Nasal Detection in Continuous Mandarin Speech

Yi-Chun Lin, Hsiao-Chuan Wang

Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
g925931@alumni.nthu.edu.tw, hcwang@ee.nthu.edu.tw

ABSTRACT

In the formant synthesis of Mandarin speech, the nasalization effect for the syllable ending or behind the nasal consonants can be generated by a model of resonator-antiresonator pair. The center frequency trajectories of resonator-antiresonator pair determine the absence or the presence of nasalization effect. This spectral property in generating the nasalization effect can be applied for detecting the existence of nasals in Mandarin speech. The Seneff auditory model is used to partition the speech signal into event segments and the spectral features in each segment is extracted. The Gaussion Mixture Model (GMM) method is used to classify nasals and non-nasal vowels. Different settings of spectral bands and feature vectors are investigated. In our experiments, the accuracy of nasal detection can reach 82%.

1. INTRODUCTION

In typical Automatic Speech Recognition (ASR), a set of features is defined to specify the characteristics of speech in each frame. This set of features is used for recognizing all speech units, such as phones or syllables. The statistical models based on this set of features are generated. This corpus-based speech recognition method can not catch the specific characteristics of each individual phone. The performance of ASR is far from human speech recognition. Toward the next generation Automatic Speech Recognition, a paradigm integrating the knowledge sources with the recognition system was proposed [1][2]. The approach is based on the concept of articulatory phonetics features and acoustic landmarks. Nasal is one of landmarks. This paper deals with the detection of nasals in continuous Mandarin speech. They include nasal consonants, /m/ and /n/, and vowels terminated in nasal endings, /n/ or /ng/.

In the synthesis of Mandarin speech, nasalization effect can be generated by a resonator-antiresonator pair [3]. When the center frequency of the resonator and the center frequency of the antiresonator are departed, a nasalization effect arises. If these two center frequencies coincide, the nasalization effect disappears. In addition, the energy of nasals concentrates in low frequencies and one formant of a nasal exists between 200Hz and 400Hz. The nasalization effect will be shown stably in frequency domain for a while [4][5]. This property can be used for detecting nasals.

2. THE NASAL DETECTION SYSTEM

The block diagram of our nasal detection system is shown in Figure 1.

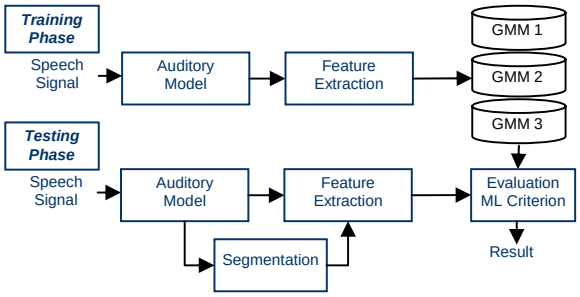

Figure 1. Nasal detection system architecture

2.1 Labeling Nasal events

At first, a set of training data should be available for labeling those nasals The nasal regions are manually located. Since nasals can be generated by a resonator-antiresonator pair in the formant synthesis method, we can mark the nasal region according to the trajectories of center frequencies of resonator-antiresonator pair [3].

When a nasal consonant exists, its subsequent vowel is nasalized. The first formant, F_1, of the vowel diverges and the center frequencies of the resonator-antiresonator pair depart to generate the nasalization effect during the first 100ms of the vowel portion. The manner of departing is according to the antecedent consonant. Then the center frequency of the antiresonator, F_{nz}, moves gradually to the center frequency of the resonator, F_{np}. Finally F_{nz} and F_{nz} are coincident so that the nasalization effect disappears. Figure 2 shows the example of trajectories of F_1, F_{np}, and F_{nz} of nasalized vowel /a/ following a nasal consonant.

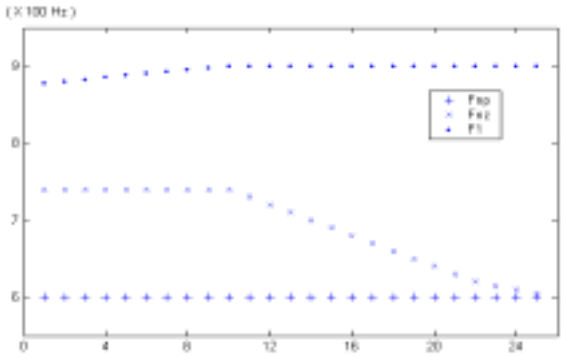

Figure 2. Trajectories of F_1, F_{np}, and F_{nz} of nasalized vowel following a nasal consonant

For the case of a vowel with nasal ending, the centers frequencies of the resonator-antiresonator pair depart gradually during the final 100ms. Figure 3 shows examples of the trajectories of F_{np}, F_{nz} (/n/), and F_{nz} (/ng/) of nasal endings .

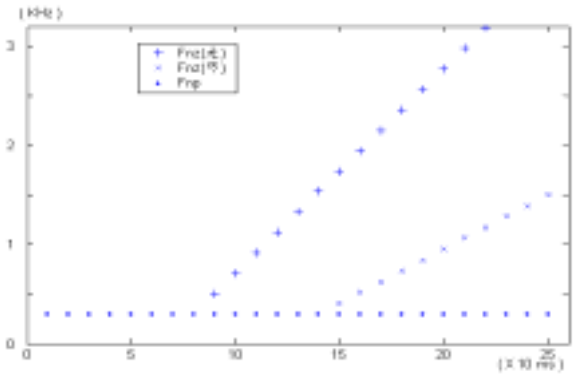

Figure 3. Trajectories of F_1, F_{np}, and F_{nz} of nasal endings

Therefore, for a syllable with nasal consonant, we mark the first 100ms of its vowel as a nasal region. For a syllable with a nasal ending, we mark the last 100ms as a nasal region.

2.2. Signal Representation

Seneff auditory model [6][7][8] suggests to transform speech signals into two sets of 40 dimensional feature vectors for segmentation and speech recognition. The auditory model is shown in Figure 4.

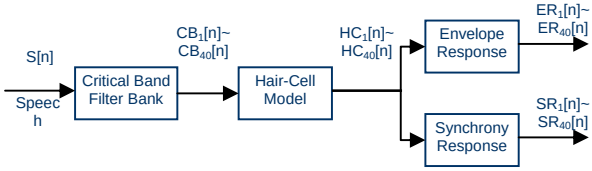

Figure 4. Seneff auditory model

The envelope response tends to enhance the onset and offsets of speech events. In order to obtain a set of reliable features, we perform the following algorithm:

(1). Record every local maximum of $HC_i[n]$.

(2). Pick the local maximum:

1. Set the first nonzero value as a flag
2. within 1.8 times pitch from the flag, only the maximal value is kept and set as new flag while the other values are set to zero
3. If the new flag value is zero, use interpolation to count the other values. Otherwise repeat step 2.

(3). Repeat steps (1), (2) until the end of file

(4). If the number of successive nonzero values is less than 28 ms, those might not be speech signals. So let them all zeros.

Figure 5 shows the speech signal in a frequency band. The output of hair cell model is in the upper part, and the envelope response is in the lower part.

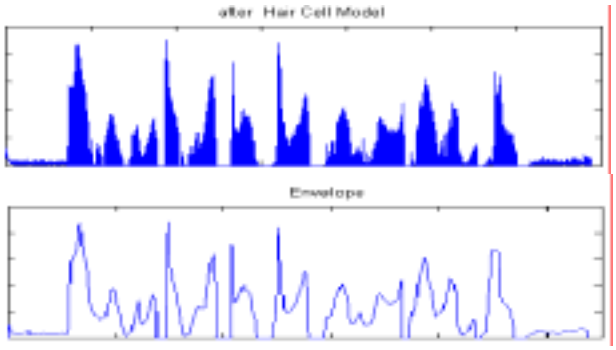

Figure 5. Envelope response in a frequency band

2.3. Segmentation

In our experiments, $ER_1[n]$~ $ER_{40}[n]$ are computed once every 4 ms, and used as the feature vector for segmentation. The method is to compare the similarities between a feature vector and its immediate neighbors. The boundaries of segments are located wherever the vector affiliation changes [9][10].

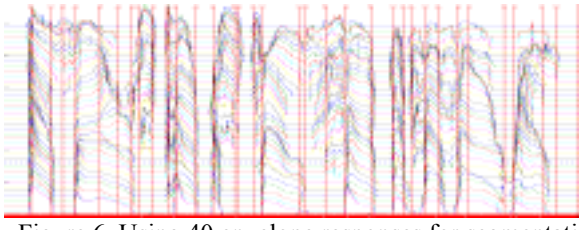

Figure 6. Using 40 envelope responses for segmentation

2.4. Feature Extraction

The experimental speech database was telephone

recorded. We examine the frequency range in 180 Hz-4kHz. One of band settings is called M-band configuration, i.e., 180 Hz– 800 Hz (M_1), 800 Hz– 2kHz (M_2), and 2kHz–4kHz (M_3) [11]. The average energy and peak amplitudes of each band are used as feature vectors for nasal detection. Since there is a formant existing in 200 Hz - 400 Hz, we particularly define an additional band in 180 Hz– 400 Hz to form another band setting, called the N-band configuration, i.e., 180 Hz– 400 Hz (N_1), 400 Hz–800 Hz (N_2), 800 Hz– 2kHz (N_3), and 2kHz– 4kHz (N_4).

According to the spectral property in generating the nasalization effect, we divide an event segment with nasal consonant or nasal ending into two halves. The energy and peak amplitudes of the second half of nasal consonants are larger than that of first half for nasal consonants. For vowels with nasal endings, the energy and peak amplitudes of the second half of nasal consonants are smaller than that of first half. Non-nasal vowels don't have such properties.

The output of synchrony response of Seneff auditory model is used for speech classification. The features are extracted as follows; Suppose that the first half of event E is from time n_0 to n_1, and the second half of event E is from time n_1 to n_2. Let $\overline{SR_i}$ denote the average of $SR_i[n_0] \sim SR_i[n_1]$. Let feature vector,

$$\vec{S} = \left[\sum_{i \in B_1} \overline{SR_i} \quad \cdots \quad \sum_{i \in B_N} \overline{SR_i} \right] \tag{1}$$

where B_j is N_j for N-band configuration and B_j is M_j for M-band configuration. This feature vector $\vec{S}$ is recognized as the average sum of synchrony responses. In addition, we compute the average differences of synchrony responses between the middle and the start of the speech event, and the average differences of synchrony responses between the terminal and middle of the speech event. Let $SR_i' = SR_i[n_1] - SR_i[n_0]$, $SR_i'' = SR_i[n_2] - SR_i[n_1]$, then we get a feature vector, $\vec{D}$

$$\vec{D} = \left[\frac{1}{K_1} \sum_{i \in B_1} SR_i' \quad \cdots \quad \frac{1}{K_N} \sum_{i \in B_N} SR' \quad \frac{1}{K_1} \sum_{i \in B_N} SR'' \quad \cdots \quad \frac{1}{K_N} \sum_{i \in B_N} SR'' \right] \tag{2}$$

where K_j is the number of the elements of B_j.

2.5. Classification

This part includes two stages: the silence detection and the ML Criterion. In the first stage, the silence threshold is the average energy of speech file during the first 20 ms. If the average energy of testing event is less than the silence threshold, the event is classified as a silence event. Otherwise the event is classified by the second stage, the ML Criterion.

In the second stage we suppose the prior probabilities of nasal model and non-nasal model are the same and evaluate the similarities of the testing event (X) with nasal model (Λ) and non-nasal model (Λ') to decide the class of X.

$$\text{score} = \log P(X \mid \Lambda) - \log P(X \mid \Lambda') \begin{cases} > \log\eta, & X \in \text{鼻音} \\ \le \log\eta, & X \in \text{非鼻音} \end{cases} \tag{3}$$

In (3), log η is a threshold, the first term is called the target term, and the second one is called the normalization term. We use obstruent model (Λ'_1) and sonorant model (Λ'_2) to set up the non-nasal model. Normalization term can be computed by three kinds of value: maximum, arithmetic mean and geometric mean [12].

3. EXPERIMENTS

The experimental data are obtained from MAT-160 database which are telephone speech in 16-bit PCM format. The training set contains 129 sentences recorded by 8 females and 11 males. It includes 165 nasal consonants, 173 nasal endings, 214 obstruents, and 170 sonorants. The testing set contains 8 sentences recorded by 4 females and 3 males. There are 12 nasal consonants and 28 vowels with nasal endings.

Experiment 1. Comparing M-band and N-band configurations

In this experiment, we examine the efficiency of using M-band and N-band configurations. Feature vectors are the output of auditory model. Neighborhood within limited bound method and majority-vote decision strategy were used. The result is shown in Figure 7. It shows that the false reject rate of N-band configuration is a little less than M-band configuration, and the false accept rate of N-band configuration reduces 15-20 %. The equal error rate of N-band configuration is about 30 %. Therefore N-band configuration is used in the following experiments.

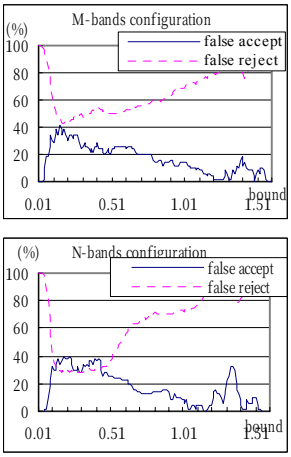

Figure 7. comparing frequency band configurations

Experiment 2. Influences of the auditory model and properties in time domains

This experiment is to investigate the influence of auditory model on nasal detection. The feature vector is average sums of synchrony responses, and the decision rule is ML Criterion. As the result given in Figure 8, the equal error rate is reduced to 25%. It shows that the performance of the feature vector extracted after all the auditory model is better. In addition, Figure 9 shows that results of using three normalization terms. The performance of geometric mean is the worst. The results of maximum and arithmetic mean are almost equal, so the normalization based on maximum is chosen for next experiment.

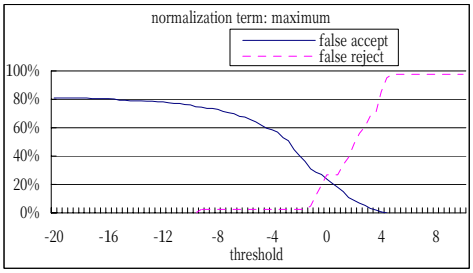

Figure 8. Error rate curves of Experiment 2

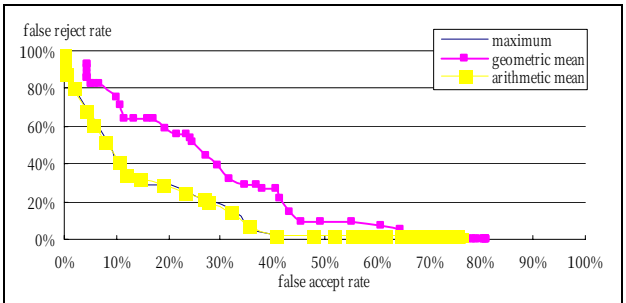

Figure 9. Detection error tradeoff curve of Experiment 2

Experiment 3. Improving feature vector

Besides the feature vector used in Experiment 2, the average differences of synchrony responses are a feature vector now. The result is shown in Figure10. Comparing with Experiment 2, the false accept rate and error reject rate are both reduced, and the best accuracy is 82%.

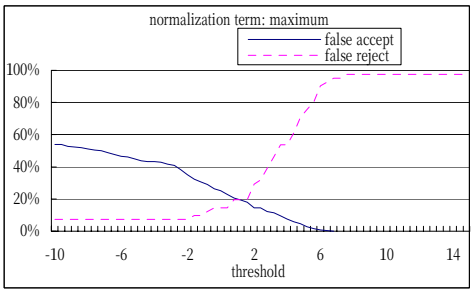

Figure 10. Error rate curves of Experiment 3

Experiment 4. Nasal model with consonants and vowels divided

In this experiment, nasal consonants and vowels with nasal endings are trained separately, and combined by maximum computation. As shown in Figure 11, the model doesn't improve the performance of nasal detection, even if the Gaussian mixture number is increased.

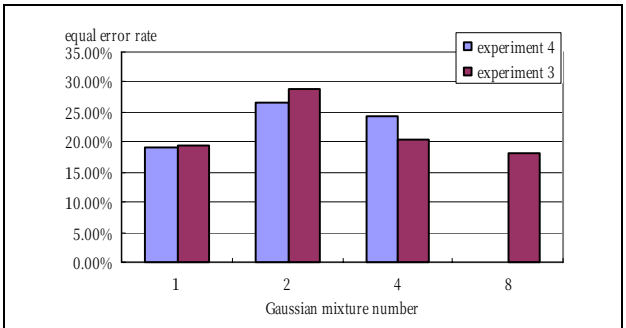

Figure 11. Detection error tradeoff curve of experiment 2

4. CONCLUSION

The application of the spectral property in generating the nasalization effect to detect the existence of nasals in Mandarin speech is presented. Our experiments show that detection accuracy is improved. Setting 180 Hz~ 400 Hz as an extra band is helpful. Using average sums and average differences of synchrony responses as the feature vector can make the performance better and reduce the false reject rate. This approach would be useful for the knowledge-based speech recognition.

5. ACKNOWLEDGEMENT

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-93-2213-E-007-019.

6. REFERENCES

[1] Chin-Hui Lee, "From Knowledge- Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition," in *International Conference on Spoken Language Processing*, *ICSLP2004*, Plenary Session , Jeju, Korea

[2] Kenneth N. Stevens, "Toward a model for lexical access base on acoustic landmarks and distinctive features," in *J. Acoust. Soc. Am*. 111 (4), pp. 1872-1891, April 2002

[3] Hong-Bin Chiou, Hsiao-Chuan Wang, Yueh-Chin Chang, "Synthesis of Mandarin Speech Based on Hybrid Concatenation," in *Computer Processing of Chinese and Oriental Languages*, Vol.5, No.3、4, pp. 217-231, November 1991

[4] John M. Howie, "Acoustical Studies of Mandarin Vowels and Tones," Cambridge University Press, 19 76

[5] Peter Ladefoged, "Vowels and Consonants An Introduction to the Sounds of Languages," Blackwell Publishers, 2001

[6] Stephanie Seneff, "A Joint Synchrony/ Mean-rate Model of Auditory Speech Processing," in *Journal of Phonetics* 16, pp. 55-76, 1988

[7] Stephanie Seneff, "A Computational Model for the Peripheral Auditory System :Application to Speech Recognition Research," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1986*, pp. 1983-1986

[8] Malcolm Slaney, "Auditory Toolbox," Technical Report #1998-010, *http://rvl4.ecn.purdue.edu/%7Emalcolm/interval/1998-010/*

[9] James R. Glass, Victor W. Zue, "Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment," S.M Thesis, Massachusetts Institute of Technology, February 1985

[10] James R. Glass, Victor W. Zue, "Detection and Recognition of Nasal Consonants in American English," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1986*, Vol. 10, pp. 1569-157

[11] Marilyn Y. Chen, "Nasal Detection Module for a Knowledge-based Speech Recognition System," in *International Conference on Spoken Language Processing, ICSLP 2000*, Vol.6, pp.636-639

[12] Gia-Tsong Lin, Hsiao-Chuan Wang, "Pattern Matching of Mandarin Speech and Its Application to Pronunciation Learning," Thesis, June 2001