J. Biomed. Eng. Res.
Vol. 20, No. 6, 591-599, 1999

# 선형 예측 모델을 이용한 비관혈적 과비음성 추정

고영일 · 김덕원* · 나동균** · 최흥식***

연세대학교 생체공학 협동과정, *연세의대 의학공학교실, **연세의대 성형외과학교실,
***연세의대 이비인후과학교실, 음성언어의학 연구소

*(1999년 7월 23일 접수, 1999년 12월 1일 채택)*

# A Noninvasive Estimation of Hypernasality using Linear Predictive Model

Y.I. Ko, D.W. Kim*, D.K. Rah**, H.S. Choi***

Graduate Program in Biomedical Engineering, Yonsei University
*Department of Medical Engineering, **Department of Plastic Surgery,
***Department of Otorhinolaryngology, The Institute of Logopedics and
Phoniatrics, Yonsei University College of Medicine

*(Received July 23, 1999. Accepted December 1, 1999)*

**요 약** : 연구개에 결함이 있는 사람의 발음은 부적절한 비음이 섞이게 되어 과비음성 비음이 되어 연구개를 복원해주는 시술을 하게 되는데, 과비음성 비음을 정량적으로 측정할 수 있다면 시술 결과를 객관화할 수 있게 된다. 현재 임상적으로 사용되고 있는 방법들은 관혈적이거나 고가의 장비를 필요로 한다. 본 논문에서는 비음의 특징인 스펙트럼에서 zero의 존재와 비강에 의한 포만트의 존재 사실, 그리고 선형 예측 모델을 이용하여 마이크로폰과 사운드 카드가 장착된 PC로 구현할 수 있는 새로운 과비음성 비음 추정 알고리즘을 제안하였다. 음성 신호의 스펙트럼에 zero가 존재하는 경우, 낮은 차수(order)의 선형 예측 모델이 그 음성을 발음한 성도 시스템에 정확히 적용되지 않는다는 점을 이용하여, 같은 음성에 대한 높은 차수의 선형 예측 모델과의 차이를 이용해서 과비음성의 정량화를 시도했다. 본 논문에서 제안된 알고리즘은 기존의 Teager Operator를 이용한 알고리즘에 비해서 Nasometer의 측정결과와 더 높은 통계적 상관관계를 보여주었다.

**Abstract** : The pronunciation of a speaker with a defective soft palate features hypernasality. The operation that recovers a defective soft palate is necessary to reduce the hypernasality. The assessment of hypernasality is needed in light of quantifying the effect of the therapy. The current clinical methods assessing hypernasality are invasive and carried out with a expensive equipment. In this paper, we proposed a new algorithm to estimate hypernasality. The implementation of algorithm requires only a microphone and a PC equipped with a sound card. The algorithm uses the fact that the low order linear predictive model which identifies a human vocal tract system is not accurate in case that the vocal tract system has zeros in its frequency response. And the zeros in frequency response of vocal tract system is one of features of hypernasality. The estimation of hypernasality is done by comparing low-order linear predictive model with high-order linear predictive model. The proposed algorithm have a higher statistical correlation with nasalance by Nasometer than the algorithm using Teager operator.

**Key words** : Soft palate, Estimation, Hypernasality, Zeros in spectrum, Linear predictive model

## 서 론

연구개에 결함이 있는 사람의 발음에는, 비강(nasal cavity)과 구강(oral cavity)이 제대로 분리가 되지 못하여 부적절한

통신저자 : 김덕원, (120-752) 서울시 서대문구 신촌동 134
Tel. (02)361-5402, FAX. (02)364-1572

비음이 섞이게 된다. 특히 구개열(Cleft Palate)이나 Velopharyngeal Incompetence(이후 VPI)의 경우가 그러한데, 연구개의 균열로 인해 비강과 구강의 분리가 완전치 못하다. 비음의 정도가 심한 경우에는 화자의 발음을 제대로 이해하기가 힘들어 원활한 의사소통에 문제를 겪게 된다. 그러므로 균열된 연구개를 복원해 주는 구개성형술(palatoplasty)을 시술하게 되는데, 만일 환자의 비음의 정도(nasality)를 정량적으로 측정할 수 있다면, 수술의 성과를 객관화할 수 있게 되어 임상의에 많은 도움을 줄 수 있다.

일반적으로 비음은 저비음성 비음(hyponasality)과 과비음성 비음(hypernasality)으로 분류된다. 저비음성 비음은 코로부터 나오는 소리가 전혀 없는 비음을 뜻한다. 손으로 코를 막고 말을 하면 들을 수 있는 비음이 바로 저비음성 비음이다. 과비음성 비음은 코로부터 나오는 소리가 정상치보다 상대적으로 큰 경우의 비음이다. 그러므로 연구개의 긴장 때문에 구강과 비강의 분리가 확실치 못해서 들리는 비음은 과비음성 비음이 된다. 본 논문에서 고려하고 있는 비음은 모두 과비음성 비음이다.

현재 임상에서 사용 중인 각종 비음의 측정법에 대해서 살펴보면 다음과 같다. 특정한 발음을 하는 동안 구개범인두부(velopharyngeal structure)를 보기 위해서 측면에서 방사선 촬영을 하는 방법이 있다[1]. 또한 multiview videofluoroscopy를 이용해서 연속된 발음을 하는 동안, 다각도로 구개범인두의 움직임을 살펴보는 방법도 있다[2]. 관혈적인 방법으로서 유연한 광섬유를 이용해서 구개범인두부의 움직임을 직접 살펴보기도 한다[1]. 이런 방법들은 구개범인두부를 직접 관찰할 수 있다는 장점이 있지만, 이런 방법들의 관찰 결과와 과비음성 비음과의 상관관계는 그리 뚜렷한 편이 아니며 정량적 데이터를 얻을 수 없다는 것과 관혈적이라는 단점이 있다[1-2].

과비음성 비음은 결국 비강과 구강의 선함이 부적절해서 나타나는 현상이므로 과비음성 비음을 측정하기 위해서 입에서 나는 소리와 코에서 나오는 소리의 비(nasal-oral ratio)를 구하는 연구가 시도되었다. Horii는 Horii Oral Nasal Coupling (HONC)라고 불리는 지표를 제안하였다[3]. 이 방법은 콧구멍 근처에 부착한 진동 가속도계의 측정치와 입 앞에 부착된 마이크로폰의 측정치의 비를 이용하는 방법이다. 이 방법의 결과와 사람이 직접 들었을 때의 과비음성 비음의 판단 결과와는 매우 높은 통계적 상관관계가 있다고 보고되었다[4].

Nasal oral ratio를 이용하는 다른 방법으로 Nasometer[5]라는 기기가 있는데 이것은 입과 코에 마이크로폰을 장착하고 각각의 음압 레벨을 측정하는 것이다. 적절한 처리과정[5]을 거친 마이크로폰에서의 측정 결과를 이용해서 식 (1)과 같이 percent nasalance를 계산하게 된다.

$$Nasalance = \frac{N}{N+O} \times 100\,\% \qquad (1)$$

위의 식에서 N과 O는 각각 코와 입에서 측정되는 신호의 크기이다. 정상적이라고 판단할 수 있는 발음의 경우, nasalance는 보통 32 % 이하로 측정된다[6]. 이 Nasometer의 결과는 사람이 직접 듣고 비음을 판단하는 경우와 높은 통계적 상관관계가 있다고 보고되고 있으며, 실제 임상적으로도 가장 많이 사용되고 있는 방법이다. 그러나 Nasometer는 매우 고가의 장비이며, 또한 측정 시에 headset을 착용해야 하는 등의 불편함이 있어서 유아들이 측정을 하는 데에 어려움이 따른다.

위에서 살펴본 바와 같이 현존하는 각종 측정 방법들은 관혈적이거나 고가의 장비를 필요로 한다. 그래서 사운드 카드와 마이크로폰이 장착된 일반적인 PC를 이용해서 비음을 추정하는 방법을 개발할 필요성이 대두되었다. 현재까지 연구된 알고리즘으로는 Cairns 등에 의해서 제안된 방법인 Teager Operator를 이용하는 알고리즘이 있다[7]. 이 알고리즘은 마이크로폰 하나를 이용해서 발음을 측정한 뒤에 적절한 신호처리를 통해서 비음의 정도를 구하게 된다[7]. 이 알고리즘은 그 방법의 특성상 음성의 포만트(formant)를 정확히 구해야 하는 문제가 있고, 측정할 음성의 pitch가 높은 경우는 신뢰할만한 결과를 얻기가 힘들다. 참고문헌 [7]의 알고리즘에 관해서는 본 논문의 실험 결과 부분에서 다시 다루기로 한다.

본 논문에서는 시스템의 all-pole 모델을 가정하는 선형 예측 모델의 특성과 비음의 스펙트럼의 특성을 이용해서 과비음성 비음을 1개의 마이크로폰과 PC를 이용해서 추정하는 알고리즘을 제시하였다.

## 사람 성도의 모델링

일반적으로 사람의 성도를 모델링하기 위한 물리적인 가정으로서 인간의 성도(vocal tract)가 에너지 손실이 없는 여러 개의 굵기가 다른 관으로 이루어져 있다고 가정한다. 또한 발음하는 음이 모음인 경우 성도를 여기(excitation)시키는 성문(vocal cord)으로부터의 신호는 impulse train이라고 가정한다. 이러한 가정 하에 불리식 방정식들을 구하고 디지털 신호처리 기법을 이용해서, 그 방정식들을 성문에서 발생된 여기 신호가 입까지 전달되는 성도 필터 시스템에 관한 관계식들로서 해석을 하면, 성문에서 입까지의 전달함수는 시간 지연을 나타내는 분자의 항을 제외하면 all pole 시스템으로서 표현되어진다[8,10]. 바로 이 사실 때문에 all-pole 시스템을 identification할 수 있는 선형 예측 모델(혹은 autoregressive 모델)이 음성 신호 분석에 널리 쓰이고 있는 것이다.

## 선형 예측 모델

선형 예측 모델에 사용되는 P차의 선형 예측 계수(linear predictive coefficient, 이하 LP 계수) $a_k$는 식 (2)와 같이 정의된다[8,10].

$$s(n) = \sum_{k=1}^{P} a_k s(n-k) + \Theta_0 u(n) \qquad (2)$$

식 (2)에서 $s(n)$은 음성신호이고, $u(n)$은 성문에서 발생되는 최대값이 1로 정규화(normalize)된 여기신호이다. $\Theta_0$ 는 여기신호의 이득(gain)이며, 동시에 선형 예측 모델을 이용해서 원래의 음성 신호를 예측했을 때, 원래 음성신호와 예측된 신호와의 mean squared residual error의 제곱근이 된다. 식 (2)를 Z-transform하면 식 (3)과 같다.

$$\Theta(z) = \frac{S(z)}{U(z)} = \frac{\Theta_0}{1 - a_1 z^{-1} - \ldots - a_p z^{-p}} = \frac{\Theta_0}{A(z)} \quad (3)$$

식 (3)에서, p차의 LP 계수는 pole이 p개인 스펙트럼을 가지는 시스템을 표현할 수 있음을 알 수 있다. 식 (3)에 $z=ej\omega$를 대입하면 여기신호의 스펙트럼 성분이 제거된, 그 음성 신호를 발음한 성도 시스템만의 스펙트럼을 얻을 수 있다.

## 비음의 특징

일반적으로 모음은 스펙트럼 상에서 포만트라고 하는 spectral peak들의 성분들에 의해서 구성된다[8]. 그러나 구강만을 통해서 발음되는 일반적인 발음들과는 달리 과비음성 비음의 경우는 비강을 통해서도 소리가 나게 되어서, 일반적인 발음과는 구별되는 다음과 같은 특징들을 가진다[7,9].

### 1) 제1 포만트의 크기 감소

거의 모든 모음에서 볼 수 있는 현상이다. 이러한 현상은 일반적으로 비강 내벽의 감쇄 특성(damping characteristics)에 의한 결과라고 생각되어진다.

### 2) Anti-resonance(zero)의 존재

일반적으로 음향신호의 스펙트럼 상에서 크기가 급작스럽게 감소(anti-resonance)하는 것은, 그 음향신호가 발생된 관에 작은 관(side-branching tube, 음성의 경우는 비강)이 연결되어 있어서 그곳으로 소리가 새어나갈 때 나타나는 현상이다.

### 3) 강한 고조파의 존재

고조파는 성도로부터 올라온 소리가 비강 내부에서 공명을 해서 나타나는 것으로 여겨진다.

### 4) 포만트의 위치이동

주파수 영역에서 포만드 위치가 이동한다. 이러한 현상은 구강에 비강이 연결됨으로써 전체적인 내부 구조에 변화가 생겨서 일어나는 현상이다.

## 본 연구에서 제안한 알고리즘

과비음성 비음은 정상적인 발음과는 달리 그 주요한 특징으로 스펙트럼 상에서 zero를 가지는데, 이것은 성도를 선형 예측 모델링하는데 가장 중요한 가정에 위배되므로 all-pole모델을 나타내는 일반적인 LP 계수로는 과비음성 비음을 정확히 나타낼 수 없다는 결론에 이르게 된다. 그러나, 이론적으로는 무한개수의 pole이 존재하면 하나의 zero를 식 (4)와 같은 관계에 의해서 표현할 수가 있다.

$$1 - z_0 z^{-1} = \frac{1}{1 + \sum_{k=1}^{\infty} z_0^{\,k} z^{-k}} \quad (4)$$

식 (4)로부터 zero인 $z_0$는 무한 차수의 분모를 갖는 all-pole 모델로서 표현할 수 있음을 알 수 있다. 그러므로 실제로 무한 차수의 all-pole 모델을 이용하지는 못해도, 일반적으로 사용되는 LP 계수의 차수인 8 12 차보다 높은 차수로 모델링한다면, 낮은 차수의 선형 예측 모델(Linear Predictive model, 이후 LP 모델)보다는 좀더 실제 성도의 스펙트럼에 가까운 LP 모델을 기대할 수 있을 것이다.

LP 모델의 차수 변화에 따른 각 차수의 LP 모델에서의 residual error의 크기를 구한 그림 1에서 볼 수 있듯이 일반적으로 정상적인 음성 신호(nasalance 25%)에 대해서는 10차 이상의 높은 차수로 LP 모델을 구하더라도 낮은 차수의 LP 모델과 큰 차이를 보이지 않는다. 반면 과비음성이 심한 음성(nasalance 50%)의 경우는 낮은 차수의 LP 모델과 높은 차수의 LP 모델의 차이는 비교적 커진다. 그러므로 결국 어떤 음성 신호에 대해서 각각 낮은 차수와 높은 차수의 LP 모델을 구한 후 그것들을 비교해보면, 차이가 심할 수록 과비음성이 심한 음성이라고 추론할 수 있다.

주어진 두 LP 모델의 차이를 나타내는 방법은 여러 가지가 있지만, 음성 인식 알고리즘 등에서 널리 쓰이는 방법으로는 LP-Cepstrum을 이용하는 방법이 있다[8,10,12]. Real LP-Cepstrum c(n), Complex LP-Cepstrum $\gamma(n)$은 각각 식 (5),(6)과 같이 정의된다.

$$\begin{aligned} \gamma(n) &= \log \Theta_0 && n=0 \\ &= -a_n + \sum_{k=1}^{n-1} \frac{k}{n} \gamma(k) a_{n-k} && n>0 \end{aligned} \quad (5)$$

$$c(n) = \begin{cases} \gamma(0) & n=0 \\ \frac{1}{2}\gamma(n) & n>0 \\ 0 & n<0 \end{cases} \quad (6)$$

식 (5)에서 $\Theta_0$는 식 (2), (3)에서의 여기신호의 이득이고, $a_k$는 M차 LP 계수이다. 식 (5)에서 k<0, k>M 에 해당하는 $a_k$는 0을 대입한다. 식 (5),(6)에서 구해지는 real LP-Cepstrum을 이용해서 LP모델의 스펙트럼의 차이를 비교할 수 있다. 즉 높은 차수의 LP 계수와 낮은 차수의 LP 계수로부터 구해지는 real LP-Cepstrum을 각각 $c_H(n)$, $c_L(n)$이라고 한다면 두 Cepstrum

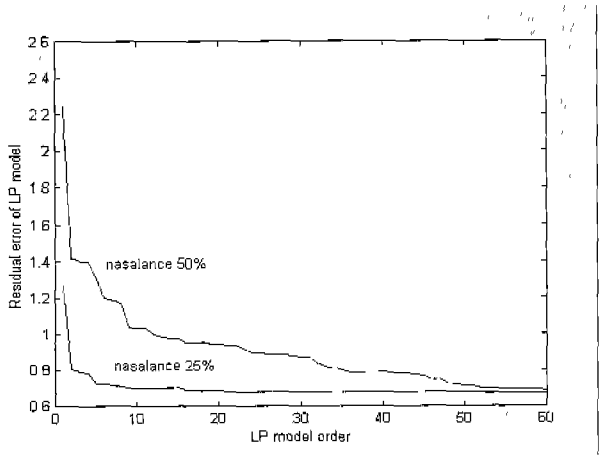


**그림 1.** Nasalance 25%, 50%인 음성 신호에 대한 LP 모델 order에 따른 residual error의 변화
**Fig. 1.** Plot of residual error corresponding to increasing order of LP model of speech signal

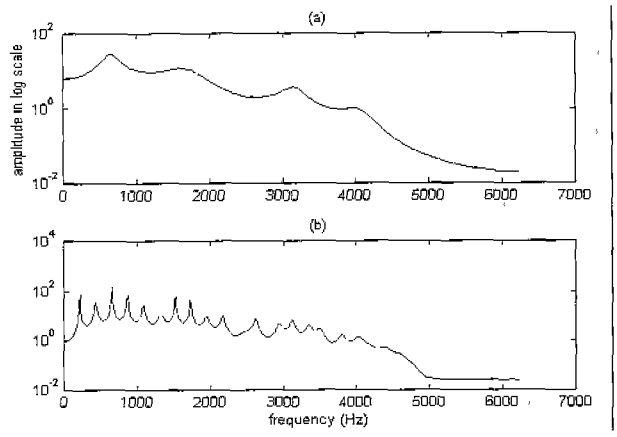


**그림 2.** Pitch 224 Hz인 음성신호에 대한 LP 모델의 스펙트럼
(a) 차수 10의 LP 모델의 스펙트럼
(b) 차수 60의 LP 모델의 스펙트럼
**Fig. 2.** Spectrum of LP medel of speech signal with pitch of 224 Hz
(a) Spectrum of order-10 LP medel
(b) Spectrum of order-60 LP medel

간의 기하학적인 거리는 식 (7)과 같은 관계가 있다[8,10,12].

$$\sum_{n=1}^{\infty}[c_H(n)-c_L(n)]^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi}[\log|A_H^{-1}(\omega)-A_L^{-1}(\omega)|]^2 d\omega \quad (7)$$

식 (7)에서 $A^{-1}(\omega)$는 식 (3)의 1/A(z)에 $z=e^{j\omega}$를 대입해서 얻는 함수이다. 식 (7)에서 볼 수 있듯이 두 Cepstrum 간의 기하학적 거리는 두 선형 예측 모델로부터 얻을 수 있는 스펙트럼의 에너지 차이를 나타내므로 두 스펙트럼의 유사성을 나타내는 좋은 척도가 된다.

## 모델 차수의 선택

본 논문에서 제시하는 알고리즘은, 주어진 음성신호에 대해서 각각 낮은 차수와 높은 차수의 LP 모델을 구한 후, 그 둘의 차이를 비교하는 것이다. 다음은 본 논문에서 사용된 LP 모델의 낮은 차수와 높은 차수의 선택 기준이다.

▶ **낮은 차수** : 일반적으로 모음의 경우는 5000 Hz 이내에 제 4 포만트까지가 유효하다고 알려져 있다[9,11]. 그러므로 켤레근까지 고려해서 이론적으로 8차의 LP모델이라면 음성 신호의 스펙트럼을 상당히 근사하게 모델링할 수 있다. 과비음성이 거의 없는 음성에 대해서는 일반적으로 널리 쓰이는 10차 정도의 낮은 차수의 LP 계수들만으로도 성도를 충분히 모델링할 수 있다. 그러므로 일반적으로 널리 쓰이는 차수인 10차를 선택하였다.

▶ **높은 차수** : 음성 신호에 LP 모델을 적용하는 경우, 성도에 가해지는 여기신호가 직교성(orthogonality)를 유지해야만 구해진 LP 모델이 정확히 성도를 모델링하고 있다고 신뢰할 수 있다[8]. 만일 여기신호의 직교성이 확보되지 못하는 경우는 구해진 LP 모델은 성도만이 아닌, 여기신호의 스펙트럼 특성까지도 포함하게 된다. 유성음의 경우는 여기신호가 impulse train의 형태를 갖게 되는데, pitch가 P Hz인 음성신호를 샘플링 속도(sampling rate) SR Hz로 샘플링하게 되는 경우는, 여기신호에서 impulse간의 간격이 SR/P 샘플이 된다. 차수 N의 LP 모델을 구하기 위해서는 lag 0 ~ N-1 까지의 음성신호의 자기상관(autocorrelation) 값이 필요한데, N-1 ≥ SR/P 인 경우는 lag SR/P에서 0이 아닌 여기신호의 자기상관 값이 존재하게 된다. 그러므로 LP 모델의 차수는 SR/P 보다 작아야 그 신뢰성을 확보할 수 있다. 그림 2에서 그 일례를 볼 수 있다. 그림 2는 pitch 224 Hz인 음성신호를 샘플링 속도 12.5 kHz로 샘플링한 음성 신호에 대한 각각 차수 10, 60의 LP 모델의 스펙트럼이다. 그림 2의 (b)에서 볼 수 있듯이, 성도의 특성뿐 아니라 여기신호의 스펙트럼 특성에 해당하는 pitch 224 Hz의 spectral peak과 그의 고조파(harmonics)들에 해당하는 spectral peak들이 그림에 나타나 있는 것을 관찰할 수 있다. 그러므로 본 논문에서 높은 차수는 SR/P를 그 한계로 설정했다.

## 이론적인 실험 및 분석

본 논문에서 제안한 과비음성 비음을 추정하는 알고리즘을 실제 음성 데이터에 적용하기 전에 이론적으로 잘 동작하는지의 여부를 살펴보기 위해서 /아/ 발음을 simulation한 파형에 대해서 적용을 시켜보았다. 여기서 simulation할 발음으로 /아/ 발음을 설정한 것은, 이후 본 논문의 알고리즘을 적용할 실제 음성 재료가 /아/ 발음들이기 때문이다.

통계적으로 미국인의 정상적인 /아/발음은 스펙트럼에서 다음과 같은 특징을 갖는다[8].

1차 포만트 : 730 Hz, 2차 포만트 : 1090 Hz, 3차 포만트 : 2600 Hz

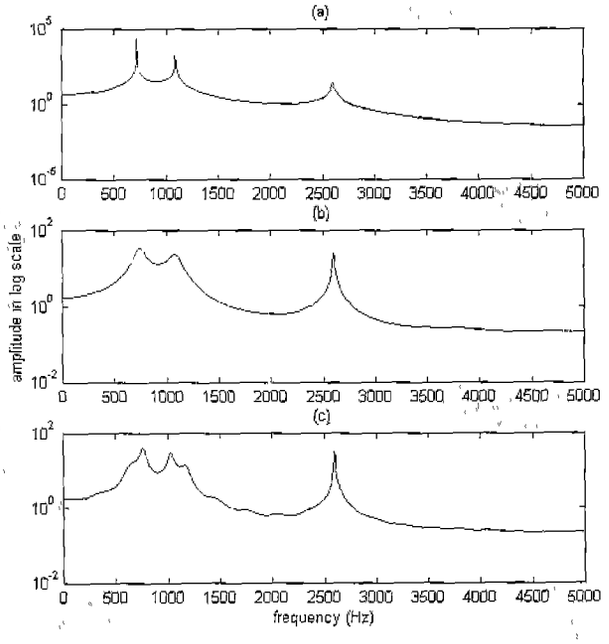


**그림 3.** k=0.0일 때의 스펙트럼
(a) H(z)의 스펙트럼, (b) LP 차수 10의 스펙트럼, (c) LP 차수 40의 스펙트럼
**Fig. 3.** Spectrum with k=0.0
(a) Spectrum of H(z), (b) Spectrum of order-10 LP model, (c) Spectrum of order-40 LP model

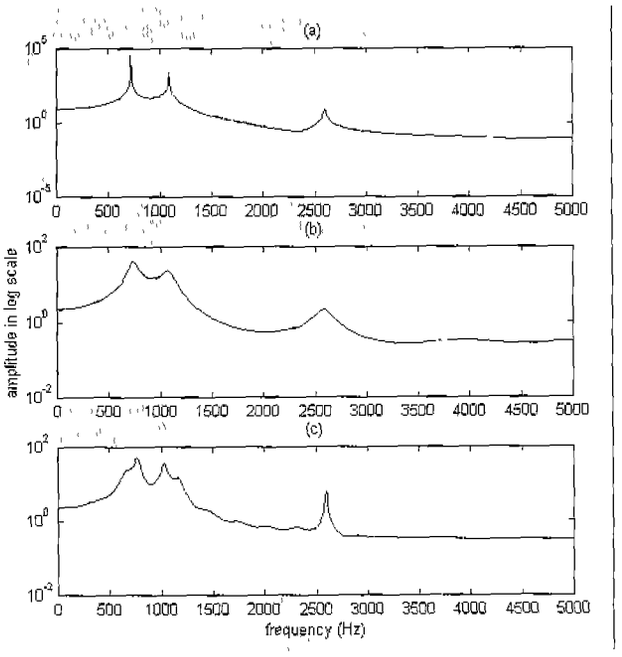


**그림 4.** k=0.9일 때의 스펙트럼
(a) H(z)의 스펙트럼, (b) LP 차수 10의 스펙트럼, (c) LP 차수 40의 스펙트럼
**Fig. 4.** Spectrum with k=0.9
(a) Spectrum of H(z), (b) Spectrum of order-10 LP model, (c) Spectrum of order-40 LP model

또한 과비음성 /아/발음의 경우, spectral zero는 2400 Hz 성도에서 생긴다. 그러므로 /아/발음을 하는 성도의 모델로서 다음과 같은 식을 세울 수 있다.

$$H(z)=\frac{(k\,e^{j\frac{2\pi f_z}{RATE}}z^{-1}-1)(k\,e^{-j\frac{2\pi f_z}{RATE}}z^{-1}-1)}{\prod_{i=1}^{3}(k_h e^{j\frac{2\pi f_h}{RATE}}z^{-1}-1)(k_h e^{-j\frac{2\pi f_h}{RATE}}z^{-1}-1)} \quad (8)$$

식 (8)에서 $f_z$ 는 2400, $f_1$, $f_2$, $f_3$는 각각 730, 1090, 2600이 된다. $k_1$, $k_2$, $k_3$은 각각 0.999, 0.99, 0.95로 사용했다. 또한 RATE 는 simulation하고자 하는 음성 신호의 샘플링 속도로서, 10 kHz를 사용하였다. 그리고 k는 0부터 2까지의 값을 갖는다. 즉 k의 값에 따라서 simulation하는 음성 신호의 과비음성의 정도가 바뀌게 되는데, 이론적으로 k=1인 경우는 zero가 unit circle 위에 생기므로, spectrum상에서 완전한 null이 되고, 가장 과비음성이 심한 음성이 될 것이다.

식 (8)에서 구한 H(z)의 여기신호로는 gain 1, pitch 200 Hz 의 impulse train을 가해주었다. 그런 후 얻은 simulation된 음성신호에 대해서 각각 차수 10, 40에 대한 LP 모델을 구한 후, 그 차이를 식 (7)에 의해서 계산하였다. 다음 그림 3, 그림 4는 각각 k가 0.0, 0.9 일 때의 식(8)의 스펙트럼과, LP 모델들로부터 얻은 스펙트럼 그림들이다.

이론적으로 k값이 0에서 1로 증가해서 zero가 unit circle에 접근할수록 식(7)에 의한 스펙트럼간의 차이는 증가하고, 다시 k가 1에서 2로 증가해서 zero가 unit circle에서 멀어지면 차이는 감소해야 한다. 다음 그림 5는 실제로 k값을 변화해가면서 식 (7)을 적용시킨 결과이다.

그림 5를 보면 전체적인 경향은 이론대로 나타남을 볼 수 있다. 또한 zero의 크기가 작은 0 - 0.6 구간과 그 이후 구간과의 비선형성을 관찰할 수 있다. 그런데, k=1 부근에서는 예상과는 정반대의 결과를 보여주고 있다. 이 현상은 본 알고리즘이 잘 적용이 되지 않는 전형적인 경우로서, /아/ 발음의 경우는 3차 포만트의 위치와, 과비음성 비음의 경우 spectral zero가 생기는 위치가 근접하기 때문에 생기는 현상이다. 즉 zero가 3차 포만트와 상쇄되어 버리는 현상이 일어나는 것이다. 이렇게 되면 실제 스펙트럼에서는 3차 포만트가 사라짐으로, 청자가 듣기에는 정상적인 /아/발음으로 들리지 않지만, 본 논문의 알고리즘이 적용된 zero는 없어져 버림으로 알고리즘 적용결과는 과비음성 비음이 적은 경우(그림 5에서 k=0.9)보다 비음이 적은 것 같은 결과가 구해진다. 이 현상은 본 논문에서 제시한 알고리즘을 실제 발음 /아/에 적용했을 때에, 결과의 정확도가 떨어지는 되는 원인이 된다. 다음 그림 6은 zero가 생성되는 위치를 1700 Hz로 변경하고 본 알고리즘을 적용시킨 결과이다. formant와 zero가 겹치지 않는 경우는 이론적으로는 잘 적용이 됨을 알 수 있다.

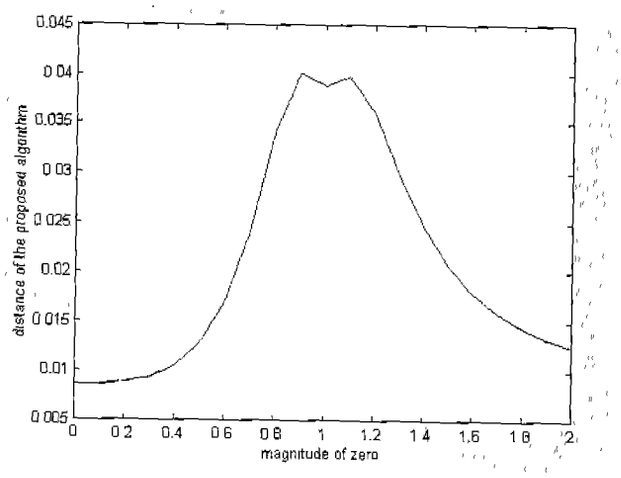


**그림 5.** Simulation된 /아/발음 음성 신호에 대한 알고리즘 적용 결과
**Fig. 5.** The result of algorithm applied to simulated speech signal /ʌ/

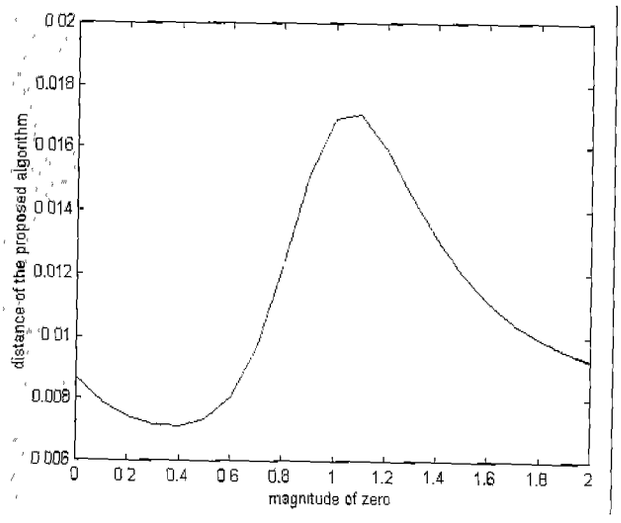


**그림 6.** Formant와 zero가 겹치지 않는 경우를 simulation한 결과
**Fig. 6.** The result of algorithm applied to simulated speech signal whose zero and formant are located far each other in spectrum

## 실제 음성 적용 결과

실제 실험에 사용된 음성 신호는 기존에 세브란스 안·이비인후과 병원의 음성 언어 치료실에서 임상적 목적으로 과거 3년간 기록을 해왔던 데이터들로서, 실제 VPI 환자들의 음성을 analog-to-digital converter(ADC)를 이용하여 PC에 저장한 데이터들이다. ADC의 샘플링 속도는 50 kHz이고, 해상도는 8 bit이었다. 실험에 사용된 음성신호의 발음은 한국어 /아/ 발음이었다. 전체 37개의 음성 재료들 중에서 제안된 알고리즘의 성능 평가를 위해서 Nasometer에 의한 nasalance가 기록되어 있는 음성 신호 재료 24개를 사용하였다. 환자들의 발음은 VPI 증상의 정중에 따라 nasalance가 비교적 고르게 분포되어 있었다. 음성 신호 재료를 발음한 환자들의 연령, 성별 구성은 표 1과 같다.

실제 음성 신호 처리 시에는 사람에 따라 각기 다른 세기로 발음된 음성 신호들에 대해 최대값을 1로 정규화했다. 또한 계산의 효율성과 일반적인 음성신호 처리 시의 데이터의 샘플링 속도가 10 kHz 정도인 것을 감안해서 차단 주파수 5000 Hz의 FIR 저대역 필터를 적용시킨 후에 4-다운 샘플링을 했다. 그래서 결국 샘플링 속도는 12.5 kHz로 만들었다.

LP 계수를 구하기 전에 음성신호는 pre-emphasis를 하였다. 음성신호 처리에 있어서 pre-emphasis가 필요한 이유는 음성이

표 1. 환자들의 성별, 연령 분포
**Table 1.** Distribution of patients related to gender and age

| 성별 / 연령 | 남성 | 여성 |
|---|---|---|
| 1-10 세 | 5 | 6 |
| 10-30 세 | 6 | 4 |
| 30 세 이상 | 1 | 2 |

성도를 통해서 발음이 된 후, 성도 특성 이외의 성분인, 입에서 공기 중으로 퍼져나갈 때 생기는 radiation effect와 여기신호의 음원 특성을 음성 신호에서 제거할 수 있기 때문이다[8,13].

본 실험에서 샘플링 속도는 12.5 kHz이므로, 일반적으로 사람이 보통의 발음을 할 때에는 pitch가 300 Hz를 넘지 않는다고 가정해서 한계 최고 차수는 40으로 제한하였다. 또한 LP 계수를 얻기 위한 음성 신호의 프레임의 크기는 375 샘플로 설정했다. 이는 샘플링 속도가 12.5 kHz일 때에 30 ms의 시간 간격에 해당하는 음성신호이다.

LP 모델간의 차이를 비교하기 위해서 구하게 되는 real LP-Cepstrum은 500개까지 구했다. 이론적으로는 식 (7)에서 n은 0부터 무한대까지 고려를 해야하지만, 실험 결과 500개 정도로도 비교적 만족할 만한 결과를 얻을 수 있었다. 그림 7은 nasalance가 낮은 발음(2 %)에 대한 10차와 34차 LP 모델의 스펙트럼의 비교 그림이며, 그림 8은 nasalance가 높은 발음(50%)에 대한 스펙트럼의 비교 그림이다.

그림 9는 24개의 음성 신호 재료들에 대해서 제안한 알고리즘을 적용시킨 결과와 Nasometer에 의해 측정된 nasalance와의 상관 관계를 보여주는 그림이다. 그림 9에서 y축은 음성 신호에 대한 nasalance이고 x축은 제안한 알고리즘으로 구한 식 (7)의 LP-Cepstrum 간의 거리이다. 알고리즘은 각각 낮은 차수는 10, 높은 차수로는 각각 34부터 최고 40차까지 적용을 시켰다. 적용 결과에 대한 Pearson's correlation은 각각 0.4575, 0.4842, 0.4988, 0.4545 이었다.

Nasalance와 본 알고리즘의 적용결과의 통계적 상관관계를 살펴보면, 본 알고리즘을 Nasometer의 대체 방법으로 사용하기에는 통계적 상관관계가 비교적 낮은 편임을 알 수 있다. 하지만 그림 9에서 볼 수 있듯이, 비록 variance는 크지만 nasalance와 본 알고리즘의 적용결과는 nasalance와 알고리즘 결과

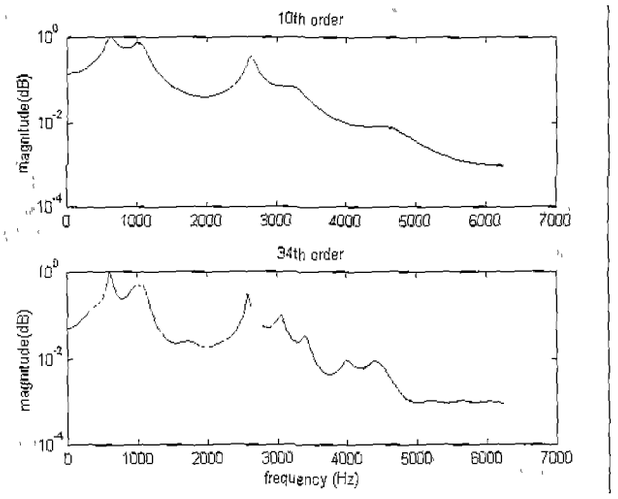


그림 7. Nasalance 2%의 /아/ 발음의 10차 및 34차 LP 스펙트럼의 비교

Fig. 7. Comparison between the order 10 and 34 LP spectrum of pronounce /ʌ/ with 2% nasalance

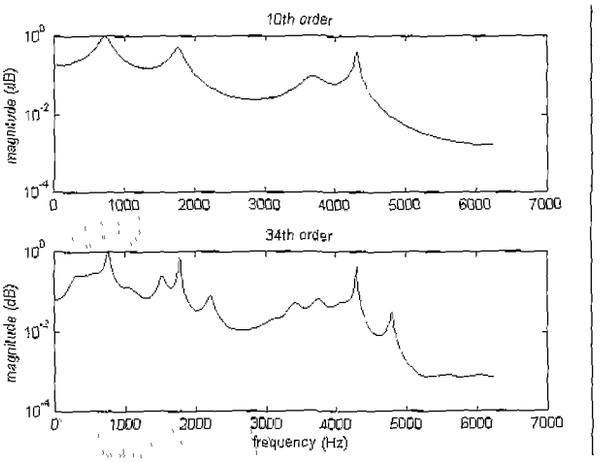


그림 8. Nasalance 50%의 /아/ 발음의 10차 및 34차 LP 스펙트럼의 비교

Fig. 8. Comparison between the order 10 and 34 LP spectrum of pronounce /ʌ/ with 50% nasalance

값이 전체적으로 비례하는 관계를 갖음을 확인할 수 있다. 그럼에도 correlation이 낮게 나온 이유보다는 nasalance 50-80 사이의 음성 데이터가 부족해서 통계적으로 전체적인 선형적 상관관계를 보장할 수 없기 때문일 것이다. 또한 variance가 커진 이유로는 과비음성 음성의 경우 생기는 스펙트럼 상의 zero의 위치가 /아/ 발음의 3차 포만트와 근접한다는 발음상의 특수성 때문일 것이다.

## 기존 방법과의 비교

순수하게 신호처리 방법을 이용해서 음성신호의 과비음성을 추정하는 방법으로는 Teager Operator를 이용하는 방법[7]이 알려져 있다. Teager Operator를 이용하는 방법(이후 Teager 방법)에 관해 간략히 정리하면 다음과 같다. 정상음과 비음을 각각 다음과 같이 모델링할 수 있다고 가정한다. 정상인의 발음은 학설에 따라서 3개 혹은 4개의 포만트로 이루어져 있다고 가정할 수 있으므로 다음과 같이 된다.

$$S_{Normal}(\omega) = \sum_{i=1}^{4} F_i(\omega)$$
$$S_{Normal}(\omega) : \textit{speech signal in frequency domain}$$
$$F_i(\omega) : \textit{ith Formant in frequency domain} \qquad (9)$$

반면 비음의 경우는 성도에 의한 포만트뿐 아니라 Antiformant(즉 zero), Nasal Formant들로 구성되어진다.

$$S_{Nasal}(\omega) = \sum_{i=1}^{4} F_i(\omega) - \sum_{k=1}^{4} AF_k(\omega) + \sum_{m=1}^{M} NF_m(\omega)$$
$$AF_k(\omega) : \textit{antiformant in frequency domain}$$
$$NF_m(\omega) : \textit{nasal Formant in frequency domain} \qquad (10)$$

또한, Teager operator는 discrete signal에 대해서 다음과 같이 정의된다.

$$\Psi_d[x(n)] = x^2(n) - x(n-1)x(n+1) \qquad (11)$$

Teager operator는 non-linear operator이므로 만일 $x(n)$이 2개의 신호의 합으로 구성되어진다면 Teager operation 결과는 다음과 같다.

$$\Psi_d[s(n)+g(n)] = \Psi_d[s(n)] + \Psi_d[g(n)] + \Psi_{cross}[s(n), g(n)] + \Psi_{cross}[g(n), s(n)]$$

여기서

$$\Psi_{cross}[g(n), s(n)] = g(n)s(n) + g(n+1)s(n-1) \qquad (12)$$

즉 신호들의 cross-term들이 생긴다.

이제 식 (9)의 정상음과 (10)의 과비음성 비음을, 1차 포만트 주파수를 중심 주파수로 하는 Band Pass Filter(BPF)로 필터링하면 1차 포만트만으로 이루어진 신호를 얻을 수 있을 것이다. 한편, 차단 주파수(cut off freq.)를 적절히 잘 선택을 해서 Low Pass Filter(LPF)를 적용하면, 정상음에 대해서는 1차 포만트만으로 이루어진 신호를 얻을 수 있는 반면, 비음에 대해서는 1차 포만트뿐만 아니라, 그 근처의 nasal formant, antiformant들도 포함된 신호를 얻을 수 있을 것이다. 그러므로, 비음에 대해서 LPF를 적용해준 음성 신호에 Teager operation을 적용시키면, antiformant, nasal formant등의 부분들에 의한 cross term들이 생겨날 것이다. 반면, 비음을 BPF를 적용해준 신호에 대해서 Teager operation하면 1차 포만트로만

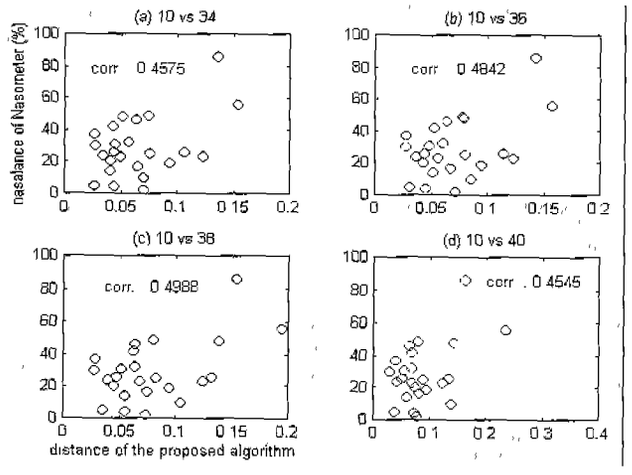


**그림 9.** 실제 음성에 적용된 본 논문의 알고리즘과 nasalance와의 상관관계
(a) 10차 vs 34차 : Pearson's correlation : 0.4575, (b) 10차 vs 36차 : Pearson's correlation : 0.4842, (c) 10차 vs 38차 : Pearson's correlation : 0.4988, (d) 10차 vs 40차 : Pearson's correlation : 0.4545

**Fig. 9.** Correlation between nasalance and proposed algorithm applied to speech signal
(a) *10 order vs 34 order* : *Pearson's correlation* : 0.4575, (b) 10 order vs 36 order : Pearson's correlation : 0.4842, (c) 10 order vs 38 order : Pearson's correlation : 0.4988, (d) 10 order vs 40 order : Pearson's correlation : 0.4545

구성되어져 있으므로 LPF를 적용해준 신호에 대한 연산 결과와는 다른 결과를 얻게 되고, 상호 상관계수(cross correlation coefficient)는 1보다 상당히 작은 값이 될 것이다. 같은 과정을 따르면 성상음의 경우는 BPF, LPF의 적용결과가 모두 1차 포만트로만 이루어져 있으므로 Teager operator의 적용 결과의 상호 상관 계수는 이상적으로 1이 될 것이다.

이 Teager 방법은 만일 비음에 의한 스펙트럼 상의 여러 특징이 LPF의 차단 주파수 바깥에서 나타나거나, 포만트들이 이동을 해서 제 2포만트와 제 1 포만트가 근접해서 나타나게 되면 부정확한 결과를 얻게 된다. 또한 Teager방법은 음성신호의 frame의 크기를 줄임으로서 스펙트럼 상에서 pitch의 영향을 없애기 때문에, pitch가 큰 경우는 알고리즘에 적용될 음성신호의 frame의 길이가 짧아져야 하므로 결과의 정확도가 떨어지는 또다른 요인이 된다. 다음 그림 10은 참고문헌 [7]에서 /아/ 발음에 대해 제시한대로, LPF의 차단 주파수를 1000 Hz로 해서 본 논문에서 제안한 방법을 적용했던 /아/ 발음 세료들에 Teager 방법을 적용시킨 결과이다. 예상대로 Teager 방법의 결과가 nasalance와 음의 상관관계를 보여주고 있다. Teager 방법 역시 variance가 상당히 큰 편인데, Pearson 상관 계수는 -0.442로, 본 논문에서 제시한 알고리즘의 결과보다는 다소 낮은 수치를 보여준다.

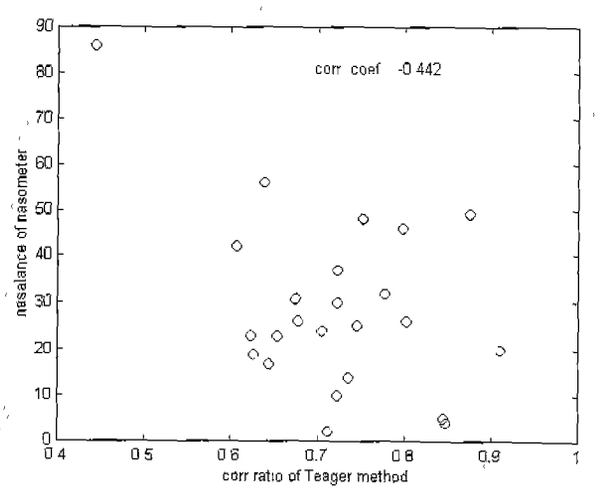


**그림 10.** 실제 음성에 적용된 Teager 방법과 nasalance와의 상관관계
**Fig. 10.** Correlation between nasalance and Teager method applied to speech signal

## 결 론

본 논문에서 제안된 알고리즘은 임상적으로 널리 사용되는 Nasometer의 지표인 nasalance와 어느 정도의 통계적인 상관관계를 가지고 있지만 정확도가 낮은 편이다. 결국 구강과 비강으로부터 나오는 음성 신호의 음압 레벨을 각각 직접 물리적으로 측정해서 비교해보는 Nasometer보다는 신뢰도가 떨어진다. 하지만, 기존에 개발된 방법인 Teager 방법에 비해서는 Nasometer와 조금 더 높은 상관관계를 보여주었다.

따라서 Nasometer의 대체를 기대할 수는 없겠지만, 본 알고리즘의 경우는 사운드 카드가 장착된 일반 PC와 마이크로폰만 있으면 소프트웨어를 이용해서 간단히 구현할 수 있다는 장점과 함께, Nasometer로 미처 측정하지 못하고 저장매체에 기록한 음성 데이터에 대해서도 적용이 가능하다는 장점이 있다. 게다가, 정상적 발음의 경우 nasalance는 32% 이하의 값을 갖는다는 사실[6]에서 알 수 있듯이, nasalance 30% 대와 그 이하의 경우는 사람이 거의 판단할 수 없을 정도로 과비음성 비음은 미약하다. 이 사실로부터 본 알고리즘은 비음의 정확한 정량화 방법으로는 다소 신뢰성이 떨어지나, 적절한 decision boundary를 설정해주면 그림 5, 6의 결과로부터 심한 과비음성 비음과 정상음에 대한 검출기 역할을 수행할 수 있을 것으로 기대된다. 그러므로 비음 측정이 아닌, 음성인식 등의 순수한 음성 신호 처리의 전처리 단계에서 사용될 수도 있을 것이다.

본 논문에서 제안한 알고리즘이 제대로 적용되지 않는 경우는, 비음의 특징 중 스펙트럼 상에서 zero없이 단순히 포만트들의 이동 현상만 있는 경우나, 혹은 zero와 포만트가 상쇄되어 버리는 경우로 추론할 수 있다. 본 연구는 nasalance가 기록된 단모음 음성 세료가 /아/ 발음뿐이었다는 한계를 가지고 있었다. 만일 과비음성 비음의 경우 zero의 생성 주파수와 포만트들

의 주파수가 겹치지 않는 발음 재료에 대해서 적용을 한다면, 그림 5, 6의 결과로부터 더 나은 결과를 얻을 수 있을 것으로 예상된다.

## 참 고 문 헌


1. J. Hirschberg, *"Velopharyngeal insufficiency"*, Folia Phoniatrica, Vol. 38, pp.221-276, 1986
2. M.L. Skolnick and E.R Cohn, *"Videofluoroscopic studies of speech in patients with cleft plate"*, Berlin:Springer-erlag, 1989
3. Y. Horii and J.Lang, *"Distributional analysis of an index of nasal coupling(HONC) in Simulated Hypernasal Speech"*, Cleft Palate J, Vol. 18, No. 4, pp279-285, 1981
4. Y. Horii, *"An accelermetric measure as a physical correlate of perceived hypernasality in speech"*, J. Speech Hear. Res, Vol. 26, pp.476-480, 1983
5. S.G Fletcher, L.E. Adams, and M.J. McCutcheon, *"Cleft plate speech assessment through oral-nasal acoustic measures"*, in Communicative Disorders Related to Cleft Lip and Plate, 3rd ed. Boston, Little and Brown, pp.246-2576, 1989
6. R.M Dalston, D.W Warren and E.T Dalston, *"Use of nasometry as a diagnostic tool for identifying patients with velopharyngeal impairment"*, Cleft Palate J., Vol. 28, pp.184-188, 1991
7. D. A. Cairns, J. H. L Hansen, and J. E. Riski, *"A noninvasive technique for detecting hypernasal speech using a nonlinear operator"*, IEEE Trans. on Biomedi cal Eng. Vol. 43, No.1, pp.35-45, 1996
8. J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen., Discrete-Time Processing of Speech Signals, Macmillan Publishing Company, 1993
9. D.R. Dickson, *"An acoustic study of nasality"*, J. Speech Hearing Res. vol.5 p.103-111 1962
10. L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signal, Prentice-Hall, 1978
11. G. E. Peterson "
*Parameters of vowel quality"*, J. Speech Hearing Res. Vol. 4, pp.10-29, 1961
12. R. J. Mammone, X. Zhang and R. P. Ramachandran, *"Robust speaker recognition"*, IEEE signal processing magazine, Vol. 13. No 5, pp.58-71, 1996