

Identification of steady-state vowels synthesized from the Peterson and Barney measurements

James Hillenbrand

Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, Michigan 49008-3825

Robert T. Gayvert

RIT Research Corporation, 75 Highpower Road, Rochester, New York 14623

(Received 28 October 1992; revised 26 February 1993; accepted 18 April 1993)

The purpose of this study was to determine how well listeners can identify vowels based exclusively on static spectral cues. This was done by asking listeners to identify steady-state synthesized versions of 1520 vowels (76 talkers \times 10 vowels \times 2 repetitions) using Peterson and Barney's measured values of $F0$ and $F1-F3$ [J. Acoust. Soc. Am. **24**, 175-184 (1952)]. The values for all control parameters remained constant throughout the 300-ms duration of each stimulus. A second set of 1520 signals was identical to these stimuli except that a falling pitch contour was used. The identification error rate for the flat-formant, flat-pitch signals was 27.3%, several times greater than the 5.6% error rate shown by Peterson and Barney's listeners. The introduction of a falling pitch contour resulted in a small but statistically reliable reduction in the error rate. The implications of these results for interpreting pattern recognition studies using the Peterson and Barney database are discussed. Results are also discussed in relation to the role of dynamic cues in vowel identification.

PACS numbers: 43.71.Es

INTRODUCTION

It is well known that a listener's impression of vowel quality is strongly correlated with the frequencies of the two or three lowest formants. The most widely cited study in this area was conducted by Peterson and Barney (1952, hereafter PB), who recorded two repetitions of ten vowels in /hVd/ context spoken by 33 men, 28 women, and 15 children (1520 signals). Acoustic measurements from narrow-band spectra consisted of formant frequencies ($F1-F3$), formant amplitudes, and fundamental frequency ($F0$). The measurements were taken at a single time slice that was considered to be "steady state." The /hVd/ signals were also presented to a large panel of listeners for identification.

The results of the measurement study showed a strong relationship between the intended vowel and the formant-frequency pattern. However, there was considerable formant-frequency variability from one speaker to the next, and there was a substantial degree of overlap in the formant-frequency patterns among adjacent vowels. The listening study showed that the vowels were highly identifiable: The overall error rate was 5.6%, and nearly all of the errors involved confusions between adjacent vowels.

Several attempts have been made to classify vowels in the PB database using various transforms of the spectral measurements. For example, Nearey *et al.* (1979) reported 81.0% classification accuracy using a linear discriminant classifier that was trained on log-transformed $F1$ and $F2$. The addition of log $F0$ and log $F3$ to the parameter set improved classification accuracy to 86.0%. Nearey *et al.* also reported that the addition of speaker-dependent normalization information greatly improved classification ac-

curacy. For example, subtraction of individual speakers' mean log $F1$ and mean log $F2$ from log-transformed $F1$ and $F2$ values of each vowel resulted in 92.0% classification accuracy (see also Gerstman, 1968; Nearey, 1978; Watrous, 1993).

Syrdal and Gopal (1986) reported 81.8% classification accuracy using a linear discriminant classifier trained on $F0$ and $F1-F3$ in Hz. Classification accuracy improved to 85.7% when the pattern classifier was trained on three bark-transformed spectral differences: $F3-F2$, $F2-F1$, and $F1-F0$ (see also Miller, 1984, 1989). However, a recent study by Hillenbrand and Gayvert (1993) used a quadratic discriminant classifier and found no advantage in category separability for nonlinear transforms such as bark, log, mel, and Koenig scales over linear frequency in Hz. There was also no improvement in classification accuracy for spectral differences as compared to absolute frequencies.

The pattern recognition results described above indicate that there are several relatively simple parameter sets that are capable of classifying the PB vowels with about 86% accuracy without the use of speaker-dependent normalizing information. It is quite important to note, however, that the performance of even the best speaker-independent parameter sets is significantly below that of human listeners. Listeners in the original PB study identified vowels with 94.4% accuracy. Since each listening session in the PB study involved random presentations from ten different talkers, it would seem that opportunities for talker normalization of the type described by Ladefoged and Broadbent (1957), Gerstman (1968), and others should have been minimal. To our knowledge, the accuracy shown by PB's listeners has not been approached by

any automatic classification algorithm using the PB data that does not make use of speaker-dependent normalizing information.

One crucial difference between the PB listening experiment and automatic classification studies of the type described above is that the human listeners had access to dynamic information (duration and spectral change), while the information provided to the classifiers consisted of the F_0 and formant pattern sampled at a single time slice. Potential limitations of this static approach were expressed in an early account of the PB data by Potter and Steinberg (1950):

It should be noted ... that we are representing a vowel by a single spectrum taken during a particular small time interval in its duration. Actually, a vowel in the word ... undergoes transitional movements from initial to final consonant ... [The] ear in identifying the word has the benefit of all the changes.

Considerable evidence has accumulated indicating that dynamic properties such as spectral change and duration play an important role in vowel perception. For example, several studies have shown very high identification rates for "silent center" stimuli—signals consisting of onglides and offglides only, with the vowel nuclei replaced with silence (e.g., Jenkins *et al.*, 1983; Nearey, 1989). These experiments imply that listeners do not rely on static targets since high intelligibility was maintained when the most stationary portion of the signal was removed. Results from Nearey and Assmann (1986) suggest that vowel identification is influenced by "vowel inherent spectral change"—the pattern of spectral change throughout the course of the vowel. Listeners were presented with brief excerpts of naturally produced vowels that were excised from "nucleus" and "offglide" portions of the signals. The excerpts were presented under three conditions: (1) natural order (nucleus followed by offglide), (2) repeated nucleus (nucleus followed by itself), and (3) reverse order (offglide followed by nucleus). Identification error rates for the natural-order signals were comparable to those for the original, unmodified vowels, while the repeated-nucleus and reverse-order conditions produced much higher error rates. Nearey and Assmann also trained a pattern recognition model with spectrum change measurements from isolated vowels. Confusion matrices produced by the pattern recognizer correlated strongly with confusion matrices produced by human listeners.

Several experiments have also examined the role of duration in vowel identification. For example, Ainsworth (1972) synthesized two-formant vowels with formant values covering the English vowel space. The vowels varied in duration from 120 to 600 ms. Results indicated that listeners were influenced by duration in a manner that was generally consistent with observed durational differences among vowels (e.g., spectrally similar vowels in the /u/-/ʊ/ region tended to be heard as /u/ at long durations and /ʊ/ at short durations). Similar results have been reported by Tiffany (1953), Bennett (1968), and Stevens (1959). Other experiments, however, have produced more equivo-

cal findings. For example, Huang (1986) presented listeners with nine-step continua contrasting a variety of spectrally similar vowel pairs at durations from 40–235 ms. While the expected duration-dependent boundary shifts occurred, duration differences much larger than those observed in natural speech were needed to move the boundaries. For duration differences that approximate those found in natural speech, boundary shifts were small or nonexistent. Huang also reported unexpected boundary shifts for lax-lax pairs (e.g., /ɪ/-/ɛ/) that do not differ in duration (Black, 1949).

Somewhat equivocal duration results were also reported by Strange *et al.* (1983). Listeners were presented with three kinds of silent center stimuli (1) durational information retained (i.e., onglides and offglides separated by an amount of silence equal to the duration of the original vowel nucleus), (2) durational information neutralized by setting the silent intervals for all stimuli equal to the shortest vowel nucleus, and (3) durational information neutralized by setting the silent intervals for all stimuli equal to the longest vowel nucleus. Results were mixed: Shortening the silent interval to match the shortest vowels did not increase error rates relative to the natural duration condition, but lengthening the intervals to match the longest vowels produced a significant increase in error rates. The authors speculated that the results for the lengthened signals may have been "... due to the disruption of the integrity of the syllables, rather than misinformation about vowel length; that is, subjects may not have perceived a single syllable with a silent gap in it, but instead, heard the initial and final portions as two discrete utterances" (Strange, 1989, p. 2140).

The purpose of the present study was to determine how well the PB vowels can be identified *based exclusively on static spectral information*. This was done by asking listeners to identify steady-state, synthesized versions of each stimulus in the PB database. The motivation for this study was based in part on the central role that the PB database has played in the evaluation of vowel recognition models. Pattern recognition studies such as Syrdal and Gopal (1986), Nearey *et al.* (1979), and Hillenbrand and Gayvert (1993) showed that there are several relatively parameter sets that allow classification of the PB vowels with accuracy in the 85%–87% range. However, it remains an open question whether human listeners will perform with similar accuracy when they are forced to rely on static spectral information that is equivalent to the data used to train and test the pattern recognition algorithms.

A second, closely related reason for conducting these tests was to gather additional information on the role of dynamic information in vowel identification. In this sense, the present experiment can be thought of as providing a source of information that is complementary to the gating studies of Jenkins *et al.* (1983), Nearey (1989), and others. The gating studies demonstrated the importance of dynamic cues in part by showing that vowel identification remained accurate when the static target was deleted. The present study was designed to determine how much infor-

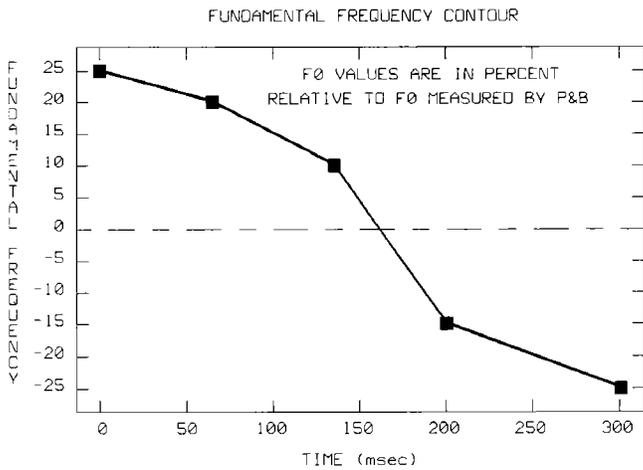


FIG. 1. Fundamental frequency pattern for signals with F_0 movement.

mation is conveyed to listeners when presented with static targets alone.

I. METHODS

A. Stimuli

A formant synthesizer (Klatt, 1980) was used to generate 300-ms, steady-state versions of all 1520 signals in the PB database (Watrous, 1991) using measured values of F_0 and F_1 – F_3 . Formant frequencies were held constant for the full duration of the stimulus. For one set of signals, F_0 was held constant at the value measured by PB. Since pilot results indicated that the error rate for these signals was relatively high, a second set of signals was generated to determine whether any improvement in identifiability would occur if a minimum amount of spectral change was introduced into the static patterns. This was done by generating an additional set of 1520 signals with the falling pitch contour shown in Fig. 1. Fundamental frequency for these signals was initiated 25% above the measured F_0 and terminated 25% below the measured F_0 .

Other synthesis parameters are summarized in Table I. Values of F_4 were set separately for each vowel and talker group based on data from Syrdal (1985). Values of F_5 and F_6 were set separately for each talker group based on data from Rabiner (1968). Formant bandwidths were calculated by a third-order polynomial that was fit to Dunn's (1961) results. Formant amplitudes were set automatically

TABLE I. Synthesis parameters for steady-state vowels synthesized from the Peterson and Barney (1952) measurements.

F_0, F_1 – F_3	Peterson and Barney (1952)
F_4	Syrdal (1985)
F_5, F_6	Rabiner (1968)
Bandwidths	Dunn (1961)
Durations	fixed at 300 ms
Formant amplitudes	adjusted automatically by series resonance synthesizer
Sample frequency	20 kHz
Overall amplitude	all signals scaled to maximum peak amplitude (12 bits)

TABLE II. Percent correct identification of steady-state synthesized versions of the Peterson and Barney vowels with and without F_0 movement. For comparison, identification results are shown for the original stimuli.

Vowel	Flat F_0	F_0 contour	Original stimuli
IY ("heed")	96.2	95.4	99.9
IH ("hid")	67.0	76.8	92.9
EH ("head")	65.8	60.9	87.7
AE ("had")	63.2	64.2	96.5
AH ("hod")	55.0	51.0	87.0
AW ("hawed")	67.2	71.6	92.8
OO ("hood")	62.0	72.8	96.5
UW ("who'd")	89.1	84.6	99.2
UH ("hud")	74.7	79.0	92.2
ER ("heard")	86.6	91.7	99.7
Total:	72.7	74.8	94.4
Men:	74.4	76.9	^a
Women:	72.2	73.8	^a
Children:	70.0	72.1	^a

^aPeterson and Barney did not report results separately for men, women, and child talkers.

by running the synthesizer in cascade mode. The stimuli were synthesized with a 20-kHz sample frequency and 12 bits of amplitude resolution. All signals were scaled to maximum peak amplitude and were ramped on and off with a 20-ms cosine function.

B. Subjects and procedures

Seventeen phonetically trained subjects served as listeners. As in the original PB listening study, stimuli were presented randomly in blocks of trials, with ten talkers per block. The choice of phonetically trained listeners was motivated by the findings of Assmann *et al.* (1982) indicating that identification errors by untrained listeners are due in large part to the listeners' uncertainty about how to map perceived vowel quality onto orthographic symbols.

Blocks of trials using stimuli with falling F_0 contours were randomly mixed with blocks using stimuli with flat F_0 contours. Stimuli were low-pass filtered at 8 kHz, amplified, and delivered over a single loudspeaker (EPI 100) at a comfortable listening level (80 dBA). Subjects entered their responses on a computer keyboard labeled with phonetic symbols for the ten vowels. Subjects were tested one at a time and were given the opportunity to hear each stimulus as many times as they wished before entering a response.

II. RESULTS

Identification rates for the flat-formant, synthesized vowels are shown in Table II, along with identification results from the original study. It can be seen that the 27.3% and 25.2% error rates for the synthesized signals are much higher than the 5.6% error rate that was reported in the original study. It can also be seen that synthesized stimuli with F_0 movement were identified slightly more accurately than signals with flat F_0 contours. This effect was shown by all 17 listeners. A t test comparing

TABLE III. Confusion matrix for synthesized vowels with flat F_0 contours.

		Listener's response									
		IY	IH	EH	AE	AH	AW	OO	UW	UH	ER
Vowel intended by speaker	IY	96.2	3.1	0.6	0.0
	IH	25.1	67.0	6.7	0.3	0.6	...	0.1	0.1
	EH	1.3	23.7	65.8	7.2	0.4	0.1	0.3	1.1
	AE	0.1	0.6	28.0	63.2	4.0	...	0.3	...	2.0	1.9
	AH	0.2	0.1	55.0	30.5	0.6	0.1	13.6	...
	AW	5.9	67.2	13.0	5.9	8.0	...
	OO	...	0.2	0.1	...	0.1	3.1	62.0	28.4	5.2	0.9
	UW	0.2	0.2	0.7	9.0	89.1	0.7	0.2
	UH	...	0.1	0.9	0.7	12.8	6.8	2.7	0.1	74.7	1.2
	ER	0.3	4.1	4.0	0.3	3.0	0.7	0.9	86.6

overall error rates for the two conditions was highly significant ($t=4.35$, $df=16$, $p<0.001$). However, the magnitude of the effect was only 2.1%, and seemed to vary considerably from one vowel to the next. We have no explanation for the differential effect of F_0 contour across vowels.

It is interesting to note that the error rates for the synthesized signals are substantially higher than the error rates reported in the pattern classification studies that were reviewed previously (e.g., Syrdal and Gopal, 1986; Nearey *et al.*, 1979; Hillenbrand and Gayvert, 1993). As we will discuss in more detail below, this finding emphasizes an important limitation of the use of pattern recognition techniques to test models of vowel perception.

For both sets of synthesized signals, the error rate was lowest for male talkers and highest for child talkers. Although the magnitude of the effect was not especially large, an ANOVA collapsed across both sets of signals showed a significant effect for talker group [$F(2,32)=5.39$, $p<0.01$]. PB did not report identification results separately for men, women, and child talkers, and we are not aware of any large-scale study comparing identification rates for naturally produced vowels across these three groups of talkers.

Error patterns for individual vowels showed some similarities to the original data (e.g., very good performance for /i/), but there were also some differences (e.g., relatively poor performance for /ɜ/). Tables III and IV show the full confusion matrix for the steady-state stimuli; the confusion matrix from the original study is shown in Table

V. Again, there are several points of similarity and some important differences. For example, the confusion matrix for the steady-state vowels shows a large number of confusions between /u/ and /ʊ/, a pattern which was not common in the original study. This difference is almost certainly due to the fact that listeners in the original study had access to durational information and, perhaps, to cues related to the pattern of spectral change over time (e.g., Peterson and Lehiste, 1960; Huang, 1986; Di Benedetto, 1989a,b). The majority of the /u/-/ʊ/ confusions involved hearing /u/ as /ʊ/. A similar tense-lax asymmetry is seen for /i/-/ɪ/ confusions, where /i/ for /ɪ/ responses predominated. These asymmetries might have occurred because the listeners tended to treat the 300-ms signals as long vowels, or it might reveal a bias against reporting lax vowels, which do not occur in isolation in English.

III. DISCUSSION

When considering these results, it is not clear whether to emphasize the 73%–75% correct identification rates or the 25%–27% error rates. On the positive side, it is quite clear that a good deal of phonetic information is preserved in the static spectral cross sections. These static patterns provided enough information for subjects to identify the majority of the stimuli, and to select an adjacent vowel category on most of the trials in which errors occurred. However, the significant increase in error rate compared to

TABLE IV. Confusion matrix for synthesized vowels with F_0 movement. Responses are in percent.

		Listener's response									
		IY	IH	EH	AE	AH	AW	OO	UW	UH	ER
Vowel intended by speaker	IY	95.4	3.7	0.3	0.6
	IH	16.4	76.8	4.6	0.1	1.1	0.5	...	0.4
	EH	0.2	31.1	60.9	5.2	0.5	0.1	0.4	1.5
	AE	0.1	0.6	27.8	64.2	3.1	0.1	0.1	...	2.0	2.0
	AH	0.1	0.1	51.0	33.6	0.6	...	14.7	...
	AW	0.1	...	3.8	71.6	13.6	5.1	5.8	...
	OO	...	0.1	0.1	2.1	72.8	20.3	4.0	0.6
	UW	0.3	14.7	84.5	0.4	...
	UH	...	0.1	0.7	0.2	10.0	5.1	3.6	0.2	79.0	1.1
	ER	0.1	2.9	1.6	0.2	2.0	0.6	0.9	91.7

TABLE V. Confusion matrix from the original Peterson and Barney (1952) study. Responses are in percent.

		Listener's response									
		IY	IH	EH	AE	AH	AW	OO	UW	UH	ER
Vowel intended by speaker	IY	99.9
	IH	...	92.9	6.8	0.3
	EH	...	2.5	87.7	9.2	0.5
	AE	2.9	96.5	0.1	0.4
	AH	0.2	87.0	9.9	0.7	...	2.2	...
	AW	5.7	92.8	0.7	...	0.6	0.1
	OO	0.2	0.5	96.5	0.9	1.7	0.2
	UW	0.8	99.2
	UH	5.3	1.2	1.0	...	92.2	0.2
	ER	0.2	99.7

the original signals indicates quite clearly that there is a great deal of information that is missing from the steady-state signals.

A small but significant improvement in identifiability was seen when a falling pitch contour was used with the flat-formant signals, although the effect was not uniform across vowels. One possible explanation for this overall improvement in identifiability is that changes over time in the frequency of individual harmonics serve to increase the definition of the spectrum envelope. A hypothesis along these lines was suggested by Carlson *et al.* (1975), who subsequently failed to find evidence for improved identifiability for signals with very slight movement in *F0* (maximum deviation=4%). Improved definition of the spectrum envelope appears not to be a likely explanation for the *F0* contour effect in the present study. Since the spectrum envelope is more poorly defined at high *F0*, it might be expected that the effect of pitch movement would be the greatest for child talkers and the least for adult male talkers. This expectation was not supported by our results, which showed no difference in the size of the *F0* contour effect for stimuli synthesized from men, women, and child talkers.

Regardless of the precise mechanism underlying the *F0* contour effect, it is important to note that the absolute magnitude of this effect is quite small. This finding suggests that the discrepancy between the identifiability of the synthesized steady-state signals and that of the original signals recorded by PB arises primarily from other sources. The most obvious possibility is that the identification of naturally produced vowels depends heavily on some combination of durational cues and cues related to the pattern of spectral change throughout the course of the utterance, as has been suggested in several previous studies (e.g., Jenkins *et al.*, 1983; Nearey, 1989; Nearey and Assmann, 1986).

While the relatively poor identifiability of the steady-state synthesized signals is almost certainly related to the absence of dynamic cues, there are at least two other factors that may also have played a role. First, the high error rate partly reflects errors in the extraction of formant frequency; that is, it is possible that some signals were misidentified because the formant frequencies were measured incorrectly in the PB study. Errors might either be spectral

(i.e., inaccurate estimates of formant frequencies) or temporal (i.e., related to uncertainties regarding the time at which measurements should be taken—see Di Benedetto, 1989a, for a discussion). It may well be that spectral errors in formant frequency measurement account for the poorer identifiability of the stimuli synthesized from tokens produced by women and children. The wide harmonic spacing of these signals would clearly make formant estimation more difficult, especially in light of the fact that the PB measurements were made from narrow-band spectra.

A second possibility is that the high error rate for the synthesized signals reflects the loss of phonetically relevant information that may occur when power spectra are reduced to formant representations. In other words, it is possible that reducing the original signals to a formant representation resulted in a loss of spectral shape cues that listeners ordinarily rely on for vowel identification. Evidence in support of this view has been presented by Bladon and Lindblom (1981), Bladon (1982), Zahorian and Jagharghi (1986, 1987), and Zahorian and Zhang (1992). However, several other studies have provided convincing evidence indicating that large changes can be made in spectral shape without affecting phonetic quality, as long as the formant frequency pattern is unchanged (e.g., Remez *et al.*, 1981; Klatt, 1982a,b; Hedlin, 1982).

There is some reason to believe that vowels that are stripped of dynamic cues are relatively difficult to identify even when stimuli are naturally produced. In these cases formant-frequency measurement error and the failure to preserve detailed spectral-shape information can be ruled out as possible explanations for high error rates. Fairbanks and Grubb (1961) recorded 1- to 2-s sustained vowels from seven highly trained male speakers. Speakers listened to several repetitions of each vowel and selected two tokens that they judged to be the best examples of the intended vowel quality. Steady-state segments, 300 ms in duration, were excised from these utterances and played to a panel of phonetically trained listeners. The error rate for these signals was 26%, very similar to the error rate reported for the steady-state synthetic signals used in the present study.

The results presented in this study emphasize an important limitation of the PB database to test models of vowel perception. The absence of duration and spectral change information from the PB database makes it impos-

sible to determine how human listeners make use of these dynamic cues in recognizing vowels. The present findings also point out a fundamental limitation of the application of pattern recognition techniques to understanding the perceptual mechanisms underlying phonetic recognition. These approaches clearly provide an important source of information since they indicate the degree to which a given set of features is reliably correlated with phonetic categories. However, as Nearey (1992) noted, these kinds of approaches "... relate only indirectly to perception, since they do not involve comparisons with ... perceptual data ..." (p. 8). Similarly, Shankweiler *et al.* (1977) noted that, "... a successful [classification] algorithm is not a perceptual strategy, only a possible strategy ..." (p. 130). In the context of the present study, this point made clear by the fact that the pattern recognition models performed considerably *better* than human listeners in classifying vowels based on static spectral cues (85%–87% for the pattern recognition models versus 73%–75% for our listeners). This finding indicates quite clearly that the statistically based pattern recognition models are sensitive to certain regularities in the spectral measurements that play little or no role in vowel identification by human listeners. In general terms, the present findings demonstrate the importance of following up pattern recognition studies with an appropriate set of perceptual tests.

ACKNOWLEDGMENTS

This work was supported by the Air Force Systems Command, Rome Air Development Center, Griffiss Air Force Base, and the Air Force Office of Scientific Research (Contract No. F30602-85-C-0008), and by a research grant from the National Institutes of Health (NIDCD 1-R01-DC01661). We are grateful to Eric Luce, Tom Ridley, and Brendon McMahon for their help with software development.

- Ainsworth, W. A. (1972). "Duration as a cue in the recognition of synthetic vowels," *J. Acoust. Soc. Am.* **51**, 648–651.
- Assmann, P., Nearey, T. M., and Hogan, J. (1982). "Vowel identification: Orthographic, perceptual and acoustic aspects," *J. Acoust. Soc. Am.* **71**, 975–989.
- Bennett, D. C. (1968). "Spectral form and duration as cues in the recognition of English and German vowels," *Lang. Speech* **11**, 65–85.
- Black, J. W. (1949). "Natural frequency, duration, and intensity of vowels in reading," *J. Speech Hear. Disord.* **14**, 216–221.
- Bladon, A. (1982). "Arguments against formants in the auditory representation of speech," edited by R. Carlson and B. Granstrom, in *The Representation of Speech in the Peripheral Auditory System* (Elsevier, Amsterdam), pp. 95–102.
- Bladon, A. and Lindblom, B. (1981). "Modeling the judgment of vowel quality differences," *J. Acoust. Soc. Am.* **69**, 1414–1422.
- Carlson, R., Fant, G., and Granstrom, B. G. (1975). "Two-formant models, pitch, and vowel perception," edited by G. Fant and M. A. A. Tatham, in *Auditory Analysis and Perception of Speech* (Academic, London), pp. 55–82.
- Di Benedetto, M-G. (1989a). "Vowel representation: Some observations on temporal and spectral properties of the first formant frequency," *J. Acoust. Soc. Am.* **86**, 55–66.
- Di Benedetto, M-G. (1989b). "Frequency and time variations of the first formant: Properties relevant to the perception of vowel height," *J. Acoust. Soc. Am.* **86**, 67–77.
- Dunn, H. K. (1961). "Methods of measuring formant bandwidths," *J. Acoust. Soc. Am.* **33**, 1737–1745.
- Fairbanks, G., and Grubb, P. (1961). "A psychophysical investigation of vowel formants," *J. Speech Hear. Res.* **4**, 203–219.
- Gerstman, L. (1968). "Classification of self-normalized vowels," *IEEE Trans. Audio Electroacoust.* **AU-17**, 78–80.
- Hedlin, P. (1982). "A representation of speech with partials," edited by R. Carlson and B. Granstrom, in *The Representation of Speech in the Peripheral Auditory System* (Elsevier Biomedical, Amsterdam), pp. 247–250.
- Hillenbrand, J., and Gayvert, R. T. (1993). "Vowel classification based on fundamental frequency and formant frequencies," *J. Speech Hear. Res.* **36** (to be published).
- Huang, C. B. (1986). "The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels," *IEEE-ICASSP*, Tokyo, Japan.
- Jenkins, J. J., Strange, W., and Edman, T. R. (1983). "Identification of vowels in 'vowelless' syllables," *Percept. Psychophys.* **34**, 441–450.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* **67**, 971–995.
- Klatt, D. H. (1982a). "Prediction of perceived phonetic distance from critical-band spectra: A first step," *Proceedings of ICASSP-82*, 1278–1281.
- Klatt, D. H. (1982b). "Speech processing strategies based on auditory models," *The Representation of Speech in the Peripheral Auditory System* (Elsevier Biomedical, Amsterdam), pp. 181–196.
- Ladefoged, P., and Broadbent, D. E. (1957). "Information conveyed by vowels," *J. Acoust. Soc. Am.* **29**, 98–104.
- Miller, J. D. (1984). "Auditory processing of the acoustic patterns of speech," *Arch. Otolaryngol.* **110**, 154–159.
- Miller, J. D. (1989). "Auditory-perceptual interpretation of the vowel," *J. Acoust. Soc. Am.* **85**, 2114–2134.
- Nearey, T. M. (1978). *Phonetic Feature Systems for Vowels* (Indiana University Linguistics Club, Bloomington, IN).
- Nearey, T. M. (1989). "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.* **85**, 2088–2113.
- Nearey, T. M., and Assmann, P. (1986). "Modeling the role of vowel inherent spectral change in vowel identification," *J. Acoust. Soc. Am.* **80**, 1297–1308.
- Nearey, T. M., Hogan, J., and Rozsypal, A. (1979). "Speech signals, cues and features," edited by G. Prideaux, in *Perspectives in Experimental Linguistics* (Benjamin, Amsterdam).
- Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.* **24**, 175–184.
- Peterson, G. E., and Lehiste, I. (1960). "Duration of syllable nuclei in English," *J. Acoust. Soc. Am.* **32**, 693–703.
- Potter, R. K., and Steinberg, J. C. (1950). "Toward the specification of speech," *J. Acoust. Soc. Am.* **22**, 807–820.
- Rabiner, L. (1968). "Digital formant synthesizer for speech synthesis studies," *J. Acoust. Soc. Am.* **24**, 175–184.
- Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. E. (1981). "Speech perception without traditional speech cues," *Science* **212**, 947–950.
- Shankweiler, D., Strange, W., and Verbrugge, R. (1977). "Speech and the problem of perceptual constancy," edited by R. Shaw and J. Bransford, in *Perceiving, Acting, and Comprehending: Toward an Ecological Psychology* (Erlbaum, Hillsdale, NJ), pp. 117–144.
- Stevens, K. N. (1959). "The role of duration in vowel identification," *Q. Progr. Rep.* **52**, Res. Lab. Electron. MIT, Cambridge, MA.
- Strange, W. (1989). "Dynamic specification of coarticulated vowels spoken in sentence context," *J. Acoust. Soc. Am.* **85**, 2135–2153.
- Strange, W., Jenkins, J. J., and Johnson, T. L. (1983). "Dynamic specification of coarticulated vowels," *J. Acoust. Soc. Am.* **74**, 695–705.
- Syrdal, A. K. (1985). "Aspects of a model of the auditory representation of American English vowels," *Speech Commun.* **4**, 121–135.
- Syrdal, A. K., and Gopal, H. S. (1986). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," *J. Acoust. Soc. Am.* **79**, 1086–1100.
- Tiffany W. (1953). "Vowel recognition as a function of duration, frequency modulation and phonetic context," *J. Speech Hear. Disord.* **18**, 289–301.
- Watrous, R. L. (1991). "Current status of the Peterson-Barney vowel formant data," *J. Acoust. Soc. Am.* **89**, 2459–2460.
- Watrous, R. L. (1993). "Speaker normalization and adaptation using

- second-order connectionist networks," *IEEE Trans. Neural Networks* **4**, 21-30.
- Zahorian, S., and Jagharghi, A. (1986). "Matching of 'physical' and 'perceptual' spaces for vowels," *J. Acoust. Soc. Am. Suppl. 1* **79**, S8.
- Zahorian, S., and Jagharghi, A. (1987). "Speaker-independent vowel recognition based on overall spectral shape versus formants," *J. Acoust. Soc. Am. Suppl. 1* **82**, S37.
- Zahorian, S., and Zhang, Z.-J. (1992). "Perception of vowels synthesized from sinusoids that preserve either formant frequencies or global spectral shape," *J. Acoust. Soc. Am.* **92**, 2414 (A).