AUTOMATIC ISOLATION OF NASAL MURMURS

H.J. WARKENTYNE

Department of Linguistics
University of Victoria
Victoria, B.C., Canada V8N 2Y2

B.C. DICKSON

Centre for Speech Technology Research
University of Victoria
Victoria, B.C., Canada V8N 2Y2

ABSTRACT

In order to determine the precise parametric values required to locate nasal murmurs in the speech signal, three routines were developed. An energy curve location routine was designed to isolate potential nasal murmurs from the data. A spectral profile-matching routine and a routine for calculating the centroid of spectral energy were then applied to the segments isolated by the energy curve location function. These operations succeeded in locating an average of 52 of 78 possible nasal murmurs for each of the ten subjects.

INTRODUCTION

The experiment reported in this paper represents a component of a research project in speaker recognition. The object was to develop a system that incorporates an automatic procedure for extracting segments from the speech signal which belong to the same phonetic class. The English nasals [m, n, ŋ] were selected as our first target since murmurs have often been shown to be significant in speaker recognition; e.g., [1], [2], [3] and [4].

The data set consisted of 88 short sentences produced by ten subjects. Each sentence contained a nasal phoneme. The phonetic context of the nasal phonemes was varied from sentence to sentence to create a wide range of environmental conditioning factors including ten vocalic environments and utterance-initial and -final positions.

Our early observations indicated that inter-speaker differences in the spectra of the nasal murmurs was a problem to be overcome before a speaker-independent nasal murmur extraction routine could be formulated. For an individual subject, detailed characterization of his nasal murmurs was required to separate them from the non-nasal segments, but this detail failed to characterize the nasal murmurs of a second subject. A series of robust parameters that isolated the nasal murmurs of all the speakers, yet did not falsely reject some murmurs on the basis of too narrow specifications, was therefore required.

As Fujimara [5] has shown, nasal murmurs can be defined in terms of three general acoustic properties independent of place of articulation, phonetic context, or individual speaker. These are: the existence of a low-frequency first formant around 300 Hz that is well isolated from any formant above it, relatively high damping factors of upper formants, and the high density or number of formants in the frequency domain, including the presence of anti-formants. The high density of weak formants should occur in the range of 300 to 2300 Hz.

Mermelstein [6] applied the above description to the development of an automatic nasal detection system for use with continuous speech recognition. He extracted four acoustic parameters using digital filtering to define four frequency bands at 0-1, 1-2, and 2-5 KHz, and a frequency centroid below 500 Hz. Digital spectra and relative intensity between frames were computed every 12.8 ms. The dynamic transition from a nasal to a vowel, or the reverse, defined by a rapid shift in the intensity, signified the probability of occurrence of a nasal murmur. The relative distribution of energy within the three frequency bands, the presence of a centroid below 500 Hz, and the dynamic shift served to indicate the presence of a nasal segment.

Our observations revealed that there was a tendency for nasal segments to be poorly defined acoustically if the duration of the murmur was less than 60 ms. It was seen that in utterance-initial environments, nasal segments were rarely accompanied by a murmur, and in utterance-final positions, a reduction in signal energy sometimes caused the murmur to be weak and irregular. Many of the phonemic nasals produced by the ten subjects appeared to be realized only on the basis of acoustic information that was a product of a transition to or from the neighbouring vowel. Because of the limited information supplied by the nasal segments in utterance-initial positions, the decision was therefore made to concentrate on isolating nasal murmurs that occur in the environment of a preceding vowel with a duration of at least 60 ms.

EXPERIMENTAL PROCEDURES

To determine the precise parametric values required to locate the nasal murmurs, three routines were developed as investigative tools on the IBM main frame. These were an energy curve location routine, a spectral profile-matching routine, and a routine for calculating the centroid of spectral energy. A high degree of flexibility was included to allow fine-tuning of values before they were incorporated into a final segment extracting system.

Energy Curve Location

To locate positions in the speech signal where overall energy of the signal dropped and maintained a steady level, a routine was first developed to convert the time-series data to an energy representation. Calculation of the signal energy was performed by passing the time series through a rectangular window and computing the mean of the squared values in the measurement interval N. The time-varying energy calculation $E(n)$ is defined by the following function:

$$E(n) = \frac{1}{N} \sum_{m=0}^{n=1} x^2(n+m)$$

where N is the number of sample points in the window.

To reduce the high amplitude of the lower frequencies in the signal, pre-emphasis was applied before energy calculation. Pre-emphasis was found to accentuate the sharp drop-off from the vowel to the nasal murmur and smooth the steady state in the energy calculation of the murmur. When pre-emphasis was applied to the signal, a greater percentage of energy in the mid-frequency range of the vowels than was present in the neighbouring nasal murmurs caused the total energy in the vowels to be accentuated.

Experimentation with the window lengths of N in the calculation of the signal energy revealed that a 20 ms window (N = 200 points, sampled at 10 K per second) yields an energy curve that is not affected by the time-varying amplitude properties of the speech signal. However, to locate the onset of the nasal murmur accurately, $E(n)$ was computed at 5 ms intervals. Thus, energy values calculated from 20 ms of time-series data were computed, advancing along the time series in 5 ms jumps.

Determination of a significant change in energy was performed in terms of ratio. For a vowel-nasal sequence, the ratio between the energy value of the triggering frame in the vowel and the lower energy value of the nasal murmur (the trigger ratio) occurring 20 ms later was found to be 1.9:1 or greater.

The following procedure was incorporated. Sequential examination of pairs of energy values representing non-overlapping 20 ms sections of the time-series data is carried out every 5 ms to locate a trigger ratio of 1.9:1 or greater. The first computed value is compared with the fifth, the second with the sixth, etc., and the ratio between the pairs is calculated. Once a trigger ratio is found, the next three consecutive pairs are examined and the pairs with the greatest ratio are selected. The high-energy value represents a 20 ms section of the time series that is the triggering frame of a potential vocalic segment. The low-energy value paired with the triggering frame represents a 20 ms portion that is potentially the start of a nasal murmur. The time co-ordinate of the low-energy frame is the start time of the steady energy level, which can be determined with an accuracy of ± 2.5ms, or half the duration of the 5 ms advance used in calculating the energy of the signal.

The steady energy level of the triggered section of the time series is also calculated by sequential examination of non-overlapping pairs of energy frames, at 5 ms intervals. The pairs must not exceed a ratio of 2:1 or fall below 0.5:1. That is, the energy value of the first frame in the steady state must not be more than double, or less than half, the energy value of the fifth frame, the same being true for the second and sixth frames, etc. The steady state must last for a minimum duration of 60 ms to be accepted. When a ratio above 2:1 or below 0.5:1 is encountered before 60 ms have elapsed, the segment is rejected. After 60 ms, these ratios are used as a shut-off and the segment of time-series data is accepted. In order to avoid acceptance of segments with gradual increases of the energy level, the segment is rejected if the value exceeded that of the triggering frame. Shut-off also occurs if the value of any frame drops below a specified value.

Profile Matching

To characterize the spectral distribution of energy common to nasal murmurs, a profile-matching routine was developed for use on the main frame. The procedure used was to select from the time-series data those sections that were isolated by energy curve location, and to create power spectra of the sections, using 50 Hz resolution and a 20 ms Hamming window advancing along the time series at 20 ms intervals. Pre-emphasis was applied to the time series. The spectra were saved on computer tape, and were later retrieved for comparison with an adjustable profile table.

The parameters incorporated for profile-matching were minimum segment duration, minimum total energy, and percentage and tolerance in up to 20 frequency buckets. The frequency buckets were defined by their upper frequency range, total percentage of all the buckets being of course 100. The routine was designed to call up the profile table upon the initiation of each operation, examine the spectrum file called up from storage, and send the results of the profile-matching to the main frame's printer. Results reported were frames matched, error vectors, and distance as a measure of closeness of fit.

As noted above, the nasal murmurs commonly show a dominance of energy in the 0-500 Hz range. To avoid the influence of individual speaker characteristics, only two frequency buckets were employed in the profile-matching routine. The first was 0-500 Hz, in which the minimum allowable percentage of energy was found to be 57% and the maximum was 99%. In the profile table this was stated as 78% of the spectral energy with a tolerance of 21%. The second frequency bucket was 500 to 5000 Hz, in which the remainder of the energy in the spectrum could be distributed. This was stated as 22% with a tolerance of 21%.

Spectral Centroid Determination

The frequency centroid of a spectrum is essentially the mean frequency of energy in the power spectrum, and is determined by the formula

$$\text{Centroid frequency} = \frac{\sum_{i=1}^{n} f_i I_i}{\sum_{i=1}^{n} I_i}$$

where f is the frequency of bin i, I is the intensity of bin i, and n is the number of the last bin that corresponds to a frequency not greater than the cut-off frequency defined for the centroid calculation. A Fortran program was written to perform this calculation, using as input the

power spectra held in storage on the main frame. An adjustment to the upper frequency *n* was included so that the centroid could be determined for any lowpass bandwidth of the power spectrum. The results of the calculations were displayed in the time domain.

When the centroid calculation was applied to the full 5 KHz passband, a large number of non-nasal segments had a frequency centroid in the range of nasal murmurs; i.e., below 600 Hz. These segments included [ɹ, l, w] voiced stops, [u] and unstressed [ɪ] and [ə], especially when the consonants combined with the vowels. Reducing the cut-off frequency to 1000 Hz eliminated most of the above unwanted segments if only segments with a centroid below 400 Hz were accepted. Although some of the non-nasal elements still exhibited low centroids, the 60 ms duration criterion succeeded in eliminating a majority of these.

Sequencing of Routines

The energy curve location routine was designed to scan the time-series data to determine areas in the speech signal where a nasal murmur was possible. Since this did not require conversion to Fourier series, it was the most economical of the routines to apply to the full data to isolate potential nasal segments. These could then be converted to the Fourier series and be processed by the profile-matching and centroid routines. The latter two routines were applied independently of each other and therefore did not require a particular order.

RESULTS

The speech data of the ten subjects under analysis contained a total of 780 post-vocalic nasal phonemes. Of these, the energy curve location routine successfully isolated 593. The routine also isolated 655 non-nasal phonetic events or sequences of events that took place in a post-vocalic environment. A further 155 non-nasal phonetic events were captured after being triggered by a high-energy non-vocalic signal, indicating the need to incorporate a subroutine that will examine the triggering frame to determine the presence of voicing.

Of the 593 potential nasal murmurs isolated by the energy curve location, the combined profile matching and centroid locating functions accepted 516. When performed independently, the profile-matching routine rejected 356 non-nasals and the centroid calculation rejected 330. The combined effect resulted in the rejection of 454 of the 655 sections of unwanted data.

For a large group of subjects, where robust parameters must be applied in order to isolate the segments, interspeaker characteristics interfere with the process of distinguishing the nasal murmurs from the non-nasals. We have found, however, that speaker-specific characteristics may be used to describe the quality of the nasal murmurs, thereby creating a criterion for rejecting most of the non-nasals. It is apparent from our observations that speaker-specific characteristics are recurrent in the majority of the nasal murmurs. A statistical approach might therefore be usefully employed to describe automatically the mean spectral characteristics of the speech events accepted by the system. A comparison could then be made of each spectral series to determine its closeness of fit to the mean, and, using a degree of tolerance or a distance metric, deviant spectral data could be rejected.

REFERENCES

[1] J.W. Glenn and N. Kleiner, "Speaker identification based on nasal phonation ", *Journal of the Acoustical Society of America 43*: 368-372, 1968.

[2] M.R. Sambur, "Selection of acoustic features for speaker identification", *IEEE Transactions in Acoustics, Speech and Signal Processing ASSP-23*: 169-176, 1975.

[3] L-S. Su, K-P. Li, and K.S. Fu, "Identification of speakers by use of nasal coarticulation", *Journal of the Acoustical Society of America 56*: 1876-1882, 1974.

[4] J.J. Wolf, "Efficient acoustic parameters for speaker recognition", *Journal of the Acoustical Society of America 51*: 2044-2056, 1972.

[5] Osamu Fujimara, "Analysis of nasal consonants", *Journal of the Acoustical Society of America 34*: 1865-1875, 1962.

[6] P. Mermelstein, "On detecting nasals in continuous speech", *Journal of the Acoustical Society of America 61*: 581-587, 1977.