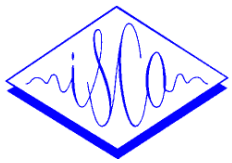

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Nasal Detection Module for a Knowledge-based Speech Recognition System

Marilyn Y. Chen

Research Laboratory of Electronics, MIT, Cambridge, MA 02139

ABSTRACT

The Lexical Access From Features (LAFF) project tries to model the representation and perception of speech by human listeners. The derivation of such a representation involves first finding certain acoustic landmarks. Based on the landmarks and the acoustic cues surrounding the landmarks, distinctive features of the speech segments may be deciphered. The present study concentrates on the nasality module that attempts to detect the presence of an underlying nasal consonant, which is almost always adjacent to a vowel. For an underlying nasal in English, the features [+voiced, +sonorant, +consonant, +nasal, -continuant] are specified. The features are then mapped into measurable acoustic properties. Normally, cues from three regions in the sound indicate the presence of a nasal consonant: (1) abrupt spectral change from the vowel to the nasal murmur, (2) vowel nasalization, and (3) nasal murmur. These cues are quantified by acoustic parameters whose values are combined to indicate the presence of a nasal. The nasality module that has been developed is a sonorant landmark detector that greatly reduces false landmark detection and distinguishes nasals from laterals by incorporating additional nasal manner cues. The module also addresses cases where one or more of the three nasal cues is absent.

1. INTRODUCTION

Phonologists have proposed that words are stored in memory as sequences of segments, each of which is represented by a set of binary distinctive features that are arranged in a hierarchical fashion. The derivation of such a segment/feature-based representation from acoustic analysis for words in running speech involves making acoustic measures in the speech signal that has been transformed into a time-frequency-amplitude domain and extracting acoustic events. These events are further categorized into landmarks, which shed light on the vocal tract and glottal characteristics. Based on the landmarks and the surrounding cues, distinctive features may be deciphered. Once the distinctive features and segments are available, the lexicon can be accessed and matched to obtain a cohort of possible words, taking into account modification of acoustic cues due to context or enhancement.

Possible modules performing specific tasks in extracting landmarks and acoustic features are: vowel landmark, glide landmark, consonant landmark, place feature, voicing feature, nasal feature, and tenseness feature [2]. Although the vowel, glide, and consonant landmarks, and certain place and voicing features have been studied, relatively little work has been done on the nasal feature. This study concentrates on the nasality module with its corresponding articulatory landmark and surrounding acoustic cues. The first step in developing the nasality module is to evaluate the performance of an existing consonant landmark detector [4], which examines the rate of change in energy in different frequency bands, in extracting vowel-nasal and nasal-vowel (VN) boundaries. Consonant landmarks occur when there is an abrupt change in the acoustics due to closure or release of a constriction. An attempt has been made to improve the sonorant landmark detector in detecting nasals by making finer evaluation of the formant amplitude changes A1 to A4 at the VN boundary. To eliminate sonorant landmarks due to laterals and false nasal landmarks, vowel nasalization and murmur detection also need to be incorporated. In a previous study [3], nasalized vowels were detected by combining vowel spectral measures: center of mass below 1 kHz, standard deviation of energy 500 Hz within the center of mass, maximum and minimum percentage of the time that an extra resonance is in the first formant region, maximum dip between the first formant and the extra resonance, and minimum amplitude difference between those two resonances. With log likelihood scores, the average vowel nasalization detection score was 78%. In addition, nasal consonants were detected using murmur spectral measures: total energy, energy stability, percentage of time a resonance is below 350 Hz, energy below 500 Hz relative to total energy, and energy below 350 Hz relative to energy between 350 and 1000 Hz. Combining with log likelihood scores results in an average murmur-detection score of 89%. In another study [7], an algorithm for identification of nasal murmur was based on duration and speaker-dependent thresholds of F1, A1 and A2 ratio, and A1 and A3 ratio from the murmur spectra. In the present study, the vowel-nasalization parameters, adjusted amplitude differences between the first formant amplitude and the nasal peak amplitude in the vowel, are based on a theoretical model of speech production [1]. The nasal consonant parameters are based on the first formant frequency and amplitude differences in the murmur.

These parameter values, along with the landmark detector result, were combined by discriminant analyses and decision tree to classify the tokens into nasal and non-nasal categories. The algorithm was developed with the first 20 sentences of the LAFF database as the training set and was tested on all of the LAFF database and another nasal-targeted database. Although the general algorithm assumes the presence of a vowel and a murmur, contexts with possible missing acoustic cues for the murmur and the vowel were also determined so that the nasality module may be restricted to the relevant component.

2. METHOD

2.1. Corpus

The two types of databases in this study are: (1) the LAFF database and (2) the nasal-targeted databases. The former consists of 100 grammatically correct sentences consisting various phonemes in several contexts, including 214 nasal consonants with 307 possible VN boundaries. For the first 20

sentences, there are 50 nasal consonants with 73 possible VN boundaries. Four speakers, two males and two females, read the sentences once, naturally but carefully.

One of the nasal-targeted databases was designed to examine the conditions under which missing nasal murmur occurs. A shortened or missing murmur is likely to occur for nasals adjacent to longer vowels or nasals preceding a voiceless obstruent [5], [6], [8], possibly due to the timing of the velum lowering, oral closure, and glottal vibration. The utterances consist of vowels (V), /ʌ, æ, e,ɑi/, adjacent to nasals (N), /m, n, ŋ/, with possible obstruent consonant (C) in nasal-vowel contexts, (e.g. "*sma*ck", "ches*tnu*t", "*mi*le", and "f*ema*le") and vowel-nasal contexts, (e.g. "c*amp*", "b*ump*y", "b*ang*", and "tr*ainer*"). These utterances were embedded in a carrier phrase, "Say _____ again". Two male and two female speakers (different from the speakers for the LAFF database) recorded the database five times.

In a separate database, sentences were developed to test systematically the environment in which syllabic nasals occur. Syllabic nasals may be caused by sustaining the oral closure or narrowing constriction from an obstruent to the nasal, despite an underlying weak vowel in-between the two consonants. The test sentences contain words with /C_1əN/ followed by a word beginning with /C_2/, C_1əN#C_2. The nasal consonant N, /m, n/, may have the same or different place of articulation as the consonant C_1, which can be a voiced or voiceless stop, fricative, affricate, or glide. The word-initial consonant C_2 can be a stop or a fricative. One male speaker recorded this database five times. The recordings were made in a sound-attenuated room with a hanging Electrovoice omnidirectional microphone, low-pass filtered at 7.5 kHz and digitized at 16 kHz.

2.2. Acoustic Analysis

The consonant landmark detector developed by Liu [4] divides the spectrogram into six frequency bands. The first derivative of the energy waveforms are generated for each band from coarse and fine processings. The landmark type-specific processing stage consists of g(lottis) landmark, s(onorant) landmark, and b(urst) landmark detectors. The s-landmark occurs in a voiced region bounded by a +g landmark (at consonant release) and a -g landmark (at consonant closure) and is caused by the release (+s) or closure (-s) of a tight constriction of a nasal or a lateral, introducing abrupt changes in bands spanning 0.8 to 5.0 kHz. The detector first determines the pivots, which are the possible candidates of the landmarks, based on the absolute peaks in the first derivative of the energy waveform in bands 2-5 that reaches a threshold. A pivot needs to pass the steady state test in the 0-600 Hz range and high-frequency abruptness test in the 1.3-8 kHz range to be labeled as a s-landmark.

In the present project, another approach to detecting the nasal landmark was developed. It is based on formant amplitudes obtained from the DFT generated with a 25.6 ms Hamming window every 10 ms from the middle of the vowel into the nasal murmur. The locations of the first four formants are determined by the peaks with the maximum harmonic amplitude in the vicinity of the formants. The formant frequency is assumed to be steady, changing from one harmonic to an adjacent one over 10 ms, even into the murmur. The change in formant amplitude between consecutive spectra is calculated for the first four formant peaks and are summed; and the most negative sum (Sum.diff) is taken to indicate the VN boundary.

The murmur parameters were measured from the spectrum farthest away from the s-landmark, up to 30 ms into the murmur. The 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 –5kHz). Each band is expected to include at least one nasal resonance. The sum of the amplitude differences between $A1_n$ and the other amplitudes (Sum.amp.diff), $A1_n$-$A2_n$, $A1_n$-$A3_n$, $A2_n$-$A3_n$, together with the frequency of the lowest peak $F1_n$ were used as the murmur parameters.

The parameters of vowel nasalization take into account the nasal peaks in the vowel, P0 and P1. The frequency of P0 is theoretically around 250 Hz; the frequency of P1 is around 1 kHz [1]. Adjustment of P0 and P1 are made based on their frequency relative to the first and second formants. The adjusted nasal peak amplitudes are subtracted from the first-formant amplitude A1. To determine if a vowel is nasalized, A1-P0a and A1-P1a, the minimum adjusted A1-P0 and A1-P1, are obtained. Not all vowels yield a useable A1-P0a and A1-P1a, depending on the closeness of F1 and F2 to the nasal peaks.

Two methods for the combination of nasal cues are discriminant analysis and decision tree. The former assigns weightings to the cues that maximally separate the groups**.** Sum.diff, A1-P0a and A1-P1a, and the murmur parameters were normalized by subtracting from the mean of the nasal group. In addition, the normalized A1-P0a and A1-P1a were divided by the corresponding range and the lower absolute value of the two was taken if both are available. The weightings for the parameters were determined from the nasals and non-nasals with s-landmarks in the training set.

Unlike discriminant analysis that uses the cues simultaneously to classify the tokens, a decision tree allows decisions to be made in logical steps. A decision tree algorithm based on membership functions, with the value 1 indicating the presence of nasal and 0 the absence of nasal, was determined from the nasals in the training set. For Sum.diff, the value is 1 for -52.4 to -0.9 dB. The function of the vowel nasalization parameters is 1 for A1-P0a $\leq$0.07 dB and for A1-P1a $\leq$10.12 dB. The function of the murmur parameters is 1 for 126$\leq$ $F1_n$ $\leq$347 Hz; for 11.8$\leq$ $A1_n$-$A2_n$ $\leq$43.3 dB; for 22.8$\leq$ $A1_n$-$A3_n$ $\leq$50.3 dB; for -9.7$\leq$ $A2_n$-$A3_n$ $\leq$33.6 dB; and for 119.5$\leq$ Sum.amp.diff $\leq$200.7 dB. The logic function with AND (Λ) and OR (V) summarizing the decision tree is:

Sum.diff Λ (**A1-P0a** $\vee$ **A1-P1a**) Λ (**$F1_n$** Λ (**$A1_n$-$A2_n$**) Λ (**$A1_n$-$A3_n$**) Λ (**$A2_n$-$A3_n$**) Λ **Sum.amp.diff**).

This function suggests that when an underlying nasal surfaces as a nasalized vowel and a murmur, it is detected as a nasal if Sum.diff, either A1-P0a or A1-P1a, and all of the murmur parameters fall within their corresponding nasal regions.

3. RESULTS

3.1. LAFF Database

For the first 20 sentences in the LAFF database, Liu's landmark detector introduced 49 s-landmark deletions (16.8%) at VN boundaries due to (a) lack of change in energy, (b) glottalization in the vowel, (c) vowel deletion, and (d) silence between the vowel and the nasal consonant. For (b) and (d), all except one case had a g-landmark present in the vicinity of the VN boundary. For (c), a vowel landmark is expected in the nasal. For 48.3% of (a), there is a detectable s-landmark on the other side of the nasal murmur; for the rest, a g-landmark would indicate the end of voicing for the murmur. There are also false s-landmark insertions in the vowel (32), vowel-glide boundary (50), and vowel-obstruent boundary (66). When the landmark detector was made less stringent by using the pivots only, 11.81% of the nasals in the other 80 sentences yielded missing pivots.

An examination of Sum.diff revealed more in-depth information on the missing s-landmarks for nasals and the non-nasal s-landmarks. Figure 1 shows the mean and standard error (SE) for Sum.diff from the training set with the number of tokens in parenthesis. The nasals with s-landmark deletion may still be distinguished from the vowels and the obstruents that introduced false s-landmarks. The means for glides and laterals with s-landmark, however, fall within the nasal range. Based on the means of the Sum.diff, a classification scheme was used for the detection of VN boundary. In order to determine if a given s-landmark is due to a nasal, one sees where its Sum.diff falls. If it is closer to the nasal range, it is likely to be a nasal; if it is closer to either the obstruent or the vowel mean, it is likely to be a non-nasal. As a result, 77.6% of non-nasals and laterals with s-landmark are classified as non-nasal; 20.7% of the nasals with too small of an energy drop for the s-landmark detector are classified as having a nasal. However, only 41.7% of the VN boundary are classified as having a nasal. The result indicates that Sum.diff is good in eliminating false and lateral s-landmarks as the nasal landmark. However, many of the true nasal landmarks are missed. A solution is to make the Sum.diff criterion more lenient while incorporating measures of vowel-nasalization and murmur criterion.

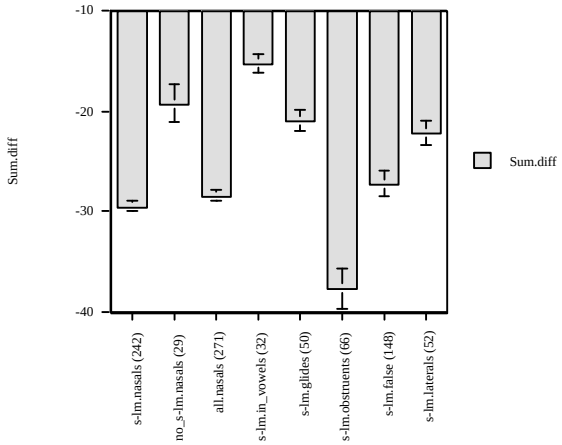

Figure 1: Sum.diff means for nasals with and without s-landmark and non-nasals with s-landmark.

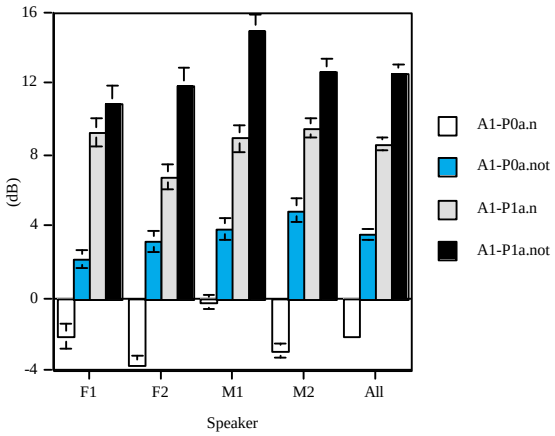

Figure 2: A1-P0a and A1-P1a of nasals and non-nasals

Figure 2 shows the mean and standard error (SE) by speaker for A1-P0a and A1-P1a for the training set. The nasals (n) include cases with and without s-landmark; and the non-nasals (not) are the cases with s-landmark detected. The separation is good for all of the speakers. Consequently, the means of A1-P0a and A1-P1a for the nasal and non-nasals were used for classification. When either A1-P0a or A1-P1a indicates nasalization, the vowel is assumed to be nasalized. With this classification technique, 86.3% of vowels adjacent to a nasal were detected as nasalized. If only one of the possible two adjacent vowels needs to indicate nasalization, 93.7% of the nasals were detected. On the other hand, 35.3% of the vowels in the non-nasal category were detected as nasal. If the vowels next to a nasal on the other side were not considered, the rate of false detection drops to 25.4%. The murmur thresholds from the nasals were applied to the non-nasal cases with s-landmark detection in the training set and yielded 45.5% false nasal detections.

The relevant nasal cues were combined by using discriminant analysis and decision tree. Discriminant analysis performed on the training set with five descriptors (Sum.diff, vowel-nasalization parameter, $F1_n$, $A1_n$-$A2_n$, and $A1_n$-$A3_n$) yielded 88% and 74% percent correct in classifying the tokens into nasal and non-nasal groups, respectively. With the above discriminant analysis applied to nasals with missing pivot in the 100 LAFF sentences, 60.1% were assigned to the nasal category, excluding the cases with missing murmur, syllabic nasal, or glottalization. The decision tree algorithm performed on the training set of cases with non-nasal sonorant landmark eliminated 92.9% of the false nasal detections. For the nasals with missing pivot in the 100 LAFF sentences, the algorithm made 91.8% nasal detections, not counting the cases with missing murmur or syllabic nasal.

For the nasal consonants in the 100 LAFF sentences, 2.2% showed missing murmur according to the spectrogram and the murmur parameters. Most of them are in the form Vnt# where V is non-high vowel (o, æ, ɑ, ɑi); the others are in the forms p#mɑi and #Nɑ. Also, 3.0% of the nasal consonants surfaced as syllabic nasals. They are found in eight words in one or more contexts; seven of them contained alveolar nasal and one with a labial nasal. The consonant preceding the schwa is a stop, fricative, or affricate. For four words, the preceding

consonant has the same place of articulation as the nasal consonant, e.g. "sudden". For two cases, the phoneme after the nasal is homorganic with the preceding consonant, e.g. "can#keep". All of the syllabic nasals have detectable murmur according to the murmur parameters.

3.2. Murmur-missing Database

With the first 20 sentences of the LAFF database as the training set, the murmur-missing database was used as the test set. For the test set, 21.1% have missing pivots, 35.9% of which also failed the Sum.diff criteria according to the means clustering. The descriptors of the discriminant analysis of the training set were applied to those cases, which are assumed to be more error-prone. The total percent correct in classifying the tokens into the nasal group is 98.4%. The decision tree algorithm derived from the training set yielded 90.2% nasal detections if the cases with missing murmur relied only on vowel nasalization.

The presence of murmur was examined by using the spectrogram, waveform, and auditory signal as well as the murmur parameters. Missing murmur was found to occur in the contexts /ænk/, /æmp/, /ænt/, /snV/ where V is /æʌ, e/, /t-mɑi/, /k-mV/ where V is /æ, e/, /dnæ/, and /tnʌ/. Data from the first 20 LAFF sentences and the murmur-missing database suggests that the murmur for /n/ may be missing if it is followed by /t, k/; if it is preceded by /t, d, s/; or if it is sentence initial. The murmur for /m/ may be missing if it is followed by /p/ or preceded by /p, t, k/. Murmur is more likely to be missing if the adjacent vowels are not high.

3.3. Syllabic Nasal Database

In systematically manipulating the context C_1əN#C_2, it was found that syllabic labial nasal tends to occur when the preceding consonant is a /f/, / /, /v/, s/, especially if C 2 is not a /p/, is a /k/, is a /f/, and is a /s/, respectively. None of the tokens with C_1 as /d, t, g, h/ introduced a syllabic labial nasal, even for C_2 labial consonants. For the alveolar nasal, a syllabic nasal is always introduced when C_1 is /d, t, z, /. It rarely occurs when the preceding consonant is /g, b, p/ or dental fricatives. When C_1 is not an alveolar consonant, C_2 having the same place of articulation as C_1 did not seem to have a prominent influence on inducing syllabic alveolar nasals.

4. CONCLUSION

The nasality module is one of the essential components of a knowledge-based speech recognition system. The generic algorithm involving formant amplitude drop between the vowel and the nasal, vowel nasalization, and murmur may suffice in most cases of nasal detection. As expected, combination of the nasal cues does better than using the individual cues by themselves. A lenient criteria for Sum.diff allows the detection of VN boundary although it introduces false s-landmarks. However, the false detections are minimized if the other two nasality cues are incorporated. Furthermore, a g-landmark or b-landmark is often in the vicinity of the nasal murmur or the VN boundary, respectively, even if a nasal landmark is not detected. In those cases where a nasal is suspected from a cohort, murmur detection may still be used. The algorithm developed from a training database was applied successfully to test databases, even with different speakers and utterances. More than 90% of the nasals with missing s-landmarks were detected as nasals using discriminant analysis or decision tree algorithm. Although in general, nasals in English occur adjacent to a vowel, there are cases where missing murmur and syllabic nasal may surface. For the former, the nasality module may be reduced to the vowel nasalization component based on the possible contexts for missing murmur found in this study. Likely contexts for syllabic nasals were also determined which can signal the reduction of the module to the murmur component.

The result of this study indicates that for any module in a knowledge-based speech recognition system, a general algorithm needs to be first implemented. As exceptions are observed, they need to be implemented as feedback to limit the algorithm to the relevant component(s). In this way, all possible manifestations of the speech sound may be captured.

5. ACKNOWLEDGEMENTS

Thanks to Melanie Matthies for her help on the discriminent analysis and Ken Stevens for his help editing this paper. This research was supported in part by NIH grant DC02978.

6. REFERENCES

1. Chen, M. Y. "Acoustic Correlates of English and French Nasalized Vowels," *J. of Acous. Soc. of Am.* 102: 2360-2370, 1997.

2. Choi, J-Y. *Detection of Consonant Voicing: A Module for a Hierarchical Speech Recognition System*. PhD thesis, MIT, Cambridge, MA, 1999.

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

4. Liu, S. A. "Landmark Detection for Distinctive Feature-based Speech Recognition," *J. of Acous. Soc. of Am.* 100: 3417-3430, 1996.

5. Lovins, J. B. "A Study of 'Nasal Reduction' in English Syllable Codas", Papers from the Fourteenth Regional Meeting, Chicago Linguistic Society, 1978.

6. Malécot, A. "Vowel Nasality as a Distinctive Feature in American English," *Language* 36: 222-229, 1960.

7. Weinstein, C. J., McCandless, S. S., Mondshein, L. F., and Zue, V. W. "A System for Acoustic-Phonetic Analysis of Continuous Speech," *IEEE Trans. ASSP* 23: 54-67, 1975.

8. Zue, V. W. and Laferriere, M. "Acoustic Study of Medial /t, d/ in American English," *J. of Acous. Soc. of Am.* 66: 1039-1050, 1979.