Automatic Classification of Nasals and Semivowels

Tarun Pruthi, Carol Y.Espy-Wilson
University of Maryland, College Park, MD 20742
E-mail: tpruthi@glue.umd.edu, espy@glue.umd.edu

ABSTRACT

In this paper, we discuss acoustic parameters and a classifier we developed to distinguish between nasals and semivowels. Based on the literature and our own acoustic studies, we use an onset/offset measure to capture the consonantal nature of nasals, and an energy ratio, a low spectral peak measure and a formant density measure to capture the nasal murmur. These acoustic parameters are combined using Support Vector Machine based classifiers. Classification accuracies of 88.6%, 94.9% and 85.0% were obtained for prevocalic, postvocalic and intervocalic sonorant consonants, respectively. The overall classification rate was 92.4% for nasals and 88.1% for semivowels. These results have been obtained for the TIMIT database, which was collected from a large number of speakers and contains substantial coarticulatory effects.

1. INTRODUCTION

We are working on an Event Based Speech recognition system (EBS) that combines knowledge-based acoustic parameters (APs) with a statistical framework for recognition [6]. In EBS, the speech signal is first segmented into the broad classes: vowel, sonorant consonant, strong fricative, weak fricative and stop. This segmentation is based on acoustic events (or landmarks) obtained in the extraction of the APs associated with the laryngeal phonetic feature *voice* and the manner phonetic features *sonorant*, *syllabic*, *continuant* and *strident* (in addition to silence). The manner acoustic events are then used to extract additional parameters relevant for the laryngeal phonetic feature *voice* and for the place phonetic features. In this project, we sought to add APs for the manner feature *nasal*.

The phonetic feature hierarchy shown in Figure 1 illustrates our recognition strategy. Nasals and semivowels are separated from all the other sounds by the phonetic features *sonorant* and *syllabic*. In this paper, we focus on the development of acoustic parameters (APs) that will help in the separation of nasal consonants from semivowels: *nasal* and *consonantal*.

Nasals as a class of phonemes have been difficult to recognize automatically, the primary reason being the presence of zeros in the nasal spectrum as a distinguishing characteristic. Zeros don't always manifest themselves as clear dips in the spectrum amplitude, given the line spectrum of periodic sounds and the possibility of pole-zero cancellations.

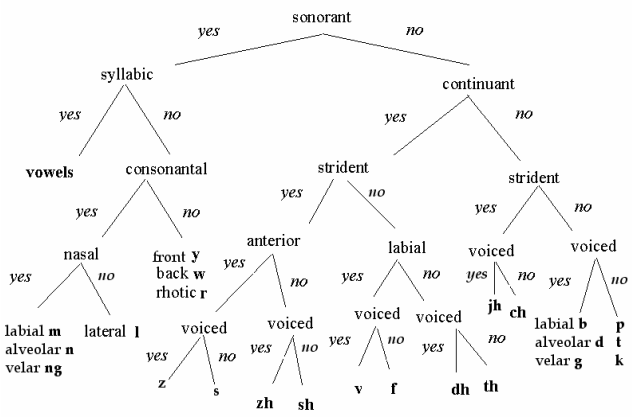

Figure 1 Phonetic Feature Hierarchy

Normally, cues from three different regions indicate the presence of a nasal consonant: abrupt spectral change between the nasal and an adjacent sonorant, vowel nasalisation, and the presence of the nasal murmur. Fujimura [1],[2] identified four properties which characterize the spectra of nasal murmurs in general: (1) the existence of a very low first formant that is located at about 300 Hz and is well separated from the upper formant structure, (2) relatively high damping factors of the formants, (3) high density of the formants in the frequency domain and (4) the existence of the zeros in the spectrum.

Several researchers have developed acoustic measures for capturing the properties of nasals [3], [4]. Experimentation with these proposed measurements led us to the development and/or selection of several acoustic parameters which include an onset/offset measure to capture the consonantal nature of nasals (abrupt spectral change that often occurs at the onset and release of nasals), and an energy ratio, a low spectral peak measure and a formant density measure to capture the nasal murmur. The four APs are combined using a separate Support Vector Machine (SVM) [7], [8] based classifier for the three different cases of prevocalic, postvocalic and intervocalic sonorant consonants. Note that we do not consider cues for vowel nasalisation in the present study.

The rest of the paper is organized as follows: In section 2 we give the details of the database used. Sections 3 talks about the experimental method, and sections 4 and 5 give the results and conclusions respectively.

ISBN 1-876346-48-5 © 2003 UAB

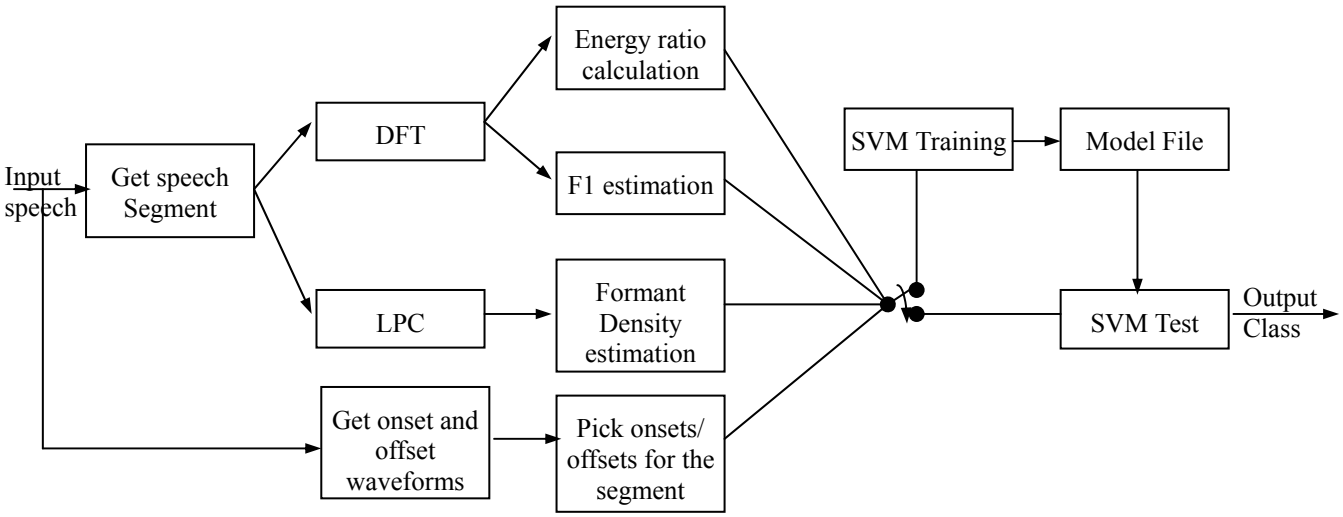

Figure 2 Flow graph for the estimation of acoustic parameters and nasal/semivowel classification

2. DATABASE

The training data consisted of 400 each of prevocalic nasals and semivowels, 400 each of postvocalic nasals and semivowels, and 500 each of intervocalic nasals and semivowels. These tokens were chosen randomly from 2586 'si' and 'sx' sentences spoken by 90 females and 235 males from the dialect regions 1-7 of the TIMIT [10] training database. The test data consisted of 504 'si' sentences spoken by 56 females and 112 males from the dialect regions 1-8 of the TIMIT test database.

3. METHOD

In our experiments, the TIMIT transcription was used to identify the nasal and semivowel boundaries, and to classify them into prevocalic (preceding a vowel), postvocalic (following a vowel) and intervocalic (vowels preceding and following). The TIMIT labels used for the various classes are given in Table 1 below.

Table 1 TIMIT labels used for different categories

Category	TIMIT label
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/

Nasal flaps (/nx/) and syllabic nasals (/em/ or /en/) were not included in this study.

A 25 ms hanning window and a frame rate of 2.5 ms were used for analysis. The energy ratio E(0-358 Hz)/E(358-5373 Hz) was obtained from the power spectrum. An estimate of F1 was obtained by finding the frequency of the peak of the log magnitude spectrum in 0-788 Hz range (as suggested in [4]). We calculated the parameters from the center 5 frames of the FFT of the segment and then averaged them. An indirect measure of the formant density was obtained by counting the average number of peaks in the LPC spectrum between 0-2500 Hz. A 30^{th} order linear predictor was used. The onsets and offsets are obtained by searching for a peak in the onset waveform (onset and offset waveforms being obtained by the procedure outlined in [5]) between the center of the sonorant consonant and the center of the following vowel for prevocalic sonorant consonant, a minimum in the offset waveform for postvocalic, and a peak in the onset and a minimum in the offset waveform for the case of intervocalic sonorant consonants. For the case of intervocalic sonorant consonants, the onset and offset are collapsed into one single parameter so as to have the same number of parameters for all the 3 cases. Our experiments indicated that this simplification does not compromise performance. A flow graph for the estimation of the APs is given in Figure 2.

The APs were normalized to have 0 mean and unit variance to reduce variation in their dynamic range and improve the training of the SVMs, which might otherwise put the largest significance on the parameter with the maximum dynamic range. The four APs are then used for training three different SVMs, one each for prevocalic, postvocalic and intervocalic sonorant consonants. We use an equal number of training data samples for the 2 classes in all cases. The experiments in this project were carried out using the SVM*light* toolkit [9], which provides very fast training of SVMs. We used only linear kernels in this case. The test data is generated in a similar manner and normalized to have 0 mean and unit variance. The test data samples are classified as belonging to class +1 (nasals) if the classifier output is positive, and as belonging to class –1 (semivowels) if the classifier output is negative.

ISBN 1-876346-48-5 © 2003 UAB

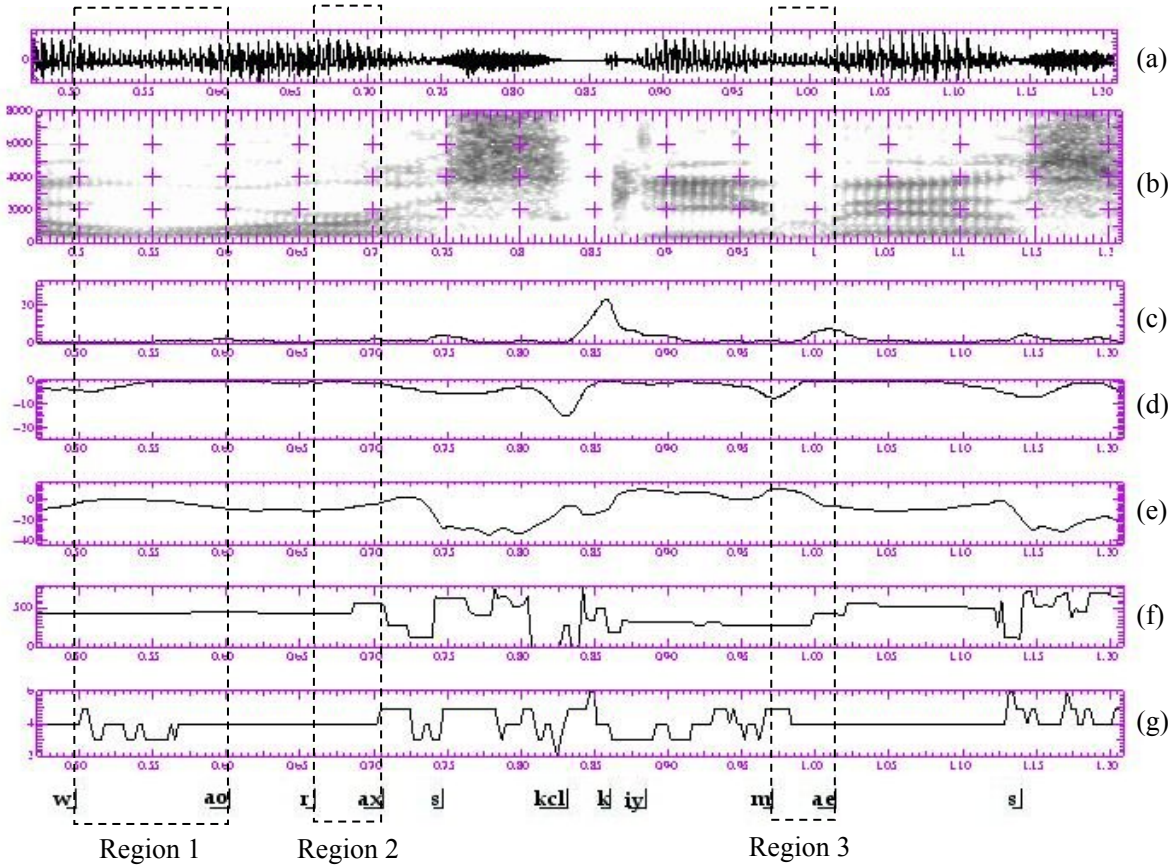

Figure 3 Acoustic parameters for an excerpt "wore a ski mask" from the file dr1.mgrl0.sx57.wav from TIMIT training database. (a) Waveform, (b) Wideband Spectrogram, (c) onsets, (d) offsets, (e) energy ratio, (f) F1 measure, (g) Formant density measure

4. RESULTS

An example of the APs extracted

In Figure 3 we give an example of the acoustic parameters that have been extracted for the distinction of nasals and semivowels. The figure shows an excerpt from the file sx57.wav spoken by speaker "mgrl0" from dialect region 1 of the TIMIT training database. The parameters have been extracted for the whole file for this example. Hence, they do not exactly correspond to the parameters that would be actually used for our purpose where we focus the calculations on segments and take averages in those segments. In fact, the parameters aren't relevant outside the marked regions (regions 1, 2 and 3). Further, it also shows that these parameters just extract estimates of F1 and formant density instead of the actual values. However, the above example should give a fair idea of what the parameters look like and their utility. Clearly the nasal region (region 3) is the only region where we see a good offset at the nasal closure and a good onset at the nasal release. Also, in this region we see a high value of the energy ratio, a very low value of F1 measure, and a high value for the formant density parameter. The energy ratio is almost always negative and F1 measure is high in region 2 corresponding to /r/. Also, there is no visible onset/offset for this region. The energy ratio is high for the case of /w/ (region 1), and it also shows a small onset and offset at the boundaries, but here the formant density measure is low and the F1 measure is high.

Results

Tables 2-4 give confusion matrices from the classification results for prevocalic, postvocalic and intervocalic sonorant consonants in the test database.

Table 2 Confusion matrix of the classification results for prevocalic sonorant consonants

	Nasals	Semivowels	% Correct
Nasals	250	28	89.93
Semivowels	117	875	88.21

Table 3 Confusion matrix of the classification results for postvocalic sonorant consonants

	Nasals	Semivowels	% Correct
Nasals	912	51	94.70
Semivowels	19	376	95.19

ISBN 1-876346-48-5 © 2003 UAB

Table 4 Confusion matrix of the classification results for intervocalic sonorant consonants

	Nasals	Semivowels	% Correct
Nasals	355	46	88.53
Semivowels	87	399	82.10

Averaging across the three classes gives classification accuracies of 88.6%, 94.9% and 85.0% for prevocalic, postvocalic and intervocalic sonorant consonants respectively, and average accuracies of 92.4 % and 88.1 % for nasals and semivowels respectively. The correct identification rate for sonorant consonants is 90.1%.

We get the best results for postvocalic sonorant consonants. A possible reason could be that we don't have any postvocalic /w/ and /y/ and these are the ones which were confusing the most with nasals. 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 for errors. This could be especially true for semivowels where often there are no apparent boundaries separating them for adjacent vowels. In this case, 1/3 of the sonorant regions was assigned to the semivowel and the remaining was assigned to the vowel [10].

5. CONCLUSIONS AND FUTURE WORK

These are very good results considering the fact that all experiments have been performed on the TIMIT database which was collected from a large number of speakers and contains substantial coarticulatory effects. In future experiments, we will integrate additional phonetic features such as *lateral* and *rhotic* to help in this distinction as well. We will also integrate these features with the existing EBS so as to go one step further down in the phonetic feature hierarchy to get to the level of distinguishing between the two classes of sonorant consonants, nasals and semivowels. Finally, we plan to develop APs for the place features *labial*, *alveolar* and *velar* for nasals.

REFERENCES

[1] O. Fujimura, "Analysis of Nasal Consonants," *J. of Acous. Soc. of Am.*, Vol. 34, No. 12, pp 1865-1875, 1962.

[2] O. Fujimura, "Formant-Antiformant structure of Nasal Murmurs," *Proceedings of the Speech Communication Seminar,* Vol 1, Stockholm: Royal Institute of Technology, Speech Transmission Laboratory, pp 1-9, 1962.

[3] J.R. Glass, *Nasal Consonants and Nasalised Vowels: An Acoustical Study and Recognition Experiment,* M.S. and E.E. thesis, MIT, Cambridge, MA, 1984.

[4] M.Y. Chen, "Nasal Detection Module for a Knowledge-based Speech Recognition System," *Proceedings of the ICSLP 2000*, Vol. IV, pp. 636-639, Beijing, China, 2000.

[5] A. Salomon, C. Espy-Wilson, O. Deshmukh, "Detection of speech landmarks from temporal information," *J. of Acoust. Soc. of Am.*, in revision.

[6] A. Juneja, C. Espy-Wilson, "Segmentation of Continuous Speech Using Acoustic-Phonetic Parameters and Statistical Learning," *Proceedings of 9th International Conference on Neural Information Processing*, Volume 2, pp. 726-730, Singapore, 2003.

[7] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 1995.

[8] Christopher J C Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Bell Laboratories, Lucent Technologies, 1998.

[9] T. Joachims, "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning," B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

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

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant 1 K02 DC00149-01A1.