physical space of vowel sounds. J. Acoust. Soc. Am. 46, 458–467.
Rosen, S., 1992. Temporal information in speech: Acoustic, auditory and linguistic Aspects. Philos. Trans.: Biol. Sci. 336, 367–373.
Rosner, B.S., Pickering, M.J., 1994. Vowel perception and production. Oxford University Press, Oxford.
Sachs, M.B., Young, E.D., 1979. Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate. J. Acoust. Soc. Am. 66, 470–479.
Sams, M., Paavilainen, P., Alho, K., Näätänen, R., 1985. Auditory frequency discrimination and event-related potentials. Electroencephalography and Clinical Neurophysiology 62, 437–448.
Savela, J., Kujala, T., Tuomainen, J., Ek, M., Aaltonen, O., Näätänen, R., 2003. The mismatch negativity and reaction time as indices of the perceptual distance between the corresponding vowels of two related languages. Cogn. Brain Res. 16, 250–256.
Sawusch, J.R., Gagnon, D.A., 1995. Auditory coding, cues, and coherence in phonetic perception. J. Exp. Psychol.: Hum. Percept. Perform. June 21, 635–652.
Schouten, B., Gerrits, E., van Hessen, A., 2003. The end of categorical perception as we know it. Speech Commun. 41, 71–80.
Schouten, M.E., van Hessen, A.J., 1992. Modeling phoneme perception. I: Categorical perception. J. Acoust. Soc. Am. 92, 1841–1855.
Shamma, S., 2001. On the role of space and time in auditory processing. Trends Cogn. Sci. 5, 340–348.
Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.
Shestakova, A., Brattico, E., Huotilainen, M., Galunov, V., Soloviev, A., Sams, M., et al., 2002. Abstract phoneme representations in the left temporal cortex: magnetic mismatch negativity study. Neuroreport 13, 1813–1816.
Shestakova, A., Brattico, E., Soloviev, A., Klucharev, V., Huotilainen, M., 2004. Orderly cortical representation of vowel categories presented by multiple exemplars. Cogn. Brain Res. 21, 342–350.
Stevens, K.N., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am. 111, 1872–1891.
Stevens, S.S., Volkman, J., 1940. The relation of pitch to frequency: a revised scale. Am. J. Psychol. 53, 329–353.
Tomiak, G.R., 1990. An evaluation of a spectral moments metric with voiceless fricative obstruents. J. Acoust. Soc. Am. 87, S106–S107.
Van Hessen, A.J., Schouten, M.E.H., 1992. Modeling phoneme perception. II: A model of stop consonant discrimination. J. Acoust. Soc. Am. 92, 1856–1868.
Versnel, H., Shamma, S.A., 1998. Spectral-ripple representation of steady-state vowels in primary auditory cortex. J. Acoust. Soc. Am. 103, 2502–2514.
Werker, J., Tees, R., 1984. Cross-language speech perception: evidence for perceptual reorganization during the first year of life. Infant Behav. Dev. 7, 49–63.
Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., et al., 1999a. Brain responses reveal the learning of foreign language phonemes. Psychophysiology 36, 638–642.
Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., Csepe, V., et al., 1999b. Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations. Cogn. Brain Res. 7, 357–369.
Zahorian, S.A., Jagharghi, A.J., 1993. Spectral-shape features versus formants as acoustic correlates for vowels. J. Acoust. Soc. Am. 94, 1966–1982.
Zeng, F.G., Nie, K., Stickney, G.S., Kong, Y.Y., Vongphoe, M., Bhargave, A., et al., 2005. Speech recognition with amplitude and frequency modulations. PNAS 102, 2293–2298.