On detecting nasals in continuous speech*

Paul Mermelstein

Haskins Laboratories, New Haven, Connecticut 06510
(Received 24 August 1976; revised 18 October 1976)

The acoustic manifestation of nasal murmurs is significantly context dependent. To what extent can the class of nasals be automatically detected without prior detailed knowledge of the segmental context? This contribution reports on the characterization of the spectral change accompanying the transition between vowel and nasal for the purpose of automatic detection of nasal murmurs. The speech is first segmented into syllable-sized units, the voiced sonorant region within the syllable is delimited, and the points of maximal spectral change on either side of the syllabic peak are hypothesized to be potential nasal transitions. Four simply extractible acoustic parameters, the relative energy change in the frequency bands 0–1, 1–2, and 2–5 kHz, and the frequency centroid of the 0–500-Hz band at four points in time spaced 12.8 msec apart are used to represent the dynamic transition. Categorization of the transitions using multivariate statistics on some 524 transition segments from data of two speakers resulted in a 91% correct nasal/non-nasal decision rate.

PACS numbers: 43.70.Sc, 43.70.Gr

INTRODUCTION

While the perceptual cues that indicate the presence of nasal murmurs in the speech signal have been carefully studied,[1,2] successful extraction of these cues for recognition of continuous speech has proved a difficult problem.[3,4] Most recognition systems classify short segments of the signal on the basis of the spectral properties of the segments without paying particular attention to the time-varying spectral characteristics of phone-sized segments. This paper reports results obtained when acoustic measurements near hypothesized transitions into and out of the nasal murmur are used to detect the class of nasals. The results suggest that improved nasal detection rates are attainable with this method.

The general framework of our approach to the recognition of continuous speech is one which first segments the signal into syllable-sized units, and then carries out a hierarchic sequence of segmentation and labeling steps for segments differentiated by voicing and manner of production.[5] Based on the structure of voicing and manner of production segments found, hypotheses for full phonetic transcription of the unit are derived from a dictionary of admissible syllabic forms. Detection of nasals is one step in this process. The syllable-sized units are detected on the basis of minima in a newly defined loudness function, a summation of the short-time energy weighted to deemphasize frequency components below 500 and above 2000 Hz. The voiced subsegment of the syllable-sized units is next delimited, and segments corresponding to the voicebar of stops and to voiced fricatives are trimmed off. The segment that remains may be entirely nasal, a syllabic nasal. Otherwise, constraints on the phonetic structure of syllables allow the existence of at most one manner of production change from nasal to nonnasal prior to the syllabic peak and one reverse change from nonnasal to nasal after the syllabic peak. The points of maximal spectral change within the delimited segments on either side of the syllabic peak are hypothesized as potential transition points between the vowel (possibly also glide or liquid) and the nasal. The regions in the acoustic signal near these points are characterized on the basis of acoustic measurements within the region as nasal or nonnasal.

We first review the acoustic cues that give rise to the nasal consonants. Acoustic measurements, selected to extract those cues, are next described. We then detail our method for detecting potential nasal transitions and formulate decision rules for their categorization. Categorization results are reported for data from 11 sentences from each of two speakers. These results, encompassing 524 transition segments, demonstrate the usefulness of this method.

I. ACOUSTIC CUES AND MEASUREMENTS

The distinctive manner feature "nasalized" pertains to both nasal vowels and nasal murmurs. This study is concerned with the transition from vowel (nasalized or not), glide or liquid to nasal murmur where the primary articulatory change is oral closure in the absence of velopharyngeal closure. Instead of searching separately for the acoustic correlates of the oral closure and velopharyngeal opening, which can be expected to show gross variations depending on the state of the other features, it appears worthwhile to look for correlates of the composite articulatory event directly.

The search for invariant acoustic cues that indicate the presence of nasal murmurs in continuous speech has a long history. Fujimura[1] reported the spectral characteristics of nasal murmurs in intervocalic contexts. He found three essential features: first, the existence of a very low first formant in the neighborhood of 300 Hz; second, the relatively high damping factors of the formants; and third, the high density of the formants in the frequency domain. Fant[2] reports that a voiced occlusive nasal (nasal murmur) is characterized by a spectrum in which the second formant is weak or absent; a formant at approximately 250 Hz dominates the spectrum but several weaker high-frequency formants occur, and the bandwidths of nasal formants are generally larger than in vowellike sounds.

A preliminary exploratory study used bisyllabic non-

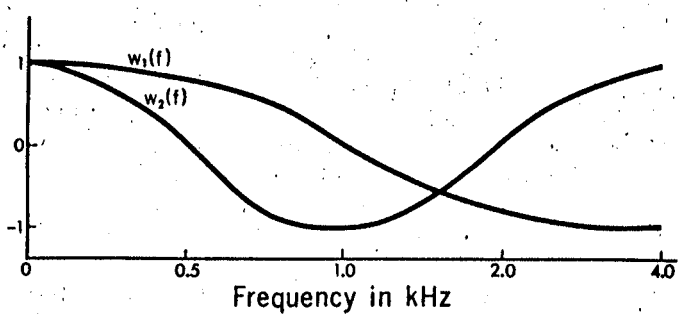

FIG. 2. Weighting functions for the determination of spectral coefficients.

The definition of spectral variation with time is guided by the following rationale. Pols[8] has computed the eigenvectors accounting for the two dimensions of maximal variance for sonorants. These roughly correspond to measures of speech spectra along the dimensions of low-frequency versus high-frequency energy, and the low-plus-high-frequency versus mid-frequency difference. In an attempt to approximate equal perceptually significant changes in vowel spectra by equal increments in our two-dimensional spectrum representation, we first transform the spectra from a linear frequency scale to a technical mel scale (linear up to 1000 Hz, logarithmic thereafter). Next we compute the first two coefficients of the Fourier cosine transform of the log power spectrum of the signal using the weighting functions shown in Fig. 2. The directions in multidimensional spectral space defined by these coefficients roughly correspond to the maximal variance directions of Pols. Our dimensions are orthogonal, and coefficient values are independent of spectrally uniform signal amplification or attenuation. Individual spectra at time k can now be represented by the coefficient pair

$$C_i(k)=\int_{f=0}^{f=4\ \mathrm{kHz}} w_i(f)\,e_k(f), \quad i=1,2$$

where the w_i are the respective mel-based cosine weighting functions and $e_k(f)$ is the measured energy function of frequency at time k.[9]

Let k_a and k_b be the boundaries of the voiced, nonfricative central section of a syllabic unit. Now define the spectral difference metric at time-sample k as

$$D(k)=\left(\sum_{i=1}^{2}\left[C_i(k+1)-C_i(k-1)\right]^2\right)^{1/2},$$

where the $C_i(k)$ are the coefficients computed for the kth spectral cross sections and unit spacing in k corresponds to a time spacing of 12.8 msec. This is the same metric as used by Itahashi[10] for phonetic segmentation, except that our definition is symmetric with time so that it is independent of movement forward or backward in time from the syllabic peak.[11]

Define the concave hull of the difference function over the interval $[k_A, k_B]$, as shown in Fig. 1, with the aid of k_P, the time frame of the syllabic peak, such that $k_A<k_P<k_B$, and

$$H(k)=\min_{k_A\le k'\le k} D(k'), \quad k_A\le k\le k_P$$

$$=\min_{k\le k'\le k_B} D(k'), \quad k_P\le k\le k_B .$$

Now find the maximum difference in the spectral difference less the hull, on either side of the syllabic peak:

$$D'(k'_Q)=\max_{k_A\le k\le k_P}\left[D(k)-H(k)\right],$$

$$D'(k'_R)=\max_{k_P\le k\le k_B}\left[D(k)-H(k)\right].$$

Then $k_Q=k'_Q+1$ and $k_R=k'_R-1$ are points of potential nasal termination or onset. Now consider the acoustic characteristics of the spectral regions on the time-negative side of k_Q and the time-positive side of k_R, namely, frames $S_n^-=S(k_Q-n)$, $S_n^+=S(k_R+n)$, $n=1,\ldots,4$. If a nasal segment forms all or part of the prevocalic or postvocalic consonantal cluster in the syllabic unit, then the spectra S_n can be expected to reflect that fact.

We now define our basic measurements. Let $\Delta E_n^{i,\pm}=E_{k\pm n}^i-E_k^i$ be the relative energy (dB) in the ith frequency band of the nth frame relative to the energy in the same frequency band at the onset or termination of the hypothesized segment. Approximate the first formant frequency at time $k+n$ by the centroid or first moment of the energy in the frequency range 0–500 Hz,

$$g_n=\sum_i f_i e_{k+n}(f_i)\Big/\sum_i e_{k+n}(f_i).$$

The f_i are power spectrum values computed with 40-Hz frequency spacing, thus the summation over i ranges over the first 12 spectral samples. The above measurements can be easily derived from the speech signal even under relatively noisy conditions and were therefore considered to be potentially robust cues for nasal segments. The question to be investigated is to what extent the parameters ΔE_n^i and g_n differentiate the nasals from the nonnasals, and thus represent useful cues for automatic nasal detection.

B. Categorization of transitions

Since we are dealing with a dynamic articulatory event, we would like to treat our parameters as multivariate in space and time. For example, transitions to obstruents are accompanied by a large energy drop and a small value for the low-frequency centroid. Liquids may show a significant drop in the 2–5-kHz band, but much less of a drop in the 1–2-kHz band. Transitions to nasals are relatively short, generally not exceeding 50 msec, thus measurements must be based on carefully selected intervals of the signal. In the absence of *a priori* information regarding the distribution of the acoustic parameters, we assumed multivariate normal distributions. Four-dimensional measurement vectors $\vec{x}^n$ were measured at four points in time, $n=1,\ldots,4$, so that they spanned 51 msec of spectral data, derived in turn from some 64 msec of waveform data. Determination of the complete covariance ma-

sense words with nasals in intervocalic environment as well as in intervocalic clusters where the nasal preceded or followed a stop consonant. Examination of spectrograms and spectral cross sections essentially confirmed Fujimura's report.[1] A low-frequency nasal resonance and drop in mid-plus-high-frequency energy (above roughly 1000 Hz) in the absence of a significant drop in low-frequency energy (below 1000 Hz) were found to be reliable cues for nasals. Suitable qualitative parameter differences were easily found by inspection. However, when the same cues were tested on continuous speech, differentiation between nasals and nonnasals proved markedly poorer. Accordingly, a new study was carried out in an attempt to quantitatively characterize these parameters in continuous speech and evaluate their utility for nasal detection.

The first parameter selected, the energy centroid in the 0–500-Hz frequency band, can be looked upon as a rough approximation to the first formant frequency. A value for this parameter near 250 Hz is a necessary but not sufficient condition for the existence of nasal murmurs because this property is shared by the first formant frequency of high vowels. The energy parameters defined below are intended to discriminate between the nasals and the high vowels. The energy centroid, although independent of overall signal level, is dependent on linear spectral distortion such as the 300-Hz high-pass filtering of telephone speech.

Fant[6] suggests that the physical phenomena underlying a particular distinctive feature need exhibit only relational invariance. For example, the weakness of a second formant may be best judged relative to the intensity of that formant in the adjacent vowel rather than in absolute terms. We employ three spectral energy parameters, all defined in relational terms with respect to the energy in the respective frequency bands prior to the transition. This definition makes the parameters independent not only of the overall signal amplitude as well as any linear spectral distortion, but corrects to a limited extent for the overall spectral shape imposed by the syllabic vowel. Since none of the parameters alone is sufficiently effective to separate the nasals, our effort has focused on the effective combination of information from several independently measured parameters in an attempt to attain classification performance superior to that obtainable by any single parameter.

II. EXPERIMENTAL PROCEDURE

Our prime interest lay in recognizing nasals in a variety of contexts such as may be encountered in free text rather than in the study of speaker-dependent variations. Therefore, previous recordings of the "rainbow passage" and five additional sentences by two male speakers were studied. The speech text, listed in the Appendix, was roughly 90 sec long and included a large number of nasals in word-initial, word-final, intervocalic, and clustered phonetic environments. The speech material was recorded in a noise-free environment at the subjects's comfortable reading rate, digitized using a 10-kHz sampling frequency and spectra

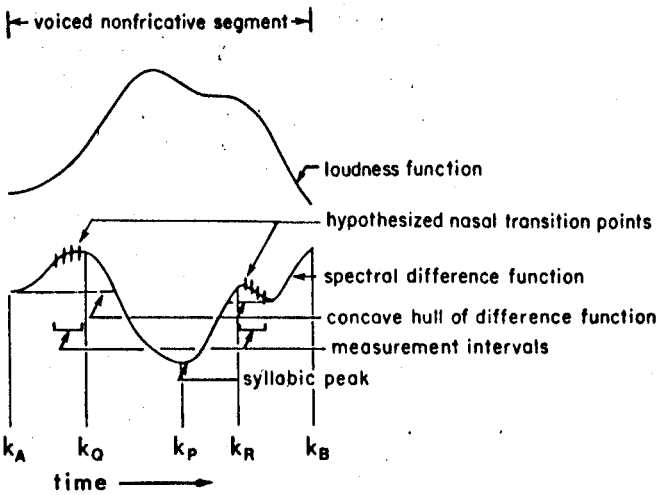

FIG. 1. Loudness and spectral difference functions for a typical syllabic segment.

were computed using a 25.6-msec Hamming time window. Adjacent spectral computations were spaced 12.8 msec in time and yielded spectra at 40-Hz frequency intervals. The material was segmented into syllabic units following procedures reported separately.[7]

A. Detection of potential nasal transitions

To test the hypothesis that points of maximal spectral change are potential nasal indicators, we need to define operationally the term "syllabic peak" and find an appropriate metric for "spectral derivative." It is our intent that the syllabic peak be located within the vocalic region of any syllable at the point of minimal spectral change so that it best reflects the color of the syllabic vowel. Having established this point of minimal spectral change within the vocalic region, the spectral derivative function within the voiced regions may be computed and the maximum points found on either side of the syllabic peak. By evaluating the acoustic information in the neighborhood of the maximal spectral changes we shall try to classify the transition as to whether it denotes the onset or termination of a nasal.

The syllabification algorithm evaluates minima in a "loudness function" (a time-smoothed, frequency-weighted energy function) as potential syllabic boundaries. Tested on roughly 400 syllables of continuous text, the algorithm results in 6.9% syllables missed and 2.6% extra syllables relative to a nominal, slow-speech syllable count. The maxima in loudness are potential syllabic peaks. Qualitative study of spectrograms augmented with loudness curves reveals that frequently the maximum in loudness occurs prior to the point in time where the formants appear to be maximally steady. Hence we construct a 6-dB loudness range below the maximal loudness level of the syllabic unit and search for the point of minimal spectral change within corresponding time interval. Figure 1 shows typical plots of loudness and spectral differences for one segment.

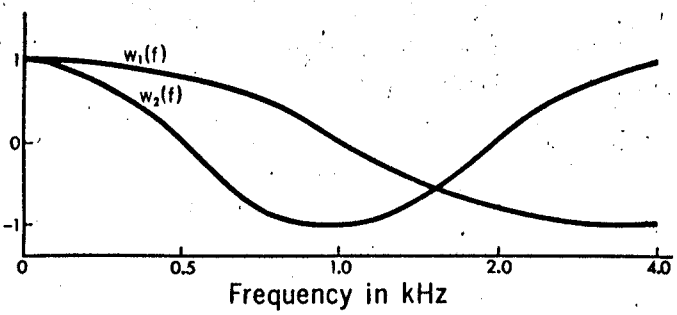

FIG. 2. Weighting functions for the determination of spectral coefficients.

The definition of spectral variation with time is guided by the following rationale. Pols[8] has computed the eigenvectors accounting for the two dimensions of maximal variance for sonorants. These roughly correspond to measures of speech spectra along the dimensions of low-frequency versus high-frequency energy, and the low-plus-high-frequency versus mid-frequency difference. In an attempt to approximate equal perceptually significant changes in vowel spectra by equal increments in our two-dimensional spectrum representation, we first transform the spectra from a linear frequency scale to a technical mel scale (linear up to 1000 Hz, logarithmic thereafter). Next we compute the first two coefficients of the Fourier cosine transform of the log power spectrum of the signal using the weighting functions shown in Fig. 2. The directions in multidimensional spectral space defined by these coefficients roughly correspond to the maximal variance directions of Pols. Our dimensions are orthogonal, and coefficient values are independent of spectrally uniform signal amplification or attenuation. Individual spectra at time k can now be represented by the coefficient pair

$$C_i(k) = \int_{f=0}^{f=4\ \mathrm{kHz}} w_i(f)\, e_k(f), \quad i = 1, 2$$

where the w_i are the respective mel-based cosine weighting functions and $e_k(f)$ is the measured energy function of frequency at time k.[9]

Let k_a and k_b be the boundaries of the voiced, nonfricative central section of a syllabic unit. Now define the spectral difference metric at time-sample k as

$$D(k) = \left(\sum_{i=1}^{2} [C_i(k+1) - C_i(k-1)]^2 \right)^{1/2},$$

where the $C_i(k)$ are the coefficients computed for the kth spectral cross sections and unit spacing in k corresponds to a time spacing of 12.8 msec. This is the same metric as used by Itahashi[10] for phonetic segmentation, except that our definition is symmetric with time so that it is independent of movement forward or backward in time from the syllabic peak.[11]

Define the concave hull of the difference function over the interval $[k_A, k_B]$, as shown in Fig. 1, with the aid of k_P, the time frame of the syllabic peak, such that $k_A < k_P < k_B$, and

$$H(k) = \min_{k_A \le k' \le k} D(k'), \quad k_A \le k \le k_P$$

$$= \min_{k \le k' \le k_B} D(k'), \quad k_P \le k \le k_B .$$

Now find the maximum difference in the spectral difference less the hull, on either side of the syllabic peak:

$$D'(k'_Q) = \max_{k_A \le k \le k_P} [D(k) - H(k)],$$

$$D'(k'_R) = \max_{k_P \le k \le k_B} [D(k) - H(k)].$$

Then $k_Q = k'_Q + 1$ and $k_R = k'_R - 1$ are points of potential nasal termination or onset. Now consider the acoustic characteristics of the spectral regions on the time-negative side of k_Q and the time-positive side of k_R, namely, frames $S_n^- = S(k_Q - n)$, $S_n^+ = S(k_R + n)$, $n = 1, \ldots, 4$. If a nasal segment forms all or part of the prevocalic or postvocalic consonantal cluster in the syllabic unit, then the spectra S_n can be expected to reflect that fact.

We now define our basic measurements. Let $\Delta E_n^{i,\pm} = E_{k\pm n}^i - E_k^i$ be the relative energy (dB) in the ith frequency band of the nth frame relative to the energy in the same frequency band at the onset or termination of the hypothesized segment. Approximate the first formant frequency at time $k+n$ by the centroid or first moment of the energy in the frequency range 0–500 Hz,

$$g_n = \sum_i f_i e_{k+n}(f_i) \Big/ \sum_i e_{k+n}(f_i).$$

The f_i are power spectrum values computed with 40-Hz frequency spacing, thus the summation over i ranges over the first 12 spectral samples. The above measurements can be easily derived from the speech signal even under relatively noisy conditions and were therefore considered to be potentially robust cues for nasal segments. The question to be investigated is to what extent the parameters ΔE_n^i, and g_n differentiate the nasals from the nonnasals, and thus represent useful cues for automatic nasal detection.

B. Categorization of transitions

Since we are dealing with a dynamic articulatory event, we would like to treat our parameters as multivariate in space and time. For example, transitions to obstruents are accompanied by a large energy drop and a small value for the low-frequency centroid. Liquids may show a significant drop in the 2–5-kHz band, but much less of a drop in the 1–2-kHz band. Transitions to nasals are relatively short, generally not exceeding 50 msec, thus measurements must be based on carefully selected intervals of the signal. In the absence of *a priori* information regarding the distribution of the acoustic parameters, we assumed multivariate normal distributions. Four-dimensional measurement vectors $\vec{x}^n$ were measured at four points in time, $n = 1, \ldots, 4$, so that they spanned 51 msec of spectral data, derived in turn from some 64 msec of waveform data. Determination of the complete covariance ma-

trix would have required the estimation of 256 covariance matrix coefficients. To reduce the computations required and to facilitate practical implementation, one desires to reduce the number of coefficients to be estimated. We have assumed that the significant covariance components are those between the respective parameters at a given point in time or in a single parameter at several points in time.

Multivariate distance-based pattern recognition techniques suffer from the disadvantage that the distance to the mean point of the distribution function corresponding to the appropriate category is generally dominated by the component of the distance along the dimension where the measurement is farthest from the estimated mean. Given a finite set of learning data, distance estimates of points where the distribution curve has a larger value can be assumed to be more reliable than the distance estimates in regions where only relatively few data points have been encountered in the learning set. Combination of all measurements into one global decision rule is therefore likely to result in a decision rule dominated by its most unreliable components. On the other hand, combination of groups of measurements into preliminary scores and combination of those scores into a final decision has the advantage that the preliminary scores can be adjusted based on our confidence in those scores.

Following Patrick,[12] for two categories, nasal (a) and nonnasal (b), with mean parameter vectors $\vec{m}_a$ and $\vec{m}_b$ and covariance matrices $\vec{\Sigma}_a$ and $\vec{\Sigma}_b$, the minimum probability of error decision rule is to decide class a if

$$\frac{P_a}{[\vec{\Sigma}_a]^{1/2}(2\pi)^{L/2}} \exp\left[-\tfrac{1}{2}(\vec{x}-\vec{m}_a)^t \vec{\Sigma}_a^{-1}(\vec{x}-\vec{m}_a)\right] > \frac{P_a}{[\vec{\Sigma}_a]^{1/2}(2\pi)^{L/2}} \exp\left[-\tfrac{1}{2}(\vec{x}-\vec{m}_b)^t \vec{\Sigma}_b^{-1}(\vec{x}-\vec{m}_b)\right].$$

P_a and P_b are the *a priori* probabilities of the nasal and nonnasal categories and L is the dimensionality of the measurement vector $\vec{x}$.

If we estimate the parameter means and covariances at the respective measurement times, we can combine this information into the following decision rule (decision rule A):

If

$$P_a \sum_n [\vec{\Sigma}_a^n]^{-1/2} \exp\left[-\tfrac{1}{2}(\vec{x}^n-\vec{m}_a^n)^t (\vec{\Sigma}_a^n)^{-1}(\vec{x}^n-\vec{m}_a^n)\right] > P_b \sum_n [\vec{\Sigma}_b^n]^{-1/2} \exp\left[-\tfrac{1}{2}(\vec{x}^n-\vec{m}_b^n)^t (\vec{\Sigma}_b^n)^{-1}(\vec{x}^n-\vec{m}_b^n)\right],$$

choose category a, otherwise choose category b. Alternatively, the initial estimates may be based on the time samples of the individual parameters. A preliminary test showed that better results were obtained by combining probabilities as derived from the multivariate measurements at different points in time than by combining probabilities resulting from each parameter measured as a multivariate function of time.

If the preliminary decision scores from selected nonoverlapping sets of measurements are combined into a final decision score, then the constraint that the unknown must belong to one of the two decision classes may be used to adjust the preliminary probability estimates. If the preliminary probabilities determined by assuming normal distributions are normalized so that the total probability that an unknown belongs to the mutually exclusive classes of nasal and nonnasal sums to unity, this reduces the effect that a measurement that is distant from both the category means can have on the total decision. If the estimates at the other time points are reliable, the contributions from those measurements will generally outweigh a single unreliable estimate. This results in decision rule B, i.e., compute n preliminary scores for each category,

$$s^n(\alpha) = P^n(x/\alpha)/[P^n(x/\alpha) + P^n(x/\beta)]$$

where α = a, b, β = b, a and $0 \le s(\alpha) \le 1$. If $P_a \sum_n s^n(a) > P_b \sum_n s^n(b)$, choose category a, otherwise b.

The effects of the *a priori* probabilities P_a and P_b may be embedded in a decision threshold θ and adjustment of θ up or down may be used to control the difference between the relative frequency of false nasal and nonnasal decisions. To obtain the results cited, the value of θ was adjusted to result in roughly equal probabilities of false nasal and nonnasal decisions.

III. RESULTS

A preliminary analysis program found the points of maximal spectral difference. On the basis of spectrographic and auditory examination, these transition points were hand labeled to indicate whether they corresponded to nasal or nonnasal transitions. Roughly 20% of the transitions were in fact to or from nasals. A single nasal segment could be manifested by two transitions if in intervocalic context, and one transition only if in a pre- or postobstruent context. Syllabic units were treated as independent information-bearing elements and each transition was classified independently. Two syllabic nasals were found in the data and these were eliminated from subsequent consideration.

Statistics were gathered separately for nasal–nonnasal and nonnasal–nasal transitions. Differences between nasals in initial and final position in the voiced sequence of the syllabic unit were not found significant, and the two classes were therefore pooled to arrive at the following results. Figure 3 gives means and standard deviation values for the measured parameters after pooling of the differently directed transition groups. One observes that the distributions of all of the parameters show considerable overlap. Only for ΔE_n^1 do we see considerable separation by categories. However, ΔE_n^1 does not separate the nonnasal sonorants from the nasals. It only serves to exclude the transitions to nonsonorants. Parameter g_n shows little separation in category means, but a large difference between the variances of the two categories. In fact, detailed examination shows the distribution of the nonnasals to be roughly bimodal; the obstruents possess rather low values of g_n, the sonorants have values higher than the mean nasal value. Clearly, the nonnasal category is not homogeneous and perhaps a representation of the

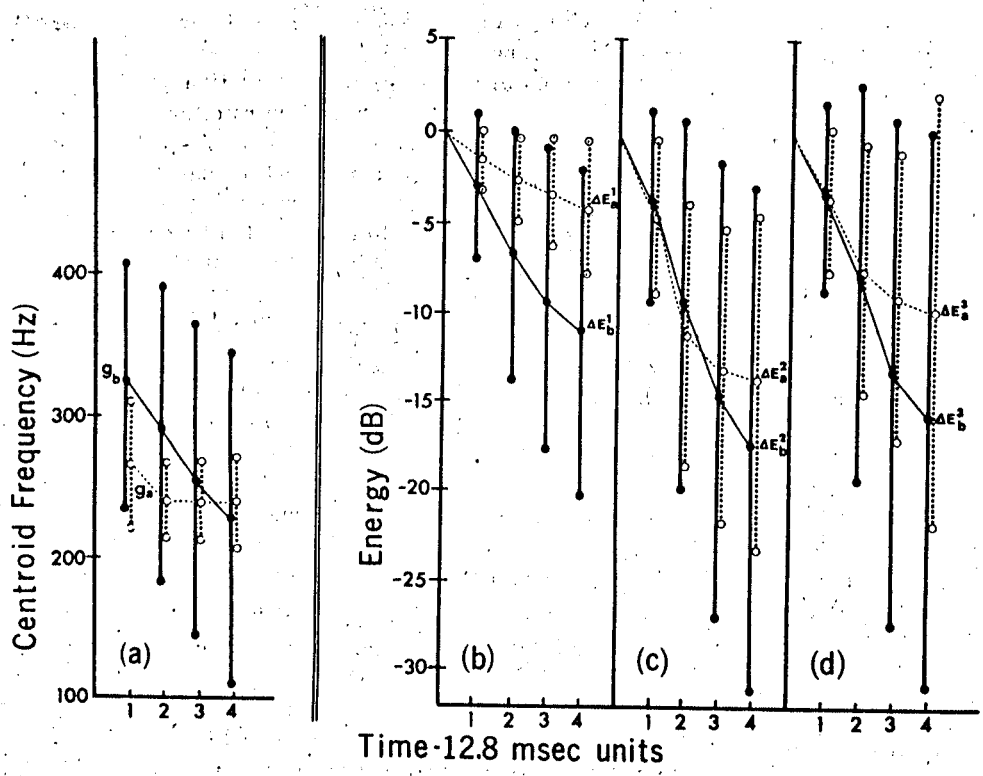

FIG. 3. Mean and standard deviation values for measured parameters: (a) centroid frequency, 0–500-Hz band; (b) relative signal energy, 0–1-kHz band; (c) relative signal energy, 1–2-kHz band; and (d) relative signal energy, 2–5-kHz band. Subscript a—nasal category, b—nonnasal category.

nonnasals in terms of a mixture of normally distributed categories would be more appropriate.

The classification results are summarized in Table I. The decision threshold for the two categories was established by pooling the parameter data of both subjects and using the same 524 transitions both to train the classifier and to test it. Comparison of results obtained with the two decision rules show a drop in the error rate from 13.9% to 9.3% when using rule B. Evidently, normalization of the conditional probabilities before combining the measurements at different points in time helps to lower the error rate.

Experiments were continued with decision rule B. Speaker to speaker variation was estimated by repeatedly testing one speaker's data against measurements derived from the other speaker's data. A total error rate of 15% was noted. The most significant differences in the nasal transition parameter data were noted in the mean centroid frequency. For $n=2$, this value was 222 ± 16 Hz for speaker LL, 256 ± 26 Hz for speaker GK. As discussed below, this parameter was found to be the most useful contributor to the total categorization score. Thus it is not surprising that the decisions are strongly dependent on small centroid frequency differences. Based on articulatory considerations, one would expect significant nasal resonant-frequency differences due to size differences between speakers's nasal cavities. The relatively small standard deviation values for the centroid-frequency parameter are more surprising and indicate its insensitivity to contextual variations. When training data and test data were separated by text material rather than speaker, the total error rate was only 11%. The higher error-rate degradation due to learning and testing on different speakers rather than different text suggests that further improvements in categorization may result through use of speaker-dependent training data.

To evaluate the relative contributions of the four measurements, a decision rule was implemented that treated each measurement as independent, normalized the conditional probabilities for each measurement, and summed to contributions from the 16 measurements.

TABLE I. Nasal/non-nasal classification results.

Training set	Test set	Decision rule	Error rate
2 speakers, 11 sentences each	same	A	13.9%
2 speakers, 11 sentences each	same	B	9.3%
2 speakers, 11 sentences each	same	16 independent parameters	13.5%
speaker B, 11 sentences	speaker A	B	15%
speaker A, 11 sentences	speaker B		
2 speakers, text 1	same speakers, text 2	B	11%
2 speakers, text 2	same speakers, text 1		

Predictably, the error rate on the total data using independent parameters was higher, 13.5% vs 9.3% using multivariate statistics. An estimate of the contribution of each measurement to the total decision score may be derived from

$$S^i = \frac{1}{J}\sum_{j=1}^{J} \beta_j [s_j^i(\mathrm{a}) - s_j^i(\mathrm{b})],$$

where β_j is +1 or −1 depending on whether or not the test item was a nasal, $s_j^i(\mathrm{a})$ and $s_j^i(\mathrm{b})$ are the respective normalized probability scores for the two categories obtained from measurement i on token j, and J is the total number of transition tokens. The low-frequency energy centroid, the one parameter dependent on the spectrum at only one moment in time, showed the highest contributions, namely, 0.71, 0.77, 0.75, and 0.74 for $n=1,\ldots,4$. The other parameters were apparently less effective, their contributions ranged from 0.53 to 0.65. Measurements at time values $n=2$, 3 were most effective, yet the others still contributed substantially to reduce the overall error rate. The one significant difference between prevocalic and postvocalic nasal transition was found in the relative effectiveness of the measurements at the distinct time values. Measurements at small-n values give relatively higher contributions for prevocalic transitions, measurements at larger-n values are more effective for postvocalic transitions. One explanation for this may be that a nasal is frequently anticipated by nasalization of the preceding vowel which causes the spectral discontinuity to be less abrupt and necessitates a longer time delay before a distinct nasal murmur can be observed.

IV. DISCUSSION

In attempting to compare our results with those of other workers, we encountered few quantitative results dealing with substantial amounts of data in the literature. Niederjohn and Thomas[2] report that the most confusion in their system is encountered between the four sounds /m, n, l, r/. However, their study included only 12 nasals from four sentences by one speaker. Weinstein *et al.*[4] report confusion statistics for consonant segment classification. If their confusion matrix is reduced to two categories, nasal and nonnasal, a misclassification rate of 21% is obtained. The test applied here to detect nasals makes use of acoustic measurements similar to that work. However, two essential differences should be noted. First, by averaging formant frequency and amplitude measurements over five points in time before classification, they implicitly assume a nasal segment model with static spectral characteristics. Our data reveal significant parameter differences with time as one moves further into the nasal segment. Second, formant computations appear not to be necessary for nasal detection. In fact, representation of the nasal spectrum by means of three formants may not be sufficiently precise. One of the differentiating characteristics of nasals is the presence of spectral zeros, the effects of which are but poorly captured in a three-formant representation. In view of our results, the use of broadband spectral information appears more robust.

Our results, when speaker-dependent parameters are used, are comparable to some results recently reported as part of larger speech recognition studies. Hess[13] reported a 90.2% recognition rate for the class of nasals in German continuous speech by one speaker. Dixon and Silverman[14] reported a 93.7% nasal recognition rate for an 8-min-long continuous English text by one speaker. The method presented here possesses advantages of simplicity in both the parameters used and the classification techniques employed—when attention may be restricted to the detection of nasals alone.

Generalization of the nasal/nonnasal discrimination to further speakers must await the collection and processing of further data. Interspeaker variation appears the most significant limitation to improvement of the calssification results. We suspect that a limited amount of unsupervised training may suffice to overcome this limitation, however, no experimental studies of this question have been carried out.

There are two important additional sources of variance in our data. The nasal spectrum depends on the color of the syllabic vowel because that is the underlying articulation on which the nasal murmur articulation is superimposed. Of course, the nasal spectrum further depends on the place of production. No attempts to use our measurements to categorize the nasal murmurs by pace of production have yet been carried out. Because good nasal/nonnasal classification is obtainable without consideration of place or production information, it appears appropriate for any complete analysis to do nasal/nonnasal classification first, followed by categorization of the nasal segments.

Most of the false indications result from the confusion of liquids, glides and semivowels with nasals. In particular, /1/ and /r/ before high vowels tend to be confused with nasals rather often. In addition, some voiced fricatives that manifest weak frication, particularly in unstressed environments can be confused with nasals. Nasals were missed most frequently when they appeared to be shortened due to a consonantal cluster context or when they appeared to be articulated as a nasal flap. In cases where nasals are shorter than 50 msec, summation of partial scores from four points in time may be inferior to a sequential classification procedure that stops consideration of new measurements whenever the partial sum of scores exceeds a given fraction of the total possible score.

V. CONCLUSION

The spectral changes manifested by the transitions to and from nasal murmurs are good cues for the recognition of the nasals as a class. Of the four measurements used, the centroid in the 0–500-Hz frequency band appears to be the most useful parameter. Use of additional measurements of energy change in three broad frequency bands allows good separation of nasals and nonnasals irrespective of context. The measurements are significantly correlated, thus resort to multivariate statistics is necessary.

It appears particularly important to treat the transi-

tion between nasal and nonnasal as a dynamic articulatory event with corresponding time-varying acoustic properties. The individual parameters show significant variation with increasing time displacement from the onset of the transition.

Maximal separation between nasals and nonnasals is not achieved at the same point in time for all the parameters. Therefore, the data must not be pooled over the separate time points of measurement. Through careful selection of the maximal spectral variation point, we achieve a time synchronization of the unknown transition with respect to the corresponding reference data, and thereby obtain improved separation between the nasal and nonnasal categories.

APPENDIX. SPEECH TEXT USED FOR DETECTION OF NASALS

When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch with its path high above and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look but no one ever finds it. When a man looks for something beyond his search, his friends say he is looking for the pot of gold at the end of the rainbow.

John and I went up to the farm in June. The sun shone all day and wind waved the grass in wide fields that ran by the road. Most birds had left on their trek south but old friends were there to greet us. Piles of wood had been stacked by the door, left there by the man who lives twelve miles down the road. The stove would not last till dawn on what he had cut, so I went and chopped more till the sun set.

*This research was supported in part by the Advanced Projects Agency of the Department of Defense under Contract No. N00014-67-A-029-002 monitored by the Office of Naval Research. This paper was presented in part at the 90th Meeting of the Acoustical Society of America, in San Francisco [J. Acoust. Soc. Am. 58, S97(A) (1975)].

[1]O. Fujimura, "Analysis of Nasal Consonants," J. Acoust. Soc. Am. 34, 1865–1875 (1962).

[2]G. Fant, "Descriptive Analysis of the Acoustic Aspects of Speech," Logos 5, 3–17 (1962).

[3]R. J. Niederjohn and I. B. Thomas, "Computer Recognition of Continuant Phonemes," IEEE Trans. Audio AU-21, 526–535, 1973.

[4]C. J. Weinstein, S. S. McCandless, L. F. Mondsheim, and V. W. Zue, "A System for Acoustic-Phonetic Analysis of Continuous Speech," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 54–67 (1975).

[5]P. Mermelstein, "A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 79–82 (1975).

[6]G. Fant, "The Nature of Distinctive Features," in *To Honor Roman Jacobson: Essays on the Occasion of his Seventhieth Birthday* (Mouton, The Hague, 1967).

[7]P. Mermelstein, "Automatic segmentation of speech into syllabic units," J. Acoust. Soc. Am. 58, 880–883 (1975).

[8]L. C. W. Pols, "Segmentation and Recognition of Mono-syllabic Words," IEEE and Air Force Cambridge Research Laboratories, Conference on Speech Communication and Processing, 1972, pp. 105–108.

[9]Preliminary evaluation shows separation of vowels comparable to a two-formant representation at a significantly reduced computational cost. If the logarithmic frequency-scale transformation were omitted, the computed coefficients would correspond to the first and second coefficients in a real cepstrum representation of the speech signal. The zeroth coefficient corresponds to the average spectrum level and is therefore not used. Truncation of the cepstrum at a point in inverse time (quefrency) lower than the pitch period yields a smoothed spectrum envelope. The first two coefficients capture the most significant aspects of the variations of spectrum envelope with frequency.

[10]S. Itahashi, S. Makino, and K. Kido, "Discrete-Word Recognition Utilizing a Word Dictionary and Phonological Rules," IEEE Trans. Audio AU-21, 239–249 (1973).

[11]Since the C_i coefficients are linear functions of the signal energy, the difference metric could be equally well defined in terms of weighted spectral differences. The C_i coefficients are initially computed for the purpose of syllabic–vowel categorization, a task not discussed in this report. The computation of the difference metric $D(k)$ is but a simple additional step.

[12]E. A. Patrick, *Fundamentals of Pattern Recognition* (Prentice–Hall, Englewood Cliffs, NJ, 1972).

[13]W. J. Hess, "A Pitch-Synchronous Digital Feature Extraction System for Phonemic Recognition of Speech," IEEE Trans. Acoust. Speech Signal Process. ASSP-24, 14–25 (1976).

[14]N. R. Dixon and H. F. Silverman, "A General Language-Operated Decision Implementation System (GLODIS): Its Application to Continuous-Speech Segmentation," IEEE Trans. Acoust. Speech Signal Process. ASSP-24, 137–162 (1976).