

REPRESENTATION OF LOW-FREQUENCY VOWEL FORMANTS IN THE AUDITORY NERVE

Tatsuya Hirahara¹, Peter Cariani², and Bertrand Delgutte^{2,3}

hirahara@idea.brl.ntt.jp peter@epl.meei.harvard.edu bard@epl.meei.harvard.edu

¹NTT Basic Research Laboratories

3-1 Morinosato-Wakamiya
Astugi, Kanagawa 243-01
JAPAN

²Eaton Peabody Laboratory

Massachusetts Eye and Ear Infirmary
Boston, MA 02114
U.S.A.

³Research Laboratories of Electronics

Massachusetts Institute of Technology
Cambridge, MA 02139
U.S.A.

ABSTRACT

We have investigated the auditory representation of vowels with low-frequency formants by recording the activity of auditory-nerve fibers in anesthetized cats in response to Japanese /i/-/e/ synthetic-vowel continua. Vowels having either low (150 Hz) or high (350 Hz) fundamental frequency F0 were varied in either first-formant frequency F1 or the level of a “crucial harmonic” near F1 to span the /i/-/e/ continuum. Two different neural representations of the stimulus spectrum in the F1 region were examined: a population rate-place profile and a population interspike interval distribution. Characteristics of both representations depend on F0. When individual harmonics are resolved by the ear, as for high F0s, first formant frequency does not have explicit correlates in either ANF rate-place patterns or interspike interval distributions. Rather, both representations show clear patterns corresponding to individual harmonics, as well as the amplitude ratios of “crucial harmonics” near F1 that determine vowel identity in psychophysical tests. When harmonics are not resolved, as for low F0s, both rate-place and population-interval profiles of individual harmonics fuse to form broader, single peaks near F1, providing an explicit neural representation of formant frequency.

1. INTRODUCTION

Formant frequencies are important for vowel identification, yet the neural representation of formants is poorly understood, particularly in low frequency regions. Formants are resonant frequencies of the vocal tract which appear as local maxima (peaks) in the envelope of the stimulus spectrum. In general, spectral energy is present at harmonics of the fundamental frequency (F0) rather than at the formant frequency. For vowel discrimination, the auditory system could use a representation based on either the spectral envelope (formant) or the fine spectral structure (harmonics). Most models of speech perception assume that vowel quality is based on a single peak at the formant

frequency in a smeared internal spectrum. However, low-frequency (<1000 Hz) harmonics are resolved by the human auditory system, so that psychoacoustic excitation patterns [1] exhibit separate peaks for individual harmonics rather than a single formant peak.

Recent psychophysical results[2] suggest that harmonic fine structure near low-frequency formants plays an important role in vowel perception. Specifically, the phonetic boundary between the Japanese vowels /i/ and /e/ appears to be primarily determined by the amplitude ratio of two crucial harmonics that are nearest the first formant frequency (F1). Normally this amplitude ratio is determined by both F0 and F1. In these experiments, two synthetic-vowel continua spanning /i/ to /e/ were constructed. One is an F1-continuum in which the frequency of a resonator was systematically shifted, thereby altering the relative amplitudes of all harmonics near F1. Another is an L(nF0)-continuum in which the amplitude of a single crucial harmonic was systematically varied. The perceptual boundary between /i/ and /e/ was found to be the same for both kinds of stimulus manipulations when expressed in terms of an amplitude ratio of the crucial harmonics (Fig. 1). Further experiments showed that changes in F0 can influence vowel quality both by determining which harmonics are crucial and by altering the amplitude ratio at the boundary.

In the present electrophysiological study, vowel stimuli from Hirahara’s psychophysical experiments[2] were used to answer two questions about how formants are represented in the auditory nerve: (1) under what conditions is formant frequency rather than fine harmonic structure explicitly represented? (2) which neural representation corresponds best to human judgments of vowel quality, a population rate-place profile or a population interspike interval distribution? We are particularly interested in high-F0 (>200 Hz) vowels whose auditory-nerve representation has not been described.

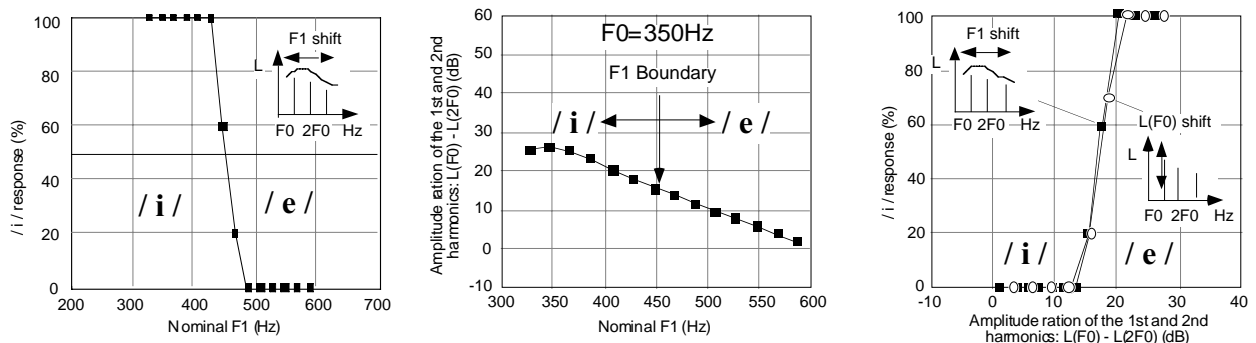


Fig. 1 Mean response curve of four subjects for the high-F0 (350Hz) F1-continuum (left). The amplitude ratio of the first and second harmonics of the stimuli covaries with F1 (middle). The perceptual vowel boundary between /i/ and /e/ are the same for the F1-continuum and the L(F0)-continuum when expressed in terms of the crucial harmonics amplitude ratio (right) [2].

2. METHODS

2.1 Stimuli

Stimuli were synthetic vowels forming continua between the Japanese vowels /i/ and /e/. All stimuli were produced by a 6-formant synthesizer using a 20 kHz sampling rate. Two types of continua were generated, each one using both a low F0 (150 Hz) and a high F0 (350 Hz). F1-continua varied the first formant frequency (F1) from 328 Hz to 528 Hz in 20-Hz steps. In these continua, amplitudes of all harmonics near F1 change as F1 increases. The phonetic boundary between /i/ and /e/ occurs roughly at F1=328 Hz for the low-F0 continuum, and at F1=448 Hz for the high-F0 continuum. Stimuli with F1 below the boundary are heard as /i/. L(nF0)-continua in which the amplitude of a single harmonic with frequency nF0 was varied in 3 dB steps over a 115 dB range that was centered at the /i/-/e/ phonetic boundary. For the low F0 continuum, the second harmonic (2F0) was manipulated, while for the high-F0 continuum, the first harmonic (F0) was manipulated. Stimuli with harmonic amplitudes above the boundary are heard as /i/.

For efficient data collection, the 11 stimuli forming each (F1- or L(nF0)) continuum were concatenated into an ascending-descending sequence. Each of the 20 stimuli in a sequence was approximately 50 msec in duration, giving a total duration of 1 sec.

2.2 Electrophysiological Recordings

The activity of auditory-nerve fibers was recorded in Dial-anesthetized cats using glass micropipettes. Stimuli were delivered through calibrated, closed acoustic assemblies at an overall level of 50 dB SPL. Times of action potentials were recorded with a precision of 1 μ sec.

Once a fiber was isolated, a threshold tuning curve and spontaneous discharge rate (SR) were determined. An accurate estimate of the fiber CF was obtained with broadband-noise stimuli using de Boer's reverse correlation technique. A rate-level function for a tone at the CF was measured to determine the

maximum discharge rate. Each vowel sequence was presented 60 times per fiber.

2.3 Data Analysis

For the rate-place analysis, normalized driven discharge rate was computed by averaging raw fiber discharge rate over the entire stimulus duration, subtracting out the spontaneous rate to obtain the driven rate, then normalizing the driven rate to a dimensionless quantity between 0 and 1 by dividing by the maximum driven rate. Rate-place profiles were formed by plotting normalized driven rate against fiber CF. A moving-window average of these profiles was obtained using 150-Hz windows.

For the interspike-interval analysis, an all-order interspike interval histogram was computed. These histograms include intervals between non-consecutive as well as consecutive spikes. Interspike intervals from all fibers with CFs below 1000 Hz were summed to form pooled interspike interval distributions.

These results are based on recordings from 235 auditory-nerve fibers in five cats.

3 RESULTS

3.1 Rate-place representation

Figure 2 shows the rate-place representation of the low-F0 F1-continuum. The left panel shows the normalized discharge rate as a function of both CF and F1 for all stimuli. Darker areas correspond to higher rates. Individual harmonics are clearly not resolved and maximum rates occur roughly when CF=F1. The plot resembles a smeared power spectrum of the vowel, with a broad peak near F1 [3][4].

Right panels are rate-place profiles for 3 vowel stimuli within the continuum, i.e. the horizontal cross sections of the left 2D diagram. Each data point represents normalized discharge rate for one auditory-nerve fiber. The solid line is a moving-average of the data points. For all 3 stimuli, the rate-place profile shows a maximum roughly centered at CF=F1. No peak is found at individual harmonics (300, 450, 600 Hz). Thus there is an explicit representation of F1 in rate-place profiles for low-F0 stimuli, when

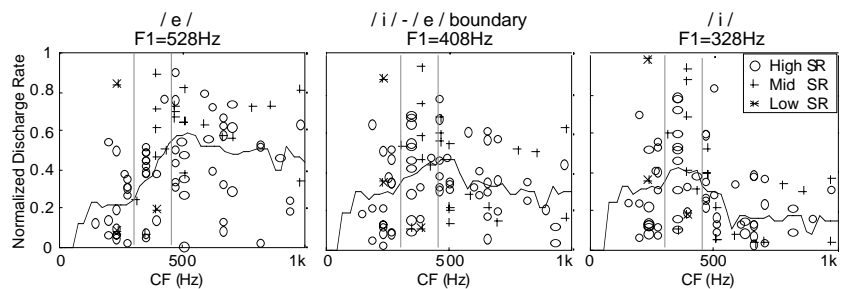
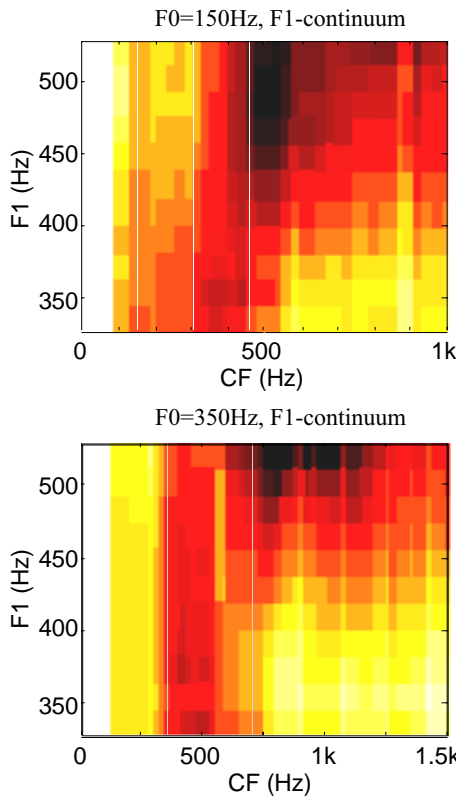


Fig. 2 The rate-place representation for the low-F0 (150Hz) F1-continuum. Peaks in the rate-place profiles occur near F1.

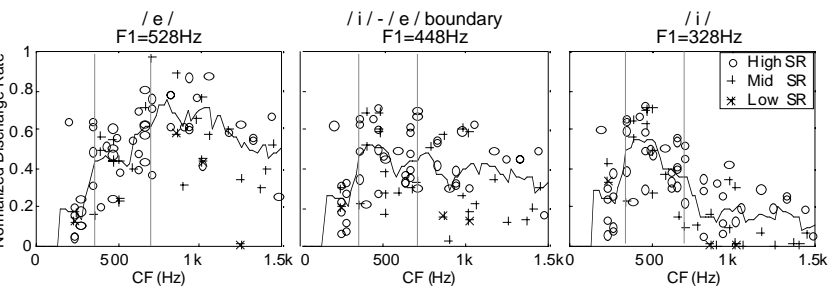


Fig. 3 The rate-place representation for the high-F0 (350Hz) F1-continuum. Peaks in the rate-place profiles occur near individual harmonics.

individual harmonics are not resolved by the ear.

Figure 3 shows the rate-place representation of the high-F0 F1-continuum. In contrast to the low-F0 continuum, population rate-place profiles for high-F0 stimuli show multiple peaks at individual harmonics rather than a single broad peak at F1. Maximum rates thus occur when $CF=nF_0$ (here 350, 700, 1050 Hz), rather than when $CF=F_1$. This implies that for higher-F0s, there is an explicit representation of individual, resolved harmonics rather than one for formant frequency F1.

3.2 Interspike-interval representation

Figure 4 shows the interspike-interval representation for the low-F0 F1-continuum. The left panel shows pooled interval distribution as a function of F1 for all stimuli in the continuum. Darker areas correspond to greater numbers of intervals. While the main modes are always at n/F_0 , secondary modes change systematically with F1. Thus, the pooled interval distribution gives an explicit representation of F1 for low F0s, when individual harmonics are not resolved in rate-place profiles.

Right panels show pooled interspike-interval distributions for 3 stimuli in the continuum, i.e. horizontal cross-sections of the left 2D diagram. For all 3 stimuli, the highest peaks in the pooled distribution are at the fundamental period $1/F_0$ and its multiples [5]. In addition, for both the /e/ stimulus and the /i/-/e/ boundary, secondary modes of the pooled distribution are found at $1/F_1$, $1/F_0 \pm 1/F_1$, $2/F_0 \pm 1/F_1$, For the /i/ stimulus, modes occur at the periods of the crucial harmonic $2F_0$, which is very close to F1.

Figure 5 shows the interspike-interval representation for the high-F0 F1-continuum. As with the low-F0 stimuli, the largest modes in the pooled interval distribution for the high-F0 stimuli are always at the fundamental periods (n/F_0). However, in contrast to the low-F0 case, secondary modes always occur at the periods of a higher crucial harmonic ($1/2F_0$). Thus, individual harmonics, rather than formant frequency, are explicitly represented in the pooled interval distribution for high F0s, when harmonics are resolved in rate-place profiles.

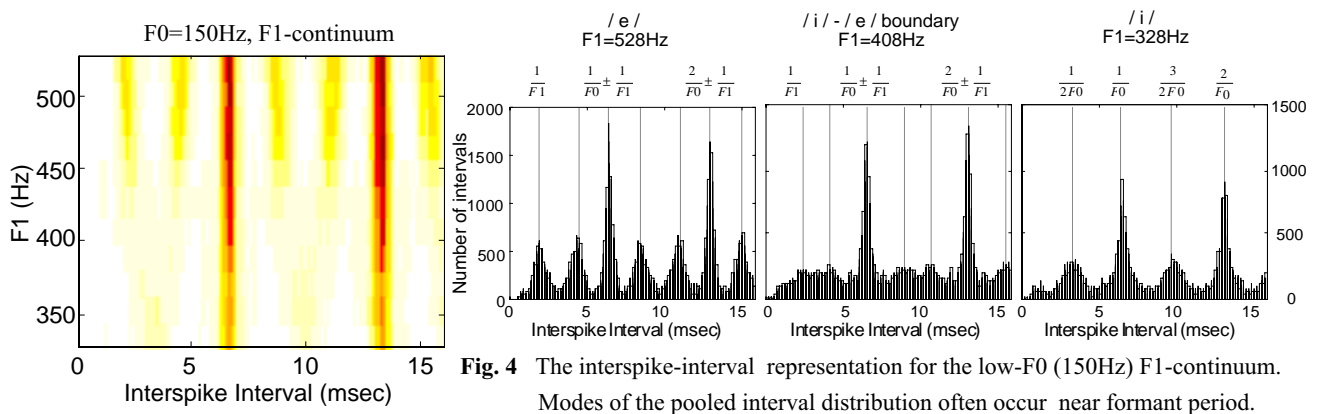


Fig. 4 The interspike-interval representation for the low-F0 (150Hz) F1-continuum. Modes of the pooled interval distribution often occur near formant period.

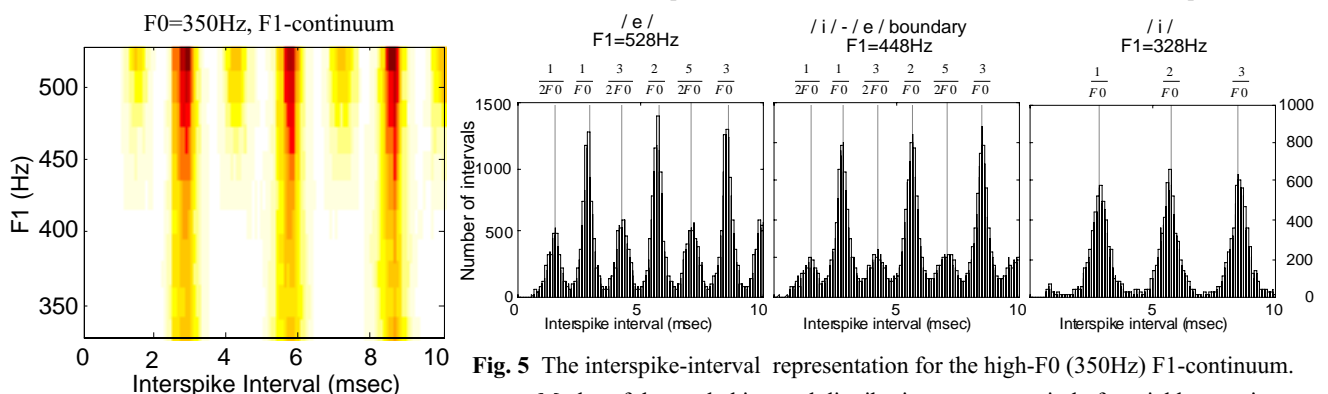


Fig. 5 The interspike-interval representation for the high-F0 (350Hz) F1-continuum. Modes of the pooled interval distribution occur at period of crucial harmonics.

Figure 6 shows interspike-interval representation for the high-F0 L(F0)-continuum. Pooled interspike-interval distributions for the high-F0 L(F0)-continuum are very similar to those for the high-F0 F1-continuum. In both cases, modes of the pooled distribution always occur at the periods of crucial harmonics. Thus, psychophysically-equivalent manipulations of changing the level of a single harmonic and shifting the first formant frequency produce pooled interval distributions that are also similar.

3.3 Comparison with psychophysics

A priori, we expect that the neural representation of vowel quality should covary with human vowel quality judgments, changing when perceived vowel quality changes, and remaining the same when vowel quality remains the same. Likewise, good neural correlates of phonetic boundaries should remain similar for phonetic boundaries that are observed using different stimulus manipulations. Human vowel quality judgments are expressed in terms of the percentage of /i/ judgments [2] for each stimulus in L(F0) and F1 continua. Neural measures of the amplitude ratio between crucial harmonics are expressed in terms of discharge rate ratios for the rate-place representation and in terms of interval ratios for the interval-based one. Discharge rate ratio is $R(F_0)/R(2F_0)$, where $R(f)$ is the average normalized discharge rate for $CF = f$. Interval ratio is $I(F_0)/I(2F_0)$, where $I(f)$ is the number of intervals at $1/f$ in the pooled interval distribution. Figure 7 plots these two neural measures against the human judgments that would be obtained for the same stimulus continuum. In the left panel, the discharge rate ratios for both continua are nearly the same at their respective phonetic boundaries (50% /i/ judgments). In the right panel, the same human vowel quality judgments (percent /i/) are plotted against the interval ratios for their corresponding stimuli. Again, for both continua the two curves are very close at the phonetic boundary (and elsewhere as well). Thus, the amplitude ratio of crucial harmonics provides a good acoustic correlate of the phonetic boundary, and both rate-place and interval-based neural measures

for this ratio provide correspondingly good neural correlates of the boundary.

4 SUMMARY AND CONCLUSIONS

1. Fundamental frequency plays an important role for the representation of low-frequency formants in the auditory nerve:

For high F0s, individual harmonics are physiologically resolved: peaks in rate-place profiles occur at the crucial harmonics, and modes of interval histograms are always at periods of specific harmonics. Thus, no explicit representation of formant frequency exists in the auditory nerve.

For low F0s, individual harmonics are not physiologically resolved: peaks in rate-place profiles occur near F1, and modes of interval histograms are often found near 1/F1. Thus, an explicit representation of formant frequency exists in the auditory nerve.

2. There exist correlates of the /i/-e/ phonetic boundary in the amplitude ratios of crucial harmonics, and these amplitude ratios have clear correlates in both rate-place profiles and patterns of interspike intervals.

These results have broader implications for vowel perception by humans. Psychophysical and physiological evidence suggests that the human ear is more frequency-selective than the cat ear [6][7]. Psychophysical measures of frequency selectivity in the human are 50-100 Hz for low frequencies. In contrast, the effective bandwidths of auditory-nerve fibers in the cat exceed 150 Hz. Interpreting the cat data in the light of these species differences leads to the following conclusions:

For all voices (men, women and children), harmonics near low-frequency formants (< 1000 Hz) are likely to be physiologically resolved. Here the amplitude ratio of crucial harmonics rather than formant frequency per se may be the key cue for vowel quality. Extrapolating from the cat data, we expect that invariant correlates of the phonetic boundary exist in both rate-place and temporal discharge patterns of the human auditory nerve.

For male voices, harmonics near higher-frequency formants

(>1000 Hz) are not likely to be physiologically resolved. Here, we expect that formant frequencies, rather than individual harmonics, are explicitly represented in both rate-place and temporal discharge patterns of the human auditory nerve.

ACKNOWLEDGMENT

The first author thanks Dr. Ken'ichiro Ishii for supporting him to carry out this work at EPL and NTT BRL. This research was supported by NIDCD grants DC02258 and DC00119.

REFERENCES

- [1] Glasberg, B. and Moore, B. (1990): Derivation of auditory filter shapes from notched-noise data, *Hearing Research* 47, 103-138
- [2] Hirahara, T. (1993) : On the role of relative harmonics level around the F1 in high vowel identification, *ARO Abstract*, 65
- [3] Sachs, M.B. and Young, E.D. (1979): Encoding of steady-state vowels in the discharge patterns of auditory-nerve fibers: representation in terms of discharge rate, *J.Acoust.Soc.Am.*, 66, 1381-1403
- [4] Sachs, M.B. (1985): Speech Encoding in the Auditory Nerve, in *Hearing Science*, Edited by C.Berlin. (Taylor & Francis, London, 1985), pp.261-307
- [5] Cariani, P. and Delgutte, B.: Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *Journal of Neurophysiology*. (in press).
- [6] Greenwood, D. (1990): A cochlear frequency-position function for several species - 29 years later, *J.Acoust. Soc.Am.* 87, 2592-2605
- [7] Pickles, J. (1979): Psychophysical frequency resolution in the cat as determined by simultaneous masking and its relation to auditory-nerve resolution, *J.Acoust.Soc.Am.* 66, 1725-1732

