Available online at www.sciencedirect.com

SCIENCE DIRECT®

Speech Communication 43 (2004) 225–239

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Acoustic parameters for automatic detection of nasal manner

Tarun Pruthi *, Carol Y. Espy-Wilson

Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA

Received 16 October 2003

Abstract

Of all the sounds in any language, nasals are the only class of sounds with dominant speech output from the nasal cavity as opposed to the oral cavity. This gives nasals some special properties including presence of zeros in the spectrum, concentration of energy at lower frequencies, higher formant density, higher losses, and stability. In this paper we propose acoustic correlates for the linguistic feature *nasal*. In particular, we focus on the development of Acoustic Parameters (APs) which can be extracted automatically and reliably in a speaker independent way. These APs were tested in a classification experiment between nasals and semivowels, the two classes of sounds which together form the class of sonorant consonants. Using the proposed APs with a support vector machine based classifier we were able to obtain classification accuracies of 89.53%, 95.80% and 87.82% for prevocalic, postvocalic and intervocalic sonorant consonants respectively on the TIMIT database. As an additional proof to the strength of these parameters, we compared the performance of a Hidden Markov Model (HMM) based system that included the APs for nasals as part of the front-end, with an HMM system that did not. In this digit recognition experiment, we were able to obtain a 60% reduction in error rate on the TI46 database.

Keywords: Nasal; Nasal manner; Acoustic parameters; Acoustic correlates; Automatic detection; Automatic speech recognition

1. Introduction

Nasals (/m/, /n/ and /ng/ for American English) are produced (Fant, 1960; Flanagan, 1965; Fujimura, 1962a,b; House, 1957; Nakata, 1959; Stevens, 1998) when the velum lowers to allow coupling to the nasal cavity and a complete closure is formed in the oral cavity. The latter feature resembles the feature characterizing the production of stop consonants and the former gives the nasals their characteristic properties and reveals itself most prominently in the spectrum of the nasal murmur (the sound produced with a complete closure at a point in the oral cavity, and with an appreciable amount of coupling of the nasal passages to the vocal tract). This coupling between the oral and nasal cavities introduces zeros in the nasal murmur spectrum. Thus, the nasal murmur spectrum cannot be exactly modeled by an all-pole transfer function, and it is difficult to study the spectral properties of nasal murmurs because the resonances have a very low amplitude caused by close-lying antiresonances and/or the lossy nasal tract. Nasals can be substantially coarticulated

* Corresponding author. Tel.: +1-301-332-4716.
E-mail addresses: tpruthi@glue.umd.edu (T. Pruthi), espy@glue.umd.edu (C.Y. Espy-Wilson).
URL: http://www.ece.umd.edu/~tpruthi.

0167-6393/$ - see front matter
doi:10.1016/j.specom.2004.06.001

with vowels—so much so, that at times the information about nasality is present only as vowel nasalization (Fant, 1960; Flanagan, 1965; Stevens, 1998). Differences in the size and shape of nasal and sinus cavities can lead to significant differences in the spectral characteristics of nasal murmurs for different speakers. A lot of other sounds can be confused with nasals. Semivowels together with nasals form the class of sonorant consonants. Besides semivowels, voice bars (low frequency periodic energy observed during the closure interval for voiced stops) and non-strident voiced fricatives produced with a weakened constriction so that they surface as a sonorant consonant, can have very similar spectral properties. Previous studies (Fant, 1960; Fujimura, 1962a,b; Dickson, 1962; Stevens, 1998) have provided us with a wealth of information about the acoustic manifestations of the feature *nasal,* but few have addressed the issue of extracting the acoustic correlates of nasality automatically. In this paper we focus on the development of relevant Acoustic Parameters (APs) that can be reliably and automatically extracted.

Cues from three different regions indicate the presence of a nasal consonant: (1) the transition region between the nasal and an adjacent vowel where there is often an abrupt spectral change, (2) the latter portion of a vowel preceding a nasal and the earlier portion of a vowel following a nasal where there might be nasalization because of the adjacent nasal, and (3) the region of the nasal murmur. The abrupt spectral change can be captured by an energy onset/offset measure. The spectral change can be an onset, an offset, or both an offset and an onset depending on whether the nasal is prevocalic (preceding a vowel), postvocalic (following a vowel) or intervocalic (between vowels) respectively. Nasalization in vowels is known to introduce additional poles and zeros in the vocal tract transfer function and cause a broadening of the vowel formants, especially in the first formant region. Several parameters have been suggested to capture vowel nasalization (Chen, 1995, 1997; Glass, 1984; Glass and Zue, 1985). For a database of 1200 words (200 words spoken by 3 male and 3 female speakers) excised from a carrier phrase, Glass and Zue (1985) were able to obtain a 74% average detection rate for nasalized vowels. Chen (1995, 1997) did a more exhaustive work on the acoustic correlates of nasality. However the parameters for vowel nasalization, as suggested by Chen, were not automatically extracted. Hence work needs to be done to automate the extraction of parameters for vowel nasalization. In the current work we have not considered vowel nasalization. Instead we have focused on the extraction of parameters related to the nasal murmur.

Fujimura (1962a,b) identified four properties that characterize the spectra of nasal murmurs: (1) the existence of a very low first formant at about 300 Hz, (2) the existence of zeros in the spectrum, (3) relatively high damping factors of the formants, and (4) high density of formants. The first and the second properties are the result of providing another path for the airflow. The frequency of the zero varies depending on the place where closure is made in the oral cavity. The third and fourth properties are the result of higher losses from the soft walls of the nasal cavity with a large surface area and a longer length of the nasal cavity, respectively. He also mentioned that the spectral characteristics of the nasal murmur are relatively stable in the low-frequency regions, and might constitute a reliable criterion to recognize the class of nasal consonants.

Weinstein et al. (1975) used formant based parameters to detect nasal consonants in sonorant regions. They obtained detection rates of 80%, 60% and 84% for prevocalic, postvocalic and intervocalic nasals respectively on a corpus of 111 sentences spoken by 6 male and 1 female speakers. Hess (1976) obtained 90% recognition rate for German nasals in continuous speech spoken by a single speaker. In this case the training and test data were the same. Dixon and Silverman (1976) were able to obtain a 92.5% recognition rate for nasals in continuous speech spoken by a single speaker.

Mermelstein (1977) proposed the use of 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 time instances spaced 12.8 ms apart to capture the dynamic transition between vowel and nasal and to categorize the transitions as belonging to nasals or non-nasals.

Using this technique, Mermelstein was able to obtain a 91% correct nasal/non-nasal detection rate on a database of 524 potential nasal transition segments from 11 sentences spoken by each of the two speakers. In this case, however, testing was done on the same database as training. When one of the speakers was used for training and the other used for testing, the correct detection rate fell to 85%. The detection rate fell to 89% when different text was used for training and testing. This variation in performance illustrates significant dependence of the parameters on speakers and the actual utterances.

Glass (1984) and Glass and Zue (1986) did some experiments on the automatic extraction of acoustic parameters for distinction between nasals and a class of impostors consisting of liquids, glides, voice bars and weak voiced fricatives. They used the following parameters: (1) difference in average energy between the consonant and adjacent vowel, (2) percentage of time a resonance was centered between 200 and 400 Hz, (3) average amplitude of the resonance in the region between 200 and 400 Hz relative to the total energy in the consonant, (4) ratio of energy between 0 and 350 Hz to energy between 350 and 1000 Hz, and (5) change in low-frequency energy throughout the consonant. These parameters attempted to capture much of the same information as suggested by Fujimura. Using a binary tree classification technique on a database of 600 sentences spoken by 30 male and 30 female speakers, Glass and Zue were able to obtain average classification rates of 79% correct for nasals, and 85% correct for the impostor class. The boundaries of the test tokens and their broad phonetic context (to classify them into prevocalic, intervocalic and postvocalic) were assumed to be known in this experiment.

Chen (2000) worked on a nasal detection module for a knowledge-based speech recognition system. The purpose was to improve the sonorant landmark detector implemented by Liu (1996) in detecting nasals, and to eliminate false nasal landmarks by incorporating parameters for nasal murmur and vowel nasalization. The nasal landmark was indicated by the most negative sum of the difference in the amplitudes of the first four formants of consecutive spectra (sum.diff) from the middle of the vowel into the nasal murmur. To compute the murmur parameters, peak amplitudes in five frequency bands were measured: $A1_n$ (0–788 Hz), $A2_n$ (788 Hz–2 kHz), $A3_n$ (2–3 kHz), $A4_n$ (3–4 kHz), and $A5_n$ (4–5 kHz). The sum of the differences in amplitudes between $A1_n$ and other amplitudes (sum.amp.diff), $A1_n - A2_n$, $A1_n - A3_n$, $A2_n - A3_n$, and the frequency of the lowest peak $F1_n$ were used as the murmur parameters. Vowel nasalization parameters $A1 - P0a$ and $A1 - P1a$ were obtained by subtracting the adjusted nasal peak amplitudes $P0a$ (peak around 250 Hz) and $P1a$ (peak around 1 kHz) from the first formant amplitude $A1$. Using discriminant analysis on the training set (20 sentences from lexical access from features (LAFF) database) with five descriptors (sum.diff, vowel nasalization parameter, $F1_n$, $A1_n - A2_n$ and $A1_n - A3_n$), classification accuracies of 88% and 74% were obtained for distinguishing between the nasal and non-nasal tokens. Again, these parameters were not automatically extracted, and they were tested on a small database with very few speakers. However, this module addressed the cases where one or more of the three cues for nasals, abrupt spectral change, vowel nasalization, and nasal murmur, are absent.

Based on the review of previous work and our own experiments, we have narrowed down on the following four APs: (1) an energy onset/offset measure to capture the consonantal nature of nasals (i.e. abrupt spectral change that occurs at the closure and release of nasal consonants), (2) an energy ratio and (3) a low spectral peak measure to capture the nasal murmur, and (4) an envelope variance parameter to capture the spectral and temporal stability during the nasal murmur. We have not developed any cues for automatic detection of vowel nasalization as of yet. As a result, we focus only on the nasals which are realized as sonorant consonants. In much of our analysis we have only focused on nasals and semivowels although voice bars and non-strident voiced fricatives might be confused with nasals at times. The reason for this is that semivowels are the only class of sounds that share the features *+sonorant*, *+voiced* and *−syllabic* with nasals. Voice bars and non-strident voiced fricatives are *−sonorant* sounds and should be distinguished by a sonorancy classifier instead of a nasal classifier. This is the approach we intend to follow when we integrate the APs for nasals proposed in this paper to our phonetic-feature-hierarchy-based speech recognition system (Juneja and Espy-Wilson, 2002, 2003). In some cases, non-strident voiced fricatives, especially /v/, might be manifested as a +*sonorant,* but it is not possible to detect such cases from the transcription provided with TIMIT (1990) database.

The rest of the paper is organized as follows: In Section 2 we give details of our parameters, the theory and motivation behind using those parameters, and the histograms depicting the distributions for each of the parameters. Section 3 gives details of our classification experiment. Section 4 gives details of the recognition experiment conducted to provide further evidence of the viability of the APs developed. This is a digit recognition experiment and so is not limited to just nasals and semivowels. Section 5 includes a discussion of the current work, comparison with previous works, and suggestions for future work. Section 6 concludes the paper.

2. Acoustic parameters (APs)

The coupling of the oral and nasal cavities during the nasal murmur causes the dominant first formant to shift to very low frequencies. Also, the closure in the oral cavity introduces zeros in the nasal murmur spectrum that occur mostly above 1000 Hz. This low frequency prominence together with the zeros leads to a significant drop in energy in the spectrum of the nasal murmur immediately above the first formant region. We use an onset/offset measure to capture the sharp decrease in energy at the vowel–nasal boundary, and a spectral peak frequency measure and an energy ratio to capture the prominent low frequency peak and the sharp decrease in energy at frequencies immediately above it. Since there are no moving parts in the nasal cavity, we expect the nasal murmur spectrum to show some amount of stability. This should be true at least for the low frequency regions of the spectrum. There might be movements at higher frequencies because of movement in the location of pole–zero pairs (Fujimura, 1962a,b). We capture this stability by a waveform envelope variance parameter.

The histograms in this section were generated by extracting APs for 600 tokens each of prevocalic, postvocalic and intervocalic nasals and semivowels chosen randomly from 2586 'si' and 'sx' sentences spoken by 90 female and 235 male speakers from the dialect regions 1–7 of TIMIT (1990) training database. The APs were normalized to have zero mean and unit variance. The histograms show these normalized values. Further, the histograms for the postvocalic cases only show the values for liquids (/l/, /r/) because there are no postvocalic glides (/w/, /y/).

2.1. Energy onsets/offsets

Fig. 1b shows the wideband spectrogram of a typical intervocalic nasal and Fig. 2b shows the wideband spectrogram of a typical intervocalic semivowel. Contrast the sharp energy offset and energy onset at the vowel–nasal and nasal–vowel boundaries with the small energy offset and energy onset at the vowel–semivowel and semivowel–vowel boundaries. As the energy onsets/offsets are only measured at the vowel–sonorant consonant boundaries, we have an onset for prevocalic sonorant consonants, an offset for postvocalic sonorant consonants and an offset and an onset for intervocalic sonorant consonants. In this work we have used the onset/offset detector implemented by Salomon et al. (2004). The strength of the onset/offset was obtained by searching for a peak/dip in the onset/offset waveform between the center of the following/preceding vowel and the center of the sonorant consonant. For the intervocalic case onset/offset strength = max(onset strength, −offset strength), where the onset strength and offset strength were obtained as described above. Our experiments indicated that this simplification does not compromise performance but provides the benefit of having the same number of parameters for all 3 cases. Fig. 3 shows the histograms for the energy onsets/offsets for the three cases of prevocalic, postvocalic and intervocalic sonorant consonants. It can be seen from the histograms that nasals, in general, have higher onsets and offsets as compared to semivowels.

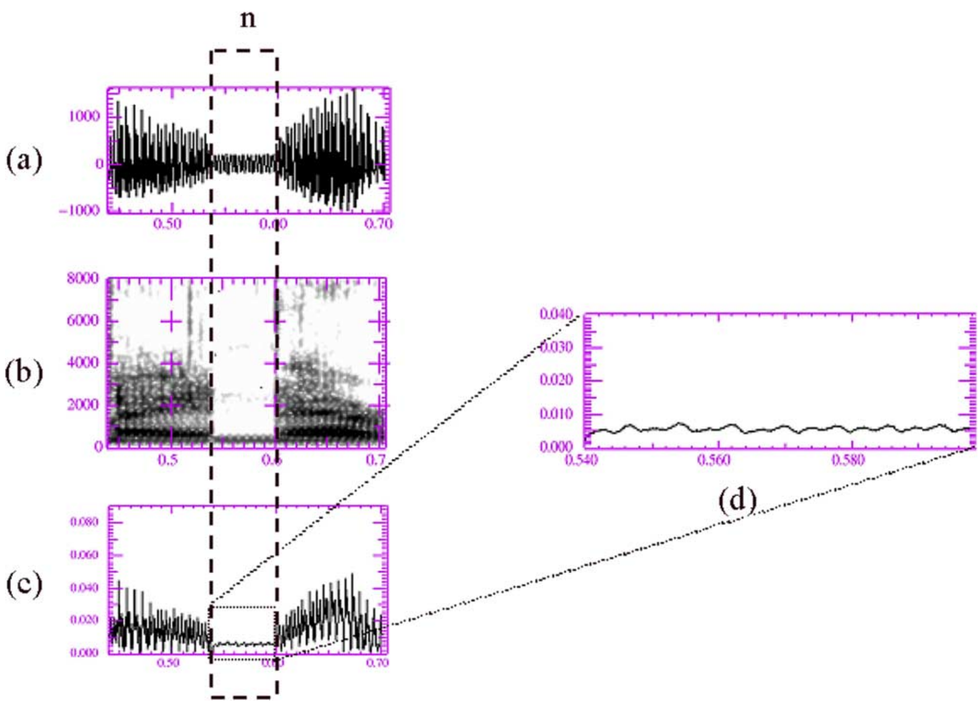

Fig. 1. A typical nasal. The figure shows underlined portion of the sentence "Women may never become completely equal to men". (a) Signal waveform, (b) wideband spectrogram, (c) Hilbert envelope, and (d) zoomed-in portion of the Hilbert envelope in the nasal region.

2.2. Energy ratio

Observe the presence of energy at very low frequency and the sudden drop in energy above it for the nasal in Fig. 1b as opposed to the semivowel in Fig. 2b with considerably more high frequency energy. We used the ratio of energies between 0–320 and 320–5360 Hz for our energy ratio parameter. These frequency values were obtained by an automatic parameter optimization technique described in detail in Bitar and Espy-Wilson (1997). Fig. 4 shows histograms for the energy ratio parameter for the three cases of prevocalic, postvocalic and intervocalic sonorant consonants. The histograms show that nasals have higher values for the energy ratio than semivowels.

2.3. Spectral peak frequency

Semivowels tend to have a much higher frequency value for the first spectral prominence as compared to nasals. This is evident from a comparison of Figs. 1b and 2b. In this case the prominence occurs at about 250 Hz for the nasal as opposed to 430 Hz for the semivowel. Our spectral peak frequency measure is the frequency corresponding to the maximum of the log magnitude FFT spectrum in 0–800 Hz range. This frequency range was also obtained by the automatic parameter optimization technique described in detail in Bitar and Espy-Wilson (1997). Chen (2000) suggested a frequency range of 0–788 Hz which is very similar to the range we obtained by our automatic parameter optimization technique. For the energy ratio and spectral peak frequency parameters we used a 512-point FFT calculated with a hanning window of size 25 ms and a shift of 2.5 ms. These two parameters were calculated for the center five frames of the nasal/semivowel segment and averaged to get the final value. Fig. 5 shows the histograms for the spectral peak frequency parameter for the three cases of prevocalic,

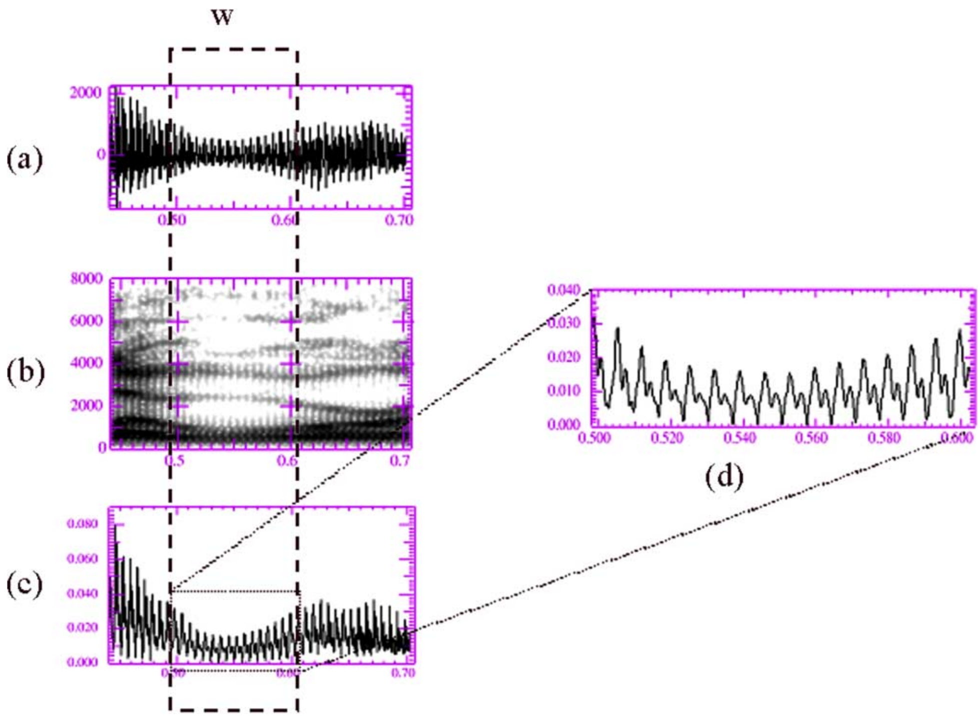

Fig. 2. A typical semivowel. The figure shows underlined portion of the sentence "The prowler wore a ski mask for disguise". (a) Signal waveform, (b) wideband spectrogram, (c) Hilbert envelope, and (d) zoomed-in portion of the Hilbert envelope in the semivowel region.

postvocalic and intervocalic sonorant consonants. The histograms show that nasals tend to have lower values for the spectral peak frequency as compared to semivowels.

2.4. Envelope variance measure

Figs. 1c and 2c show the Hilbert envelope of the signal, and Figs. 1d and 2d show zoomed-in portions of the envelope during the nasal region and the semivowel region respectively. It is clear from these figures that nasals tend to have a much more constant envelope as compared to semivowels. Thus, the energy stability during the nasal murmur can be captured by a parameter that aims to quantify the variance of the envelope. To get this stability measure, we used the standard deviation of the Hilbert envelope of the signal waveform. Fig. 6 shows the histograms for the envelope variance parameter for the three cases of prevocalic, postvocalic and intervocalic sonorant consonants. The histograms show that nasals have a much smaller standard deviation in the waveform envelope when compared to semivowels.

3. Classification experiment

This experiment evaluated the efficiency of the proposed APs on a classification task. Given a set of nasal and semivowel segments, the system had to assign a class to each of the segments.

3.1. Database

The training data consisted of 600 instances each of prevocalic, postvocalic and intervocalic nasals and semivowels totaling to 1800 nasal and 1800 semivowel tokens. These tokens were chosen randomly from 2586 'si' and 'sx' sentences spoken by 90 female and 235 male speakers from the dialect regions 1–7 of the TIMIT (1990) training

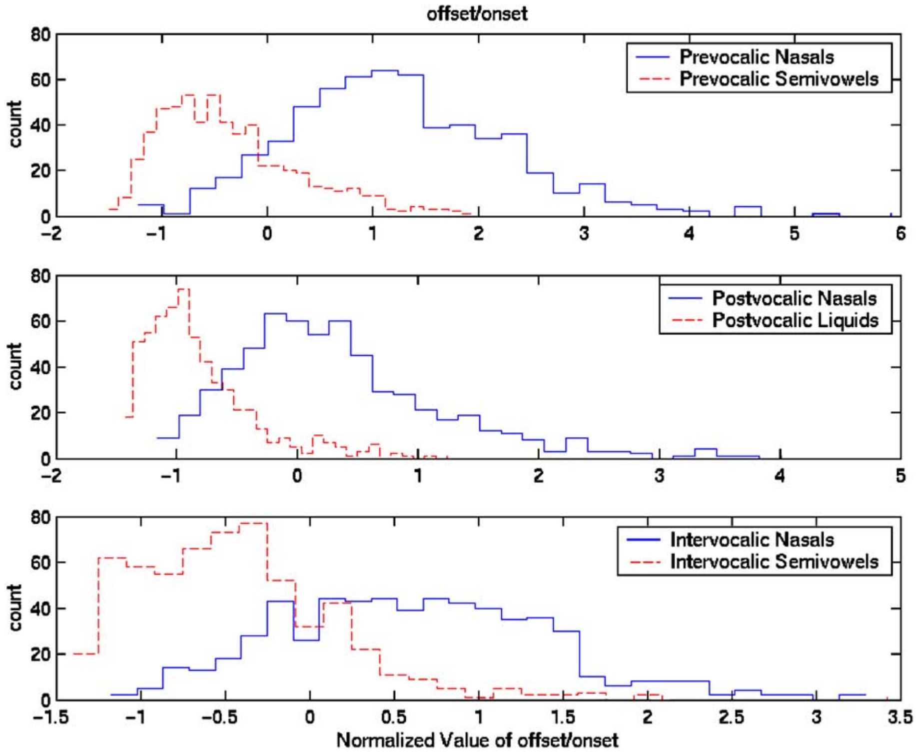

Fig. 3. Histograms for the energy offsets/onsets.

database. The test data consisted of 504 'si' sentences spoken by 56 female and 112 male speakers from dialect regions 1–8 of the TIMIT test database.

3.2. Method

In this experiment, we used the TIMIT transcription to get the nasal and semivowel boundaries, and to classify them into prevocalic, postvocalic and intervocalic categories. The TIMIT labels used for the various classes are given in Table 1. Nasal flaps (/nx/) and syllabic nasals (/em/,/en/) were not included in this study.

The APs were normalized to have zero mean and unit variance, and were used to train three different Support Vector Machines (SVMs) (Burges, 1998; Vapnik, 1995), one each for prevocalic, postvocalic and intervocalic sonorant consonants. The experiments were carried out using the *SVMlight* toolkit (Joachims, 1999). We used only linear kernels. The test data samples were classified as belonging to class +1 (nasals) if the classifier output was positive, and as belonging to class −1 (semivowels) if the classifier output was negative.

3.3. Results and conclusions

Tables 2–4 give confusion matrices of the classification results for prevocalic, postvocalic and intervocalic sonorant consonants in the test database. Averaging across nasals and semivowels gives correct identification rates of 89.53%, 95.80% and 87.82% for prevocalic, postvocalic and intervocalic sonorant consonants respectively. Weighted average of these values gives overall correct identification rate of 91.52% for sonorant consonants. Averaging across the three classes of sonorant consonants gives average accuracies of 94.27% and 89.11% for nasals and semivowels respectively.

The best results were obtained for postvocalic sonorant consonants. One reason for this is that

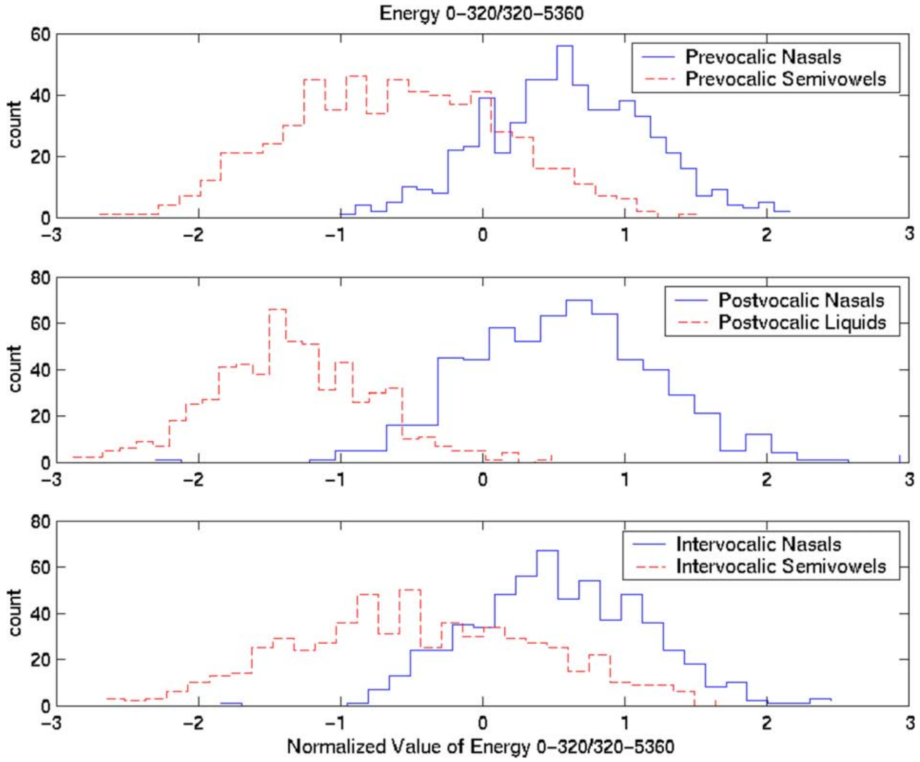

Fig. 4. Histograms for the energy ratio.

we do not have the glides /w/ and /y/ in postvocalic contexts and these phonemes constitute a major source of errors in prevocalic and intervocalic contexts. This is not unexpected because /w/ has a very low Fl and F2 leading to a concentration of energy in the low frequency regions, and /y/ has a very low Fl and a high F2 leading to a sudden drop in energy at frequencies immediately above the first formant region. Two of our parameters, spectral peak frequency measure and energy ratio, capture exactly this information and therefore can create confusions between nasals and glides. Almost all the nasals which are misclassified have a very low value for the energy ratio either because of the high frequency of the prominent spectral peak in the 0–800 Hz region, or simply because of the presence of a lot of energy at frequencies above 320 Hz in the nasal murmur. The presence of a significant amount of energy in the nasal murmur also leads to smaller onsets and offsets in energy at the nasal boundaries. Thus, our parameters fail to classify such cases properly. An example of a nasal with significant amount of energy at high frequencies is given in Fig. 7.

The nasal /ng/ has a tendency to have a lot more energy at high frequencies as compared to /m/ and /n/. The reason for this could be that the first antiresonance of /ng/ occurs at a frequency above 3000 Hz (Fujimura, 1962a,b; Stevens, 1998), which is much higher than the frequency of the first antiresonance for either /m/ or /n/. Further, the movement of the tongue body at the velar closure/release is much slower than the tongue tip movement for an alveolar closure/release or the lip movement for a labial closure/release (Stevens, 1998). Thus, the onsets and offsets are not as sharp for /ng/ as for /m/ or /n/. This could explain the poorer recognition rates of /ng/ as compared to /m/ and /n/.

Mermelstein (1977) reported that /l/ and /r/ before high vowels are often confused with nasals. This is also true for the results obtained in this

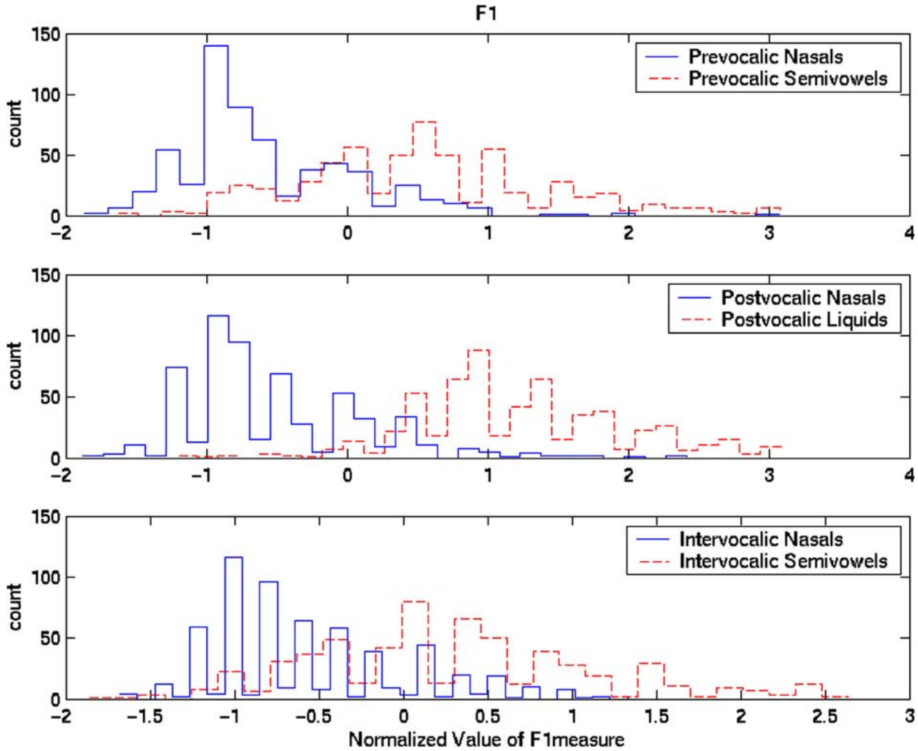

Fig. 5. Histograms for the spectral peak frequency.

study. For the case of prevocalic semivowels, 31 out of the 47 incorrectly classified /l/'s occur before the high vowels /ih/, /iy/, /uh/ and /uw/, and 4 out of 8 incorrectly classified /r/'s occur before high vowels. For the intervocalic case, however, only 12 out of 34 incorrectly classified /l/'s and 3 out of 13 incorrectly classified /r/'s occur before high vowels. Another reason for misclassifications is that in a lot of these cases the nasals/semivowels occur before unstressed vowels. In this context, the phonemes fail to show many of their characteristic properties suggesting that they are not fully articulated. Espy-Wilson (1992) also noted that the onsets of prevocalic sonorant consonants before unstressed vowels are much less than those of prevocalic sonorant consonants before stressed vowels. Further, onsets and offsets of intervocalic sonorant consonants in a falling stress are smaller than those of intervocalic sonorant consonants in a rising stress environment.

Since we are relying heavily on the transcription provided with the TIMIT database to classify sonorant consonants as prevocalic, postvocalic and intervocalic, and to get the phoneme boundaries, it is a potential source of errors. One of the reasons for errors based on the transcription is that the transcription might be incorrect. An example of incorrect transcription in TIMIT is given in Fig. 8. The spectrogram of the transcribed nasal only shows frication, and the transcribed region is also heard as just frication. A nasal is audible in the neighborhood, but not where it is marked.

In some cases, nasals which are realized completely as nasalized vowels are marked in the transcription as sonorant consonants. This also constitutes an error. An example of such a case is given in Fig. 9. There is no energy offset/onset at the boundaries of the transcribed intervocalic nasal. The two arrows in the figure show two low frequency peaks running through the transcribed nasal from, and into the preceding and following vowels. The presence of an extra peak in the first formant region is just the cue described by Chen

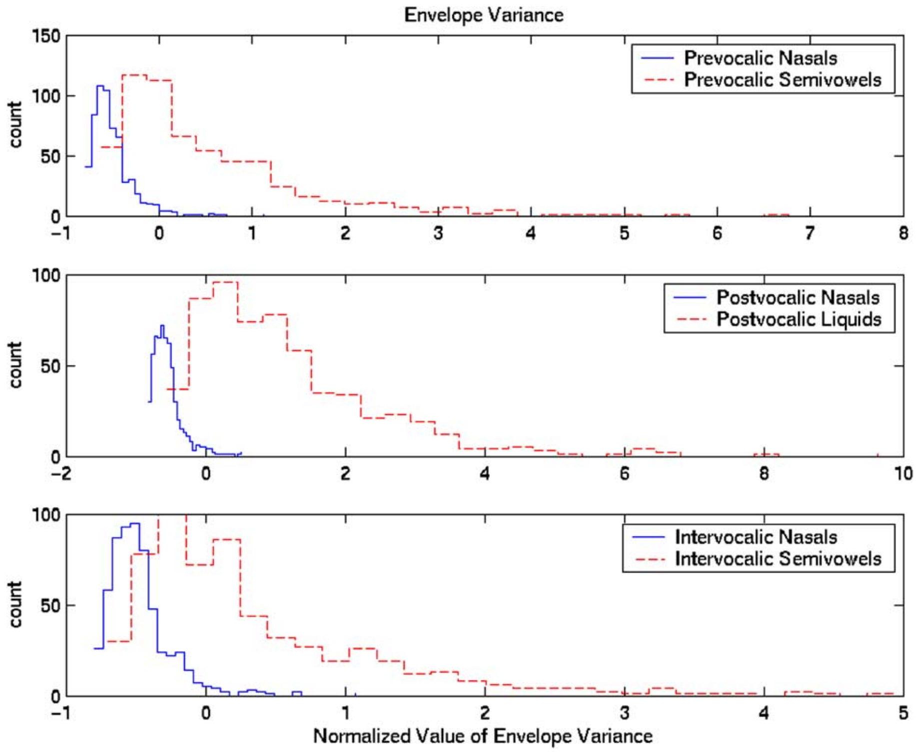

Fig. 6. Histograms for the envelope variance parameter.

Table 1
TIMIT labels used for different categories

Category	TIMIT labels
Vowels	/iy/, /ih/, /eh/, /ey/, /ae/, /aa/, /aw/, /ay/, /ah/, /ao/, /oy/, /ow/, /uh/, /uw/, /ux/, /er/, /ax/, /ix/, /axr/, /ax-h/
Nasals	/m/, /n/, /ng/
Semivowels	/l/, /r/, /w/, /y/

Table 2
Confusion matrix for prevocalic sonorant consonants

Label	Nasals	Semivowels	% Correct
m	149	10	93.71
n	110	9	92.44
l	47	250	84.18
r	8	345	97.73
w	36	212	85.48
y	23	71	75.53
Nasals	259	19	93.17
SVs	114	878	88.51

(1995, 1997) for vowel nasalization. Our parameters currently do not handle vowel nasalization. Thus, such cases contribute to errors in the current scenario.

The APs might be calculated at wrong locations if the boundaries of phonemes are not exact. This could be especially true for semivowels, particularly in the intervocalic cases, where often there are no apparent boundaries separating them from adjacent vowels. For these cases, 1/3 of the sonorant region was assigned to the semivowel and the remaining was assigned to the vowel in TIMIT (1990). An example of inexact boundaries is given in Fig. 10. In this case, the beginning of the /y/ is marked in the middle of the frication for the preceding voiced fricative /v/.

Table 3
Confusion matrix for postvocalic sonorant consonants

Label	Nasals	Semivowels	% Correct
m	186	8	95.88
n	598	27	95.68
ng	137	7	95.14
l	8	217	96.44
r	7	163	95.88
Nasals	921	42	95.64
SVs	15	380	96.20

Table 4
Confusion matrix for intervocalic sonorant consonants

Label	Nasals	Semivowels	% Correct
m	166	16	91.21
n	185	10	94.87
ng	17	7	70.83
l	34	166	83.00
r	13	145	91.77
w	21	89	80.91
y	7	11	61.11
Nasals	368	33	91.77
SVs	75	411	84.57

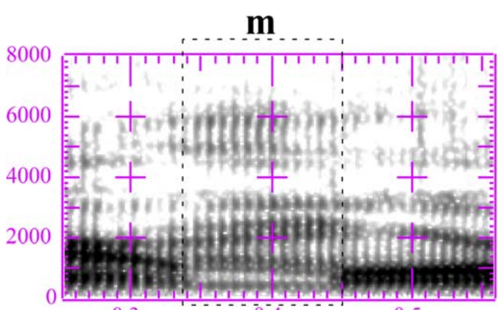

Fig. 7. Wideband spectrogram of a case where the nasal has a lot of energy at higher frequencies. The figure shows underlined portion of the sentence "There are more obvious nymphomaniacs on any private-eye series" from TIMIT.

4. Recognition experiment

The use of the APs developed in our lab in a Hidden Markov Model (HMM) back-end environment was demonstrated in an earlier paper by Deshmukh et al. (2002). It was mentioned in that paper that one reason for the poorer performance

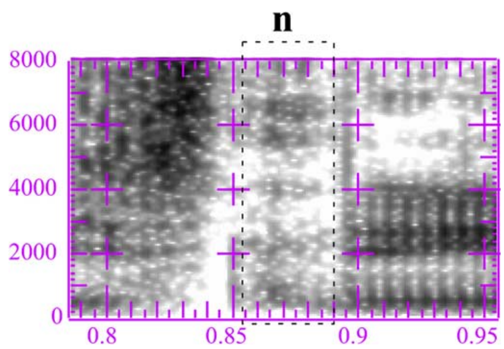

Fig. 8. Wideband spectrogram of a case where the transcription is wrong. The figure shows underlined portion of the sentence "These needs usually concern the reduction of guilt and some relief of tension" from TIMIT.

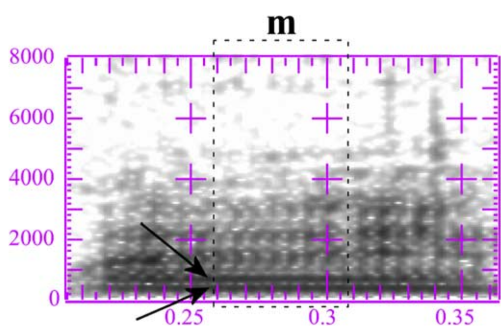

Fig. 9. Wideband spectrogram of a case where the nasal is realized as a nasalized vowel. The figure contains the underlined portion of the sentence "Women may never become completely equal to men" from TIMIT.

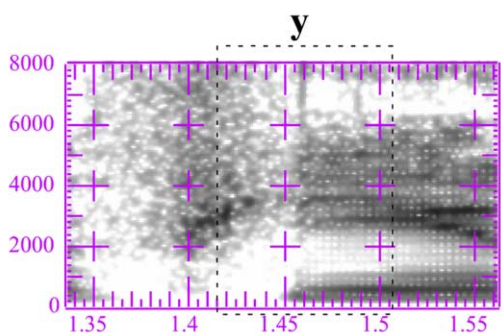

Fig. 10. Wideband spectrogram of a case where the boundaries are not correctly marked. The figure shows underlined portion of the sentence "It suffers from a lack of unity purpose and respect for heroic leadership" from TIMIT.

of our APs as compared to Mel Frequency Cepstral Coefficients (MFCCs) in an experiment with training and testing on adults, was the absence of APs for nasals and other critical features. The purpose of this experiment was to substantiate the claims made in that paper and to quantify the reduction in confusions between English digits (particularly the digits 1 and 9) by adding the APs for nasals proposed in this paper to our old set of APs.

4.1. Database

The TI46 (Liberman, 1993) database was used for this experiment. The training set had 160 utterances of each digit: 10 repetitions by each of the 8 male and 8 female speakers. The test set had 256 utterances of each digit: 16 repetitions by each of the 8 male and 8 female speakers.

4.2. Method

For the purposes of this experiment, all the APs were extracted at each frame of the speech signal and fed into an HMM back-end for recognition of digits. We used the mean of the Hilbert envelope in each frame to convert the segment based parameter to a frame based one. We did not add our F1 measure because the old set of APs already had a F1–F0 parameter which essentially captures the same information that our F1 measure tries to extract. Further, the old set of APs had onset and offset parameters (Bitar, 1997) which were replaced by the new onset and offset parameters (Salomon et al., 2004) used in this work. In all, we had a set of 30 APs instead of the old set of 28. We used a 3-state (+2 terminal states) HMM with 8 Gaussian mixtures in each state for each digit. The models were left-to-right models with no skips and equiprobable transitions. Each mixture was initialized to have zero mean and unit variance. All mixture weights were initialized to the same value. HTK (Young, 1995) was used for these experiments.

4.3. Results and conclusions

The confusion matrix for digits one, five and nine obtained with our old set of APs is shown in Table 5. We only show the results for one, five and nine because those are the only digits where reduction in confusions might be expected because of APs for nasals. Notice that these digits did not have any confusions with any other digits. Table 6 shows the confusion matrix with our new set of APs. A comparison of the results in Tables 5 and 6 shows the marked reduction in the confusions between the digits one, five and nine obtained by adding the APs for nasals to our old set of APs.

The addition of these parameters also reduced some confusions for the digits two, three, four and eight. Of these digits, only the digit four had

Table 5
Confusion matrix with old set of 28 APs

	One	Five	Nine	Total	% Correct
One	250	0	5	255	98.0
Five	0	254	0	254	100.0
Nine	4	2	248	254	97.6

Table 6
Confusion matrix with new set of 30 APs

	One	Five	Nine	Total	% Correct
One	254	0	1	255	99.6
Five	0	254	0	254	100.0
Nine	0	1	253	254	99.6

confusions with one and five which were eliminated. However, the reduction in confusions for two, three, four and eight could be solely because of our new improved onset and offset parameters which have a wide application besides detection of nasals. For example, we expect these onsets and offsets to give us better boundaries for all consonants, not just nasals. The total-word-correct scores increased from 98.82% for the old set of APs to 99.53% for the new set of APs. This corresponds to a 60% reduction in error rate.

5. General discussion

In this paper we have developed APs to distinguish the class of nasals from other classes of sounds. Results of our experiments demonstrate that we are moving in the right direction in terms of capturing the acoustic properties of the feature *nasal*. Moreover, we are doing very well given the fact that our classification experiment was performed on a continuous speech database with a large number of speakers and contextual variations. Unlike all previous research (Chen, 2000; Glass, 1984; Glass and Zue, 1986; Mermelstein, 1977) which focused only on capturing spectral properties of nasals, we used a temporal parameter to capture the stability during the nasal murmur. An interesting observation is that we claim to use a parameter for stability during the nasal murmur region whereas Mermelstein (1977) was in favor of capturing the dynamic transition during the murmur. A close inspection of his results shows that the centroid frequency in 0–500 Hz band, and the relative signal energy in 0–1 kHz band were stable across the murmur duration. This is in consonance with the observation we made earlier that we expect the spectrum in the nasal murmur region to be stable at least in the low frequency regions.

Although we are able to get very good classification rates with these parameters, we believe that there is still scope for improvement. As was highlighted in Section 3.3, a majority of misclassifications of nasals occur when nasals have a lot of high frequency energy. Although, a lot of these cases can be perceived as nasals, our parameters failed to classify such cases properly. This shows that we are still missing out on some relevant information. We already know from Fujimura's (1962a,b) work that we expect to have a higher density of formants and larger bandwidths of formants in the nasal murmur region because of the longer length of the nasal cavity and higher losses in the nasal tract. We still do not have any parameters to capture this information. We originally had a parameter to capture the density of formants (Pruthi and Espy-Wilson, 2003), but we removed it from consideration here because of the inherent problems in reliably extracting formant-based parameters. We might add that parameter later when we are able to find ways to extract the formant density information with sufficient accuracy. Also, we do not have any parameters to capture vowel nasalization. Vowels are almost always nasalized when adjacent to nasals (although there can be variations because of context, speaker or language) (Stevens, 1998). So, nasalization can be an important clue for the presence of nasals especially when the nasals are articulated only as nasalization in vowels. Chen (1995, 1997) had suggested parameters to capture vowel nasalization, but those parameters were not automatically extracted. So, work needs to be done to automate the extraction of these parameters.

In the future, we would also like to do some experiments to see if our APs are invariable to changes in speakers, their age and gender. We expect them to be invariant to speaker changes, because the speakers in the test database in our classification experiment were completely different from those in the training database. But it needs to be seen how these parameters will perform if they are trained for females and tested for males or vice versa. Similarly, it needs to be seen how these parameters perform if they are trained on adults and tested on children or vice versa. It would also be interesting to see if the APs, as trained for American English, are able to detect nasals in another language, may be one with a larger number of nasals. That would be a very good test to show that these parameters really capture the acoustic correlates of nasality independent of the actual phoneme or the language.

6. Conclusions

In conclusion, we have developed acoustic correlates for the linguistic feature *nasal* to distinguish nasals as a class from other classes of sounds. These parameters can be extracted automatically and reliably from the acoustic signal. An interesting feature in this work was the use of a temporal parameter to capture stability in the nasal murmur. This AP is very different from the spectral parameters that have been the focus of all previous studies. The recognition experiment described in this paper illustrates that we are moving in the right direction in terms of capturing the properties of nasals. Although we were able to obtain very high classification rates, we believe there is still scope for improvement. One area which needs work is development of APs to capture nasalization in vowels. We believe this can potentially give a big improvement in the detection of nasals especially because at times nasals might be articulated only as vowel nasalization. In the future, we would also like to work on improving the classification rates for the nasal /ng/.

Acknowledgements

This work was supported in part by NIH grant 1 K02DC00149-01A1. We would also like to thank Steven Greenberg and an anonymous reviewer for their helpful comments on an earlier version of this paper.

References

Bitar, N.N., 1997. Acoustic analysis and modeling of speech based on phonetic features. Ph.D. Thesis, Boston University.

Bitar, N.N., Espy-Wilson, C.Y., 1997. The design of acoustic parameters for speaker-independent speech recognition. In: Proceedings of Eurospeech. pp. 1239–1242.

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Chen, M.Y., 1995. Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers. Journal of the Acoustical Society of America 98, 2443–2453.

Chen, M.Y., 1997. Acoustic correlates of english and french nasalized vowels. Journal of the Acoustical Society of America 102, 2360–2370.

Chen, M.Y., 2000. Nasal detection module for a knowledge-based speech recognition system. In: Proceedings of the ICSLP 2000, vol. IV. Beijing, China, pp. 636–639.

Deshmukh, O., Espy-Wilson, C.Y., Juneja, A., 2002. Acoustic-phonetic speech parameters for speaker-independent speech recognition. In: Proceedings of ICASSP 2002. pp. 593–596.

Dickson, D.R., 1962. Acoustic study of nasality. Journal of Speech and Hearing Research 5 (2), 103–111.

Dixon, N.R., Silverman, H.F., 1976. A general language operated decision implementation system (GLODIS): Its application to continuous-speech segmentation. IEEE Transactions ASSP 24, 137–162.

Espy-Wilson, C., 1992. Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English. Journal of the Acoustical Society of America 92 (2), 736–757.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton, The Hague, Netherlands.

Flanagan, J.L., 1965. Speech Analysis Synthesis and Perception. Springer-Verlag, New York.

Fujimura, O., 1962a. Analysis of nasal consonants. Journal of the Acoustical Society of America 34 (12), 1865–1875.

Fujimura, O., 1962b. Formant-antiformant structure of nasal murmurs. In: Proceedings of the Speech Communication Seminar, vol. 1. Stockholm, Royal Institute of Technology, Speech Transmission Laboratory, pp. 1–9.

Glass, J.R., 1984. Nasal consonants and nasalised vowels: An acoustical study and recognition experiment. M.S. and E.E. Thesis, MIT, Cambridge, MA.

Glass, J.R., Zue, V.W., 1985. Detection of nasalized vowels in American English. In: Proceedings of ICASSP, pp. 1569–1572.

Glass, J.R., Zue, V.W., 1986. Detection and recognition of Nasal consonants in American English. In: Proceedings of ICASSP. pp. 2767–2770.

Hess, W.J., 1976. A pitch-synchronous digital feature extraction system for phonemic recognition of speech. IEEE Transactions ASSP 24, 14–25.

House, A.S., 1957. Analog studies of nasal consonants. Journal of Speech and Hearing Disorders 22, 190–204.

Joachims, T., 1999. Making large-scale SVM learning practical. In: Schlkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods—Support Vector Learning. MIT-Press.

Juneja, A., Espy-Wilson, C., 2002. Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning. In: Proceedings of 9th International Conference on Neural Information Processing, vol. 2. Singapore, pp. 726–730.

Juneja, A., Espy-Wilson, C., 2003. Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. In: Proceedings of International Joint Conference on Neural Networks. Portland, Oregon.

Liberman, M., 1993. T146-Word, Available from <http://www.ldc.upenn.edu>.

Liu, S.A., 1996. Landmark detection for distinctive feature-based speech recognition. Journal of the Acoustical Society of America 100, 3417–3430.

Mermelstein, P., 1977. On detecting nasals in continuous speech. Journal of the Acoustical Society of America 61 (2), 581–587.

Nakata, K., 1959. Synthesis and perception of nasal consonants. Journal of the Acoustical Society of America 31, 661–666.

Pruthi, T., Espy-Wilson, C.Y., 2003. Automatic classification of nasals and semivowels. In: Proceedings of 15th International Congress of Phonetic Sciences (ICPhS). Barcelona, Spain.

Salomon, A., Espy-Wilson, C., Deshmukh, O., 2004. Detection of speech landmarks: Use of temporal information. Journal of the Acoustical Society of America 115 (3), 1296–1305.

Stevens, K.N., 1998. Acoustic Phonetics. MIT.

TIMIT Acoustic-Phonetic Continuous Speech Corpus, National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-5050651996, October 1990.

Vapnik, V.N., 1995. Nature of Statistical Learning Theory. Springer-Verlag.

Weinstein, C.J., McCandless, S.S., Mondshein, L.F., Zue, V.W., 1975. A system for acoustic-phonetic analysis of continuous speech. IEEE Transactions ASSP 23, 54–67.

Young, S., 1995. The HTK Book. Cambridge Research Lab: Entropics, Cambridge, England, Available from <http://htk.eng.cam.ac.uk/>.