The internal representation of vowel spectra investigated using behavioral response-triggered averaging

W. Owen Brimijoin[a] **and Michael A. Akeroyd**
MRC Institute of Hearing Research (Scottish Section), Glasgow Royal Infirmary, 16 Alexandra Parade, Glasgow G31 2ER, United Kingdom
owen@ihr.gla.ac.uk, maa@ihr.gla.ac.uk

Emily Tilbury and Bernd Porr
University of Glasgow, School of Engineering, 72 Oakfield Avenue, Glasgow G12 8LT, United Kingdom
e.tilbury@edu.salford.ac.uk, bernd.porr@glasgow.ac.uk

Abstract: Listeners presented with noise were asked to press a key whenever they heard the vowels [a] or [i:]. The noise had a random spectrum, with levels in 60 frequency bins changing every 0.5 s. Reverse correlation was used to average the spectrum of the noise prior to each key press, thus estimating the features of the vowels for which the participants were listening. The formant frequencies of these reverse-correlated vowels were similar to those of their respective whispered vowels. The success of this response-triggered technique suggests that it may prove useful for estimating other internal representations, including perceptual phenomena like tinnitus.

PACS numbers: 43.71.Es, 43.71.An, 43.71.Rt [SGS]
Date Received: October 23, 2012 **Date Accepted:** January 4, 2013

1. Introduction

A vowel can be mostly defined by the frequencies of its first two formants (Peterson and Barney, 1952). Since the absolute frequencies of the formants for a given vowel vary from men to women, talker to talker, and utterance to utterance it is unclear what criteria listeners use to identify vowels. Much of the previous work on identifying these criteria has been based on masking experiments (Moore and Glasberg, 1983; Sidwell and Summerfield, 1985); the current study used a technique based on the timing of responses to an ongoing random noise stimulus. This reverse-correlation method, borrowed from auditory neurophysiological studies and vision research (Gold *et al.*, 1999; Gosselin and Schyns, 2003), does not provide listeners with any auditory exemplars, and so is argued to be an unbiased means of determining the criteria listeners use to identify vowels.

Random stimuli have been used for decades as a means of examining sensory receptive fields (de Boer and Kuyper, 1968). By recording the timing of the responses to a time-varying random stimulus, it is possible to identify the stimulus properties that most frequently precede a response. In the field of auditory neurophysiology, such reverse-correlation techniques are used to estimate the spectrotemporal receptive field, which is thought to illustrate the way in which a cell's response to complex sound changes as a function of time (Eggermont *et al.*, 1981) and is thought to be related to an "optimal" stimulus, one capable of driving a neuron at a high rate of activity (deCharms *et al.*, 1998).

[a]Author to whom correspondence should be addressed.

This study adapted the reverse-correlation method from neurophysiology but used a noise stimulus to investigate the perceptual phenomenon of vowel identity. Reverse correlation has been used previously in an auditory behavioral study (Shub and Richards, 2009); unlike that study, however, listeners in the current experiment were not provided with actual signals embedded in the noise. The stimuli used were white noise on long-term average, but their instantaneous spectral content varied in a controlled manner. Listeners were asked to respond with a key press when they perceived a particular vowel in the noise (either [a] or [i:]). The average of the stimulus spectra immediately prior to each key press is taken as the measure of a listener's internal vowel representation.

2. Methods

Using inverse Fourier transforms, we generated 120-s noise stimuli that changed randomly in their frequency content every 0.5 s (each stimulus, containing 240 "frames," is referred to as a "block"). For each frame, the spectrum was divided into 60 logarithmically spaced bins from 0.1 to 22 kHz. The levels in each bin were assigned one of six values (0, −4, −8, −12, −16, −20 dB) [Fig. 1(a)]. The level values applied to the frequency bins in each of the 240 frames were assigned in a pseudorandom shuffling process to ensure that over the course of the whole stimulus all the frequency bins contained the same total energy, i.e., all bins had 40 repetitions of each of the 6 levels. This process created a noise whose long term average was perfectly uniform, but whose spectrum at any given instant was shaped in a defined manner [Fig. 1(b)]. Each full signal was presented to listeners over headphones (ES55, Audio Technica, Tokyo Japan) at a comfortable listening level [∼75 dB SPL (sound pressure level)].

Listeners were university students with self-reported normal hearing, ranging in age from 18 to 24 years old, and were all Scottish native speakers of English. Data were collected from a total of 18 listeners (10 male and 8 female). Each listener completed one hour of listening consisting of 30 blocks, each block consisting of a newly generated 120-s random noise. For 15 of the blocks, listeners were asked to press a key whenever they heard the vowel [a], but in the other 15 blocks they were asked to respond when they heard the vowel [i:]. These two vowels were chosen because they are spectrally dissimilar from one another, having a peak and a trough, respectively, at around 1000 Hz.

Listeners received no training in the task, but were familiarized with the stimulus in the instructions they received. They were told they would hear a rushing sound and were asked to respond as quickly as they could whenever they heard the vowel in the stimulus. These instructions were given verbally by one of the authors (a Scottish

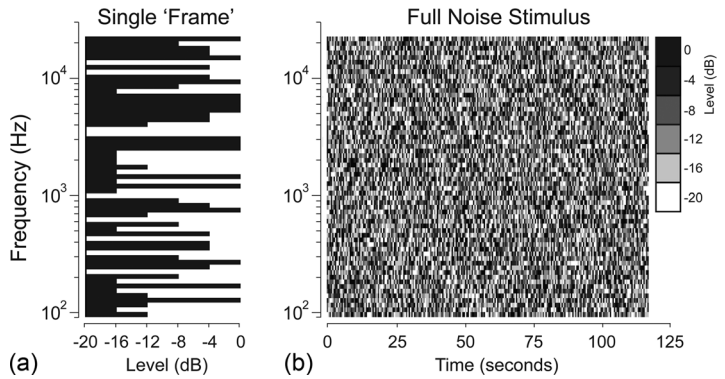

Fig. 1. (a) Amplitude spectrum for a given 0.5-s frame, showing the spectrum divided into 60 logarithmically spaced bins from 0.1 to 22 kHz. The individual bins were each randomly assigned one of six levels from −20 to 0 dB. (b) Schematic spectrogram of an example 120-s noise stimulus, created by combining 240 sequential frames of 0.5-s each. The final stimuli were continuous noises whose frequency content changed every 0.5 s but whose long-term spectra were uniform.

female) and were accompanied by onscreen text reminders of which vowel they were listening for on a given block. The stimulus properties and the times of each key press were recorded in MATLAB (Mathworks, Natick MA) and stored for subsequent analysis. Frequency/time spectrograms of the stimulus (dB values in 60 frequency bins and 300 time bins) in the 3.0 s window immediately prior to each key press were computed and summed for each block. These were then summed over all 15 blocks for each vowel, and then normalized to span a range from 0.0 to 1.0.

3. Results

A spectrogram of the reverse-correlated [a] vowel is shown in the Fig. 2(a), which we will refer to as a "vowel primitive." This figure represents data averaged from all [a] responses from all 18 listeners, each responding on average 384 [± 43.0 SEM (standard error of the mean)] times. The brightest region of the vowel primitive indicates that energy around 1 kHz was often found roughly 0.5 s prior to a response. But there was rarely energy found at this frequency 1 s prior to a response, suggesting that listeners were most likely to respond after a sudden increase in signal level at 1 kHz. The time bins from −1.5 to −3.0 s contained no discernible features, indicating that signals more than 1.5 s prior had no impact on the likelihood of response. Figure 2(b) plots the average spectrum across 0.4 to 0.1 s prior to each response as a solid black line. There are two distinct peaks in energy, one at 1030 Hz and a second at 1370 Hz. The dotted gray line shows the spectrum of a synthetic vowel [a] created with a Klatt synthesizer

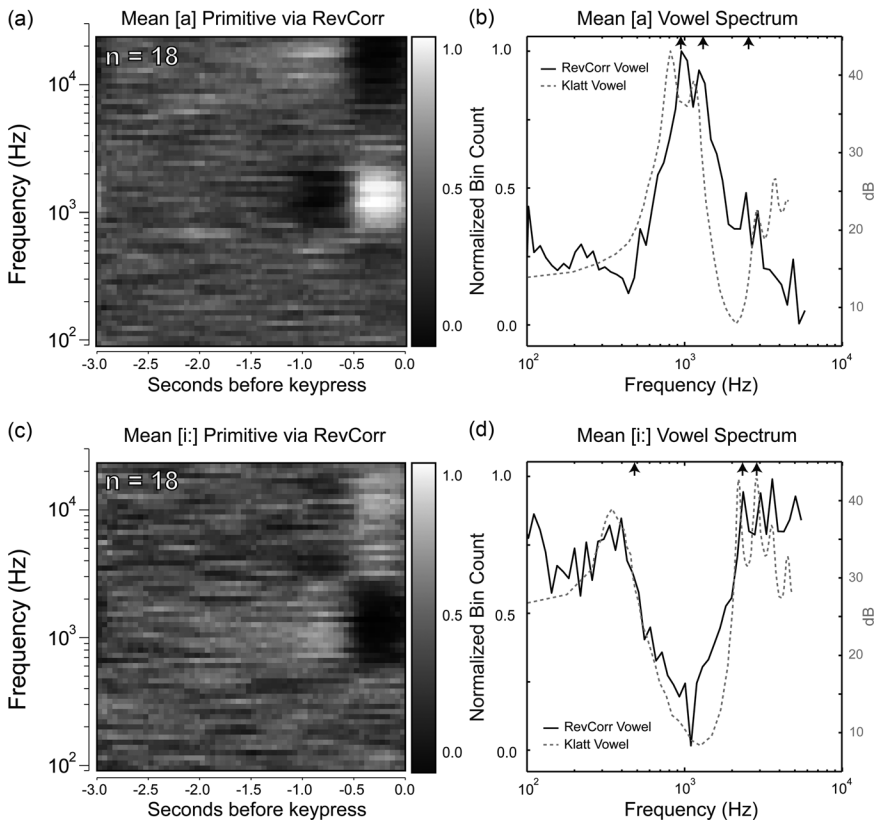

Fig. 2. (a) This [a] vowel primitive represents the mean reverse-correlated vowel spectrogram for $n = 18$ listeners. (b) The mean spectrum between 0.4 and 0.1 s prior to response (solid line) and the spectrum of a Klatt synthesized vowel (dotted line). The black arrows indicate the first three formant frequencies for male whispered [a] vowels. (c) The [i:] vowel primitive spectrogram. (d) The mean [i:] vowel spectrum (solid line), Klatt spectrum (dotted line), and whispered vowel formant frequencies (black arrows).

(Klatt, 1980). The spectra are similar to one another, but the vowel primitive is shifted to higher frequencies.

The same analysis was applied to the blocks during which participants were listening for [i:]. These data are shown in Fig. 2(c). This spectrogram shows average data from all 18 listeners, here with a mean of the total number of responses of 317 (± 43.7 SEM). The difference in response rate for [a] versus [i:] vowels was not significant ($t_{[36]} = 1.03$, $p = 0.311$). The bright region above 2500 Hz and extending up above 6000 Hz indicates that this vowel primitive was defined by an abundance of high frequency energy and a lack of low frequency energy. Again there was an increment in energy prior to response, this time above 2500 Hz. The average spectrum (across 0.4 to 0.1 s prior to responses) is plotted in Fig. 2(d). As was the case with [a], the spectrum of the vowel primitive [i:] is similar to a synthetic [i:], but shifted up in frequency.

4. Discussion

Despite the fact that no actual [a] or [i:] vowel sounds were intentionally embedded in the random noise stimulus, listeners responded at moments that corresponded to stimulus conditions that were evocative of the sounds for which they were listening. These responses were frequent and consistent enough to result in average spectrograms that had features that closely resembled classical vowel formants. This finding suggests that the criteria listeners use to identify vowels are resilient to noise, and underscores the extent to which the brain functions as a highly sensitive pattern detector compensating for missing and distorted information.

The upward shift of formant frequencies is likely due to the fact that the noise stimulus did not contain any amplitude modulations mimicking the glottal pulse rate of the human voice. Thus, the vowel sounds that participants were listening for were those of a whispered voice. It is known that the formant frequencies found in whispered vowels are higher than those of voiced vowels; the black arrows in Figs. 2(b) and 2(d) represent the frequencies of the first three formants found in whispered vowels (Jovičić, 1998) and are similar to the formant frequencies of the measured vowel primitives in the current data.

We did not present individual reverse-correlated formant frequencies here because in some listeners (three for [a] and six for [i:]), due to high variability in responses, it was not possible to identify formant peaks in those listeners' vowel primitives. Indeed response variability was a concern that could only be overcome by greatly increasing the number of blocks. It is a question for future research whether there is a difference in mean formant frequencies for male versus female listeners, a possibility raised by our data but currently unsupported due to variance.

Response-triggered averaging is not often used in behavioral auditory studies (cf. Shub and Richards, 2009), but has been used in behavioral vision research since Ahumada (1996) used pairs of noisy images to estimate vernier acuity. Typically these studies used discrete trials, whereas our use of an ongoing free-response paradigm is closest in implementation to that of Ringach (1998), and our use of random noise rather than signals embedded in noise is most akin to that of Gosselin and Schyns (2003). The current technique has a number of possible applications apart from uncovering features of vowel representations. It should be possible, for example, to use this method to reveal a patient's tinnitus spectrum, a percept that is notoriously difficult to estimate. Also, examining the way in which a person's internal representation of sounds changes with presbycusis could yield insights into aging and hearing impairment.

ACKNOWLEDGMENTS

This work was supported by the Medical Research Council (Grant Number U135097131), the Chief Scientist Office of the Scottish Government, and the University of Glasgow. The authors would also like to thank David McShefferty and William Whitmer for invaluable comments on drafts of this manuscript.

References and links

Ahumada, A. J. (1996). "Perceptual classification images from vernier acuity masked by noise," Perception **25**, ECVP Abstract Supplement.
de Boer, R., and Kuyper, P. (**1968**). "Triggered correlation," IEEE Trans. Biomed. Eng. **15**, 169–179.
deCharms, R. C., Blake, D. T., and Merzenich, M. M. (**1998**). "Optimizing sound features for cortical neurons," Science **280**, 1439–1443.
Eggermont, J. J., Aertsen, A. M., Hermes, D. J., and Johannesma, P. I. (**1981**). "Spectro-temporal characterization of auditory neurons: Redundant or necessary," Hear. Res. **5**, 109–121.
Gold, D. L., Bennett, P. J., and Sekuler, A. B. (**1999**). "Identification of band-pass filtered faces and letters by human and ideal observers," Vision Res. **39**, 3537–3560.
Gosselin, F., and Schyns, P. G. (**2003**). "Superstitious perceptions reveal properties of internal representations," Psychol. Sci. **14**, 505–509.
Jovičić, S. (**1998**). "Formant feature differences between whispered and voiced sustained vowels," Acta Acust. **84**, 739–743.
Klatt, D. H. (**1980**). "Software for cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.
Moore, B. C., and Glasberg, B. R. (**1983**). "Masking patterns for synthetic vowels in simultaneous and forward masking," J. Acoust. Soc. Am. **73**, 906–917.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in the study of vowels," J. Acoust. Soc. Am. **24**, 175–184.
Ringach, D. L. (**1998**). "Tuning of orientation detectors in human vision," Vision Res. **38**, 963–972.
Shub, D. E., and Richards, V. M. (**2009**). "Psychophysical spectro-temporal receptive fields in an auditory task," Hear. Res. **251**, 1–9.
Sidwell, A., and Summerfield, A. Q. (**1985**). "The effect of enhanced spectral contrast on the internal representation of vowel-shaped noise," J. Acoust. Soc. Am. **78**, 495–506.