Analysis of nasal consonants using perceptual linear prediction

Yingyong Qi
Department of Speech and Hearing Sciences and Department of Electrical and Computer Engineering,
University of Arizona, Tucson, Arizona 85721

Robert A. Fox
Division of Speech and Hearing Sciences, Ohio State University, Columbus, Ohio 43210

(Received 31 May 1991; accepted for publication 19 November 1991)

Until recently, speech analysis techniques have been built around the all-pole linear predictive model. This study examines the effectiveness of using the perceptual linear predictive method for analyzing nasal consonants. Six speakers (three men and three women) produced 300 CV syllables with initial nasal consonants /m/ and /n/. A threshold-based boundary detection algorithm was developed to extract nasal segments from the CV contexts. Poles of a fifth-order perceptual linear predictive model were calculated and the frequency of the second pole was used to characterize the place of articulation of nasal consonants. Results indicated that the frequency for the second transformed pole was significantly lower for /m/ than for /n/ and was independent of factors such as vowel context and gender of the speaker. A nasal identification rate of 86% was obtained based on the frequency of the second pole. The use of the perceptual linear predictive method may thus overcome some difficulties associated with analyzing nasal consonants.

PACS numbers: 43.71.Es, 43.71.Cq, 43.72.Ar

INTRODUCTION

The progressive phonetic refinement of broad categories of speech sounds (stop, fricative, and sonorant) is an important step in implementing speech understanding and/or speech recognition systems (Reddy, 1976); however, knowledge about the acoustic cues that allow for reliable extraction of phonetic characteristics of speech from the acoustic waveform is limited (Harrington, 1988). For example, robust features that enable automatic within-class differentiation of nasal consonants have not been well established.

The acoustic structure of nasal consonants has long been predicted by the acoustic theory of speech production (Fant, 1960; Fujimura, 1962; Flanagan, 1972). The presence of a side-branching resonator (the blocked oral cavity) will introduce an antiresonance (zero) in the spectrum of nasal consonants. Theoretically, the antiresonance can be used to identify the place of articulation of nasal consonants because the frequency of the antiresonance is determined by the dimension of the side-branching resonator. For example, the frequency of the antiresonance should be lower for /m/ than for /n/ because the blocked oral cavity attached to the pharyngeal-nasal passage is longer for /m/ than for /n/. Such differences, however, are difficult to detect using conventional techniques of spectral analysis. The presence of the spectral zero introduces nonlinear equations to parametric methods of spectral analysis (Kay, 1987). The harmonic structure, the absence of a prominent spectral valley, and the significantly damped high-frequency energy in nasal spectra are the sources of some of the difficulties associated with analyzing nasal consonants using nonparametric (periodogram) methods.

Fujimura (1962) examined the acoustic structure of nasal consonants and measured the frequency of the antiresonance using the method of analysis by synthesis. This study provided, for the first time, valuable experimental evidence on the predicted acoustic characteristics of nasal consonants. As pointed out by the author, however, "the procedure ... is by no means simple or mechanical ... and a machine could hardly replace the human operator at the stage of this study."

Kurowski and Blumstein (1987) reported that the place of articulation of nasal consonants in CV syllables could be differentiated by the pattern of spectral energy change in the vicinity of the oral release once the spectra were transformed using an algorithm based on the psychoacoustic characteristics of human auditory system. Utilizing critical band analysis based on the Bark scale (Zwicker, 1970), Kurowski and Blumstein examined the spectra of two glottal pulses of the nasal murmur immediately preceding oral release of the nasal stop closure and the first two glottal pulses after the nasal release. The spectral energy increase was reported to be larger between 5–7 Bark than between 11–14 Bark for the nasal /m/; whereas the spectral energy increase was reported to be larger between 11–14 Bark than between 5–7 Bark for the nasal /n/. Although the auditory-based spectral transformation provides a reasonable alternative for analyzing the nasal consonants, these features were inevitably context dependent because spectral information of the following vowel was an important part of the measurement. In addition, there was a certain degree of arbitrariness in the selection of the frequency range for comparing energy increase around nasal release since the selection was neither systematically optimized given the transformed spectrum nor directly prescribed by theory.

Hermansky (1990) proposed a new approach, called...
perceptual linear predictive (PLP) method, for the spectral
transformation based on characteristics of the auditory sys-
tem. While the underlying principles of the spectral transfor-
mation was similar to those previously reported (Kewley-
Port and Luce, 1984; Kurowski and Blumstein, 1987), the
final all-pole modeling of the transformed spectrum paramete-
trically specified the spectrum and, thus, simplified the ex-
traction of acoustic characteristics of speech signals. Her-
mansky (1990) demonstrated that a fifth-order PLP
spectral transformation improved the performance of a tem-
plate-based speech recognition system.

In light of the newly proposed PLP method of spectral
analysis, the present study was undertaken to determine
whether acoustic characteristics of nasal consonants could
be found that were independent of the following vowel con-
text in a CV syllable and to evaluate whether place of articu-
lation could be automatically identified on the basis of these
features. The paper is organized as follows. Section I in-
cludes a brief rationale for using the PLP method and the
implementation details of the PLP analysis. Section II con-
tains a description of the context-independent acoustic char-
acteristics of nasal consonants revealed by the PLP analysis
and the results of nasal identification based on these features.
Section III is the discussion and conclusion.

I. METHOD

The integration of spectral energy as part of the audi-
tory-based spectral transformation provides a unique advan-
tage in analyzing the spectral characteristics of nasal con-
sonants. As stated earlier, the absence of a prominent
spectral valley and the reduced spectral energy in the mid- to
high-frequency ranges make it difficult to accurately locate
the center frequency of the antiresonance and to use this
frequency to distinguish nasal consonants. There are, how-
ever, less-intensive spectral perturbations at different fre-
quency ranges for different nasal consonants due to the side-
branch (oral) coupling. For example, there is a relatively
broad energy reduction in the mid-frequency (1000-2000
Hz) range within the nasal murmur of nasal /n/ that is not
as apparent for the nasal /m/. When a nonuniform integra-
tion of power is undertaken as part of the spectral transfor-
mation, some otherwise unreliable energy concentrations
are expected to manifest themselves as spectral peaks. These
spectral peaks could be used to differentiate the place of ar-
ticulation of nasal consonants. Because the spectral integra-
tion in the PLP transformation is based on the characteris-
tics of the auditory system and spectral peaks in the
transformed spectrum have important perceptual implica-
tions, it is reasonable to assume that integrated spectral
peaks in the transformed spectrum could be used to differen-
tiate the place of articulation of nasal consonants. Thus in-
stead of attempting to measure the difficult to locate antires-
onance directly, poles of the PLP transformed spectra are
used to characterize and differentiate nasal consonants.

Hermansky (1990) has claimed that a fifth-order PLP
model is most effective in a template-based speech recogni-
tion system. The order of the model implies that two pairs
of conjugate poles and one real pole are available for each signal
frame. Each pair of the conjugate poles provides the frequen-
cy and bandwidth of one equivalent spectral peak. The real
pole specifies the overall spectral tilt of the transformed
spectrum. Because nasal spectra are dominated by only one
low-frequency pole (250-300 Hz) (Fujimura, 1962; Fujii-
mura and Lindqvist, 1971), the 5th-order PLP model is
adopted here for the analysis. However, the real pole is not
considered in the final analysis because spectral tilt can easi-
ly be influenced by factors such as recording conditions
and glottal characteristics of a particular speaker and is often
considered as a phonetically nonessential factor (Klatt,
1982).

A. Subjects and recordings

Six native speakers of American English (three men and
three women) provided the speech samples. Each speaker
produced ten CV syllables with initial nasal consonants /m/
and /n/ followed by five different vowels /i/, /a/, /o/, /a/,
and /u/. Each syllable was repeated five times. The syllables
were randomly ordered. The recordings were made in a quiet
room with background noise at about 30 dB (SPL). The
microphone was placed approximately 10 cm from the
mouth of the speaker. Recorded tokens were low-pass fil-
tered (\(f_c = 4.5\) kHz) and digitized at a sampling rate of
10 kHz with a 16-bit quantization level.

B. Nasal segment selection

A 25.6-ms segment of the nasal consonant before the
oral release into the following vowel (i.e., within the nasal
murmur) was selected for the PLP analysis. The boundary
between the nasal consonant and the following vowel was
automatically located from the acoustic waveform. The ex-
act algorithm is as follows. The signal was bandpass filtered
(1-3.5 kHz, fourth-order Butterworth filter) to enhance the
energy discontinuity of the boundary since energy for the
nasal murmur is mostly in the frequency range below about
800 Hz (Mermelstein, 1977). The upper frequency limit for
the bandpass filter was used to eliminate unpredictable high-
frequency influences. The boundary was determined using a
short-term amplitude signal calculated from the filtered
waveform. A hamming window, with a window size of 10 ms
and a window step of 2 ms, was used in the calculations. The
short-term amplitude was calculated as the absolute sum of
sampled values within the window

\[
A_i = \sum_{n=0}^{N-1} |s_n h_n|,
\]

where \(A_i\) is the amplitude signal of the \(i\)th frame, \(s_n\) is
the bandpass filtered waveform, \(h_n\) is the Hamming window,
and \(n\) is the discrete time index. The boundary was automati-
cally detected and marked as the time location where the
increase of short term amplitude, \(A_{i+1} - A_i\), exceeded a
threshold \(A_F\) with the condition that \(A_f\) was less than another
threshold \(A_T\). The threshold \(A_T\) was adaptively established
as a fixed percentage of the average amplitude of two con-
secutive frames

\[
A_T = 0.25 (A_{i+1} + A_i) / 2.
\]

The second threshold \(A_T\) was used to avoid detecting large
energy fluctuations within the vowel segment. When more
than one time location was marked within a single syllable because of the finite window length, the last mark was chosen as the boundary between the nasal and the vowel because the threshold $A_T$ minimized the possibility that a boundary could be marked within the vowel segment. The boundary marks were visually checked using time-synchronized spectrographic displays. All the marked boundaries were well aligned with the onset of high-frequency energy in the spectrograph. The marked original and filtered waveforms and the spectrograms of the original signals are illustrated in Fig. 1 for the sample syllables /ma/ and /na/.

C. PLP spectral analysis

PLP spectral analysis was implemented using the following procedure: (a) 16 simulated critical-band masking patterns separated at a 1-Bark interval (covering the total frequency range of 50–4674 Hz) were each calculated as a 128-point weighting function and were pre-emphasized by a simulated equal loudness curve [see Fig. 2(a)]; (b) a 128-point power spectrum was calculated for each selected nasal segment using a 256-point FFT; (c) a weighted sum of the power spectrum was calculated for each weighting function and the summation was compressed by the cubic-root intensity-loudness relationship obtaining 16 filter outputs; (d) 16 autocorrelation coefficients were derived from the filter outputs using a 32-point IDFT, 5 of which were used to solve the Yule–Walker equations to obtain the fifth-order PLP coefficients. Finally, roots of the fifth-order PLP model were calculated to represent the nasal consonants. The process of the calculation is illustrated in Fig. 2(b).

FIG. 1. The original waveform (top window), the filtered waveform (middle window), and spectrograms of the original waveform (bottom window) for the syllable /ma/ (left) and /na/ (right). The identified boundary marks between the nasal and the vowel are shown in each window.

D. Statistical analysis

A mixed-design analysis of variance (ANOVA) was undertaken for each frequency and bandwidth of the two spectral peaks to assess the effects of place of articulation, vowel context, gender of speaker, as well as their interactions. Because the identification of each individual speaker was also a factor that was nested within each gender group, its interaction with those main factors was treated as an error term in the subsequent testing of hypothesis (Fisher, 1946).

The frequency $f_i$ and bandwidth $b_i$ of each transformed pole were obtained first in the Bark scale using the formula

$$f_i = \frac{16 \cdot \arctan(z_i)}{\pi}, \quad (3)$$

$$b_i = 32 \cdot \ln |z_i| / \pi, \quad (4)$$

where $z_i$ represent the $i$th conjugate pole ($i = 1,2$). The $f_i$ and $b_i$ were then transformed back into linear scale using the inverse relationship between frequency and Bark

$$f_i = 600 \sinh(f_b/6), \quad (5)$$

where $f_i$ is frequency in Hz and $f_b$ is frequency in Bark (Schroeder, 1977).

II. RESULTS

The distribution of the calculated poles were shown in Fig. 3 for the nasal consonants produced by male and female speakers. The average transformed spectra for the nasal consonants /m/ and /n/ produced by the three male and three female speakers are presented in Fig. 4 together with the overall average transformed spectra. The corresponding average FFT spectra are presented in Fig. 5.

The nasal /m/ and /n/ exhibited quite different characteristics in both the pole distribution plots and the average PLP spectra. The difference between nasal /m/ and /n/ in the average FFT spectra, however, was rather distributed and difficult to quantify. The second peak of the PLP spec-
FIG. 3. Pole distributions for nasal consonants produced by male (top) and female (bottom) speakers. Small circle depicts /m/ and plus sign depicts /n/.

The average frequencies of the spectral peaks were $f_1 = 290 \text{ Hz}$ and $f_2 = 1742 \text{ Hz}$ for /m/ and $f_1 = 305 \text{ Hz}$ and $f_2 = 2062 \text{ Hz}$ for /n/. The average bandwidth of the spectral peaks were $b_1 = 305 \text{ Hz}$ and $b_2 = 563 \text{ Hz}$ for /m/ and $b_1 = 305 \text{ Hz}$ and $b_2 = 479 \text{ Hz}$ for /n/. Means and standard
Fig. 5. Average FFT spectra for nasal consonants produced by (a) male speakers, (b) female speakers, and (c) all speakers. Solid line represents nasal /m/ and dotted line represents nasal /n/.

deviations for the frequency and bandwidth of transformed poles were summarized in Table I for each CV combination. Average frequencies for the first and second poles and the standard deviation for the second pole were illustrated in

<table>
<thead>
<tr>
<th>Syllable</th>
<th>( f_1 ) (Hz) Mean</th>
<th>( f_2 ) (Hz) Mean</th>
<th>( b_1 ) (Hz) Mean</th>
<th>( b_2 ) (Hz) Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>mi</td>
<td>300.3</td>
<td>23.4</td>
<td>1846.2</td>
<td>201.8</td>
</tr>
<tr>
<td>ma</td>
<td>293.2</td>
<td>25.6</td>
<td>1764.8</td>
<td>210.9</td>
</tr>
<tr>
<td>mo</td>
<td>279.2</td>
<td>19.8</td>
<td>1530.4</td>
<td>229.0</td>
</tr>
<tr>
<td>m3</td>
<td>288.1</td>
<td>20.1</td>
<td>1692.8</td>
<td>213.7</td>
</tr>
<tr>
<td>ma</td>
<td>291.3</td>
<td>25.3</td>
<td>1776.5</td>
<td>224.9</td>
</tr>
<tr>
<td>n/</td>
<td>310.3</td>
<td>23.7</td>
<td>2107.7</td>
<td>170.0</td>
</tr>
<tr>
<td>ne</td>
<td>303.0</td>
<td>22.7</td>
<td>1989.5</td>
<td>155.3</td>
</tr>
<tr>
<td>no</td>
<td>304.5</td>
<td>24.2</td>
<td>2035.4</td>
<td>132.9</td>
</tr>
<tr>
<td>ni</td>
<td>302.1</td>
<td>22.3</td>
<td>2123.3</td>
<td>135.6</td>
</tr>
</tbody>
</table>

Fig. 6(a) for each CV syllable and in Fig. 6(b) for each individual speaker.

A /m-n/ classification test was undertaken based on the frequency of the second pole. The nasal segments were iden-
were established from randomly selected tokens (one-third the remaining two-thirds of the tokens). For comparison, the nasal segments were also classified using the cepstral group-delay spectral distance measure

$$d_{GD} = \sum_{i=1}^{n} p \cdot (c_{iR} - c_{iT})^2,$$

where $c_{iR}$ and $c_{iT}$ were the cepstral coefficients of the reference and test PLP spectra, respectively (Yegnanarayana and Reddy, 1979) and $p$ was equal to the order of the PLP model. A correct nasal identification rate of 87% was obtained.

III. DISCUSSION AND CONCLUSION

The acoustic theory of speech production predicts that the place of articulation of nasal consonant is characterized by antiresonance as well as heavily context-dependent resonances. Effective methods for measuring the antiresonance frequency, however, are not available. The present study essentially uses the broad range spectral characteristics closely related to the antiresonance to characterize the place of articulation of nasal consonants. Results demonstrated that differences of energy distribution in the middle frequency range between /m/ and /n/ (see Fig. 5) were clearly reflected in the second transformed spectral peak (see Fig. 4). The frequency of the second transformed spectral peak has been effectively used to identify the place of articulation of nasal consonants. These results conform with the conjecture that the spectral perturbations introduced by the antiresonance are quantitatively more convenient and effective than the exact antiresonance frequency itself in characterizing the place of articulation of nasal consonants once appropriate spectral transformation is undertaken.

Auditory-based spectral transformations have been applied to speech analysis in a number of studies (e.g., Kewley-Port and Luce, 1984; Kurowski and Blumstein, 1987; Hermansky, 1986). While the implementation details of the transformation might be different among the studies, these studies have demonstrated some of the advantages in speech analysis of applying the two basic operations, the amplitude-weighted integration and frequency warping. The PLP all-pole modeling of the transformed spectra (Hermansky, 1990) provides an effective algorithm of applying the transformation to speech analysis. The poles of the PLP all-pole model were used in this study to characterize and identify the place of articulation of nasal consonants. Since the transformed poles can be automatically derived from the recorded acoustic waveform and the identification rate based on the poles was reasonably high, it appears that a perceptually based analysis procedure for automatic identification of place of articulation of nasal consonants has been validated.

The differences revealed in the transformed spectra between /m/ and /n/ are interpretable on the basis of the underlying mechanism of nasal production and the properties of the perceptually based transformation. As stated earlier, the center frequency for antiresonance of nasal consonant is difficult to locate directly. However, there are observable spectral perturbations at different frequency ranges for different nasal consonants (see Fig. 4). For example, nasal /n/ may have a higher, less sharply tuned antiresonance than /m/ (Fujimura, 1962) and, thus, the energy reduction associated with antiresonance is located higher in frequency and has wider effect for /n/ than for /m/. This difference could be further deepened by the auditory-based transformation because the integration interval in the transformation is smaller for low frequencies than for high frequencies making distributed spectral perturbation at higher frequencies more significant. Both factors may have contributed to the differences between /m/ and /n/ in the second pole of the transformed PLP spectra.

The nasal identification rate based on the frequency of the second transformed pole was comparable to that based on cepstral distance—a measure which uses information from the whole spectrum (Gray and Markel, 1976). However, the features characterized by the transformed poles have more direct perceptual implication than the cepstral distance measure. Hermansky (1990) claimed that the transformed spectral characteristics for vowels support the concept of effective second formant frequency (Fant and Risberg, 1962) and auditory normalization (Bladon and Lindblom, 1981). The fact that the frequency of the second transformed pole is the dominant acoustic feature for differentiating nasal consonants again supports the claimed importance of the effective second formant frequency. The gender-independent nature of the second spectral peak also supports the hypothesis that perceptual normalization may be performed at the auditory level.

With the established correlation between nasal production and its transformed spectrum, a universal or absolute criterion could be developed for nasal classification. Using such a criterion, the identification of place of articulation of nasal consonants would only involve one-pass or bottom-up process. In contrast, the identification of nasal consonants based on Euclidian spectral distance will always depend on the estimation of prototypes for each sound category and a decision cannot be made until cross comparisons with each prototype are completed. Such a process offers few insights about the coding and decoding activities involved in human speech communication. The present work represents one of the effort that probes the relevance of speech production to perception mechanisms.

In this work, however, only the nasal murmur was used to identify the place of articulation of nasal consonants. Although a reasonably high identification rate (86%) was achieved using the PLP method, this rate is far from perfect especially since chance performance is 50%. Information present in the transition period between the nasal and the following vowel could conceivably be used to increase the identification rate. It remains to be determined how the transitional information could be incorporated into the PLP analysis without resorting to the conventional computation methods in pattern recognition.

The continuous movement of articulators in speech production often imposes contextual effect on the predicted acoustic features of individual speech sound. While a num-

FIG. 7. (a) Waveform of the nasal–vowel combination, /n/ and /i/, in CV (upper) and VC (lower) context and their corresponding transformed running spectra in CV (upper) and VC (lower) context. (b) Waveform of the nasal–vowel combination, /m/ and /æ/, in CV (upper) and VC (lower) context and their corresponding transformed running spectra in CV (upper) and VC (lower) context.

of studies have supported the assertion that context independent or invariant acoustic features for place of articulation exist despite acoustic variations introduced by production dynamics (Stevens and Blumstein, 1978; Kewley-Port, 1982; Kurowski and Blumstein, 1987), several problems remain to be resolved before final conclusions can be reached. The context-independent nature of the acoustic characteristics for nasal consonants in CV context demonstrated in this study lends some support to the acoustic invariant hypothesis. Automatic identification of nasals in a VC context based on the same features, however, has only achieved slightly better than chance nasal identification rate. The reduced high-frequency energy in the vowel segment because of nasalization have smeared the discontinuity between vowel and nasal consonant in the VC combinations (Stevens et al., 1987). Furthermore, it appears that the second half of the VC segment became a nasalized version of the leading vowel and the desired nasal murmur could not be identified either from the filtered version of the signal or from a spectrogram. Consequently, significant contextual effect is present until the very last 20 to 30 ms where the energy of the signal is close to zero. The absence of discontinuity, particularly in the high-frequency range, and the substantial contextual effect (see Fig. 7) seem to be the major difficulties in reliably obtaining the reported features. Similar difficulties have been reported in previous studies (Kurowski and Blumstein, 1987). Thus it cannot be claimed that the features revealed here represent the universal, context-independent features for both syllable initial and syllable final nasals.

In summary, this study established a procedure for analyzing the acoustic characteristics of nasal consonants using the PLP method. The second transformed pole was used to automatically identify the place of articulation of nasal con-
sonants in a CV context. Reasonably high identification rates were achieved.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Osamu Fujimura for many suggestions and comments on earlier versions of this paper. This work is supported, in part, by a grant from U.S. West Telecommunication Inc.


