Automatic Detection of Landmark for Nasal Consonants from Speech Waveform[•]

Limin Du

Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China,

Tel: (8610)6262-9250 Fax: (8610)6262-9250 email: dulm@public.east.cn.net

Kenneth Noble Stevens

Dept Electrical Engineering and Computer Science, Massachusetts Institute of Technology, MA02139, USA,

Tel: (617)253-3209 email: Stevens@speech.mit.edu

ABSTRACT

A knowledge-based approach towards automatically detecting nasal landmarks (/m/, /n/, and /ng/) from speech waveform is developed. The acoustic characteristics Fn1 locus calculated on each frame of speech waveform as the mass center of spectrum amplitude in the vicinity of the lowest spectral prominence between 150-1000Hz, and A23 locus calculated on the same speech frame as a band energy between 1000-3000Hz were incorporated together to construct the nasal landmark detector, which alarms at the instants of closure and release of nasal murmur. Experiment observations on the acoustic characteristics of Fn1 and A23 and the nasal consonant landmark detection results on the VCV database are also presented.

1 INTRODUCTION

All languages consist of a finite number of basic linguistic units called phonemes which are distinguished and mutually exclusive. However, when phonemes are concatenated together in time to produce speech, the physical slices of each phonemes are overlapped and the acoustic characteristics of a given phoneme are modified by the immediate phonetic environment. The acoustic variations from this contextual differences together with inter-speaker differences and intra-speaker difference can influence the observed acoustic pattern of all phonemes, and complicate speech recognition.

There are two major techniques for computer to recognize the basic linguistic units in the presence of all the variations and ultimately interpret the utterances of human beings. In one major technique, a speech recognition system is built with a statistical approach by training sample utterances. The system tries to "learn" the acoustic characteristics of given phonemes from the training sample utterances that cover all the possible variations, and then uses this information to help "understand" what the speaker says in the actual recognition process. Another major technique is the knowledge-based approach. Here, knowledge about the speech perception process and the acoustics and physiological aspects of speech production is explicitly built into the recognition system by system designer. From this information, the system tries to model the production of each sound. During the process of speech recognition, the system tries to match every produced sound to one particular model from all of the possible models involved in speech production for the particular language of interest.

It is believed that a more robust speech recognition system should be built by using statistical approach to deal with variations that cannot be described, and by incorporating knowledge sources explicitly and in a meaningful manner to improve the system and to reduce the amount of training utterances needed by pure statistical approach. The knowledge necessary to design such a system needs to be comprehensive and the desired acoustic parameters needs to be extractable.

In this paper, a knowledge-based approach towards automatically detecting nasal landmarks (/m/, /n/, and /ng/) from speech waveform is developed. Previous works in the area of

[•] This work was supported by the National Natural Science Foundation of China and the Foundation of the Institute of Acoustics, Chinese Academy of Sciences.

landmark detection was performed by Sharlen Liu [1995] and Water Sun [1996]. Liu's work focused on finding and locating landmarks for abrupt consonants, and Sun's work focused on finding and recognize landmark for glides (/w/ and /j/).

2. ACOUSTIC CHARACTERISTICS FOR NASALS LANMARK DECTTION

The acoustic characteristics of nasal consonants and nasalized vowels has been studied by experimental observations and theoretic analysis in the past by several authors, including House [1957], Fujimura [1962,1971], Glass [1984], and Stevens [Forthcoming]. The most significant acoustic characteristics of nasal consonants including: (1) The spectrum at frequencies below the lowest spectral prominence shows very little chang in amplitude as the supraglottal constriction is closed and the released. (2) An extra peak appears above the first formar frequency immediately before the closure, and a similar pea appears briefly immediately following the release. When th first formant frequency of the vowel is high, the extra pea might appear below the first formant. (3) During the murmu the low-frequency spectrum is dominated by the peak at abou 250 Hz, and there are no other spectral peaks in the vicinity c this prominence. A second resonance appears in the spectrum i the vicinity of 1000 Hz, and this peak is continuous with th extra prominence that occurs in the vowel spectrum precedin and following the nasal murmur. (4) There is a rather rapi change in the frequency of the low-frequency prominence at th closure and at the release. (5) There is an 15-20 dB change i spectrum amplitude in the vicinity of the second or third formar of the vowel preceding and following the nasal murmur.

For automatic detection of nasal consonant landmark, the ke issue is how to construct a parameter model which can captur robustly the significant cues of the closure and release instant of nasal murmur. Experimental observations were conducted t examine the availability of the mentioned acousti characteristics of nasal consonants. It is found that the low-frequency spectrum prominence, Fn1, occurring at below about 1000Hz shows consistently a rather rapid change at both closure and release of a nasal murmur, inspiring the use of the spectrum prominence locus as one of the parameters of the detector. It is also found that at both closure and release of a nasal murmur, the change of the spectrum prominence locus is accompanying with the change of spectrum amplitude in the vicinity of 2000Hz, A23. These two acoustic parameters are more correlated to the nasal landmark event.

3. CALCULATION OF Fn1 AND A23

First, measurements have to be performed on the data to obtain the parameters needed. A broadband spectrogram is calculated with 6 ms Hanning widows every 2 ms. Each 6 ms frame is zero-padded out to 512 point before DFT is taken. To reduce the effect of noise, the spectrogram is the smoothed with an 10 ms moving averager. This duration was chosen because the finite movement rate of the tongue, the uvula, and the vocal cavity is slow enough that abrupt changes within 10 ms can be deemed to be noise.

There are several methods to calculate the low-frequency spectrum prominence Fn1. The choice depends on the detection requirements. LPC root solving method may obtain the prominence accurately in most voiced intervals, but it suffers from noise and speech segments with fast moving poles. Instead of LPC root solving, two mass center methods were tested to calculate the spectrum prominence.

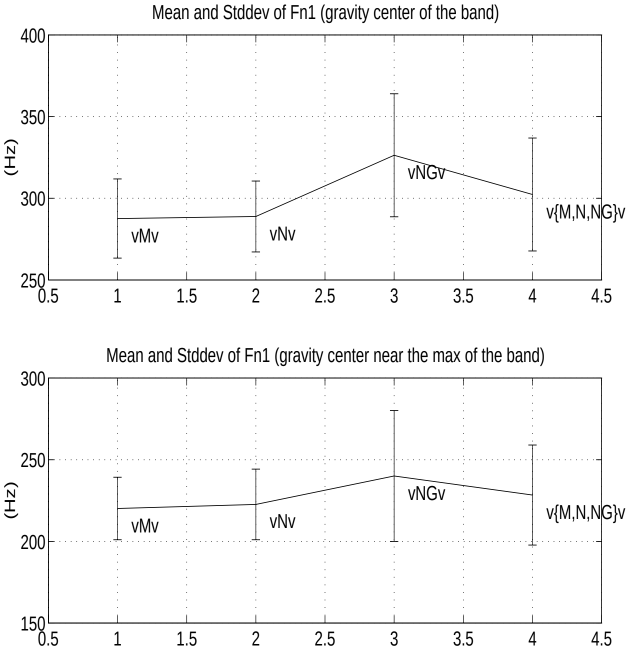

Figure 1: The Mean and the stand deviation of the mass center of nasal VCV triphones. .Figure 1(a) shows the mean and stddev of the mass center (MC) of a given band. Figure 1(b) is different from Figure 1(a) in such a way that the mean and stddev of the mass center were calculated within a more narrowed bank centered at the max of the given band , i.e. max mass center (MMC). The value of Each point of the experiment observation was calculated among all vowels regarding a given nasal consonant. In the set vMv, vNv, vNGv, and v{M,N,NG}v , v?v represents all the triphones of consonant ? in the VCV database

Experiment observation shows that the MMC method is more exactly in locating the spectral prominence of murmur than MC method. However, the continuance of MC method is much better than MMC method.

In this research, parameter Fn1 is chosen and calculated on eac frame as the mass center of spectrum amplitude in the vicinity (the lowest spectral prominence between 150-1000Hz bar which covers the first formant range of most men and womei and the lowest spectral prominence of nasal murmur.

Parameter A23 is calculated on each frame as a band energ between 1000-3000 Hz which is intended to capture the chang of spectrum amplitude in the vicinity of the second or thi formant of the vowel preceding and following the nasal murmu

4. DETECTING NASAL LANDMARK VIA THE RATE OF Fn1 AND A23

The detector of landmark for nasal consonants receives Fn A23 and their backward (minus for closure and plus for releas rates or differences as its input, combing with absolu thresholds of the rates of Fn1 and A23 at the instants of nas closure and release as its references. When the rates of Fn1 ar A23 exceed corresponding thresholds, the locus of Fn1 was use to identify further if the murmur spectral prominence exists. Th combination of Fn1 and A23 into the detector may reduce the rate of false alarms.

The interval and threshold of differences were obtained fror experiment observations on the existing vowel-consonant-vow (VCV) database created within the Speech Communicatio Group at MIT. This database has three speakers, two male (DV and KS) and one female(CB). Each speaker produced a possible combinations of the vowel-consonant-vowel utteranc for the vowels /aa/, /ae/, /ah/, /eh/, /ih/, /iy/, /oh/, /uh/ and /uw and the consonants /b/, /ch/, /d/, /dh/, /dj/, /f/, /g/, /h/, /j/, /k/, /l /m/, /n/, /ng/, /p/, /r/, /s/, /sh/, /t/, /th/, /v/, /w/, /z/, and /zh/. Th speech wavwform was sampled in 16KHz in 16 bit accurac The instants of closure and release of the nasal subset of VC database were hand-labeled through listening to the utterance and visually observing the spectrograms. The regions of closur and release were examined to determine the rate of loci of Fn and A23.

Figure 2(a) and 2(b) shows the mean and stddev of Fn1 rate a the closure and release respectively for all vowels.

Figure 3(a) and 3(b) shows the mean and stddev of A23 rate a the closure and release respectively for all vowels.

From the experiment observations, the threshold of the rates c Fn1 and A23 for alarming the landmarks of nasal closure anu release was obtained, thus finishing the design of the landmark detector for nasal consonants.

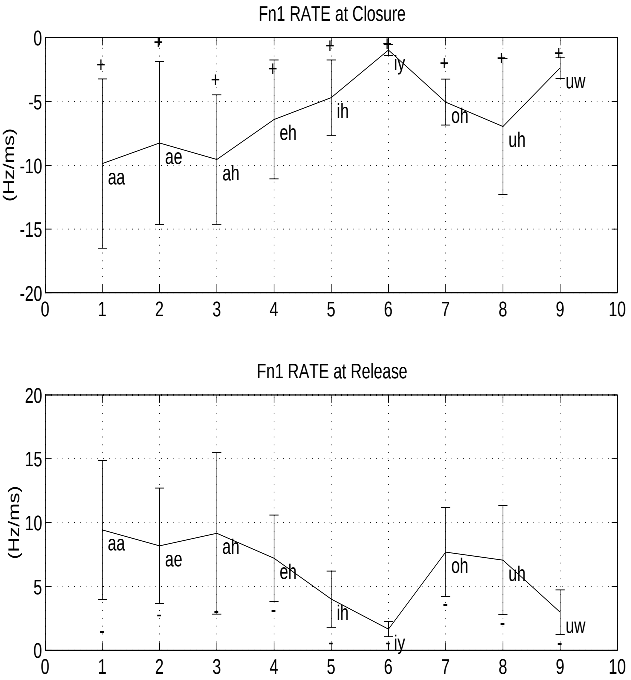

Figure 2: The Mean and the stand deviation of Fn1 rate of nasal VCV triphones. The value of Each point of the experiment observation was calculated among all nasals consonants regarding a given vowel.

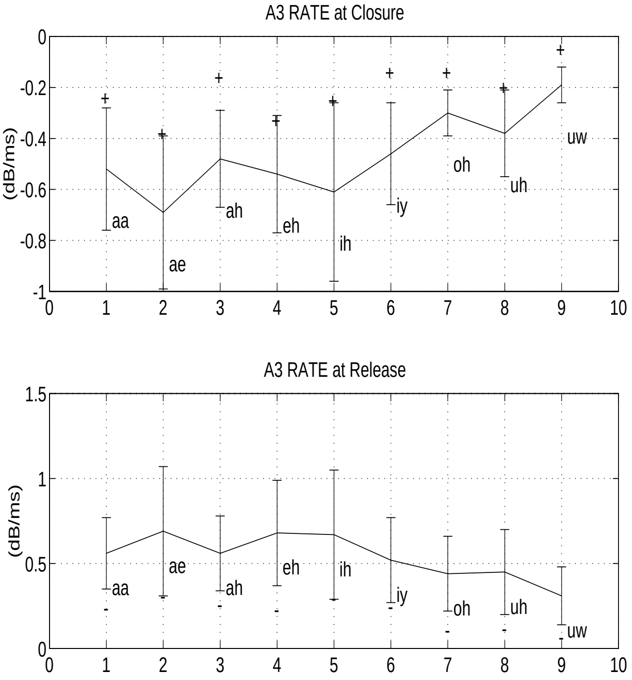

Figure 3: The Mean and the stand deviation of A23 rate of nasal VCV triphones. The value of Each point of the experiment

observation was calculated among all nasals consonants regarding a given vowel.

5. RESULTS

The nasal consonant landmark detection results obtained in the VCV database were summarized in Figure 4(a), 4(b), 4(c), and 4(d) in the form of error distribution and accuracy distribution at closure and release respectively.

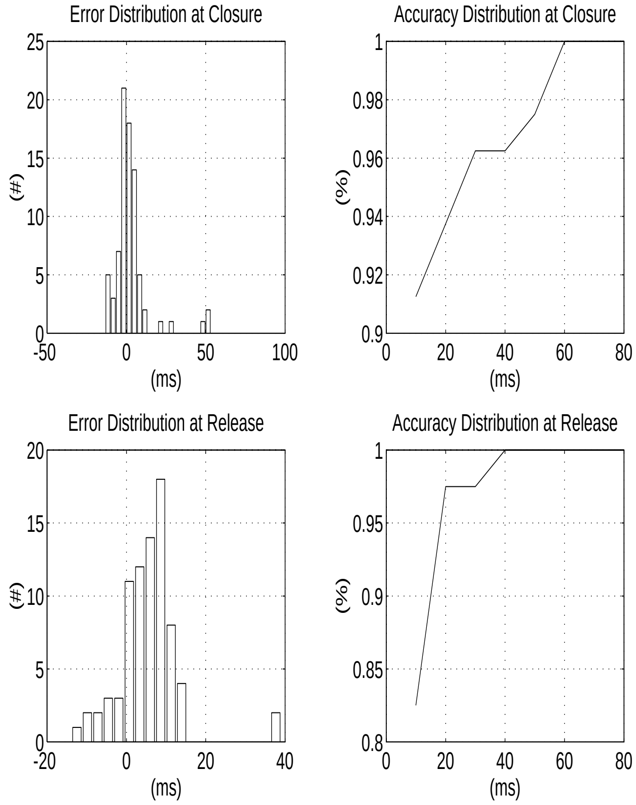

Figure 4: The nasal consonant landmark detection results. Error distribution displays the histogram of alarmed instants in terms of the error to the hand-labeled references of each landmark instant. Accuracy distribution counts for the accumulated detection accuracy in terms of the error to the hand-labeled references of each landmark instant.

For intervals of 20 ms, 40 ms and 60 ms apart from the hand-labeled instants of closure and release, the accuracy of detection was 94%, 96%, and 100% at closure and 97%, 100%, 100% at release respectively.

6. DISCUSSION

Fn1, the locus of the lowest spectral prominence in the range that covers the first formant and the lowest spectral prominence of nasal murmur, is estimated by the trace of a mass center of a smoothed spectral amplitude. Although the locus is not as accurate as LPC root solving in estimation of the pole of the transfer function, it avoids the ill problem of root solving and is reliable for tracing the rapid change in the frequency of the low-frequency prominence at the closure and release of nasal consonants. A23, the locus of the energy between 1000-3000 Hz, is another useful cue to trace the change in spectrum amplitude of the closure and release of the nasal murmur. However the acoustic correlates, to the events of nasal closure and release, of A23 rate is much less than that of Fn1 rate. Further interesting work should investigate integration of the nasal landmark detector presented in this paper with previously developed landmark detectors for abrupt consonants and for glides, so that more speech knowledge may be explicitly incorporated together and provide a landmark pre-processor for speech recognition.

7. REFERENCES

1. Liu, S., A., "Landmark detection for distinctive feature-based speech recognition," Massachusetts Institute of Technology, PhD's Thesis, 1995

2. Sun, W., "Analysis and interpretation of glide characteriestics in pursuit of an algorithm for recognition," Massachusetts Institute of Technology, Master's Thesis, 1996

3. House, A. S., "Analog studies of nasal consonants", Journal of Speech and Hearing Disorders, Vol.22, pp. 190-204, 1957

4. Fujimura, O., "Analysis of nasal consonants," Journal of the Acoustical Society of America, Vol. 34, No. 12, pp. 1865-1875, 1962

5. Fujimura, O., Lindqvist, J., "Sweep-tone measurements of the vocal tract characteristics," Journal of the Acoustical Society of America, Vol. 49, No. 2, pp. 541-558, 1971

6. Glass, J., R., "Nasal consonants and nasalized vowels: an acoustic study and recognition experiment," Massachusetts Institute of Technology, Master's Thesis, 1984

7. Steven, K., N., Acoustic Phonetics, (Forthcoming)