Acoustic Parameters for the Automatic Detection of Vowel Nasalization

Tarun Pruthi, Carol Y. Espy-Wilson

Institute of Systems Research and Dept. of Electrical and Computer Engg.
University of Maryland, College Park, MD 20742, USA.
{tpruthi, espy}@umd.edu

10.21437/Interspeech.2007-40

Abstract

The aim of this work was to propose Acoustic Parameters (APs) for the automatic detection of vowel nasalization based on prior knowledge of the acoustics of nasalized vowels. Nine automatically extractable APs were proposed to capture the most important acoustic correlates of vowel nasalization (extra pole-zero pairs, F1 amplitude reduction, F1 bandwidth increase and spectral flattening). The performance of these APs was tested on several databases with different sampling rates and recording conditions. Accuracies of 96.28%, 77.90% and 69.58% were obtained by using these APs on StoryDB, TIMIT and WS96/97 databases, respectively, in a Support Vector Machine classifier framework. To our knowledge these results are the best anyone has achieved on this task.

Index Terms: nasal, nasalization, acoustic parameters, landmark, speech recognition.

1. Introduction

Nasalization in very simple terms is the nasal coloring of other sounds. Nasalization occurs when the *velum* (a flap of tissue connected to the posterior end of the hard palate) drops to allow coupling between the oral and nasal cavities. When this happens, the oral cavity is still the major source of output but the sound gets a distinctively nasal characteristic.

Coarticulatory nasalization of the vowel preceding a nasal consonant is a regular phenomenon in all languages of the world. The coarticulation can, however, be so large that 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) is completely deleted and the cue for the nasal consonant is only present as nasalization in the preceding vowel. This is especially true for spontaneous speech. Thus, for example, nasalization of the vowel might be the only feature distinguishing ”cat” from ”can't”. Hence, the automatic detection of vowel nasalization is an important problem. A vowel nasalization detector is also essential for speech recognition in languages with *phonemic nasalization* (i.e. there are minimal pairs of words in such languages which differ in meaning with just a change in the nasalization in the vowel), and therefore, should be an important part of a landmark-based speech recognition system (like [1]). Further, it was suggested in [2] that detection of vowel nasalization is important to give the pronunciation model the ability to learn that a nasalized vowel is a high probability substitute for a nasal consonant.

Therefore, the aim of this study was to develop Acoustic Parameters (APs) for the automatic detection of vowel nasalization based on prior knowledge about the acoustics of nasalized vowels. The rest of the paper is organized as follows: Section 2 gives a brief description of the databases used. Section 3 gives a description of the proposed APs, and section 4 describes the methodology in detail. Section 5 presents results on the various databases of American English described in Section 2. Section 6 summarizes the most important points of this study and presents future directions.

2. Databases

2.1. StoryDB

Acoustic recordings of carefully articulated isolated words with seven vowels /aa, ae, ah, eh, ih, iy, uw/ in nasalized and non-nasalized contexts were obtained for one American English speaker. The sampling rate was 16 KHz. A total of 896 words were recorded: 7 vowels x 8 words/vowel x 4 conditions (standing and supine, with and without the application of Afrin) x 4 repetitions. The database was divided equally into train and test databases. All words were manually segmented to mark the beginning and ending of the vowels in consideration. For the purposes of testing the proposed APs it was assumed that every vowel before a nasal consonant is nasalized. Thus, this database was the simplest test case.

2.2. TIMIT

TIMIT [8] contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers (438 males, 192 females) from 8 major dialect regions of the United States. The speech files were sampled at 16 KHz and divided into training and testing sets. While using this database, it was assumed that all vowels preceding nasal consonants are nasalized. The set of nasal consonants included /m/, /n/, /ng/ and /nx/. Further, all syllabic nasals (/em/, /en/ and /eng/) were considered to be nasalized vowels. Vowels were considered to be oral/non-nasalized when they were not in the context of nasal consonants or syllabic nasals. This definition, would, however, classify vowels in words like /film/ as oral vowels even though the vowel (being in the same syllable as the nasal consonant) would most likely be nasalized due to anticipatory coarticulation with the syllable-final nasal consonant /m/. Since, such cases may be somewhat ambiguous, they were removed from consideration by not considering vowels as oral when the second phoneme after the vowel was a nasal consonant. This condition is similar to that imposed by [7]. In the case of vowels following nasal consonants, nasalization might not be very strong. Hence, these cases were also removed from consideration.

2.3. WS96/97

WS96 and WS97 databases are parts of the switchboard corpus [9] which were phonetically transcribed in workshops at Johns Hopkins University in 1996 and 1997 respectively. These

Table 1: A list of the acoustic correlates of vowel nasalization and the APs used to capture them.

Acoustic Correlate	Proposed APs
Extra peaks at low frequencies and the relative amplitudes of these peaks as compared to the first formant amplitude	• $sgA1 - P0$, where $A1$ is the amplitude of the first formant, and $P0$ is the amplitude of an extra peak below $F1$. The prefix sg implies that a combination of cepstrally smoothed spectra (s) and group delay spectra (g) was used to find the exact location of the extra peaks. $F1$ was obtained by using the ESPS formant tracker [3]. • $sgA1 - P1$, where $P1$ is the amplitude of an extra peak above $F1$. The APs, $sgA1 - P0$ and $sgA1 - P1$ are automatically extractable versions of the APs proposed by [4]. • $sgF1 - F_{P0}$, where F_{P0} is the frequency of the extra peak below $F1$. • $teF1$, correlation between the teager energy profile [5] of speech passed through a narrowband filter (bandwidth = 100 Hz) and a wideband filter (bandwidth = 1000 Hz) centered around $F1$.
Extra peaks across the spectrum	• $nPeaks40dB$ counts the number of peaks within 40dB of the maximum dB amplitude in a frame of the spectrum.
Reduction in $F1$ amplitude	• $a1 - h1max800$ is the difference between $A1$ and the amplitude of the first harmonic ($H1$). The value of $A1$ was estimated by using the maximum value in 0-800 Hz. • $a1 - h1fmt$ is the same as the previous AP except that $A1$ is now estimated by using the amplitude of the peak closest to $F1$ obtained by using the ESPS formant tracker. The APs $a1 - h1max800$ and $a1 - h1fmt$ are automatically extractable versions of the $A1 - H1$ parameter proposed by [6].
Increase in $F1$ bandwidth	• $F1BW$ is the bandwidth of $F1$.
Spectral flattening at low frequencies	• $std0 - 1K$ is the standard deviation around the center of mass in 0-1000 Hz. This AP not only captures the spectral flatness in 0-1KHz, but also captures the effects of the increase in $F1$ bandwidth and the reduction in $F1$ amplitude. This AP was proposed by [7].

databases consist of telephone bandwidth spontaneous speech conversations recorded at a sampling rate of 8 KHz. Nasalization was marked in these databases with a diacritic. A vowel was marked as nasalized if the duration of nasalization during the vowel region was appreciable, irrespective of the presence of a nasal consonant adjacent to it. The combined WS96 and WS97 database was divided into training and testing databases by alternately assigning the files to train and test directories. Thus, there were a total of 2553 conversations in the training set, and 2547 conversations in the test set.

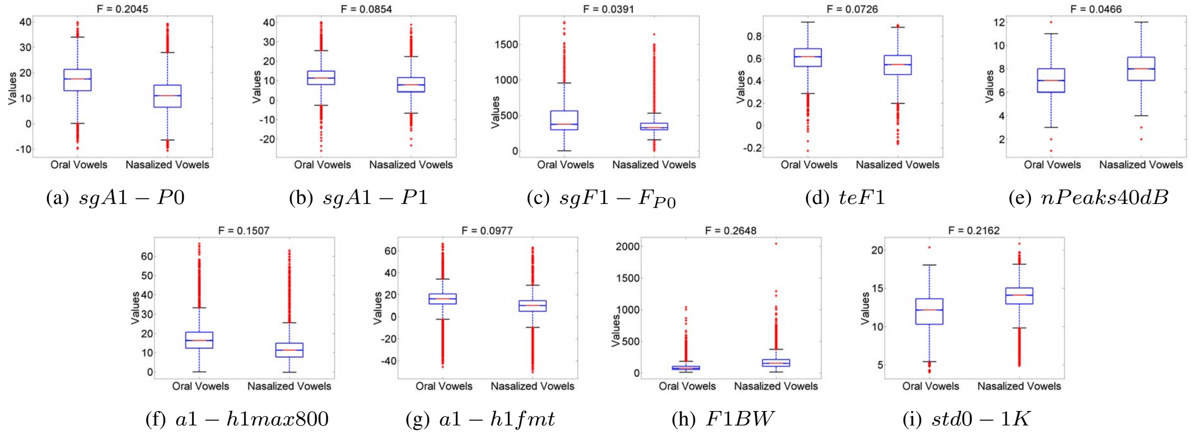
3. Acoustic Parameters

The most important acoustic correlates of nasalization that have been cited in past literature include the introduction of pole-zero pairs in the first formant ($F1$) region and across the spectrum due to the asymmetry of the nasal passages and coupling to the nasal cavity and the paranasal sinuses, reduction in $F1$ amplitude because of proximity to zeros and because of an increase in the bandwidths of formants due to losses in the nasal cavity, and spectral flattening in the low frequency region because of the introduction of several pole-zero pairs in the $F1$ region (c.f. [10, 11, 12]). An earlier study by the authors presented a detailed description of the acoustic characteristics of nasalization and an analysis of the reasons behind the introduction of those characteristics [13]. The APs proposed in this study are based on the knowledge gained through the analysis presented in [13]. The proposed APs try to capture all of the characteristics of nasalization mentioned above.

Table 1 summarizes the acoustic correlates and the APs proposed to capture each of those correlates. All of these APs are extracted automatically from speech. This set of 9 APs will, henceforth, be referred to as the *tf9* set. A more detailed description of these APs and the procedure to extract them is available in [14]. Box and whisker plots of the proposed APs, shown in Figure 1, highlight the discrimination capability of each of these APs. These plots are based on oral and nasalized vowel segments extracted from the TIMIT training database. The normalized F-ratios for each AP, obtained through ANOVA, are also shown on top of each figure. The normalization of the F-ratios was done by dividing F by the total degrees of freedom (= number of samples + 1). The normalized F-ratios give a measure of the relative discriminating capability of the APs. A comparison of the box plots and the F-ratios for the proposed APs reveals that $std0 - 1K$ is the most discriminative AP, and $sgF1 - F_{p0}$ is the least discriminative AP for this task.

4. Method

In this experiment, the transcription provided along with the databases was used to isolate the vowel regions and the APs were calculated only in these vowel regions. The APs were normalized to have zero mean and unit variance, and were used to train a common Support Vector Machine (SVM) classifier



(a) $sgA1 - P0$ (b) $sgA1 - P1$ (c) $sgF1 - F_{P0}$ (d) $teF1$ (e) $nPeaks40dB$

(f) $a1 - h1max800$ (g) $a1 - h1fmt$ (h) $F1BW$ (i) $std0 - 1K$

Figure 1: Box and whisker plots for the proposed APs (based on TIMIT training database).

Table 2: Classification Results for oral vs nasalized vowels using the $gs6$ set.

	Tr: TIMIT, Te: TIMIT			Tr: WS96/97, Te: WS96/97		
	Linear (%)	RBF (%)	Tokens	Linear (%)	RBF (%)	Tokens
Oral Vowels	65.38	68.71	14136	56.45	57.12	12373
Nasalized Vowels	77.84	76.24	4062	62.82	64.16	1119
Chance Norm. Acc.	71.61	72.48	18198	59.63	60.64	13492

for all vowels to distinguish between oral and nasalized vowels. The experiments were carried out using the SVM*light* toolkit [15] with both Linear and Radial Basis Function (RBF) kernels.

The training data was collected by considering every oral and nasalized vowel in succession (ground truth for each database decided by the procedure described in Section 2), and selecting only the middle 1/3rd of the frames for oral vowels and the last 1/3rd of the frames for nasalized vowels. This 1/3rd selection rule minimizes the possibility of the inclusion of ambiguous oral or nasalized vowel frames in the training data. Once the pool of data had been collected, an equal number of oral and nasalized vowel frames were randomly selected from this set to ensure that frames from all different vowels were included in the training set. It must be noted, that frames extracted from syllabic nasals were not included in the training set, but they were tested in the performance evaluation.

Once the SVM outputs were obtained for the training samples, the outputs were mapped to pseudo-posteriors with a histogram. If $N(g, d = +1)$ is the number of training examples belonging to the positive class for which the SVM discriminant had a value of g, the histogram posterior estimate is given by:

$$P(d = +1/g) = \frac{N(g, d = +1)}{N(g, d = +1) + N(g, d = -1)} \qquad (1)$$

Histogram counts were always obtained by using the same number of samples for the positive and negative classes, so that the pseudo-posterior $P(d/g)$ is proportional to the true likelihood $P(g/d)$. Given that the pseudo-posteriors are proportional to the true likelihoods, and assuming frame independence, the probability for a segment to belong to the positive class can be obtained by multiplying the pseudo-posteriors for each frame in the segment. Thus, a vowel segment was declared as nasalized if

$$\prod_{i=frame_1}^{i=frame_n} P_{nasal}(i) > \prod_{i=frame_1}^{i=frame_n} P_{oral}(i) \qquad (2)$$

where, $P_{nasal}(i)$ = Probability that the i^{th} frame is nasalized. and, $P_{oral}(i)$ = Probability that the i^{th} frame is non-nasalized.

5. Results

Tables 2, 3 and 4 compare the results obtained by using the *gs6* set (author's implementation of the 6 APs proposed in [7]), the *mf39* (set of 39 standard Mel-Frequency Cepstral Coefficients, MFCCs) and the *tf9* set, respectively, in the current experimental framework. The results for the *gs6* set and the *mf39* set form the baseline results. These tables show the results for StoryDB, TIMIT and WS96/97 with Linear and RBF SVM classifiers. Table 2 does not show the results for StoryDB because the *gs6* set was a set of segment based APs (that is, one set of 6 APs for the whole segment), and not enough vowel segments were available in StoryDB for accurate training of the SVM classifiers. The *mf39* and *tf9* sets were frame based APs (that is, one set of 39/9 APs for each frame). The bottom row of the tables gives the chance normalized accuracy which is obtained by averaging the accuracies of the classifier for the two classes of oral and nasalized vowels.

A comparison of the results suggests that:

1. The performance of the *tf9* set is the best in all but one cases. The *mf39* set outperforms the *tf9* set for TIMIT with RBF SVMs.
2. The performance of the *gs6* set is the worst in all cases where it was tested except when Linear SVM classifiers were used for TIMIT where it outperformed the *mf39* set.
3. The performance of all the sets of APs falls as the complexity of the database increases from StoryDB to WS96/97.
4. The performance of the *mf39* set improves significantly with RBF SVM classifiers. However, even with the RBF SVM classifiers, the performance of this set is not very good for the spontaneous speech database WS96/97.
5. The performance of the *tf9* set is very balanced across the oral and nasalized vowel classes, especially so for linear classifiers. On the other hand, the performance of the *gs6* and *mf39* sets differs widely across the oral and nasalized vowel classes. For example, for the *gs6* set there is

Table 3: Classification Results for oral vs nasalized vowels using the $mf39$ set.

	Tr: StoryDB, Te: StoryDB			Tr: TIMIT, Te: TIMIT			Tr: WS96/97, Te: WS96/97		
	Linear (%)	RBF (%)	Tokens	Linear (%)	RBF (%)	Tokens	Linear (%)	RBF (%)	Tokens
Oral Vowels	62.50	97.32	112	76.87	90.32	14136	77.26	80.13	12373
Nasalized Vowels	68.75	94.35	336	43.55	69.50	4062	44.68	48.61	1119
Chance Norm. Acc.	65.62	95.83	448	60.21	79.91	18198	60.97	64.37	13492

Table 4: Classification Results for oral vs nasalized vowels using the $tf9$ set.

	Tr: StoryDB, Te: StoryDB			Tr: TIMIT, Te: TIMIT			Tr: WS96/97, Te: WS96/97		
	Linear (%)	RBF (%)	Tokens	Linear (%)	RBF (%)	Tokens	Linear (%)	RBF (%)	Tokens
Oral Vowels	83.93	95.54	112	75.75	81.18	14136	68.27	73.56	12373
Nasalized Vowels	96.13	97.02	336	71.42	74.62	4062	66.85	65.59	1119
Chance Norm. Acc.	90.03	96.28	448	73.58	77.90	18198	67.56	69.58	13492

a difference of about 12% in the accuracies for oral and nasalized vowels for the TIMIT database with a linear SVM classifier. The differences are much more significant for the *mf39* set. In fact, the accuracy of the *mf39* set for nasalized vowels is even below the chance accuracy of 50% for three cases (for TIMIT and WS96/97 with linear classifiers, and for WS96/97 with RBF classifier).

6. Summary and Future Work

In this paper, nine knowledge-based APs were proposed for the task of classifying vowel segments into oral and nasal categories automatically. These APs were tested in an SVM classifier framework on three different databases with different sampling rates, recording conditions and a large number of male and female speakers. Accuracies of 96.28%, 77.90% and 69.58% were obtained by using these APs on StoryDB, TIMIT and WS96/97 respectively with an RBF kernel SVM. These results were compared with baseline results obtained by using two other sets of APs for this task in the current experimental framework. Comparison with the baseline results showed that the proposed APs not only formed the most compact set (9 APs as opposed to 39 MFCCs), but also gave the best performance on this task. In future, we plan to incorporate this nasalization detector into the landmark-based speech recognition system proposed in [1]. This would enable the landmark-based system to classify vowels into oral and nasalized classes, hence, moving it one step closer to a complete system.

7. Acknowledgments

Work supported by NSF Grant BCS0236707.

8. References

[1] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. dissertation, University of Maryland, College Park, MD, USA, December 2004.

[2] M. Hasegawa-Johnson et al, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *Proceedings of ICASSP*, 2005, pp. 213–216.

[3] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *J. Acoust. Soc. Am.*, vol. 82, no. S1, p. S55, 1987.

[4] M. Y. Chen, "Acoustic correlates of English and French nasalized vowels," *J. Acoust. Soc. Am.*, vol. 102, no. 4, pp. 2360–2370, 1997.

[5] D. A. Cairns, J. H. L. Hansen, and J. E. Riski, "A non-invasive technique for detecting hypernasal speech using a nonlinear operator," *IEEE Transactions on Biomedical Engineering*, vol. 43, no. 1, pp. 35–45, 1996.

[6] M. K. Huffman, "The role of F1 amplitude in producing nasal percepts," *J. Acoust. Soc. Am.*, vol. 88, no. S1, p. S54, 1990.

[7] J. R. Glass and V. W. Zue, "Detection of nasalized vowels in American English," in *Proceedings of ICASSP*, 1985, pp. 1569–1572.

[8] TIMIT, "TIMIT acoustic-phonetic continuous speech corpus, national institute of standards and technology speech disc 1-1.1, NTIS Order No. PB91-5050651996, october 1990," 1990.

[9] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proceedings of ICASSP*, 1992, pp. 517–520.

[10] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Netherlands: Mouton, 1960.

[11] S. Maeda, "Acoustic cues for vowel nasalization: A simulation study," *J. Acoust. Soc. Am.*, vol. 72, no. S1, p. S102, 1982c.

[12] S. Hawkins and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal-nasal distinction for vowels," *J. Acoust. Soc. Am.*, vol. 77, no. 4, pp. 1560–1575, 1985.

[13] T. Pruthi, C. Espy-Wilson, and B. H. Story, "Simulation and analysis of nasalized vowels based on MRI data," *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. 3858–3873, 2007.

[14] T. Pruthi, "Analysis, vocal-tract modeling and automatic detection of vowel nasalization," Ph.D. dissertation, University of Maryland, College Park, MD, USA, January 2007.

[15] T. Joachims, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999, ch. Making large-Scale SVM Learning Practical.