REPRESENTATION OF THE FIRST FORMANT IN SPEECH RECOGNITION AND IN MODELS OF THE AUDITORY PERIPHERY

Dennis H. Klatt

Room 36-523, Massachusetts Institute of Technology, Cambridge MA 02139, USA

Abstract. The frequency and amplitude of the first formant are not easy to measure as fundamental frequency (f0) varies in speech. Perceptual data indicate that the auditory system is not bothered by changes to f0, but processing strategies used in speech recognition, such as linear prediction, filterbank analysis, and the synchrony spectrum are seriously perturbed as f0 varies. The irrelevant variation makes it difficult/unreliable to perform phonetic comparisons between similar vowels based on simple ideas of pattern similarity. Of the possible solutions to this problem considered here, the one of greatest practical attraction is to implement a synchrony spectrum representation of vowel-like speech sounds, and a "learned pattern equivalence" approach to vowel phonetic-quality equivalence across different fundamental frequencies.

DFT magnitude spectra (25.6 ms Hamming window) of the lowest 1 kHz of a series of 5 kHz synthetic vowels are shown in Figure 1. All synthesis parameters have been held constant across stimuli except for the fundamental frequency of voicing (f0), which has been assigned a different constant value for each stimulus. The stimuli were devised to illustrate the problem of estimating the frequency (F1) and level (A1) of the first formant as fundamental frequency changes.

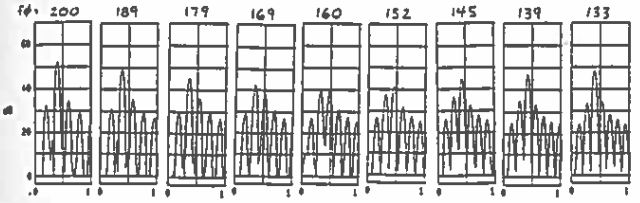

Figure 1. DFT magnitude spectra of 9 synthetic vowel stimuli varying only in f0.

The first formant frequency is 400 Hz in each synthetic waveform, and the first formant bandwidth is 50 Hz. These values, as well as the chosen frequencies and bandwidths of higher formants (F2=1800 Hz, B2=140, F3=2900, B3=240, F4=3800, B4=350), are typical for a vowel such as in the word "bit" (Klatt, 1980). Fundamental frequencies were selected in equal logarithmic steps from 133 Hz to 200 Hz. For the lowest fundamental, the third harmonic is exactly aligned with the 400 Hz first formant frequency; for the highest fundamental in the set of stimuli, the second harmonic is exactly aligned with the first formant frequency. For stimuli with intermediate values of fundamental frequency, no harmonic is exactly aligned with F1, and one has to interpolate by eye to determine the probable location of the first formant. This interpolation is not easy to perform automatically, as will become clear when we discuss the performance of various popular algorithms for formant estimation. There is a tendency for the first formant frequency estimate to be biased toward the frequency of the most intense harmonic, resulting in an error of up to plus-or-minus 8 percent for this stimulus set (Table 1).

Furthermore, the amplitudes of harmonics close to F1 are considerably less intense for intermediate stimuli of the stimulus set. The harmonic amplitudes are determined by the transfer function of the vocal tract, which peaks rather sharply at 400 Hz. If no harmonic is near F1, the strongest harmonic can be attenuated by up to 9 dB, resulting in a spectral peak that is attenuated by as much as 6 dB (filter banks) or 8 dB (linear prediction), which agrees with theory (Fant and Liljencrants, 1962) and measurements of real speech (Fintof, Lindblom and Martony, 1962). The formant amplitude misestimates of linear prediction are a result of misestimating formant bandwidths by a considerable factor (Atal and Schroeder, 1975).

STIM	f0	F1	HARMON	FB	LP
A	200	400	400	400	400
B	189	400	378	382	389
C	179	400	358	367	384
D	169	400	338	371	398
E	160	400	amb.	401	425
F	152	400	456	430	436
G	145	400	435	430	432
H	139	400	417	417	423
I	133	400	399	400	400
MAX ERROR:			+16% -15%	+7% -8%	+9% -4%

Table 1. First formant frequency predictions of nearest harmonic hypothesis (HARMON), peak location in wide-bandwidth filter bank (FB), and linear prediction spectrum (LP). Error increases if f0 is increased or BW1 is decreased.

According to one theory (HARMON in Table 1), the first formant is perceived to be the frequency of the strongest harmonic, at least for fundamental frequencies such that the ear can resolve individual harmonics (Chistovich, 1971).

According to a second theory, the formant peak is found by smoothing the spectrum in frequency such that individual harmonics are not seen (Chistovich et al., 1979). This proposal is similar in effect to earlier models which proposed to weight the importance of two strong harmonics according to the relative strength of their auditory representations (Carlson, Fant and Granstrom, 1975). In order to test the predictions of this theory, a particular smoothing algorithm was chosen -- the dft spectrum was smoothed by a 300-Hz wide Gaussian filter. As can be seen from Table 1, the energy smoothing model predicts that the perceived formant frequency will be somewhere between the "true" 400 Hz synthetic formant and the strongest harmonic. The amount of formant shift with changes to fundamental frequency is, however, quite large (see also Lindblom, 1962; Monsen, 19xx). Stimuli C and F differ by 63 Hz according to this model, which is 16 percent of F1. This difference would be easily audible because the JND for F1 is about 3% (Flanagan, 1955; Mermelstein, 1978). Thus Stimuli C and F should be heard as different vowels (/i/ and /I/) if this model were an accurate predictor of perceptual formant shifts with changes in formant/harmonic relationships. Apparently, the problem with the energy smoothing model is that a harmonic changes amplitude very rapidly as it slides down the skirt of a formant with a narrow (50 Hz) bandwidth. As soon as a harmonic is reduced by 4 to 6 dB below an adjacent harmonic, it hardly influences the location of the peak in the energy-smoothed spectrum.

According to a third theory, linear prediction spectra (autocorrelation form, 14-pole, 25.6 ms Hamming window) can extract F1 as the peak in the LP spectrum. Linear prediction fits an all-pole model to the waveform (Atal and Hanauer, 1971; Markel, 1972) or spectrum (Makhoul, 1975), thereby providing a method for effectively interpolating between harmonic locations to infer formant peaks. It is a particularly good model to apply to these stimuli since they were generated by an all-pole synthesizer and have virtually no noise or voicing source irregularities. The predictions of the linear prediction model are shown in the final column of Table 1. Linear prediction is not much better in performance than simple energy smoothing: there is a 52 Hz swing in the predicted F1 from stimulus C to F, which is a 13 percent change. Also, there is a slight bias toward overestimating F1 because the first harmonic amplitude is attenuated by the first difference analysis calculation. The reason that linear prediction does no better than the energy smoothing model is that the autocorrelation method uses a window of several pitch periods in duration, which means that the model must try to predict not only the damped vocal tract response to the first excitation at the beginning of the window, but also the time and magnitude of additional later glottal excitations and damped responses to them (Atal and Schroeder, 1975).

Perceptual Data. Does the human perceptual apparatus employ processing strategies which make all of these stimuli sound like exactly the same vowel (F1 the same) with the same loudness (vocal effort the same)? Naively, one might expect that if these stimuli are played in succession, one would hear not only a change in pitch, but also changes in loudness, spectral tilt, and vowel quality.

(1) First Formant Amplitude and Perceived Loudness. To see whether formant amplitude changes produce loudness differences across stimuli, Stimulus E was synthesized in its standard form and with 1,2,...6 dB added to the voicing sound source intensity. This set of stimuli was compared with both Stimuli A and I in unaltered form, using an "AX" randomized sequence in which subjects made a forced choice as to whether the first or second member of the pair was louder. Results from four listeners indicate a perceptual equal-loudness crossover at 2.0 dB. Thus when the pair of harmonics straddling F1 are 8 dB less intense (Stimulus E) than the single harmonic identical to F1 (Stimulus I), one must increase the level by only 2 dB to match subjective loudness.

Normally, it is said that loudness of a vowel depends primarily on the energy at F1, since this is usually the most intense part of the spectrum. We see that this is not the entire story because Stimuli E and I differ by 6 to 9 dB (depending on how energy near F1 is estimated), whereas an increase of only 2 dB makes these stimuli sound equally loud. Other possible determinants of vowel loudness are (1) the intensities of harmonics below F1, (2) energy in higher formants, (3) spectral tilt, and (4) the inferred shape of the vocal tract transfer function, i.e. the transfer function peak height instead of physical energy present at F1. Any one of these other potential cues could account for our loudness judgement results.

The variation in spectral amplitude of F1 as f0 is changed may be just as serious a deficiency of these spectral representations as mislocations of F1 in frequency. Any speech recognition device employing a distance metric that is sensitive to differences in relative formant amplitudes, such as the Itakura (1975) linear-prediction minimum prediction residual, or a filter-bank-based Euclidean metric (Plomp, 1970), will see considerable differences as f0 varies, even though the vowel is phonetically constant. This irrelevant variability can swamp out an ability to make fine phonetic distinctions in any current recognition device employing filter banks or linear prediction representations.

(2) First Formant Frequency and Perceived Vowel Quality. What kind of a perceptual effect on vowel quality is to be expected when f0 is changed? One possibility is that the auditory system somehow is able to extract the true F1, so vowel quality is unaffected. A second possibility is that the auditory system is fooled, or partially fooled, in exactly the same way as our processing schemes. A third possibility, one that somewhat confounds the choice between these alternatives, is that a change in f0 automatically invokes a kind of vowel-normalization process such that vowels spoken at higher f0 are assumed to come from shorter vocal tracts (Miller, 1953; Fujisaki and Kawashima, 1968; Carlson, Granstrom and Fant, 1970; Schwartz, 1971; Slawson, 1968; Traunmuller, 1982; Syrdal, 1985). A listening test was devised to distinguish among these alternatives (Klatt, 1985). Results showed convincingly that the auditory system is able to recover the true F1 with no bias toward the strongest harmonic, but there is also an automatic normalization process which makes it seem as if the vocal tract is shorter as f0 increases.

DISCUSSION

Our perceptual results are consistent with those of an excellent earlier paper that addressed the same issues (Carlson et al., 1975). They too found a regular shift in phonetic perception consistent with the view that f0 affects expectations of the vocal tract length of a talker. The authors examined their data to determine whether any phoneme boundary shifts could be attributed to perceptual biases toward the strongest harmonic, or toward a weighted mean of 2 or more harmonics. The weighting scheme that they employed was not the same as ours in that it did not weight harmonics according to their energy, and they did not examine an f0 range where harmonic biases go in an opposite direction from normalization biases, but the conclusions were the same -- there was no evidence of a bias toward the strongest harmonic as opposed to F1 (see also Florin, 1979; Assmann and Nearey, 1983; Darwin and Gardner, 1985).

So far this has been a largely negative paper: we have isolated defects in most speech processing algorithms that lead to unnecessary spectral confusions, but we have not provided any solutions. Three possible solutions are considered next.

Pitch-Synchronous Short-Window Analysis. If the analysis window is shorter than a single pitch period (e.g. windowed dft with a fixed 2 to 4 ms Hamming window, or covariance linear prediction during the inferred closed phase of glottal period) one can estimate the natural damped response of the vocal tract transfer function in the absence of excitations (Atal and Hanauer, 1971). This type of model is attractive, but is not easy to implement in a practical speech analysis system in such a way as to avoid occasional gross errors. If the window is misplaced, some very irregular spectra can be generated. The greatest problem with this kind of model is finding the time of glottal closure. Misplacements are particularly probable for high pitches and in noise. Until such time as analyses of this type can be made to mimic human perception consistently, we will have reason to doubt the validity of the technique as a speech analysis tool. An alternative might be to attempt to model the vocal tract transfer function using linear prediction, while simultaneously modeling the glottal waveform by some other appropriate representation (Milenkovic, 1986).

Auditory Modeling: Synchrony Detection. Sachs et al (1982) have shown that a measure of the tendency of neural firings to be synchronous with aspects of the basilar membrane displacement waveform has important advantages for speech processing. The synchrony measure is far less sensitive to changes in intensity of a vowel than are the average firing rate data. Synchrony data are also more immune to background noise and reverberation distortions (Allen, 1985), and they are not strongly affected by spectral tilt and formant amplitude variation (Srulovicz and Goldstein, 1983) which agrees with data on phonetic perception (Klatt, 1982). Processing schemes based on synchronous responses are reviewed in Carlson and Granstrom (1982), Delgutte (1984) and Seneff (1984). Thus it is of interest to determine whether any of these measures of synchronous response contains a representation of F1, and if so, is the estimate biased toward the strongest harmonic?

An answer comes directly from the Sachs et al. data, and from theoretical analysis of the waveforms observed at the outputs of the low-frequency critical band filters in this type of model. Physiological data and current models agree that the auditory system resolves individual harmonics near F1 for stimuli such as our family of synthetic vowels. Nowhere in the neural pattern are there time intervals between firings that are the inverse of F1. Only intervals related to harmonics are present. There is essentially only a sine wave at the outputs of these simulated mechanical filters because of a kind of FM capture effect that makes the strongest harmonic dominate the synchrony response in any channel (Allen, 1985). It will therefore be up to the central nervous system to figure out the first formant frequency from the relative proportions of fibers responding to each of the harmonics (and perhaps the relative phases of synchrony across channels). We can say little about the existance or details of such a calculation at this point.

Spectral Pattern Equivalence Sets. One interesting alternative that is not usually considered in speech recognition devices is that the harmonic pattern in the synchrony response is not processed centrally to recover an estimate of F1, but rather serves as a pattern vector in its raw form [Dick Lyon (personal communication) has expressed a similar

viewpoint]. The CNS would then have to learn pattern equivalence sets across different fundamental frequencies, even though there may not be striking pattern similarity for equivalent vowel tokens. The total number of patterns in such a system would be much larger than the largest current vector quantization pattern set, but the approach, given sufficient labeled training data (see e.g. Kopek, 1985 for one of a number of possible implementation methods), could potentially overcome a number of other puzzling aspects of cross-speaker variability, as well as some of the distortions to a normal formant shape caused by (1) truncation effects (Fant and Ananthapadmanabha, 1982), (2) other source-tract interactions (Fant, 1985), (3) breathy-normal-creaky voice quality variations (Fant et al., 1985), and (4) vowel nasalization (Hawkins and Stevens, 1985). These four factors can introduce additional errors in algorithms designed to measure formant frequencies based on the detection of spectral peaks, and forcefully call into question the desirability of simple-minded approaches to the extraction of the frequency of F1 from speech waveforms (Bladon, 1982), although there can be no question of the importance of changes in F1 for vowel perception (Klatt, 1982).
[This research was supported by ARPA.]

REFERENCES

Allen, J. (1985), "Cochlear Modeling," IEEE ASSP Magazine, Jan., 3-29.

Assmann, P.F. and Nearey, T.M. (1983), "Perception of Height Differences in Vowels", J. Acoust. Soc. Am. 74, S89 (A).

Atal, B.S. and Hanauer, S.L. (1971), "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am. 50, 637-655.

Atal, B.S. and Schroeder, M.R. (1975), "Recent Advances in Predictive Coding: Applications to Speech Synthesis," in G. Fant (Ed.) Speech Communication, Uppsala, Sweden: Almqvist and Wiksell, Vol. I, 27-31. [Reprinted in Markel, J.D. and Gray, A.H. (1976), Linear Prediction of Speech, New York: Springer-Verlag, 188-189.]

Bladon, A. (1982), "Arguments against Formants in the Auditory Representation of Speech", in R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System, Amsterdam: Elsevier Biomedical Press, 95-102.

Carlson, R., Granstrom, B. and Fant, G. (1970), "Some Studies Concerning Perception of Isolated Vowels", Speech Transmission Laboratories Quarterly Progress and Status Report 2-3, Royal Institute of Technology, Stockholm, 19-35.

Carlson, R., Fant, G., and Granstrom, B. (1975), "Two-Formant Models, Pitch, and Vowel Perception", in G. Fant and M.A.A. Tatham (Eds.), Auditory Analysis and Perception of Speech, New York: Academic Press, 55-82.

Carlson, R. and Granstrom, B. (1982), "Towards an Auditory Spectrograph," in R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System, Amsterdam: Elsevier Biomedical.

Chistovich, L.A. (1971), "Problems of Speech Perception," in L.L. Hammerich, R. Jakobson and E. Zwirner (Eds.), Form and Substance, Copenhagen: Akademisk Forlag, 83-93.

Chistovich, L.A., Sheikin, R.L., and Lublinskaja, V.V. (1979), "Centers of Gravity and Spectral Peaks as Determinants of Vowel Quality", in B. Lindblom and S. Ohman (Eds.), Frontiers of Speech Communication Research, London: Academic, 143-158.

Darwin, C.J. and Gardner, R.B. (1985), "Which Harmonics Contribute to the Estimation of First Formant Frequency?", Speech Communication 4, 231-235.

Delgutte, B. (1984), "Speech Coding in the Auditory Nerve II: Processing Schemes for Vowel-Like Sounds", J. Acoust. Soc. Am. 75, 879-886.

Fant, G. (1985), "The Voice Source: Theory and Acoustic Modeling", in I.R. Titze and R.C. Scherer (Eds.), Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control, xx.

Fant, G. and Ananthapadmanabha, T.V. (1982), "Truncation and Superposition", Speech Transmission Labs QPSR 2-3, Royal Institute of Technology, Stockholm, 1-17.

Fant, G. and Liljencrants, J. (1962), "How to Define Formant Level: A Study of the Mathematical Model of Voiced Sounds," Speech Transmission Labs QPSR-2, Stockholm, Sweden: Royal Institute of Technology, 1-8.

Fant, G., Lin, Q.G. and Gobl, C. (1985), "Notes on Glottal Flow Interaction," Speech Transmission Labs QPSR 2-3, Royal Institute of Technology, Stockholm, 21-45.

Fintof, K., Lindblom, B. and Martony, J. (1962), "Measurements of Formant Level in Human Speech," Speech Transmission Labs QPSR-2, Stockholm, Sweden: Royal Institute of Technology, 9-17.

Flanagan, J.L. (1955), "A Difference Limen for Vowel Formant Frequency", J. Acoust. Soc. Am. 27, 613-617.

Floren, A. (1979), "Why Does [aa] Change to [ao] when FO is Increased?", PERILUS I, Institute of Linguistics, Univ. Stockholm, 13-23.

Fujisaki, H. and Kawashima, T. (1968), "The Roles of Pitch and Higher Formants in the Perception of Vowels," IEEE Trans. AU-16, 73-77.

Itakura, F. (1975), "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans ASSP-23, 57-72.

Hawkins, S. and Stevens, K.N. (1985), "Acoustic and Perceptual Correlates of the Non-Nasal/Nasal Distinction for Vowels," J. Acoust. Soc. Am. 77, 1560-1575.

Klatt, D.H. (1980), "Software for a Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am. 67, 971-995.

Klatt, D.H. (1982), "Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step", Proc. ICASSP-82, 1278-1281.

Klatt, D.H. (1985), "The Perceptual Reality of a Formant Frequency," J. Acoust. Soc. Am. 78, S81 (A).

Kopek, G.E. (1985), "Formant Tracking Using Hidden Markov Models," ICASSP-85, 1113-1116.

Lindblom, B. (1962), "Accuracy and Limitations of Sonagraph Measurements," Proc. 4th Int. Congr. Phonetic Sci., The Hague: Mouton, 188-202.

Makhoul, J. (1975), "Spectral Linear Prediction: Properties and Applications," IEEE Trans ASSP-23 283-296.

Markel, J.D. (1972), "Digital Inverse Filtering: A New Tool for Formant Trajectory Estimation," IEEE Trans AU-20, 129-137.

Mermelstein, P. (1978), "Difference Limens for Formant Frequencies of Steady State and Consonant-Bound Vowels", J. Acoust. Soc. Am. 63, 572-580.

Milenkovic, P., (1984), "Model Reference Glottal Inverse Filter of High FO Voice", J. Acoust. Soc. Am. 76, S2 (A).

Miller, R.L. (1953), "Auditory Tests with Synthetic Vowels," J. Acoust. Soc. Am. 25, 114-121.

Plomp, R. (1970), "Timbre as a Multidimensional Attribute of Complex Tones," in R. Plomp and G. Smoorenburg (Eds.), Frequency Analysis and periodicitY Detection in Hearing, Leiden: Sijthoff, 397-411.

Sachs, M.B., Young, E.D. and Miller, M.I. (1982), "Encoding of Speech Features in the Auditory Nerve, in R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System, Amsterdam: Elsevier Biomedical, 115-130.

Seneff, S. (1984), "A Synchrony Model for Auditory Processing of Speech", in J. Perkell and D.H. Klatt, (Eds.) Variability and Invariance of Speech Processes, Hillsdale, NJ: Erlbaum, xx-xx.

Schwartz, R.M. (1971), "Automatic Normalization for Recognition of Vowels of All Speakers", S.B. Thesis, MIT, Cambridge.

Slawson, A.W. (1968), "Vowel Quality and Musical Timbre as Functions of Spectrum Envelope and Fundamental Frequency", J. Acoust. Soc. Am. 43, 87-101.

Srulovicz, P. and Goldstein, J.L. (1983), "A Central Spectrum Model: A Synthesis of Auditory Nerve Timing and Place Cues in Monaural Communication of Frequency Spectrum", J. Acoust. Soc. Am. 73, 1266-1276.

Syrdal, A.K. (1985), "Aspects of a Model of the Auditory Representation of American English Vowels", Speech Communication 4, 121-135.

Traunmuller, H. (1982), "Perceptual Dimension of Openness in Vowels", J. Acoust. Soc. Am. 69, 1465-1475.