Interaction between formant and harmonic peaks in vowel perception.

Hector Raul Javkin, Hynek Hermansky and Hisashi Wakita

Speech Technology Laboratory, Santa Barbara, California

ABSTRACT

The listener of a voiced vowel receives a signal consisting of formant-modulated harmonics. How this information is used in deriving both vowel timbre and resulting vowel identity is still not well understood. The suggestion by Klatt (1985,1986), that listeners perceive the actual resonance peaks, is contradicted by many works, including Mushnikov and Chistovich (1971) and Carlson, Granstrom and Fant (1975) who proposed weighted averages of neighboring harmonic peaks as the correlates of perceived vowel quality. Our perceptual experiments and re-analysis of the formant difference limen experiments of Flanagan (1955) and Nord and Sventelius (1979), support an interaction between formants and harmonic peaks in vowel perception.

INTRODUCTION

Although the influence of fundamental frequency on the perception of vowels is by now generally accepted [1,2,7,9, etc.], Klatt [5,6] has recently suggested that subjects respond to formant peaks without being affected by the location of the harmonic peaks determined by the fundamental, although he did find evidence for a normalization related to F0.

We found surprising support for the role of harmonic peaks in vowel perception in difference limen data shown in figures 1 [3], and 2 [8]. Both works provide the original measurement points along with the interpolated sensitivity curves. The measurements required a generous amount of interpolation to obtain smooth curves. If one takes into account the frequencies of the harmonic peaks in examining the published graphs, the origin of some of the outlying points can be hypothesised. The experiments were carried out with analog circuitry which might have produced some errors in fundamental frequency setting. If we allow for slight deviations of F0 values, almost all outlying points can be hypothetically attributed to the harmonic peak spacing. In Hermansky and Javkin [4] we reported on

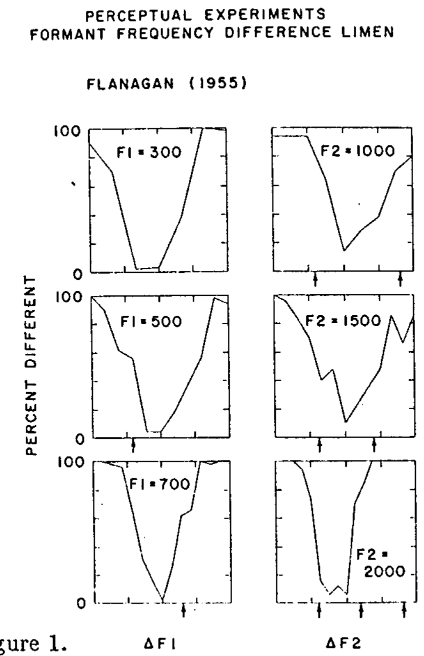

Figure 1.

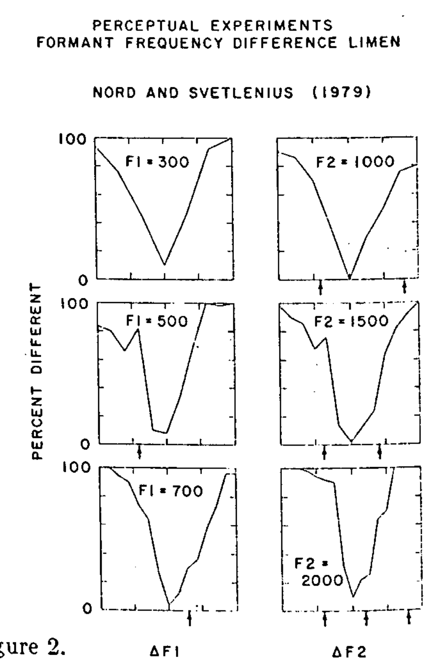

Figure 2.

further difference limen experiments on selected vowels that generally confirmed this hypothesis. Figure 3 shows the results for one of these experiments, for a vowel with formants at 500, 2000, 2500, 3500 and 4500 Hz. Bandwidths were 50, 90, 120, 150 and 180 for these formants. The fundamental period was varied between T0 = 8.5 and T = 8.0 msecs. in 0.1 msec. increments. Dashed lines in the figure connect points with equal formant frequency deviation from the reference vowel. Asymmetries, resulting from different distributions of harmonic peaks depending on the fundamental, are quite substantial. Figure 4 shows the results of the experiment with the same vowel but with the fundamental period T0 = 4.2 msec (with consequently wide harmonic spacing). Here the sensitivity curve shows an irregular (non-unique value) portion, similar to those observed in Flanagan's [3] data, coincident with a harmonic.

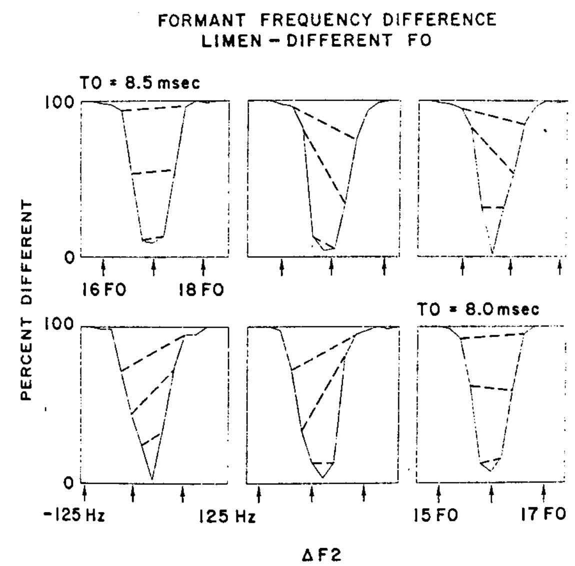

Figure 3.

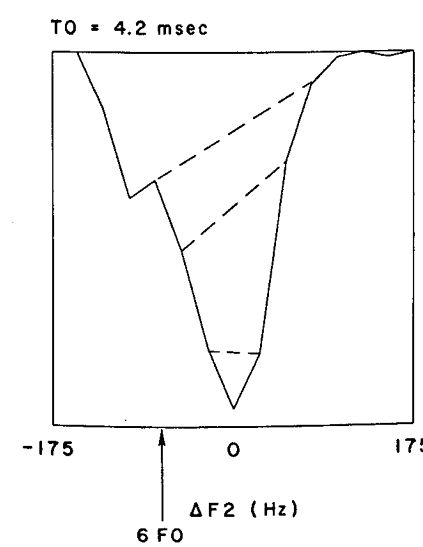

Figure 4.

The results provide further evidence for the hypothesis that the human auditory system tends to shift the formant peak estimate towards the nearest harmonic peak, but do not provide a basis for quantifying this shift. Carlson, Fant and Granstrom [1] attempted to model human listeners' perception of formant peaks and proposed the idea of "most important frequency" or MIF, which determines the weighted means of the two most prominent harmonics by an equation which can be written as follows:

$$\mathrm{MIF} = \frac{f_m W_m + f_n W_n}{W_m + W_n}$$

f_m is the frequency of the most prominent harmonic, f_n is the frequency of the next most prominent harmonic, and W_m and W_n are the weights given the respective harmonics. Carlson et al made the weights equal to the amplitude of the harmonics in the Sone space, so that S_m and S_n were used for W_m and W_n.

This formula suggests that listeners will accurately find the peaks when a formant lies between two harmonics, but will be less accurate when a formant coincides with or is close to a harmonic. An average of the two strongest partials, even one that is weighted towards the stronger, contains at least some contribution of the second strongest partial, and pulls the model's calculation of the formant away from the formant peak. If listeners can accurately find the formant peak when it coincides with a harmonic, the model will differ from their responses. The hypothesized estimation contains another, implicit hypothesis. Taking an arithmetic weighted mean in the sone space makes the assumption that listeners evaluate the amplitude of two neighboring harmonics for the purpose of peak location in the same way that they evaluate the amplitude of sounds presented separately. That is to say, they evaluate the relative amplitudes of the two without an interaction that increases the perceived amplitude of one or diminishes the perceived amplitude of the other.

Carlson et al conducted a perceptual experiment with F0 values from 100 to 160 Hz in 15 Hz steps and F1 values ranging from 250 to 350 Hz in 25 Hz steps. Their results showed that their hypothesis worked the best among those examined, although its prediction is quite different from the perceptual data when F0 = 100 and F1 = 300, i.e. when the formant coincided with one of the harmonics.

A less compressed scale such as magnitude or intensity will increase the contribution of the stronger partial and can increase the correlation between the output of Carlson et al's equation and their experimental data. It should be noted, however, that using a less compressed scale is functionally similar to using the same scale but with the addition of some form of peak enhancement.

MATCHING EXPERIMENT

To test whether a different scale would yield results closer to those of human listeners, a matching experiment was conducted. Our aim here was to avoid the effects of categorization that occur in vowel perception and investigate the psychoacoustic effects. Accordingly, single-formant stimuli with a single resonance driven by a pulse train with a flat spectrum were synthesized. F0 was kept constant at 200 Hz. One set of stimuli had peaks ranging from 600 to 800 Hz in 20 Hz increments, the other set had the same increments, but ranging from 2000 to 2200 Hz. Both sets were prepared with three bandwidths, of 50, 100, and 150 Hz, for a total of 66 stimuli. Durations of the single-formant stimuli were 500 msec with 40 msec leading and 70 msec trailing edges, while the tones had a 500 msec duration but 60 msec leading and 120 msec trailing edges. The inter-stimulus interval was 200 msec.

The presentation of the stimuli and the recording of responses were performed by a computer with a 16-bit digital-to-analog converter using a sample rate of 10 kHz with the output appropriately filtered. The stimuli were presented in different quasi-random orders to different subjects, who listened through earphones inside a sound-treated room. Subjects set the loudness of presentation to a comfortable level, and the level was checked visually after each subject completed the experiment. None of the subjects reported any hearing pathology. For each trial, subjects heard one of the single-formant stimuli followed by a sine wave. Their task was to match the timbre of the first stimulus by adjusting the frequency of the sine wave, using keys on a computer terminal. Their responses were limited to between 550 and 850 Hz for the F1 range stimuli and between 1950 and 2250 Hz for the F2 range stimuli. They could make adjustments for as long as they wished and heard a repetition of the two stimuli after each adjustment. When they indicated satisfaction with a match, their last adjusted value was automatically recorded in a computer file and the next trial began.

RESULTS OF MATCHING EXPERIMENT

Twelve subjects participated in the experiment. The task proved quite difficult for some subjects and two were eliminated after complaining of the difficulty and giving over a third of the responses at the response limits. The results for the different bandwidths did not differ significantly but were noisy. The results were band-limited to within 150 Hz (approximately two standard deviations) of the presented stimuli in order to limit somewhat the distorting effects of outliers. This meant that, for example, responses greater than 750 Hz to a 600 Hz stimuli were dropped from the data. Because of the band limitations in the presented stimuli and in the possible subject responses, points far away from the stimuli would severely distort the means. In addition, given a fundamental frequency of 200 Hz, a response of more than 750 Hz to a stimulus with a formant at 600 Hz might be the result of approaching the harmonic at 800 Hz. The results for the three bandwidths were combined in table 1, showing the results of the experiment for stimuli in the F1 range, and in table 2, showing the results for stimuli in the F2 range.

Table 1.

	Stimuli										
	600	620	640	660	680	700	720	740	760	780	800
subj	608	615	627	640	656	693	698	730	743	778	779
sone	640	648	661	674	687	700	713	726	739	752	760
mag	622	630	645	663	681	701	720	738	756	771	778
int	603	606	616	635	664	702	738	767	785	794	797

Table 2.

	Stimuli										
	2000	2020	2040	2060	2080	2100	2120	2140	2160	2180	2200
subj	2005	2010	2035	2047	2056	2118	2100	2140	2188	2165	2200
sone	2054	2059	2067	2077	2087	2098	2109	2120	2130	2138	2143
mag	2037	2043	2053	2067	2082	2098	2114	2128	2142	2153	2159
int	2010	2014	2024	2040	2065	2095	2126	2153	2172	2183	2188

Figures 5 and 6 graph the same results. Subjects' responses are represented by a solid line. The predictions of MIF calculated in sones are represented by a line of long dashes; the predictions calculated in magnitude are represented by a line of short dashes; and the predictions of MIF in the intensity space are shown by alternating short and long lines.

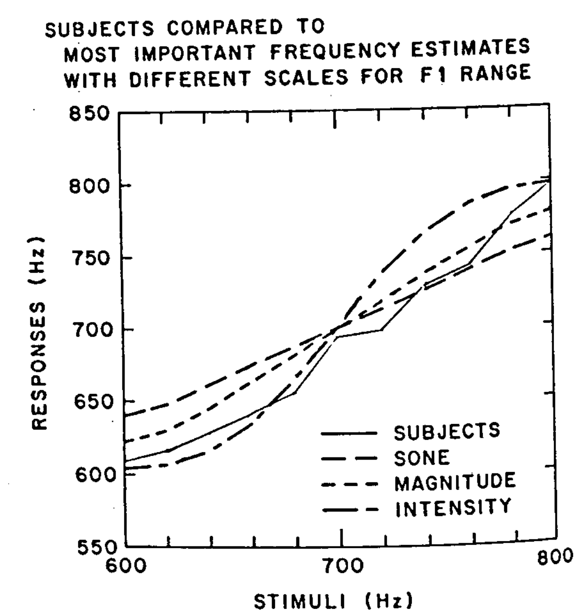

Figure 5.

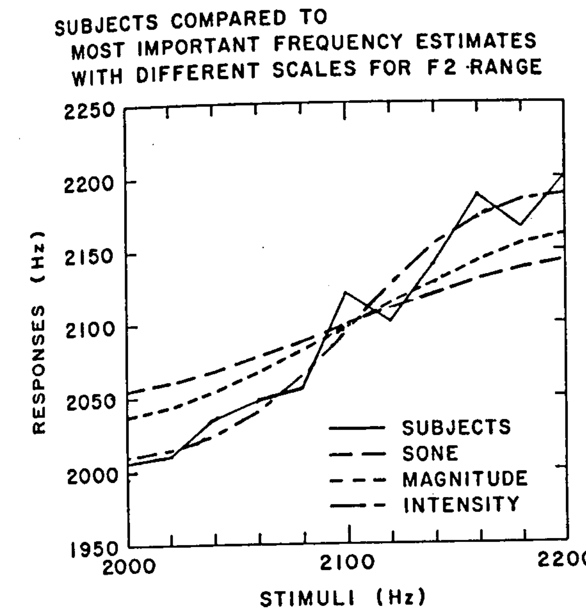

Figure 6.

The results for both sets are similar at their end-points, although the data for the F2 range shows a less smooth pattern than the data for the F1 range. Both show a tendency for stimuli with harmonics close to the formant peak to attract responses and also for responses to show a "plateau" when the formant is equidistant between harmonics. The responses would approximate a straight line if subjects were responding to the location of the formant peak without regard to the location of harmonics, so that the experiment confirms the effect of harmonic peaks. Nevertheless the experiment does not confirm the predictions of Carlson et al.

CONCLUSIONS

It is clear from the experiments reported here, as well as the vast majority of the experimental literature, that the location of harmonics plays a role in the perception of vowels, and, more specifically, that harmonic peaks which coincide or nearly coincide with formants tend to attract judgments of formant location. This effect appears to be too strong to be represented by a weighted average of the two most prominent harmonics in the loudness space. Such an average can be improved by using a different scale, effectively expanding the differences in amplitudes. Although the results reported here are still somewhat sketchy and must be considered with caution, they support the idea that such an expansion is necessary to describe the response of the human auditory system.

REFERENCES

1. Carlson, R., Fant, G., Granstrom B. 1975. Two formant models, pitch and vowel perception - in Auditory Analysis and Perception of Speech (G. Fant & M.A.A. Tatham, eds.) Academic Press, London.
2. Chistovich, L.A. and Chernova, E.I. 1986. Identification of one- and two-formant steady-state vowels: a model and experiments. Speech Communication 5:3-16.
3. Flanagan, J.L. 1955. A difference limen for vowel formant frequency. Jour. Acoust. Soc. Am. 27.3:613-617.
4. Hermansky, H. and Javkin, H.J. 1986. Evaluation of ASR front-ends using synthetic speech - Paper presented at the 112th Meeting of the Acoustical Society of America, Anaheim, California.
5. Klatt, D. 1985. The perceptual reality of a formant frequency. Jour. Acoust. Soc. Am. 78, Suppl. 1:S81.
6. Klatt, D. 1986. Representation of the first formant in speech recognition and in models of the auditory periphery - Proceedings of the Montreal Symposium on Speech Recognition, McGill University, July, 1986.
7. Mushnikov, V.N., Chistovich, L.A. 1971. Method for the experimental investigation of the role of component loudnesses in the recognition of a vowel. Akusticheskii Zhurnal 17.3:405-411.
8. Nord, L., Sventelius, E. 1979. Analysis and perception of difference limen data for formant frequencies. STL-QSPR 3-4/1979:60-72.
9. Traunmuller, H. 1981. Perceptual dimension of openness in vowels. Jour. Acoust. Soc. Am., 69.5:1465-75.

ACKNOWLEDGEMENTS

We would like to thank Ted Applebaum, Jared Bernstein, Gregory De Haan, Brian Hanson, Dennis Klatt, Katia McClain, Lucio Mendes, Paul Neyrinck, Ben Reaves and Kathy Sangster for valuable discussions and help in various aspects of this paper.