VOWEL RECOGNITION BASED ON "LINE-FORMANTS" DERIVED FROM AN AUDITORY-BASED SPECTRAL REPRESENTATION*

Stephanie Seneff

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

A new approach to vowel recognition is described, which begins by reducing a spectrographic representation to a set of straight-line segments that collectively sketch out the formant trajectories. These "line-formants" are used for recognition by scoring their match to a set of histograms of line-formant frequency distributions determined from training data for the 16 vowel categories in the recognition set. Speaker normalization is done by subtracting F_0 from line-formant frequencies on a Bark scale. Although the formants are never enumerated or tracked explicitly, the frequency distributions of the formants are the main features influencing the recognition score. Recognition results are given for 2135 vowels extracted from continuous speech spoken by 292 male and female speakers.

INTRODUCTION

The formant frequencies are probably the most important information leading to the recognition of vowels, as well as other sonorant and even possibly obstruent sounds. Therefore, researchers have spent a considerable amount of effort designing robust formant trackers, which attempt to associate peaks in the spectrum with formant frequencies, using continuity constraints to aid in the tracking of the formants. Once the formant tracks are available, it then becomes possible to identify directions and degree of formant movements, features that are important in recognizing diphthongs, semivowels, and place of articulation of adjacent consonants.

It is impossible to design a "perfect" formant tracker. The most serious problem with formants is that when they are wrong there are often gross errors. Therefore, we have decided to adopt a somewhat different approach, one that can lead to information about formant movements without explicitly labelling the formant numbers. The method also collapses the two stages of formant tracking and track interpretation (e.g., "rising formant") into a single step. The outcome is that a spectrographic representation is reduced to a skeleton sketch consisting of a *set* of straight-line segments, which we call "line-formants," that collectively trace out the formant tracks. The recognition strategy then involves matching all of the line-formants of an unknown segment to a set of templates, each of which describes statistically the appropriate line-formant configurations for a given phonetic class (which could be as detailed as "nasalized /æ/" or as general as "front vowel"). Usually the number of line-formants for a given speech segment is considerably larger than the number of formants, because in many cases several straight-line segments are required to adequately reflect the transitions of a single formant.

*This research was supported by DARPA under Contract N00039-85-C-0254, monitored through Naval Electronic Systems Command.

SIGNAL PROCESSING

Spectral Representation

The system makes use of two spectrogram-like representations that are based on our current understanding of the human auditory system. These have been described in detail previously [1,2], and will only be discussed briefly here. The analysis system consists of a set of 40 critical band filters, spanning the frequency range from 160 to 6400 Hz. The filter outputs are processed through a nonlinearity stage that introduces such effects as onset enhancement, saturation and forward masking. The outputs of this stage are then processed through two independent analyses, each of which produces a spectrogram-like output. The "Mean Rate Spectrogram" is related to mean rate response in the auditory system, and is used for locating sonorant regions in the speech signal. The "Synchrony Spectrogram" takes advantage of the phase-locking property of auditory nerve fibers. It produces spectra that tend to be amplitude-normalized, with prominent peaks at the formant frequencies. The amplitude of each spectral peak is related to the amount of energy at that frequency *relative* to the energy in the spectral vicinity. The line-formant representation is derived from this Synchrony Spectrogram.

Line-formant Processing

The line-formants are obtained by first locating sonorant regions, based on the amount of low frequency energy in the Mean Rate Spectrogram. Within these sonorant regions, a subset of robust peaks in the Synchrony Spectrogram is selected. Peaks are rejected if their amplitude is not sufficiently greater than the average amplitude in the surrounding time-frequency field. For each selected peak, a short fixed-length line segment is determined, whose direction gives the best orientation for a proposed formant track passing through that peak, using a procedure as outlined in Figure 1. The amplitude at each point on a rectangular grid within a circular region surrounding the peak in question is used to update a histogram of amplitude as a function of the angle, θ. Typical sizes for the circle radius are 20 ms in time and 1.2 Bark in frequency. The maximum value in the histogram defines the amplitude and corresponding θ for the proposed track, as marked by an arrow in Figure 1*c*.

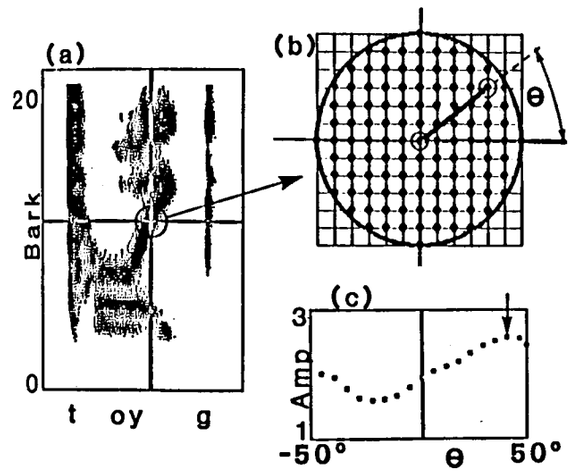

Figure 1: Schematic illustration of process used to determine an orientation for a formant passing through a peak. (a) Synchrony Spectrogram with cross-bars indicating a referenced peak. (b) Schematic blow-up of region around the peak, outlining procedure to generate a histogram of amplitude as a function of angle. (c) Resulting histogram for the example in part *a*.

At each time frame several new short segments are generated, one for each robust spectral peak. A short segment is then merged with a pre-existing partial line-formant whenever the two lines have a similar orientation, and the distance between each endpoint and the other line is sufficiently small. The merging process is accomplished by creating a weighted-average line-formant that incorporates the new line. If a given new segment is sufficiently unique, it is entered as a new partial line-formant.

The resulting *Skeleton Spectrogram* for the /ɑ/ in the word "shock" is illustrated in Figure 2*a*, along with a *Schematized Spectrogram* in Figure 2*b*, included to facilitate visual evaluation. The latter is constructed by replacing each line-formant with a time sequence of Gaussian-shaped spectral peaks with amplitude equal to the line's amplitude. The corresponding Synchrony Spectrogram is shown in Figure 2*c*, with line-formants superimposed. For direct comparison, Figure 2*d* shows a Synchrony Spectral cross section at the time of the vertical bar, on which is superimposed a cross section of the Schematized Spectrogram. For this example, we see that peak locations and amlitudes in the vowel are accurately reflected. In addition, formant transitions appropriate for the palatal fricative on the left and the velar stop on the right are also captured.

RECOGNITION EXPERIMENT

Thus far, we have focused our studies on speaker-independent recognition for 16 vowels and diphthongs of American English in continuous speech, restricted to obstruent and nasal context. The semivowel context is excluded because we believe that in many cases vowel-semivowel sequences should be treated as a single phonetic unit much like a diphthong.

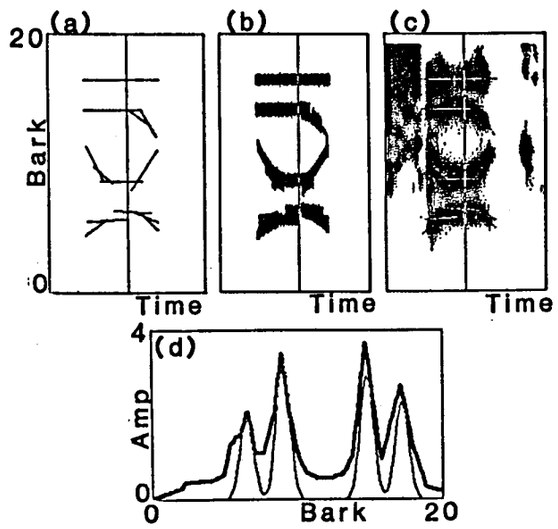

Figure 2: Sample line-formant outputs: (a) Skeleton Spectrogram for word "shock," (b) Corresponding Schematized Spectrogram, (c) Synchrony Spectrogram with line-formants superimposed, (d) cross-sections from *b* and *c* at the cursor, superimposed.

Speaker Normalization

Our first task was to devise an effective speaker-normalization procedure. Many investigators have noted the strong correlation between formant frequencies and F_0 [3]. The relationship is clearly nonlinear – the second formant for female /i/ is higher on average by several hundred Hz, whereas the F_0 difference is on the order of 100 Hz. However, on a Bark (critical band) scale the male-female difference in F_2 for /i/ becomes much more similar to that in F_0. Thus we decided to try a very simple scheme – for each line-formant, subtract from the line's center frequency the *median* F_0 over the duration of the line, on a Bark scale.

We found this normalization procedure to be remarkably effective, as illustrated in Figure 3. Part *a* shows a histogram of the center frequencies of all of the lines for 35 male and 35 female /æ/ tokens. Part *b* shows the same data, after median F_0 has been subtracted from each line's center frequency. The higher formants emerge as separate entities after the F_0 normalization. The normalization is not as effective for F_1, but the dispersal in F_1 is due in part to other factors such as vowel nasalization. A valid question to ask is the following: if it is supposed that speaker normalization can be accomplished by subtracting *a factor times* F_0 from all formant frequencies, then what should be the numerical value of the factor? An answer can be obtained experimentally using autoregressive analysis. We defined $F_n' = F_n - \alpha F_0$ to be the normalized formant frequency for each line. Using vowels for which the formants are well separated, we associated a group of lines with a particular formant such as F_2. The goal was to minimize total squared error for each remapped formant among all speakers, with respect to α. The resulting estimated value for α was 0.975, providing experimental evidence for the validity of the proposed scheme.

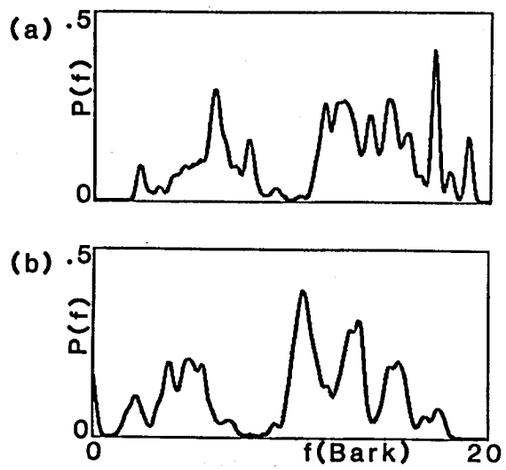

Figure 3: Histograms for center frequencies of all line-formants for 35 female and 35 male tokens of /æ/, (a) without F_0 normalization, and (b) with F_0 normalization.

Scoring Procedures

Our goal in developing a recognizer for the vowels was to emphasize the formant frequency information without ever explicitly identifying the formant numbers. We wanted to avoid traditional spectral template-matching schemes, because they depend too heavily on irrelevant factors such as the loudness or the overall spectral tilt. On the other hand, we did not want to specify, for example, the distance between F_2 and a target F_2, because this relies on accurately enumerating the formants.

We decided to construct histograms of frequency distributions of spectral peaks across time, based on data derived from the line-formants. The scoring amounts to treating each histogram as a probability distribution, and matching the unknown token's line-formants against the appropriate distributions for each vowel. To construct the histograms for a given vowel, all of the line-formants in a training set were used to generate *five* histograms intended to capture the distributions of the formants at significant time points in the vowel. All lines were normalized with respect to F_0, which was computed automatically using a version of the Gold-Rabiner pitch detector [4]. Each line-formant's contributions to the histograms were weighted by its amplitude and its length.

Only left, center and right frequencies of the lines were used in the histograms. The left frequency of a given line-formant falls into one of two bins, depending upon whether or not it is near the beginning of the vowel. Right frequencies are sorted similarly, with a dividing point near the end of the vowel. Center frequencies are collected into the same histogram regardless of their time location. Such a sorting process results in a set of histograms that reflects general formant motions over time. For example, the F_2 peak in the histograms for /e/ shifts upward from left-on-left to center to right-on-right, reflecting the fact that /e/ is diphthongized towards a /y/ off-glide, as illustrated in Figure 4.

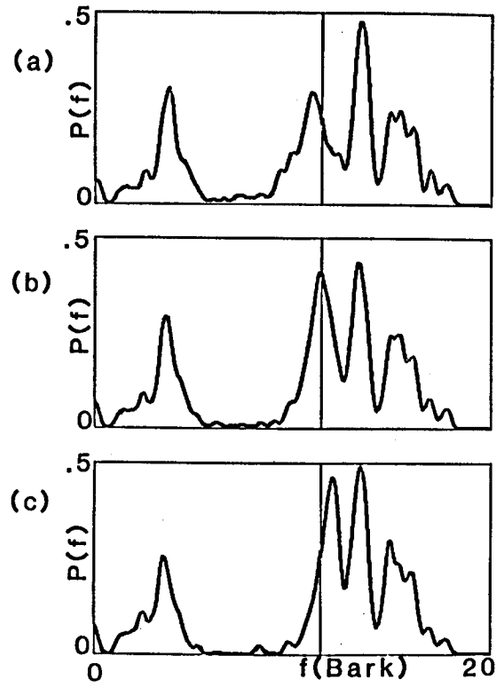

Figure 4: Histograms for (a) left-on-left, (b) center, and (c) right-on-right line-formant frequencies for 128 tokens of /e/, F_0 normalized.

2135 Vowels, 288 Speakers

u	i	ɪ	e	ɛ	æ	aʷ	aʸ	ɑ	ʌ	ɔ	ɔʸ	o	ʊ	u	ɝ
90	220	268	128	153	155	131	92	103	147	156	96	83	114	96	103

Table 1: Distributions of vowels in recognition experiment

To score an unknown token, the left, center, and right frequencies of all of its lines are matched against the appropriate histograms for each vowel category, which are treated as probability distributions. The score for the token's match is the weighted sum of the log probabilites for the five categories for all of the line-formants. The amplitude of the line does not enter into the match, but is used only as a weight for the line's contribution to the score. This strategy eliminates the problem of mismatch due to factors such as spectral tilt or overall energy.

Recognition Results

The vowels used for recognition were extracted from sentences in the TIMIT database [5]. The speakers represented a wide range of dialectical variations. A total of 2135 vowel tokens spoken by 206 male and 82 female speakers were used as both training and test data, using a jackknifing procedure. The distributions of vowels are shown in Table 1. Each speaker's vowel tokens were scored against histograms computed from all of the line-formants *except* those from that speaker. The scoring procedure was as discussed above, with histograms defined for sixteen vowel categories. The endpoints for the vowels were taken from the time-aligned phonetic transcription.

	u	i	ɪ	e	ɛ	æ	aʷ	aʸ	ɑ	ʌ	ɔ	ɔʸ	o	ʊ	u	ɝ
u	49	15	11	6	3	1								4	8	3
i	11	70	5	8	1					1			1	2	1	
ɪ	11	10	32	16	11	5				2				6	5	2
e	5	9	5	60	7	6	1	1		1		3		2		
ɛ	3	1	12	14	37	16	1	1		6			2	3	1	4
æ	1	1	1	10	9	59	4	7		4		1		1	2	1
aʷ					2	13	39	6	13	7	7	2	7	3		2
aʸ					1	7	4	58	15	2	1	8	2	2		
ɑ						3	7	15	40	4	27		5			
ʌ	1	1	1		7	6	2	5	17	39	2	2	5	6	5	1
ɔ							3	3	29	1	46	4	12	1	1	1
ɔʸ						1	1	10	1		3	67	8		6	2
o						4	7	1	5	6	14	4	53		5	1
ʊ	20	1	11		4	1			1	5	3	3	9	28	11	4
u	11	2	2	2	1	1	1	1	2	4	1	2	9	17	40	3
ɝ	8	.	4	3	3		1	1	1	3	1	1	5	5	3	62

Table 2: First choice confusion matrix for the vowels
Row = Labeled Category, Column = Recognized Category.

A matrix of first-choice confusion probabilities is given in Table 2, in terms of percent correct in the phonetic category. For the most part, confusions are reasonable. We feel encouraged by this performance, especially considering that multiple dialects and multiple contexts are included in the same histogram.

Figure 5 summarizes recognition performance in terms of percentage of time the correct answer is in the top N, for all speakers, and for male and female speakers separately. Recognition was somewhat worse for females, who represented only 25% of the population. Also shown are the recognition results for female speakers when the F_0-normalization scheme is omitted, both in collecting the histograms and in scoring. Significant gains were realized as a consequence of the normalization. The performance for the male speakers without F_0 normalization however (not shown) did not change.

FUTURE PLANS

We believe that recognition performance can be improved by extensions in several directions. One is to divide each vowel's histograms into multiple subcategories, based on both general features of the vowel and coarticulation effects. General categories, useful for the center-frequency histogram, would include "nasalized," "Southern accent," or "fronted." Left- and right-context place of articulation, such as "velar," could be used to define corresponding histogram subcategories. We also plan to explore an alternative recognition strategy for explicitly matching each *line-formant* against a set of template *line-formants* describing a particular phonetic category, instead of reducing the line to three "independent" points. We believe that such an approach will better capture the fact that a given left frequency and a given right frequency are connected. Finally, we plan to gradually expand the scope of the recognizer, first to vowels in all contexts and then to other classes such as semivowels.

REFERENCES

[1] Seneff, S. (1986) "A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research," ICASSP Proceedings, Tokyo, Japan, 37.8.1-37.8.4.

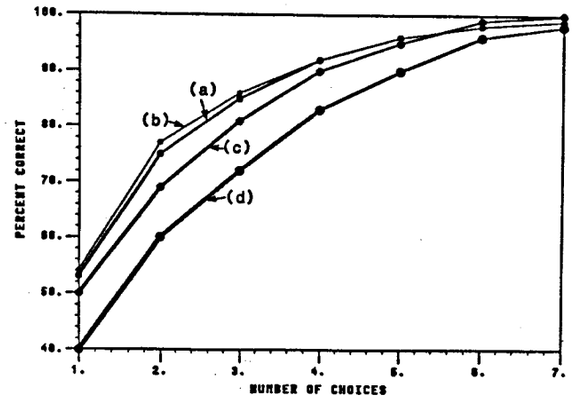

Figure 5: Recognition results expressed as percent of time correct choice is in top N, for the following conditions: (a) all speakers, (b) males only, (c) females only, and (d) females without F_0 normalization.

[2] Seneff, S. (1988) "A Joint Synchrony/Mean-rate Model of Auditory Speech Processing," Journal of Phonetics, Special Issue on Representation of Speech in the Auditory Periphery, to appear in Jan.

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

[4] Gold, B. and L.R. Rabiner (1969) "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am. 46, 442-448.

[5] Lamel, L. F., R. H. Kassel, and S. Seneff (1986) "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," Proceedings of the DARPA Speech Recognition Workshop Palo Alto, CA., Feb 19-20, 100-109.