On Detecting Nasals in Continuous Speech*

Paul Mermelstein

ABSTRACT

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 extractable 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 percent correct nasal/nonnasal decision rate.

INTRODUCTION

The search for invariant acoustic cues that indicate the presence of nasal murmurs in continuous speech has a long history. Fujimura (1962) 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 (1962) 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 vowel-like sounds. These cues appear to be generally adequate for human identification of nasal segments in spectrograms; however, precise quantitative data are unavailable. This work represents an attempt to evaluate the contributions of a set of simply extractable acoustic measurements to the detection of nasals.

*A version of this paper was presented at the 90th annual meeting of the Acoustical Society of America, San Francisco, Calif., 3-7 November 1975. [J. Acoust. Soc. Am. (1975), Suppl. 58, S97(A).]

Acknowledgment: 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.

[HASKINS LABORATORIES: Status Report on Speech Research SR-44 (1975)]
We approach the problem as a typical categorization task in automatic phonetic transcription. A general framework for the solution of this task has been presented previously (Mermelstein, 1975a). The point of view adopted there emphasizes the hierarchical nature of speech perception and structures the automatic phonetic analysis tasks analogously. Detection of nasals is one stage of that process. The continuous speech signal is first segmented into syllable-sized segments called syllabic units. The syllabic unit 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. We first delimit the voiced sonorant portion of the syllable and look for points of maximal spectral change within the delimited interval on either side of the syllabic peak. These points of maximal spectral change are hypothesized as potential transition points between the vowel (possibly also glide or liquid) and the nasal. The detected transitions are categorized on the basis of acoustic measurements in the transition region.

What are the acoustic cues that allow the listener to establish that a particular syllable contains a nasal segment? A preliminary exploratory study used bisyllabic nonsense 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 (1962) report. Reliable cues for nasals are found to be a low-frequency nasal resonance and a drop in mid-+ high-frequency energy (above roughly 1000 Hz) in the absence of a significant drop in low-frequency energy (below 1000 Hz). 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 remarkably poorer. Accordingly, a new study was carried out in an attempt to characterize quantitatively these parameters in continuous speech and evaluate their utility for nasal detection.

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 actual spectrum of the nasal murmur is known to vary with the syllabic vowel as well as the place of oral closure (Fujimura, 1962). Since place-of-production discrimination can be expected to be highly manner-of-production dependent, we chose first to detect nasal murmurs as a class and subsequently to discriminate among them.

The first parameter selected, the energy centroid in the 0–500-Hz frequency band, can be looked on 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 (1967) 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.

**EXPERIMENTAL PROCEDURE**

In order to examine nasals in a wide variety of vocalic and consonantal contexts, previous recordings of the "rainbow passage" and six additional sentences by two speakers were studied. The speech material was recorded at the subjects' comfortable reading rate, digitized using a 10-kHz sampling frequency, and spectra were computed using a 25.6-msec Hamming time window. Adjacent spectral computations were spaced 12.8 msec apart and yielded results with 40-Hz frequency spacing. The material was segmented into syllabic units following procedures reported separately (Mermelstein, 1975b).

To test the hypothesis that points of maximal spectral change are potential nasal indicators, we need to define operationally the term "syllabic peak" and 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 within the voiced regions on either side of the syllabic peak can be computed and maxima found. 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 program evaluates minima in a "loudness function" (a time-smoothed, frequency-weighted energy function) as potential syllabic boundaries. 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 time at which 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 the corresponding time interval. Figure 1 shows typical plots of loudness and spectral differences for one segment. The definition of spectral variation with time is guided by the following rationale. Pols (1972) 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- and high-frequency versus midfrequency difference. In an attempt to approximate perceptually equal 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 Figure 2. The directions in multidimensional spectral space defined by these coefficients roughly correspond to the maximal variance directions of Pols (1972). 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

\[
C_i(k) = \begin{cases} 
  f = 4 \text{ kHz} & w_i(f) e_k(f), \ i = 1, 2 \\
  f = 0
\end{cases}
\]

where the \( w_i \) are the respective weighting functions and \( e_k(f) \) is the measured energy as a function of frequency at time \( k \).\(^1\)

Let \( k_n \) 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 \( k^{th} \) spectral cross sections and unit spacing in \( k \) corresponds to a time spacing of 12.8 msec. This is the same metric as that used by Itahashi, Makino, and Kodo (1973) 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.\(^2\)

Define the concave hull of the difference function over the interval \( k_a-k_p-k_b \) by

---

\(^1\)If the logarithmic frequency-scale transformation were omitted, the computed coefficients would correspond to the first and second coefficients in a real cepstrum (cosine transformation of power spectrum) 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. Preliminary evaluation shows separation of vowels in the two-coefficient space comparable to a two-formant representation at a significantly reduced computational cost.

\(^2\)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.
Figure 1: Loudness and spectral difference functions for a typical segment.

Figure 2: Weighting functions for the determination of spectral coefficients.
\[ H(k) = \min_{k_a \leq k' \leq k} D(k'), \quad k_a \leq k < k_p \]
\[ \quad = \min_{k < k' \leq k_b} D(k'), \quad k_p \leq k < k_b \]

where \( k_p \) is the time frame of the syllabic peak. Now find the maximum difference in the spectral difference less the null, on either side of the syllabic peak

\[ D'(k' \ 'p) = \max_{k_a \leq k < k_p} \left[ D(k) - H(k) \right] \]
\[ D'(k' \ 'q) = \max_{k_p \leq k < k_b} \left[ D(k) - H(k) \right] \]

Then \( k_q = k_q' - 1 \) and \( k_r = k_r' + 1 \) are points of potential nasal onset or termination. Now consider the acoustic characteristics of the spectral regions on the time-positive side of \( k_q \) and the time-negative side of \( k_r \). For notational convenience, let \( \alpha \) be \(-1\) near \( k_r \), \(+1\) near \( k_q' \) and consider the spectral frames \( S_n(k_q + \alpha n), S_n(k_r + \alpha n), n = 1, \ldots, 4 \). If there is a nasal segment in time-final position in the syllabic unit, then the spectra \( S_n \) can be expected to reflect that manner category.

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

\[ g_n = \sum_i f_i e_{k+n}(f_i) / \sum_i e_{k+n}(f_i) \]

Our analysis system outputs filter spectra at 40-Hz intervals, thus the summation over \( i \) ranges over the first 12 spectral samples. The above measurements can be seen to be simply derivable 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 \( \Delta E_n^i \) and \( g_n \) differentiate the nasals from the nonnasals, and thus represent useful cues for automatic nasal detection.

The information contributed by the various parameters at the respective points in time may be combined in diverse ways. The simple statistical model used for our initial attempts at classification treated the cues as independent time-varying quantities. The relative likelihood that the transition belongs to the nasal or nonnasal class is computed by summing the relative likelihood scores arrived at from each parameter at each point in time. However, the parameters are in fact correlated. For example, transitions to obstruents are accompanied by a large energy drop and a low low-frequency centroid. Furthermore, liquids show a significant drop in the 2-5-kHz band, but much less of a drop in the 1-2-kHz band. An improved statistical model uses multivariate statistics on all parameters at the respective points in time. Since the time course of parameter
variation may be different for the various parameters, separate likelihood scores are computed at each point in time and summed to result in a composite score.

In the absence of a priori information regarding the distribution of the acoustic parameters, we assumed multivariate normal distributions and estimated the parameter mean vectors $\mathbf{m}_a^n$, $\mathbf{m}_b^n$ and the parameter covariance matrices $\mathbf{\Sigma}_a^n$ and $\mathbf{\Sigma}_b^n$ for the nasal and nonnasal transitions, respectively. The superscript $n$ denotes the point in time for the parameter measurement. Data were collected for $n = 1, \ldots, 4$, a time frame of 51 msec of spectral data, which in turn is derived from some 64 msec of waveform data.

Following Patrick (1972), the minimum probability of error decision rule for two categories when each has a Gaussian distribution with estimated mean and covariance matrix is to decide Class "a" if

$$
\frac{P_a}{\sqrt{\frac{1}{2}}} \exp \left[ -\frac{1}{2} \begin{pmatrix} \mathbf{x} - \mathbf{m}_a \end{pmatrix}^t \mathbf{\Sigma}_a^{-1} \begin{pmatrix} \mathbf{x} - \mathbf{m}_a \end{pmatrix} \right] >
$$

and

$$
\frac{P_b}{\sqrt{\frac{1}{2}}} \exp \left[ -\frac{1}{2} \begin{pmatrix} \mathbf{x} - \mathbf{m}_b \end{pmatrix}^t \mathbf{\Sigma}_b^{-1} \begin{pmatrix} \mathbf{x} - \mathbf{m}_b \end{pmatrix} \right]
$$

where $P_a$ and $P_b$ are the a priori probabilities of the respective categories, and $L$ is the dimensionality of the measurement vector $\mathbf{x}$. To improve the reliability of our decisions as well as to make them insensitive to small registration errors in the time signal, we wish to combine information from parameter measurements at several points closely spaced in time. Since we are dealing with a dynamic articulatory event, parameter statistics must be assumed nonstationary. Instead of treating our parameters as multivariate in space and time—an operation that would require significantly more parameter estimation data than available—indepen-dent estimates of parameter means and covariances were carried out at each measurement time. One decision rule that combines this information (decision rule A) is the following: if

$$
\frac{P_a}{\sqrt{\frac{1}{2}}} \exp \left[ -\frac{1}{2} \begin{pmatrix} \mathbf{x} - \mathbf{m}_a \end{pmatrix}^t \mathbf{\Sigma}_a^{-1} \begin{pmatrix} \mathbf{x} - \mathbf{m}_a \end{pmatrix} \right] >
$$

and

$$
\frac{P_b}{\sqrt{\frac{1}{2}}} \exp \left[ -\frac{1}{2} \begin{pmatrix} \mathbf{x} - \mathbf{m}_b \end{pmatrix}^t \mathbf{\Sigma}_b^{-1} \begin{pmatrix} \mathbf{x} - \mathbf{m}_b \end{pmatrix} \right]
$$

choose category $a$, otherwise choose category $b$. Here the superscript $n$ denotes the measurement time.

Alternatively, we may wish to normalize the relative probabilities before summation over the different measurements. For this decision rule (decision rule B), define a category score
\[ s^n(\alpha) = \frac{p^n(x|\alpha)}{p^n(x|\alpha) + p^n(x|\beta)}, \quad 0 \leq s(\alpha) \leq 1; \quad \alpha = a, b; \quad \beta = b, a \]

and if \( P_a \sum_n s^n(a) > P_b \sum_n s^n(b) \), choose \( a \), otherwise choose \( b \).

The effects of the a priori probabilities \( P_a \) and \( P_b \) may be embedded in a decision threshold \( \Theta \) and adjustment of \( \Theta \) 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, a value of \( \Theta \) was used that results in roughly equal probability of false nasal and nonnasal decisions.

**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. A single nasal segment could be manifested by two transitions if in intervocalic context, and one transition only if in a pre- or post-obstructed 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-nasal and nonnasal-nasal transitions. Differences between nasals in initial and final position in the voiced sequences 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 \( \Delta E_n^1 \) do we see considerable separation by categories. However, \( \Delta E_n^1 \) 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 have 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 nonnasals in terms of a mixture of normally distributed categories would be more appropriate.

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. A misclassification rate of 13.9 percent was observed for decision rule A. The procedure was repeated with decision rule B and resulted in 9.3 percent errors. 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 percent was noted. The most significant differences in the nasal transition parameter data
Figure 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; (d) relative signal energy, 2-5-kHz band. Subscript a - nasal category, b - nonnasal category.
were noted in the mean centroid frequency. For \( n = 2 \), this value was \( 222 \pm 16 \) Hz for speaker LL, \( 256 \pm 26 \) 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' nasal cavities. The relatively small standard deviation values for the parameter are more surprising and indicate the insensitivity of this parameter to contextual variations. When training data and test data were separated by text material rather than speaker, the total error rate was only 11 percent. 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 measurement 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. Predictably the error rate on the total data using independent parameters was higher: 13.5 vs. 9.3 percent, using multivariate statistics. An estimate of the contribution of each measurement to the total decision score may be derived from

\[
S_i = \frac{1}{J} \sum_{j=1}^{J} \beta_j \left[ s_j^i(a) - s_j^i(b) \right]
\]

where \( \beta_j \) is +1 or -1 depending on whether or not the test item was a nasal, \( s_j^i(a) \) and \( s_j^i(b) \) 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—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 requires a longer time delay before a distinct nasal murmur can be observed.

**DISCUSSION**

In attempting to compare our results with those of other workers, we encountered few quantitative results in the literature. Weinstein, McCandless, Mondsheim, and Zue (1975) report confusion statistics for consonant segment classification. If their confusion matrix is reduced to two categories—nasal and nonnasal—a misclassification rate of 21 percent is obtained. The test applied there to detect nasals makes use of information similar to that in our 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 poorly captured in a three-formant representation. In view of our results, the use of broad-band spectral information appears more robust.

Generalization of the nasal/nonnasal discrimination to further speakers must await the collection and processing of further data. Interspeaker variation appears to be the most significant limitation to improvement of the classification 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 of the nasal. No attempts to use our measurements to categorize the nasal murmurs by place 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, /ɾ/ 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 owing 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.

CONCLUSIONS

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 transition 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.

REFERENCES