JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 19, No. 5 : 1243~1250 May. 2015

# 모의 음성 모델을 이용한 효과적인 구개인두부전증 환자 음성 인식

성미영[1] · 권택균[2] · 성명훈[2] · 김우일[1*]

# Effective Recognition of Velopharyngeal Insufficiency (VPI) Patient's Speech Using Simulated Speech Model

**Mee Young Sung[1] · Tack-Kyun Kwon[2] · Myung-Whun Sung[2] · Wooil Kim[1*]**

[1*]Department of Computer Science & Engineering, Incheon National University, Incheon 406-772, Korea
[2]Otorhinolaryngology, Seoul National University College of Medicine, Seoul 110-744, Korea

## 요 약

본 논문에서는 VPI 환자 음성을 정상인 음성으로 복원하기 위한 기술의 단계로서 효과적인 VPI 음성 인식 기술을 소개한다. 소량의 VPI 환자 음성을 모델 적응에 효과적으로 사용하기 위해 정상인의 모의 음성을 이용하여 화자 적응을 위한 사전 모델로 이용하는 기법을 제안한다. MLLR 기법을 이용한 화자 적응을 통해 평균 83.60%의 인식률을 보이고, 모의 음성 모델을 화자 적응의 사전 모델로 이용함으로써 평균 6.38%의 인식률 향상을 가져온다. 음소 인식 평가 결과는 제안한 화자 적응 방식이 대폭적인 음성 인식 성능 향상을 가져오는 것을 증명한다. 이러한 결과는 본 논문에서 제안하는 모의 음성 모델을 이용한 화자 적응 기법이 대량의 VPI 환자 음성을 취득하기 어려운 조건에서 보다 향상된 성능의 VPI 환자 음성 인식기를 구축하는데 효과적임을 입증한다.

## ABSTRACT

This paper presents an effective recognition method of VPI patient's speech for a VPI speech reconstruction system. Speaker adaptation technique is employed to improve VPI speech recognition. This paper proposes to use simulated speech for generating an initial model for speaker adaptation, in order to effectively utilize the small size of VPI speech for model adaptation. We obtain 83.60% in average word accuracy by applying MLLR for speaker adaptation. The proposed speaker adaptation method using simulated speech model brings 6.38% improvement in average accuracy. The experimental results demonstrate that the proposed speaker adaptation method is highly effective for developing recognition system of VPI speech which is not suitable for constructing large-size speech database.

**키워드** : 구개인두부전 (VPI), 음성 인식, 화자 적응, 모델 적응, 모의 음성

Key word : VPI (Velopharyngeal Insufficiency), Speech recognition, Speaker adaptation, Model adaptation, Simulated speech

**Received** 29 January 2015, **Revised** 10 February 2015, **Accepted** 26 February 2015

*** Corresponding Author** Wooil Kim(E-mail:wikim@inu.ac.kr, Tel:+82-32-835-8459)
Department of Computer Science and Engineering, Incheon National University, Incheon 406-772, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2015.19.5.1243 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright Ⓒ The Korea Institute of Information and Communication Engineering.

## Ⅰ. 서 론

구순구개열은 얼굴의 선천성 기형 중 빈도가 높은 장애의 하나로 알려져 있으며, 선천적으로 구순 (입술) 또는 구개 (입천장)가 갈라져서 구강과 비강이 연결된 상태를 말한다. 우리나라의 경우 약 700명의 신생아 중 1명 꼴로 발생하는 비교적 유병률이 높은 질환이다[1,2]. 정상인이 발성할 때 연구개가 비강과 구강을 차단시켜 비음을 막는 것과 달리, 구순구개열 환자는 경구개 또는 연구개가 갈라져 있거나 연구개가 선천적으로 짧기 때문에 성도로부터 나온 공기 흐름이 비강과 구강을 동시에 공명하게 됨으로써 발성 및 조음 장애가 발생한다.

이러한 증상을 구개인두부전증(Velopharyngeal Insufficiency, VPI)이라고 한다. VPI 환자는 발성 및 조음 장애로 인해 타인과의 소통에 어려움을 겪고, 저연령층의 경우 언어 및 지능 발달에 장애를 가져올 가능성이 높다. 따라서 VPI 환자가 정상 상태의 음성을 발성할 수 있도록 도와주는 의료적 시술이 필요하며, 공학 기술의 발달에 따라 정상인의 음성으로 복원 및 향상을 위한 연구가 요구된다.

본 논문은 VPI 환자의 음성 향상 및 복원 기술 개발을 위한 과정의 하나로서 VPI 환자의 음성을 자동으로 인식하는 연구 결과를 다룬다. 본 연구의 사전 연구로서, VPI 환자의 음성 처리 연구를 위해 공동 음성 데이터베이스 구축을 위한 발음 목록 설계 및 수집 환경 조성 등을 실시하였고, 수집된 일부 음성에 대해 비음도를 측정하였다[2]. 또한, VPI 환자의 발음과 이를 정상인으로부터 실험적으로 유사하게 발생시킨 모의 음성의 분석을 실시하였다[3]. 본 연구에서는 VPI 환자 음성을 자동으로 인식하기 위해 대표적인 음성 인식 기법을 채용하여 음성 인식기를 구축한다. 음향 모델로는 은닉 마르코프 모델 (Hidden Markov Model, HMM)을 사용하고 특징 추출 기법으로 MFCC (Mel-Frequency Cepstral Coefficients)를 사용한다.

VPI 환자의 음성은 정상인의 음성과 비교하여 음향적 왜곡 정도가 크기 때문에 정상인의 음성 데이터를 이용하여 구축된 음성 인식기는 많은 인식 오류를 발생시킨다. 따라서 음성 인식기의 음향 모델을 VPI 환자의 음성과 유사하게 변화시켜야하는데, 본 논문에서는 음성 인식 분야에서 일반적으로 사용되고 있는 화자 적응 (Speaker Adaptation) 기술을 적용하여 인식 성능을 평가한다. 또한 정상인이 실험 장치를 이용하여 발음한 모의 환자 발음을 화자 적응에 효과적으로 이용할 수 있는 가능성을 타진하고자 한다.

본 논문은 다음과 같이 구성된다. II 장에서는 본 연구에서 목표로 하는 VPI 음성 향상 시스템을 간략하게 소개한다. III 장에서는 본 논문에서 사용한 음성 데이터베이스의 수집 과정에 대해 설명하고, IV 장에서는 본 논문에서 채용한 화자 적응을 위한 음향 모델의 적응 기법에 관해 소개한다. V 장에서 실험 과정과 결과를 설명하고 VI 장에서 논문의 결론을 맺는다.

## Ⅱ. VPI 음성 향상 시스템

본 논문에서는 VPI 환자의 음성을 정상인 음성으로 향상 및 복원하는 시스템의 처리 단계의 하나로, VPI 환자의 음성을 자동으로 인식하는 연구에 관한 내용을 소개한다. 본 연구의 VPI 음성 향상 시스템에서는 자동 인식된 결과로부터 음소 정보를 취득하고, 취득된 음소 정보를 이용하여 음질 향상 및 복원 처리를 실시한다.

따라서 입력된 VPI 환자 음성으로부터 정확한 음소 정보를 취득하는 것이 음성 복원 성능을 결정하는 중요한 요소 기술 중 하나이다. 본 논문에서는 VPI 음성 인식의 초기 연구로서 PBW (Phoneme Balanced Words) 452[4] 데이터베이스와 같이 제한된 어휘 안에서 VPI 음성을 인식하는 연구를 수행하였다. 그림 1은 본 논문에서 목표로 하는 VPI 음성 향상 시스템의 블록다이어그램을 나타낸다.

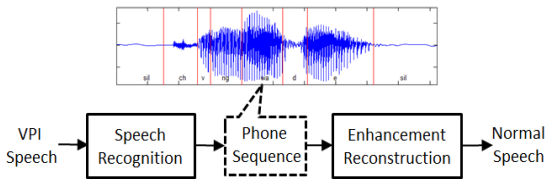


**그림 1.** 구개인두부전증 환자 음성 향상 및 복원 시스템의 블록다이어그램
**Fig. 1** Block diagram of enhancement/reconstruction system for VPI patient's speech

## Ⅲ. 음성 데이터베이스 수집

본 절에서는 본 연구에서 사용한 음성 데이터베이스의 설계 및 수집 과정을 설명한다[3].

### 3.1. 단어 목록 선정

본 연구에서는 VPI 환자 음성 인식 실험 및 연구를 위해 음성 인식 분야에서 단어 인식 시스템 구축을 위해 많은 연구자들이 사용하고 있는 한국어 PBW452 데이터베이스[4]의 단어 목록을 사용하였다. PBW452 단어 목록에서 VPI 환자의 발음 오류의 양상이 주로 나타나는 50개의 단어를 언어치료사가 선정하여 발음 목록으로 사용하였다. 표 1은 본 연구에서 발음 목록으로 사용한 50개의 단어 리스트를 나타낸다.

**표 1.** PBW452 목록으로부터 선정된 50개의 단어 목록[3]
**Table. 1** 50-word list selected from PBW452 list[3]

| |
|---|
| 컴퓨터, 계획하고, 뒷받침해, 쾌적한, 그쪽으로, |
| 과부가, 분석됐다, 감수성이, 변증법적, 부딪혀, |
| 톡톡히, 빼앗기고, 깨끗이, 세탁비누, 수필집, |
| 바쁘게, 보궐선거, 뜻밖에, 두번째, 깨닫게, |
| 표적수사, 펼치고, 코방귀를, 국세청, 집계됐다, |
| 뒤쪽에, 뱃속에, 재검토, 끌어안고, 불빛이, |
| 솜씨가, 붙잡고, 종소리가, 수법을, 교통사고, |
| 아카데미, 스포츠, 생태계, 손쉽게, 백화점, |
| 수수께끼, 첫번째, 총체적, 편집국장, 플라스틱, |
| 뽑았다, 깨닫고, 재조정, 덮었다, 불법복제 |

### 3.2. 수집 대상

음성 녹음에 협조가 잘 되고, 발화 목록에 따른 발음이 적절히 이루어지도록 만 10세 이상의 VPI 환자와 정상 발음을 가진 성인을 대상으로 하였다. 모집과정에서 정상 모의 환자와 VPI 환자의 녹음 의지를 확인한 후 피험자 동의서를 받았다. 대부분의 발화자들은 구개구순열 수술 후 언어 치료를 위해 외래에 정기적으로 내원하는 환자로 구성되었다.

### 3.3. 수집 방법

녹음은 주변 환경 소음을 최대한 피하기 위해 외래 진료 후 언어치료실 한 곳을 지정하여 시행하였다. 녹음 과정에는 음성 언어치료사, 의공학과 연구원, 이비인후과 의사가 참여하였고, 언어치료사의 주도하에 발화자가 긴장하지 않도록 최대한 편한 분위기를 조성하였다. 입으로부터 약 40 cm 떨어진 위치에 고감도 마이크를 설치하여 녹음을 실시하였다. 마이크로부터 나오는 음성신호를 다목적 USB 녹음장치인 U46XL (SuESI Audiotechnik GmbH, Leonberg, Germany)을 이용하여 주파수 44.1 kHz, 양자화 비트수 16 비트로 디지털화하고 Cubase LE5 소프트웨어를 이용하여 녹음 파일을 취득하였다.

### 3.4. 모의 음성 수집

정상인으로부터 실험적으로 VPI 환자의 발음을 유발하기 위해 1mm 내경을 가지는 고무관 (Nelaton Catheter)을 사용하였다 [2]. 넬라톤 카테터를 양측 비강을 통해 넣고 긴장도가 없는 상태에서의 위치를 지혈겸자 (Hemostatic Clamp)로 표시해 놓고, 통증을 유발하지 않는 선에서 최대의 긴장도가 생성되는 위치를 표시하였다. 카테터가 최대 긴장도의 위치에 있을 때를 VPI 모의 환자 중증 (Severe) 상태로 정의하였고, 긴장도가 없는 위치와 최대 긴장도 위치의 중간에 있을 때 VPI 발음이 녹음된 것을 경도 (Mild) 상태로 정의하였다.

## Ⅳ. 음향 모델 적응 기법

VPI 환자의 음성은 정상인의 음성과 비교하여 음향적 특성이 많이 다르다. 사전 연구에서 VPI 환자 음성과 정상인 음성의 비교 및 분석을 통해 VPI 환자 음성모음의 경우 제 1, 제 2 포먼트 주파수의 위치가 정상인의 음성과 크게 다른 것을 관찰하였다[3]. 특히 VPI 환자 음성의 제 2 포먼트의 위치가 대폭적으로 낮아지는 것이 분석되었으며, /ㅓ/, /ㅗ/, /ㅡ/ 모음의 경우 정상인의 /ㅗ/ 발음의 제 1, 제 2 포먼트 위치가 유사하게 나타난다. 실제 듣기 평가 실험에서도 /ㅓ/, /ㅡ/ 발음이 /ㅗ/ 발음으로 잘못 인식되어 높은 오인식률을 나타냈다. 또한 VPI 환자의 /ㅣ/, /ㅐ/ 발음의 경우 제 2 포먼트의 위치가 대폭적으로 하락하는데, /ㅣ/ 발음은 오인식이 매우 높은 대표적인 모음이다.

이와 같이 상이한 음향 특성을 갖는 VPI 환자의 음성을 컴퓨터가 자동으로 인식하기 위해서는 VPI 환자의 음성을 대량으로 취득하여 음성 인식기의 음향 모델 훈련

에 사용해야 하지만, 현실적으로 대량의 VPI 환자의 음성을 수집하는 것은 쉽지 않은 작업이다. 따라서 본 논문에서는 음성 인식 분야에서 일반적으로 활용되는 모델 적응 기법을 이용하여 VPI 환자 발음을 효과적으로 반영할 수 있는 음향 모델을 생성한다. 또한, 상대적으로 취득이 용이한 정상인의 VPI 환자 모의 음성 데이터를 모델 훈련에 효과적으로 이용할 수 있는 방안을 제시한다.

음향 모델 적응 기법에서는 음성 인식 시스템의 음향 모델이 훈련된 환경과 실제 인식 시스템이 적용되는 테스트 환경의 음향적 조건이 동일할 때 최고의 성능을 가지는 것을 가정한다. 여기에서 음향적 조건은 음성이 입력되는 환경을 말하는데, 화자 (Speaker), 녹음 장치, 배경 잡음 등의 요소가 포함된다. 따라서 실제 환경과 동일한 음향적 특성을 가지는 음성 데이터를 취득하고, 이를 적응 과정에 이용하여 적응된 모델 파라미터를 예측하는데 사용한다. 모델 적응에 사용하는 데이터는 실제 환경과 유사한 환경에서 모델 적응을 위해 별도로 수집을 하거나 실제 입력 음성을 사용한다. 모델 적응 기술은 소량의 음성 데이터로부터 신뢰성이 있는 모델 파라미터를 효과적으로 예측하는 것을 목표로 하며 대표적인 적응 기술로 MAP (Maximum A Posteriori) 기반 적응 기법[5], MLLR (Maximum Likelihood Linear Regression) 기반 적응 기법[6] 등이 있다.

### 4.1. MAP 기반 적응 기법

MAP 예측 기반의 적응 기법[5]에서는 사후 확률 (A Posterior Probability), 즉 모델 적응에 사용되는 음성 데이터 $x$가 주어졌을 때 해당 모델 $\lambda$의 확률 값을 최대로 하는 모델 파라미터를 예측하며 다음과 같이 나타낼 수 있다.

$$\begin{aligned}\lambda_{MAP} &= \arg\max_{\lambda} f(\lambda|x) \\ &= \arg\max_{\lambda} f(\lambda) f(x|\lambda)\end{aligned} \quad (1)$$

MAP 기반 적응 기법에서는 모델에 대한 사전 확률 (Prior Probability) 정보와 반복적인 EM (Expectation Maximization) 알고리즘을 적용함으로써 "불완전한 (Incomplete)" 데이터로부터 모델 파라미터를 예측할 수 있다. 각 모델이 가우시안 혼합 모델로 표현된다고 가정하면, 각 가우시안 요소의 평균 (Mean) 파라미터는 MAP 기반 적응 기법에 따라 다음의 식으로 갱신된다.

$$\hat{\mu_k} = \frac{\tau_k \mu_k + \sum_{t=1}^{T} \zeta_t(k) x(t)}{\tau_k + \sum_{t=1}^{T} \zeta_t(k)} \quad (2)$$

식 (2)에서 $\mu_k$, $\tau_k$, $\zeta_k(t)$는 각각 적응 이전의 초기 모델의 $k$번째 가우시안 요소의 평균 벡터, 적응 계수, 입력 데이터 $x(t)$가 주어졌을 때 $k$번째 가우시안 요소가 발생할 확률을 나타낸다. 따라서 적응 계수 $\tau_k$가 크면 사전 모델에 의존적인 적응 파라미터가 얻어지고, 반대로 적응 계수가 작은 값이면 관찰 값 즉 적응 데이터에 의존적인 파라미터 값을 얻을 수 있다. 적응 계수가 0과 같은 값이면 적응 데이터만을 이용한 최대 우도 (Maximum Likelihood, ML) 기반의 예측 기법과 같아진다.

### 4.2. MLLR 기반 적응 기법

또 다른 대표적인 적응 기법인 MLLR 기반의 예측 기법[6]에서는 모델 파라미터의 "변환 (Transformation)"에 의해 새로운 파라미터를 얻는다. MLLR 기법은 MAP 기법에 비해 비교적 소량의 데이터로부터 효과적인 모델 적응 성능을 얻을 수 있는 것으로 알려져 있다. 적응 데이터로부터 선형 Regression 변환을 위한 행렬을 추정하고, 이를 이용하여 다음과 같은 식으로 평균 벡터를 갱신할 수 있다.

$$\hat{\mu_k} = W\mu_k \quad (3)$$

위 식에서 $\mu_k$는 바이어스 요소를 포함한 확장된 평균 벡터 (n+1차원의 벡터)이고, 파리미터 변환 행렬 $W$는 n×(n+1) 크기를 갖는 행렬이다. 행렬 $W$는 적응 데이터의 우도 (Likelihood)를 최대화하는 EM 알고리즘을 통해서 얻을 수 있다.

## Ⅴ. 실험 및 결과

### 5.1. 실험 환경 및 음성 인식 시스템

본 논문에서는 VPI 음성의 인식 성능 평가를 위해 음성 인식 분야에서 널리 사용되는 HTK[7]와 PBW452 음성

데이터베이스를 이용하여 기본 음성 인식 시스템을 구축하였다. 원래의 PBW452 데이터베이스는 총 남녀 71명이 2번씩 발음한 452단어의 발음으로 이루어져 있으나, 본 논문에서는 연구용으로 배포된 버전의 일부를 사용하였다. 본 논문에서는 남자 8명이 2번씩 발음한 452단어 총 7,232 발음을 이용하여 452개의 단어를 인식하는 기본 음성 인식기를 구축하였다. 단어의 시작과 끝의 묵음 구간 모델을 포함하여 총 42개의 문맥 독립형 음소 모델을 사용하였다. 각 HMM은 각 문맥 독립형 음소 모델을 나타내며, 3개의 상태(State)로 구성되고 각 상태는 8개의 요소로 구성된 가우시안 혼합 모델을 출력 확률 함수로 갖는다. ETSI 표준 방식의 MFCC 특징 추출 기법[8]을 채용하여 13차 Static 특징(c0~c12)과 미분 계수를 포함한 총 39차원의 특징 벡터를 추출하며, 특징 추출 단계에 켑스트럼 평균 정규화 (Cepstral Mean Normalization, CMN) 기법을 적용하였다. 기본 인식 시스템의 성능 평가를 위해 훈련 세트와 중복되지 않은 PBW452 평가용 버전의 남성 화자 5명의 총 2,260 발음을 사용하였으며, 깨끗한 환경의 음성 데이터에 대해 98.89%의 단어 인식률을 갖는다.

### 5.2. 화자 적응 인식 테스트

표 2는 앞에서 설명한 기본 인식기를 이용하여 각 화자의 음성에 대해 음성 인식 평가를 실시한 결과이다. 참고로 각 정상인의 정상 발음은 PBW452 평가용 버전의 테스트 세트의 인식 성능과 유사한 95% 이상의 인식률을 보였으며, 이와 같은 결과는 본 실험에 참여한 정상인의 발음은 표준 발음에 가깝고 녹음 환경 역시 성능 평가에 큰 영향을 주지 않는다는 것을 의미한다. 정상인의 모의 발음 (모의환자 1-3, S1-3) 모두 아무런 처리를 하지 않았을 때는 약 20% 대의 낮은 인식률을 보였으며, 환자 1과 2 (P1, P2) 모두 2.68%와 39.33%로 낮은 인식률을 보였다.

화자 적응은 앞 절에서 설명한 MLLR 기반의 모델 적응 기법과 MAP 기반의 기법을 각각 적용하여 인식 성능을 평가하였다. 소량의 음성 데이터 세트를 효과적으로 화자 적응 실험에 이용하기 위해 각 화자의 데이터베이스를 모델 적응용과 테스트용 데이터들이 겹치지 않도록 조합된 3종류의 세트로 구성하였다. 즉, 3번 반복으로 구성된 각 화자의 데이터에서 1번째 발음을 테스트할 때에는 2번째, 3번째 발음을 화자 적응에 사용하고, 2번째 발음을 테스트 할 때에는 1번째, 3번째 발음을 화자 적응에 사용하였다. 마찬가지로 3번째 발음을 테스트할 때에는 1번째, 2번째 발음을 화자 적응에 사용하였다. 화자 적응 실험 결과 MLLR, MAP 적응 기법 모두 대폭적인 인식 성능 향상이 있었다. 모든 화자에 대해 MLLR 기법이 MAP 기법에 비해 높은 성능 향상을 보였으며, 모의환자 1 (S1, 즉, 정상인 1의 모의발음)을 제외하고 MLLR 기법이 70~80% 인식률의 대폭적인 성능 향상을 가져오는 것을 확인할 수 있었다. 이러한 결과는 VPI 환자 음성의 정확한 인식을 위해서는 해당 환자 음성을 이용한 화자 적응이 필수적임을 의미하며, 실제 VPI 환자의 조건과 같이 대량의 데이터 수집에 한계가 있을 경우에는 MLLR 모델 적응 기법이 화자 적응에 효과적임을 입증한다.

**표 2.** 화자 적응 인식 실험 결과 (단어 인식률, %)
**Table. 2** Speech recognition result of speaker adaptation (word accuracy, %)

| Speaker | No Processing | Speaker Adaptation | |
|---|---|---|---|
| | | MLLR | MAP |
| S1 | 22.67 | 54.00 | 52.67 |
| S2 | 20.41 | 84.69 | 76.53 |
| S3 | 21.33 | 74.67 | 62.67 |
| P1 | 2.68 | 79.87 | 64.43 |
| P2 | 39.33 | 87.33 | 84.00 |

### 5.3. 모의 발음 모델을 이용한 화자 적응 실험

표 3은 정상인의 모의 음성을 이용하여 모의 음성 모델을 생성하여 이를 기반으로 화자 적응을 실시한 후, 환자 1과 2 (P1, P2) 음성에 대한 인식 성능 평가 결과이다. 즉, PBW452 데이터를 이용하여 구축된 기본 인식기를 기반으로 모의 음성을 이용하여 모델 적응을 실시함으로써 화자 적응을 위한 사전 모델을 생성하였다. 모의 음성 모델을 생성하기 위한 모델 적응 과정에는 정상인의 모의 발음 전체(즉, 모의환자 1-3 전체, 즉 S1-3)를 사용하였다. MAP 기법 또는 MLLR을 각각 적용하여 모의 음성 모델을 만들고, 5.2절에 설명한 화자 적응 과정과 동일하게 MLLR 기법을 적용하여 화자 적응을 실시하였다. 표 3의 실험 결과로부터, MLLR을 이용한 모의 음성 모델 생성은 인식 성능에 전혀 영향이 없고, MAP를 이용하여 생성한 모의 음성 모델은 인식 성능 향상에 도움을 준 것을 알 수 있다.

특히 환자 1(P1)의 경우 MAP 적응 기법을 적용하여 모의 음성 모델을 화자 적응의 사전 모델로 이용함으로써 79.87%에서 91.28%로 대폭적인 단어 인식률의 향상을 보였다. 이와 같은 결과는 데이터 확보에 한계를 가지는 VPI 환자 음성 인식 시스템 구축을 위해, 상대적으로 확보가 용이한 정상인의 모의 음성을 이용하여 음향 모델을 구축함으로써 인식 성능을 향상 시킬 수 있음을 의미한다.

**표 3.** 모의 환자 음성 모델을 이용한 화자 적응 인식 결과 (단어 인식률, %)
**Table. 3** Speech Recognition result of speaker adaptation using simulation speech model (word accuracy, %)

| | | P1 | P2 |
|---|---|---|---|
| MLLR (Result of Table 2) | | 79.87 | 87.33 |
| Improved Initial Model | **MAP-MLLR** | **91.28** | **88.67** |
| | MLLR-MLLR | 79.87 | 87.33 |

### 5.4. 음소 인식 성능 평가

그림 2와 3은 환자 1과 2의 음소 인식 성능 평가 결과이다. 본 실험에서는 음소 인식 평가를 위해 음소 인식기를 별도로 구성하지 않고, 앞에서 설명한 음성인식 평가에 사용한 단어 인식 시스템을 사용하여 음소 인식 성능을 평가하였다.

즉, 모음 /ㅏ/에 대한 음소 인식 성능 평가를 위해서 모음 /ㅏ/가 들어간 간 단어의 음소 /ㅏ/를 다른 모음들로 치환하여 총 7개의 단어 후보 리스트로 생성하여 인식 평가를 실시하였다. 예를 들어 VPI 환자 음성 녹음에 사용한 50단어 목록 중 “계획하고”라는 단어의 경우, 모음 /ㅏ/의 평가를 위해 7개의 단어 후보, 자음 /ㄱ/의 평가를 위해서는 18개의 단어 후보 리스트를 각각 생성하였다. 음소별 발음을 별도로 수집하여 인식기를 구축하고 평가하는 방법도 있으나 각 음소의 발음의 길이가 지극히 짧고, 독립된 음소의 발음은 자연스러운 발음을 유도하기가 어렵기 때문에 현실적인 방법이 되지 못한다. 또한, 음소 네트워크를 구성하여 무제한 음소 인식기를 구축하는 방법도 있으나 발음 환경에 따라 대량의 삽입 (Insertion) 오류가 발생하기 때문에 정확한 대체 (Substitution) 오류를 측정하기 어려운 단점이 있다.

**표 4.** 음소 인식 성능 평가를 위한 후보 단어 리스트의 예
**Table. 4** An example of word candidates for phone recognition performance

| | |
|---|---|
| 계획**하**고<br>모음 / ㅏ / | 계획**하**고, 계획**허**고, 계획**호**고, 계획**후**고, 계획**흐**고, 계획**히**고, 계획**해**고 |
| **계**획하고<br>자음 / ㄱ / | **계**획하고, **녜**획하고, **뎨**획하고, **례**획하고, **몌**획하고, **볘**획하고, **셰**획하고, **졔**획하고, **쳬**획하고, **켸**획하고, **톄**획하고, **폐**획하고, **혜**획하고, **꼐**획하고, **뗴**획하고, **뼤**획하고, **쎄**획하고, **쪠**획하고 |

그림 2와 3은 표 2의 아무런 처리도 적용하지 않은 조건 (No Processing)과 표 3의 모의 음성 모델을 사용한 화자 적응 실험 결과 (MAP-MLLR)를 모음과 자음에 대해 비교한 것이다. 즉, Baseline 그래프는 No Processing을, Adaptation 그래프는 MAP-MLLR 성능을 나타낸다. 각 그림에서 실선과 점선으로 나타낸 평행선은 각 실험에서의 모음 또는 자음 평균 인식률을 나타낸다. 음성 인식 성능 평가와 동일한 실험 데이터를 사용했지만, 앞에서 설명한 음소 인식 평가의 조건이기 때문에 절대적인 인식 성능 수치에서는 단어 인식률과 상당히 다른 값을 나타낸다는 측면을 고려해야 한다. 하지만, 앞의 단어 인식 성능 결과의 경향과 유사하게 화자 적응을 통해 모음 및 자음 모두 대폭적인 인식 성능 향상이 있음을 확인할 수 있다.

비록 2명의 환자 데이터이기 때문에 일반적인 성능 규칙을 발견하기는 힘들지만, 모음에서는 환자 1과 2에 공통적으로 / ㅣ / 발음의 대폭적인 성능을 관찰할 수 있다. 이와 같은 현상은 선행 연구의 결과[3]에서 그 원인을 찾을 수 있다. 즉, / ㅣ / 발음의 경우 VPI 환자의 발음은 청취자 듣기 평가 실험에서 대부분 /ㅡ/ 발음으로 오인식되어 5%이하의 인식률을 나타내었다. 포먼트 분석 실험 결과 / ㅣ / 발음의 경우 VPI 환자와 정상인의 모의 발음에서 공통적으로 제 2 포먼트의 주파수가 정상인의 발음에 비해 급격하게 낮아지게 되는 것을 확인할 수 있는데, 그 하락 폭이 다른 모음에 비해 크다. 따라서 본 논문에서 제안한 모의 음성 모델을 이용한 화자 적응 기법이 / ㅣ / 발음을 보다 효과적으로 모델링함으로써 인식 성능을 향상시켰음을 보여주는 결과이다. 자음에 대해서는 /ㄴ/, /ㅂ/, /ㅅ/, /ㅊ/, /ㅋ/, /ㅍ/, /ㄲ/, /ㄸ/, /ㅃ/, /ㅉ/ 등의 발음에 대해 환자 1과 2에 공통적인 성능 향상을 관찰할 수 있었다. 이러한 자음들은 선행 연

구의 실제 듣기 평가 실험에서 VPI 환자 음성 중 오인식이 많이 발생한 음소이며, 본 논문에서 제안한 화자 적응 기법이 VPI 음성의 자음에 대해 인식 성능을 향상시키는 효과를 가지는 것을 나타낸다.

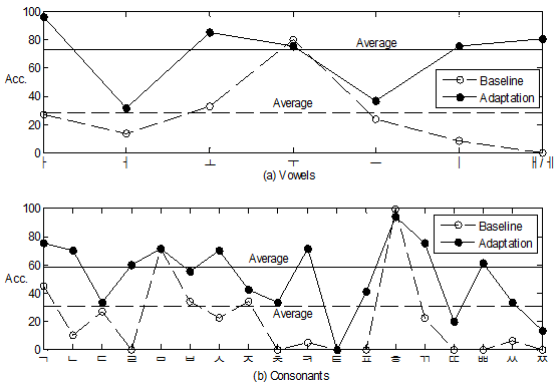


**그림 2.** 환자 1의 음소 인식 결과 (단어 인식률, %)
**Fig. 2** Phone recognition result of patient 1 (word accuracy, %)

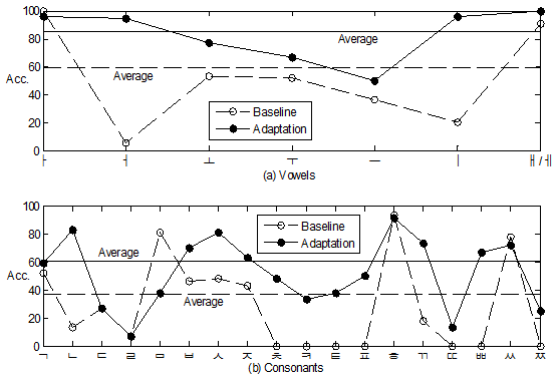


**그림 3.** 환자 2의 음소 인식 결과 (단어 인식률, %)
**Fig. 3** Phone recognition result of patient 2 (word accuracy, %)

## Ⅵ. 결 론

본 논문에서는 VPI 환자 음성을 정상인 음성으로 복원하기 위한 기술의 단계로서 효과적인 VPI 음성 인식 기술을 소개하였다. VPI 음성 인식을 위해 HMM 기반의 음성 인식기를 구축하고, 향상된 음성 인식을 위해 VPI 환자 음성에 대한 화자 적응 기법을 적용하였다. 소량의 VPI 환자 음성을 모델 적응에 효과적으로 사용하기 위해 정상인의 모의 음성을 이용하여 화자 적응을 위한 사전 모델을 생성하였다. 화자 적응 실험을 실시한 결과로 MLLR 기법이 MAP 기법에 비해 높은 인식 성능을 보였고, MLLR 기법을 이용한 화자 적응을 통해 평균 83.60%의 인식률을 보였다. 모의 음성 모델을 화자 적응의 사전 모델로 이용함으로써 평균 6.38%의 인식률 향상을 가져왔다. 음소 인식 평가에서도 제안한 화자 적응 방식이 대폭적인 음성 인식 성능 향상을 가져오는 것을 확인할 수 있었다. 이러한 결과는 본 논문에서 제안하는 모의 음성 모델을 이용한 화자 적응 기법이 대량의 VPI 환자 음성을 취득하기 어려운 조건에서 보다 향상된 성능의 VPI 환자 음성 인식기를 구축하는데 효과적임을 입증한다.


## 감사의 글

이 논문은 미래창조과학부 공공복지안전연구사업(No. 2013-2244) 지원에 의하여 연구되었음.


## REFERENCES


[1] S. G. Fletcher, “Theory and instrumentation for quantitative measurement of nasality,” *Cleft Palate Journal,* vol. 7, pp. 601–609, 1970.

[2] J.-E. Lee, et al., “Research on Construction of the Korean Speech Corpus in Patient with Velopharyngeal Insufficiency,” *Korean Journal of Otorhinolaryngol - Head & Neck Surgery*, vol.55, no.8, pp.498-507, 2012 .

[3] M. Sung, et al., “Analysis on Vowel and Consonants Sounds of Patient’s Speech with Velopharyngeal Insufficiency (VPI) and Simulated Speech,” *Journal of Korea Institute of Information and Communication Engineering,* vol.18, no.7, pp.1740-1748, July, 2014.

[4] B.-W. Kim, et al., “A Study on the Design and the Construction of a Korean Speech DB for Common Use,” *Journal of the Acoustic Society of Korea*, vol.16, no.4, pp.35-41, 1997.

[5] J. L. Gauvain and C. H. Lee, “Maximum a Posteriori

Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Proc.*, vol.2, no.2, pp.291-298, 1994.

[6] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMMs," *Computer Speech and Language*, 9, pp.171-185, 1995.

[7] http://htk.eng.cam.ac.uk/

[8] ETSI standard document, ETSI ES 201 108 v1.1.2 (2000-04), Feb. 2000.

**성미영(Mee Young Sung)**

1990년 프랑스 INSA de Lyon 컴퓨터공학 박사
1990년 ~ 1993년 한국전자통신연구소 선임연구원
1993년 ~ 현재 인천대학교 컴퓨터공학부 교수
2001년 ~ 2002년 미국 카네기 멜론 대학교 교환교수
2008년 ~ 2009년 미국 UC 버클리 대학교 교환교수
※관심분야 : 멀티미디어, 가상현실, 햅틱스, 음성인식

**권택균(Tack-Kyun Kwon)**

2006년 서울대학교 의과대학 의학박사
2003년 ~ 2004년 미국 피츠버그 의과대학 Voice Clinic Fellow
2012년 ~ 2013년 미국 샌디에고 대학 Clinical Research 석사과정
2012년 ~ 현재 서울대학교 의과대학 이비인후과학 부교수
※관심분야 : 음성수술, 음성질환진단, 음성분석, 임상시험 및 연구

**성명훈(Myung-Whun Sung)**

1991년 서울대학교 의과대학 의학박사
1990년 ~ 1999년 서울대학교 의과대학 이비인후과 조교수
1993년 ~ 1995년 미국 피츠버그 의과대학 Research Fellow
1999년 ~ 2004년 서울대학교 의과대학 이비인후과 부교수
2004년 ~ 현재 서울대학교 의과대학 이비인후과 교수
※관심분야 : 두경부 종양, 음성수술, 음성질환, 임상시험 및 연구

**김우일(Wooil Kim)**

2003년 고려대학교 전자공학과 공학박사
2004년 ~ 2005년 미국 카네기 멜론 대학교 박사후 연구원
2005년 ~ 2012년 미국 텍사스 주립대 (University of Texas at Dallas) 연구원 및 연구교수
2012년 ~ 현재 인천대학교 컴퓨터공학부 조교수
※관심분야 : 신호처리, 패턴인식, 음성인식, 휴먼 컴퓨터 인터페이스