Journal of Speech and Hearing Research, Volume 36, 694–700, August 1993

Vowel Classification Based on Fundamental Frequency and Formant Frequencies

James Hillenbrand
Western Michigan University, Kalamazoo

Robert T. Gayvert
RIT Research Corporation Rochester, NY

A quadratic discriminant classification technique was used to classify spectral measurements from vowels spoken by men, women, and children. The parameters used to train the discriminant classifier consisted of various combinations of fundamental frequency and the three lowest formant frequencies. Several nonlinear auditory transforms were evaluated. Unlike previous studies using a linear discriminant classifier, there was no advantage in category separability for any of the nonlinear auditory transforms over a linear frequency scale, and no advantage for spectral distances over absolute frequencies. However, it was found that parameter sets using nonlinear transforms and spectral differences reduced the differences between phonetically equivalent tokens produced by different groups of talkers.

KEY WORDS: vowel recognition, speech recognition, speech acoustics, pattern recognition

A problem of longstanding interest in experimental phonetics concerns the information that is used by human listeners in vowel identification. It has long been recognized that perceived vowel quality is strongly correlated with the frequencies of the two or three lowest formants (see Jenkins, 1987; Miller, 1989; Nearey, 1989, Strange, 1989; for reviews). However, it is equally well known that the acoustic properties of vowels vary depending on the individual talker, the rate of speech, and the phonetic context in which the vowel occurs (e.g., Gay, 1978; Lindblom, 1963; Peterson & Barney, 1952; Shankweiler, Strange, & Verbrugge, 1977). One result of these complex mapping relationships between spectral patterns and perceived vowel quality is that a good deal of emphasis has been placed on understanding the normalizing mechanisms that might be involved in adjusting for these talker- and context-dependent variations in spectral patterns.

Numerous models have been proposed for classifying vowels based on some combination of fundamental frequency (F0) and formant frequencies. As Disner (1980) pointed out, the goal of these models has generally been to (a) maximize differences between vowel categories, and (b) minimize differences in the same vowel spoken by different talkers, particularly those differences associated with vocal-tract length variability. Miller and his colleagues (Fourakis & Miller, 1987; Miller, 1984; 1989) proposed a vowel classification scheme based on perceptual "target zones" in a three-dimensional space consisting of log-transformed spectral distances: logF3-logF2, logF2-logF1, and logF1-SR, where SR ("sensory reference") is a transform of the speaker's F0 ($SR = 168/(F0/168)^{1/3}$).[1] Miller (1989) tested his normalization

[1]The figure of 168 Hz in the sensory reference formula is the geometric mean of the fundamental frequency (GMFO) calculated across all talkers in the Peterson and Barney (1952) database. In some applications, Miller (1989) calculates GMFO as a running geometric average of fundamental frequency throughout the course of each individual token. It is important to note that the running average is calculated over the course of the

0022-4685/93/3604-0694$01.00/0

scheme with F0 and formant measurements from 435 utterances that were taken from several different databases (see Appendix B of Miller, 1989, for a description of the vowel database). The classification scheme involved determining whether individual tokens fell within the boundaries of the appropriate target zones. The boundaries of the target zones were drawn by hand for optimal fit to the data. Miller reported 93.0% classification accuracy using this scheme.

A normalization scheme proposed by Syrdal (1985) is very similar to the Miller model except that spectral distances are represented on the critical-band-based bark scale rather than a logarithmic scale. The three parameters in the Syrdal model are B3-B2 (F3 in bark minus F2 in bark), B2-B1, and B1-B0. Syrdal & Gopal (1986) evaluated the model with F0 and formant measurements from the Peterson & Barney (1952) database (76 talkers, 10 vowels, two repetitions of each vowel). Linear discriminant analysis was used to classify each token in the database. Syrdal & Gopal reported significantly better classification performance for a discriminant model that was trained on the three bark spectral distances (85.7% correct classification) than a model trained on the absolute frequencies of F0-F3 in Hz (81.8% correct classification).

The vowel classification results reported by Syrdal & Gopal (1986) address the first of Disner's (1980) two requirements for normalization schemes: maximizing differences between phonetically distinct vowel categories. Syrdal (1985) evaluated the second requirement—minimizing differences in the same vowel spoken by different talkers—by determining how well the discriminant model classified tokens on the basis of talker group (i.e., men vs. women vs. children). Syrdal argued that an ideal normalization algorithm should minimize differences between vowels produced by different talkers. Consequently, a classifier should find it difficult to differentiate among tokens on the basis of talker group. Syrdal's results showed that a linear discriminant classifier was quite good at identifying tokens by talker group when trained on F0-F3 in Hz (89.6% correct). However, performance fell to 41.7% when the classifier was trained on bark spectral differences. Although this is better than the 33.3% that would be expected by chance, it is quite clear that the normalization algorithm greatly reduces differences among tokens produced by different groups of talkers. There is also the possibility that the above-chance performance of the pattern recognizer reflects dialect differences among men, women, and child talkers (Byrd, 1992; Syrdal, 1985).

Several closely related classification schemes were proposed and tested by Peterson (1961) and Nearey (1978; 1992; Nearey, Hogan, & Rozsypal, 1979). As Miller (1989) noted recently, many of these models are variations of "formant ratio" or "relative resonance" theory, a very old idea suggesting that vowels with similar qualities have similar formant ratios (e.g., Lloyd, 1890). A long-recognized weakness of formant ratio theory is that there are many pairs of distinct vowels with similar formant ratios (e.g., /ɑ/-/ɔ/, /u/-/ʊ/). In the Miller and Syrdal models this problem is addressed by including the spectral distance between F0 and F1 as a parameter **Hillendbrand & Gayvert:** *Vowel Classification* (see also Traunmüller, 1981). In this way, a vowel such as /ɑ/ can be distinguished from /ɔ/ by virtue of a greater spectral distance between F0 and F1. Fujisaki & Kawashima (1968) have also suggested that F3 and the upper formants might serve a similar role in distinguishing vowels with similar formant ratios.

The purpose of the present study was to compare several different normalization schemes on the basis of the two criteria proposed by Disner: (a) maximization of vowel-category separability, and (b) minimization of within-vowel-category differences among talkers. To simplify this discussion, we will refer to the first of these criteria as "category separability" and the second as "within-category variability." Comparing different normalization schemes is difficult on the basis of existing literature for several reasons, including (a) not all investigators have made use of the same database of spectral measurements, (b) a variety of techniques have been used to measure category separability, and (c) within-category variability has either not been evaluated, or the methods that have been used to measure variance minimization have varied considerably from one study to the next. For example, direct comparisons between the Miller and Syrdal models is not possible based on existing literature because (a) the databases used for testing the two models are quite different, (b) the linear discriminant analysis technique used by Syrdal (1985; Syrdal & Gopal, 1986) is quite different from the method used by Miller (1989), which relies on hand-drawn boundaries, and (c) no direct comparisons have been made of the within-category variance minimization properties of the two algorithms. The present study was designed to address these problems by evaluating a variety of normalization schemes using the same database, the same pattern-classification technique, and the same method of measuring within-category variability.

Method

The statistical technique that was used for classification was the maximum likelihood distance measure (Johnson & Wichern, 1982). This method is a quadratic version of the linear discriminant analysis technique used by Syrdal (1985) and Syrdal & Gopal (1986). Both methods involve the computation of a variance-normalized distance measure between the feature vector for a given token and the center (or centroid) of each of the training categories. For example, the feature vector might consist of values of F1-F3 for a particular token, and the training categories would consist of statistics (the average feature vector and the covariance matrix) for each of 10 vowel categories. Distances are computed to the centroids of each training category, and the token is assigned to the category that resulted in the shortest distance. Overall classification accuracy is determined simply by calculating the percentage of tokens that were assigned by the classifier to the category that was intended by the talker. The distance measure itself is nothing more than an extension of a *z* score

individual token, not across tokens produced by the talker. This means that the Miller model, like that of Syrdal (1985), is an intrinsic normalization scheme. In the case of the present study, the running average did not come into play since the Peterson and Barney database contains a single measure of fundamental frequency for each token.

into multivariate space. A *z* score, of course, is the difference between an individual score and the mean, divided by the standard deviation. The same logic is used in the distance measure that forms the basis of discriminant analysis, except that (a) the difference between an individual score and the mean is replaced by the difference between a feature vector for an individual token and the average feature vector for the category, and (b) the standard deviation is replaced by the covariance matrix.

The difference between the linear and quadratic versions of this technique is that linear discriminant analysis uses a single covariance matrix that is pooled across all training categories, whereas the quadratic method uses a separate covariance matrix for each training category. On a practical level, the quadratic method offers two potential advantages over the linear method: (a) since separate covariance matrices are used for each training category, the quadratic method is able to account for any differences that might exist from one category to the next in the size or orientation of the category, and (b) since the formula is quadratic, the decision surfaces are curved rather than linear.

The database used for the classification studies consisted of the 1,520 spectral measures from Peterson & Barney (1952) using the data archive described by Watrous (1991). This database consists of F0 and F1-F3 measurements from two repetitions of 10 vowels in /hVd/ context recorded from 33 men, 28 women, and 15 children. The F0 and formant measurements were sampled at times that were judged by Peterson and Barney to be the most steady.

Results

Category Separability

Table 1 shows classification results for several combinations of parameters. Results are shown for linear frequencies in Hz and for several nonlinear transforms, including (a) log frequencies, (b) a bark-scale transform using the formula from Syrdal & Gopal (1986), (c) a mel-scale transform using the technical approximation from Fant (1973), and (d) a Koenig-scale transform (Koenig, 1949). As can be seen in Figure 1, the bark, mel, and Koenig transforms are quite similar to one another. The bark scale is approximately linear below about 500 Hz and approximately logarithmic above 500 Hz, the mel scale is approximately linear below about 1000 Hz and approximately logarithmic above 1000 Hz, and the Koenig scale is exactly linear below 1000 Hz and exactly logarithmic above 1000 Hz. Entries under the column labeled "LOG" used the log of Miller's sensory reference instead of the log of the fundamental frequency for parameter sets including F0. The numbers in Table 1 are overall classification accuracies for each parameter set. Results here and throughout the paper are based on the "jackknife" method in which individual tokens are removed from the training statistics before the distance measures are calculated (Johnson & Wichern, 1982).

TABLE 1. Overall classification accuracy using various combinations of parameters.

Parameter Set	Transform				
	LINEAR	LOG	BARK	MEL	KOENIG
F1, F2	74.9	75.2	76.1	75.5	76.1
F1, F2, F3	83.6	83.6	83.6	83.5	83.6
F0, F1, F2	85.9	84.0	85.0	85.8	85.0
F0, F1, F2, F3	86.6	86.1	86.6	86.6	86.6
F1-F0, F2-F1, F3-F2	85.5	86.2	86.8	85.2	86.8

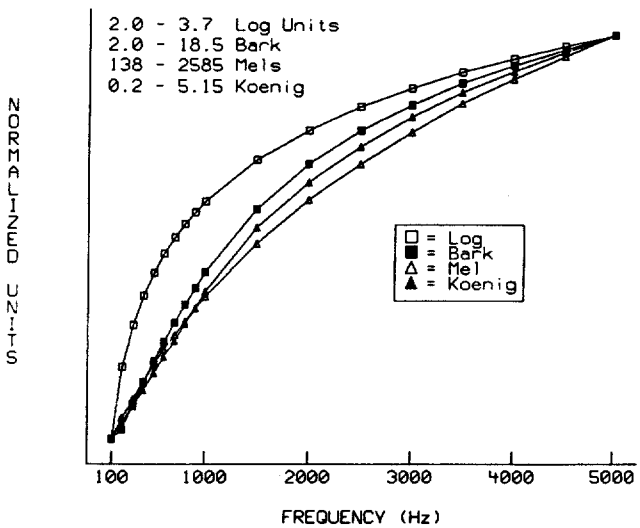

FIGURE 1. Comparison of log, bark, mel, and Koenig scales.

One point that emerges quite clearly from Table 1 is that error rates are relatively high when classification is based on F1 and F2 alone. This was true for linear frequency and for all of the nonlinear auditory transforms. Although the results are not shown in Table 1, we also tested several parameter sets that used F2′ in place of F2, using the technical approximation from Carlson, Fant, and Granstrom, 1975. F2′ is a weighted combination of F2 and higher formants. The F2′ or "effective second formant" concept is based on the suggestion of Delattre, Liberman, Cooper, and Gerstman (1952) that the auditory system averages formants that are relatively close in frequency. Although there is good psychophysical evidence in support the F2′ concept (e.g., Carlson et al., 1975; Chistovich & Lublinskaya, 1979; Chistovich, Sheikin, & Lublinskaya, 1979), our results did not show any improvement in category separability when F2′ was substituted in place of F2.

It can also be seen in Table 1 that the addition of either F0 or F3 to the two lowest formants results in a substantial improvement in performance. These findings would seem to be consistent with the suggestions of Miller (1989) and Fujisaki & Kawashima (1968) regarding the normalizing role of these spectral features. However, the confusion matrices, presented in the Appendix, show an across-the-board improvement in classification performance rather than improvements related exclusively or primarily to vowel pairs with similar formant ratios, such as /ɑ/-/ɔ/ and /u/-/ʊ/.

The last row of entries in Table 1 is for parameter sets that make use of spectral distances rather than absolute frequencies. The second column of log spectral distances is the Miller model and the third column of bark spectral differences

is the model proposed by Syrdal. The main point to be made about these results is that *there was no advantage for any of the nonlinear transforms over a linear frequency scale, and no advantage for spectral distances over absolute frequencies.* These findings are in conflict with Syrdal & Gopal's (1986) linear discriminant classification results, which showed significantly better classification accuracy for bark spectral differences as compared to absolute linear frequencies. Since Syrdal & Gopal also used the Peterson & Barney (1952) database, the discrepancy between the two sets of results is due to the use of a quadratic classification technique in the present study.

The 86.8% classification accuracy for the Miller model is substantially lower than the 93.0% accuracy reported by Miller (1989). Both the database of spectral measurements and the classification method used by Miller differed from the present study, so it is difficult to determine what combination of these two factors accounts for the discrepancy between the two findings.

Collapsed across all parameter sets, classification accuracy was somewhat lower for the child talkers (78.9%) than the men (83.3%) or women (86.5%). A oneway ANOVA for talker group was significant [F (2, 87) = 9.7, $p < 0.01$]. Neuman-Keuls post hoc tests showed that the children differed from both groups of adult talkers. This finding might be due either to an increase in formant frequency measurement error for tokens produced by child talkers, or possibly to a larger number of production errors by the children. Since Peterson & Barney did not report listening test results separately for men, women, and child talkers, it is not clear whether tokens produced by the children were less identifiable than those produced by the adults. We are not aware of any large-scale study that has compared the identifiability of vowels naturally produced by men, women, and children.

Within-Category Variability

The results presented thus far address the category-separability criterion but not the within-category variability criterion. Recall that the method used by Syrdal (1985) to evaluate within-category variability involved the use of a linear discriminant classifier that was trained to recognize talker group rather that vowel category, on the assumption that an optimal classifier should perform close to chance in differentiating vowels spoken by men, women, and children.

Figure 2 shows talker-group classification results for various combinations of parameters using a quadratic classifier. The results indicate that (a) considerably more talker-group information is preserved by absolute frequencies (transformed or untransformed F0-F3) than spectral-distance representations, and (b) more talker-group information is preserved by linear spectral distances than spectral distances represented on some kind of nonlinear auditory scale. Both findings are consistent with Syrdal's (1985) linear discriminant analysis results. Differences among the nonlinear transforms are slight. Overall talker-group classification accuracy is somewhat better in the present study than in Syrdal, which can be attributed to the use of the quadratic classifier.

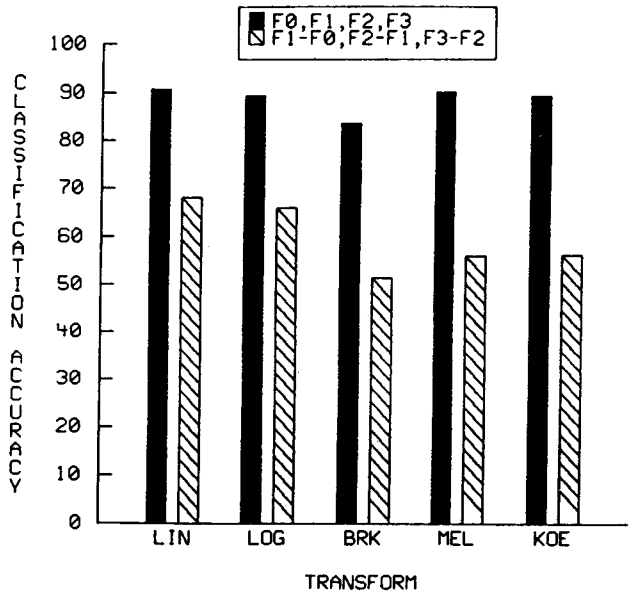

FIGURE 2. Overall accuracy of a quadratic discriminant classifier in identifying talker group (men vs. women vs. children). The classifier was trained on talker-group categories collapsed across all 10 vowels.

It should be noted that the talker-group classification results presented in Figure 2, following the method used by Syrdal, are based on training categories in which all 10 vowels were combined. In other words, the pattern classifier attempted to recognize whether each token was spoken by a man, a woman, or a child *based solely on group statistics collapsed across all vowels.* An alternate approach to this problem would involve separate talker-group training categories for each vowel. In our view, this method gets closer to the within-category variability issue since classification performance under these conditions gives an indication of the degree to which tokens differ as a function of talker group *when each talker is producing the same vowel.*

Figure 3 shows talker-group classification results based on separate man, woman, and child training categories for each vowel. Each of 10 separate classification tests consisted of 66 tokens produced by men, 56 tokens produced by women, and 30 tokens produced by children. The classification rates that are shown in the figure are averages across the 10 vowel categories. As with the previous set of talker-group classification results, these findings indicate that more talker-group information is preserved by the representations that use absolute frequencies than those using spectral differences. Talker-group differences are also larger for linear spectral differences than spectral differences based on the nonlinear transforms. It can also be seen that talker-group classification accuracy is considerably higher for this method than the previous method in which vowel categories were collapsed. Talker-group classification accuracy for the nonlinear spectral difference parameter sets averaged 13.2% higher for the method based on separate vowel training categories than the pooled method. This suggests that a good deal more talker information is preserved by these representations than would be indicated by the results in Figure 2, and by results

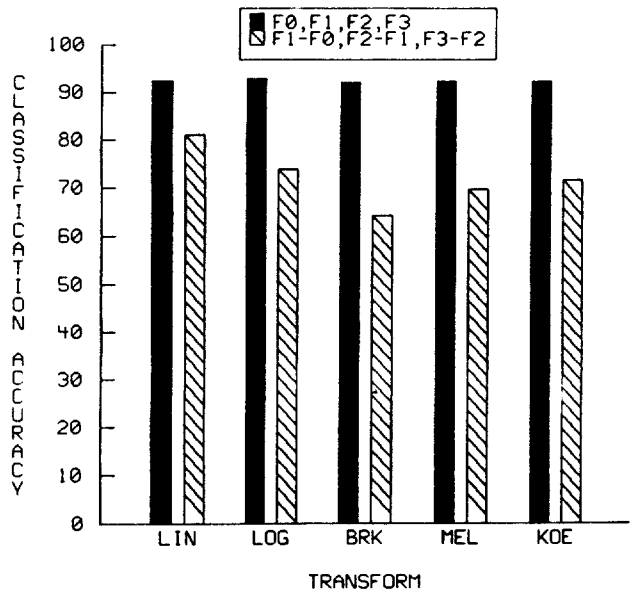

FIGURE 3. Overall accuracy of a quadratic discriminant classifier in identifying talker group (men vs. women vs. children). The classifier was trained on separate talker-group categories for each vowel. Results shown in the figure are averages across the 10 vowels.

reported previously by Syrdal (1985). In our view, the results based on separate training categories for each vowel better reflect the degree of talker-dependent information that is preserved by the normalization schemes. It should be noted, however, that these conclusions are based on the assumption that tokens of a particular vowel produced by men, women, and children are phonetically equivalent. Although specific evidence was not cited, Syrdal (1985) suggested that there may be ". . . linguistically relevant dialectical differences between the speech of men, women, and children" (p. 130). Whether these dialect differences are sufficient to account for the 64–74% talker-group classification rates seen in the present study for the nonlinear spectral difference parameters sets remains an open question.

Discussion

The primary difference between the present findings and those reported previously by Syrdal (1985; Syrdal & Gopal, 1986) is that we found no advantage in category separability for any of the nonlinear auditory transforms over linear frequency, and no advantage in category separability for spectral differences as compared to absolute frequencies. It is quite clear that these nonlinear transforms make much better sense than linear frequency in terms of what is known about the physiology and psychophysics of the auditory system. However, that fact by itself does not mean that these transforms will necessarily solve problems related to vowel separability. We also found very similar performance across the four nonlinear auditory transforms that were tested. This is perhaps not surprising given the considerable similarity of these transforms. However, several other procedures are available for evaluating these transforms. A recent study by Nearey (1992) using a variety of generalized linear modeling techniques found relatively small but consistent advantages for the log scale.

Despite the failure to find advantages in category separability for nonlinear auditory transforms or spectral differences, there were clear advantages for these representations in the reduction of within-category variability because of differences in talker group. Although we found that a good deal more talker-group information is preserved by these representations than was suggested by Syrdal's (1985) findings, it is clear that both auditory transforms and spectral difference representations reduce differences between the same vowel spoken by men, women, and children. Other advantages for these kinds of representations have been noted. For example, Syrdal (1985; Syrdal & Gopal, 1986) showed that high vowels can be separated from mid and low vowels with a high degree of accuracy based on a simple three-bark B1-B0 criterion (see also Traunmuller, 1981), and that front vowels can be separated from back vowels based on a three-bark B3-B2 criterion. These kinds of regularities, which agree well with certain aspects of spectral integration in vowel perception (e.g., Chistovich & Lublinskaya, 1979; Chistovich, et al., 1979), cannot be derived in any simple way from absolute linear frequency representations.

Although results such as those reported here and in previous pattern recognition studies are clearly relevant to vowel perception, there are crucial aspects of this problem that cannot be addressed with this family of techniques. As Nearey (1992) noted, these "data analytic" methods can only measure the degree of correspondence between intended vowels and a particular set of features. As such, these methods can suggest logically possible perceptual strategies, but other information is required to determine whether listeners actually adopt a proposed strategy. It is worth noting that the 13–14% error rates shown by the best of the parameter sets tested in this study are considerably higher than the 5.6% error rate shown by Peterson & Barney's (1952) listeners. There are several plausible explanations for this discrepancy in error rates. One possibility that has received only sporadic attention is that phonetically relevant information is lost when vowel spectra are reduced to formant representations, as has been suggested by Bladon (1982; Bladon & Lindblom, 1981; see also Zahorian & Jagharghi, 1986, 1987). It also seems clear that at least part of this discrepancy is due to the fact that listeners in the original study had access to durational information and the pattern of spectral change throughout the course of the utterance. A significant body of evidence has accumulated suggesting that these dynamic properties play an important role in vowel perception (e.g., Bennett, 1968; DiBenedetto, 1989a; 1989b; Hillenbrand & Gayvert, 1993; Jenkins, Strange, & Edman, 1983; Nearey, 1989; Nearey & Assman, 1986; Stevens, 1959; Tiffany, 1953). Despite this evidence, the specific relations between vowel identification and dynamic cues are not well understood. A significant challenge for future research will be to reach a clearer understanding of the mechanisms that are involved in mapping dynamic spectral cues onto perceived vowel quality.

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).

References

Bennett, D. C. (1968). Spectral form and duration as cues in the recognition of English and German vowels. *Language and Speech, 11,* 65–85.

Bladon, A. (1982). Arguments against formants in the auditory representation of speech. In R. Carlson & B. Granstrom (Eds.), *The representation of speech in the peripheral auditory system* (pp. 95–102). Amsterdam: Elsevier Biomedical Press.

Bladon, A., & Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *Journal of the Acoustic Society of America, 69,* 1414–1422.

Byrd, D. (1992). *Sex, dialects, and reduction. 1992 ICSLP Proceedings,* 827–830.

Carlson, R., Fant, G., & Granstrom, B. C. (1975). Two-formant models, pitch, and vowel perception. In G. Fant & M. A. A. Tathum (Eds.), *Auditory analysis and perception of speech* (pp. 55–82). London: Academic Press.

Chistovich, L. A., & Lublinskaya, V. V. (1979). The 'Center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research, 1,* 185–195.

Christovich, L. A., Sheikin, R. L., & Lublinskaya, V. V. (1979). Centers of gravity and spectral peaks as the determinants of vowel quality. In B. Lindblom & S. Öhman (Eds.), *Frontiers of speech communication research* (pp. 143–158). London: Academic Press.

Delattre, P. C., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel colour: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word, 8,* 195–210.

Di Benedetto, M-G. (1989a). Vowel representation: Some observations on temporal and spectral properties of the first formant. *Journal of the Acoustical Society of America, 86,* 55–66.

Di Benedetto, M-G. (1989b). Frequency and time variations of the first formant: Properties relevant to the perception of vowel height. *Journal of the Acoustical Society of America, 86,* 67–77.

Disner, S. F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America, 76,* 253–261.

Fant, G. (1973). *Speech sounds and features.* Cambridge, MA: MIT Press.

Fourakis, M., & Miller, J. D. (1987). Measurements of vowels in isolation and in sentence context. *Journal of the Acoustical Society of America, (Suppl. 1), 81,* S17 (A).

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Trans. Audio Electroacoustics.* AU-16, 73–77.

Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America, 63,* 223–230.

Hillenbrand, J., & Gayvert, R. T. (in press). Identification of steady-state vowels synthesized from the Peterson and Barney measurements. *Journal of the Acoustical Society of America.*

Jenkins, J. J. (1987). A selective history of issues in vowel perception. *Journal of Memory and Language, 26,* 542–549.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in 'vowelless' syllables. *Perceptual Psychophysiology, 34,* 441–450.

Johnson, R. A., & Winchern, D. W. (1982). *Applied multivariate statistical analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America, 35,* 1773–1781.

Lloyd, R. J. (1890). *Some researches into the nature of vowel-sound.* Liverpool, England: Turner and Dunnett.

Miller, J. D. (1984). Auditory processing of the acoustic patterns of speech. *Archives of Otolaryngology, 110,* 154–159.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85,* 2114–2134.

Miller, R. L. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America, 18,* 114–121.

Nearey, T. M. (1978). *Phonetic feature systems for vowels.* Bloomington, IN: Indiana University Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85,* 2088–2113.

Nearey, T. M. (in press). Applications of generalized linear modeling to vowel data. *Proceedings of the 1992 International Conference on Spoken Language Processing.*

Nearey, T. M., Hogan, J., & Rozsypal, A. (1979). Speech signals, cues and features. In G. Prideaux (Ed.), *Perspectives in experimental linguistics.* Amsterdam: Benjamin.

Nearey, T. M., & Assmann, P. (1986). Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America, 80,* 1297–1308.

Peterson, G., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Peterson, G. E. (1961). Parameters of vowel quality. *Journal of Speech and Hearing Research, 4,* 10–28.

Shankweiler, D., Strange, W., & Verbrugge, R. (1977). Speech and the problems of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and comprehending: Toward an ecological psychology.* Hillsdale, NJ: Lawrence Erlbaum.

Stevens, K. N. (1959). The role of duration in vowel identification. *Quarterly Progress Report 52,* Research Laboratory of Electronics, MIT.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America, 85,* 2135–2153.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America, 74,* 695–705.

Syrdal, A. K. (1985). Aspects of a model of the auditory representation of American English vowels. *Speech Communication, 4,* 121–135.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory represenation of American English vowels. *Journal of the Acoustical Society of America, 79,* 1086–1100.

Tiffany, W. (1953). Vowel recognition as a function of duration, frequency modulation and phonetic context. *Journal of Speech & Hearing Disorders, 18,* 289–301.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America, 69,* 1465–1475.

Zahorian, S., & Jagharghi, A. (1986). Matching of 'physical' and 'perceptual' spaces for vowels. *Journal of the Acoustical Society of America, (Suppl. 1), 79,* S8 (A).

Zahorian, S., & Jagharghi, A. (1987). Speaker-independent vowel recognition based on overall spectral shape versus formants. *Journal of the Acoustical Society of America, (Suppl. 1), 82,* S37 (A).

Received September 29, 1992
Accepted March 17, 1993

Contact author: James Hillenbrand, Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, MI 49008.

Appendix

Sample confusion matrices for parameter sets involving bark-transformed frequencies are shown in Tables A-1 (B1, B2), A-2 (B1-B3), and A-3 (B0-B2). The general pattern of confusions shown in these tables is quite similar to that seen with the other transforms and with linear frequency in Hz. Note the across-the-board improvement in classification accuracy with the addition of either the fundamental frequency or the third formant to the two lowest formants.

TABLE A-1. Confusion matrix for discriminant analysis using B1 and B2. The row labels indicate the vowel that was intended by the talker and the column labels indicate the vowel that was classified by the discriminant analysis algorithm.

	/i/	/ɪ/	/ɛ/	/æ/	/ɝ/	/ʌ/	/ɑ/	/ɔ/	/ʊ/	/u/
/i/	142	10	0	0	0	0	0	0	0	0
/ɪ/	11	126	12	0	3	0	0	0	0	0
/ɛ/	0	21	103	3	25	0	0	0	0	0
/æ/	0	0	14	120	11	7	0	0	0	0
/ɝ/	0	3	15	7	98	0	0	0	25	4
/ʌ/	0	0	0	8	3	121	8	9	3	0
/ɑ/	0	0	0	0	0	15	120	17	0	0
/ɔ/	0	0	0	0	0	6	17	121	7	1
/ʊ/	0	0	0	0	27	10	1	3	90	21
/u/	0	0	0	0	12	3	0	1	21	115

TABLE A-2. Confusion matrix for discriminant analysis using B1, B2, and B3. The row labels indicate the vowel that was intended by the talker and the column labels indicate the vowel that was classified by the discriminant analysis algorithm.

	/i/	/ɪ/	/ɛ/	/æ/	/ɝ/	/ʌ/	/ɑ/	/ɔ/	/ʊ/	/u/
/i/	144	8	0	0	0	0	0	0	0	0
/ɪ/	11	126	15	0	0	0	0	0	0	0
/ɛ/	0	21	118	5	8	0	0	0	0	0
/æ/	0	0	16	129	4	1	0	0	2	0
/ɝ/	0	2	15	4	129	0	0	0	1	1
/ʌ/	0	0	0	4	0	129	10	8	1	0
/ɑ/	0	0	0	0	0	12	127	13	0	0
/ɔ/	0	0	0	0	0	3	17	123	6	3
/ʊ/	0	0	0	0	1	6	0	4	120	21
/u/	0	0	0	0	0	0	0	1	26	125

TABLE A-3. Confusion matrix for discriminant analysis using B0, B1, and B2. The row labels indicate the vowel that was intended by the talker and the column labels indicate the vowel that was classified by the discriminant analysis algorithm.

	/i/	/ɪ/	/ɛ/	/æ/	/ɝ/	/ʌ/	/ɑ/	/ɔ/	/ʊ/	/u/
/i/	144	7	1	0	0	0	0	0	0	0
/ɪ/	11	127	14	0	0	0	0	0	0	0
/ɛ/	0	20	124	7	1	0	0	0	0	0
/æ/	0	0	14	132	3	3	0	0	0	0
/ɝ/	0	2	6	1	135	2	0	0	6	0
/ʌ/	0	0	0	2	4	130	12	4	0	0
/ɑ/	0	0	0	2	0	12	130	7	1	0
/ɔ/	0	0	0	0	0	2	16	124	6	4
/ʊ/	0	0	0	0	10	5	0	4	115	18
/u/	0	0	0	0	4	1	0	2	14	131