ISSN 2005-8063
2017. 3. 31.
Vol.9 No.1
pp. 41-47

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2017.9.1.041

기본주파수와 성도길이의 상관관계를 이용한 HTS 음성합성기에서의 목소리 변환*

Voice transformation for HTS using correlation between fundamental frequency and vocal tract length

유 효 근 · 김 영 관** · 서 영 주 · 김 회 린

Yoo, Hyogeun · Kim, Younggwan · Suh, Youngjoo · Kim, Hoirin

Abstract

The main advantage of the statistical parametric speech synthesis is its flexibility in changing voice characteristics. A personalized text-to-speech(TTS) system can be implemented by combining a speech synthesis system and a voice transformation system, and it is widely used in many application areas. It is known that the fundamental frequency and the spectral envelope of speech signal can be independently modified to convert the voice characteristics. Also it is important to maintain naturalness of the transformed speech. In this paper, a speech synthesis system based on Hidden Markov Model(HMM-based speech synthesis, HTS) using the STRAIGHT vocoder is constructed and voice transformation is conducted by modifying the fundamental frequency and spectral envelope. The fundamental frequency is transformed in a scaling method, and the spectral envelope is transformed through frequency warping method to control the speaker's vocal tract length. In particular, this study proposes a voice transformation method using the correlation between fundamental frequency and vocal tract length. Subjective evaluations were conducted to assess preference and mean opinion scores(MOS) for naturalness of synthetic speech. Experimental results showed that the proposed voice transformation method achieved higher preference than baseline systems while maintaining the naturalness of the speech quality.

Keywords: voice transformation, HMM-based speech synthesis, STRAIGHT vocoder, fundamental frequency, vocal tract length

1. 서론

텍스트 음성 변환(text-to-speech, TTS)시스템이란 입력으로 들어오는 텍스트 또는 일련의 문자열을 자연스러운 음성을 내는 음성 신호로 바꿔주는 기술을 말한다. 일반적인 TTS 시스템은 앞단의 텍스트 분석부와 뒷단의 음성 합성부로 구성된다 (Tokuda *et al.*, 2013). 텍스트 분석부에는 일반적으로 문장 분할(sentence segmentation), 단어 세분화(word segmentation), 텍스트 정규화(text normalization), 품사 태깅(part-of-speech tagging), 서기소음소 변환(grapheme-to-phoneme conversion)과 같은 다수의 자연어 처리 기법(natural language processing, NLP)이 활용된다. 즉 텍스트 분석 부분은 일련의 단어열을 입력받아 다양한 언어

* 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구입니다.(지원번호 : 14-814-09-012, 사용자 디지털 감성 DNA에 기반한 디지털생명체 기술 개발)

** 한국과학기술원, cleanthink@kaist.ac.kr, 교신저자

Received 31 January 2017; Revised 15 March 2017; Accepted 21 March 2017

문맥(linguistic context)을 갖는 음소열을 출력하는 역할을 한다. 다음으로 이어지는 음성 합성부는 앞서 출력된 언어 문맥 정보를 담은 음소열을 입력으로 받아 합성 음성 파형을 출력한다. 이 부분은 일반적으로 운율 예측 및 음성 파형 생성을 포함하고 있으며, 본 논문은 뒷단의 음성 합성 부분에 해당하는 연구이다.

음성 합성 방법에는 여러 가지 기술들이 활용되고 발전되어 왔다. 그 흐름을 살펴보면 1960 년대 이후에는 조음기관 합성기(articulatory synthesis)나 포먼트 합성기(formant synthesis)와 같은 규칙기반(rule-based)의 방법으로 음성 파형을 생성했다. 점차 컴퓨터 연산능력이 발전하고 다량의 음성데이터를 수집하면서 1990 년대에 들어서부터 데이터 기반(data-driven)의 접근방식이 행해졌다. 데이터 기반의 음성 합성 방법은 크게 두 가지 방식이 있는데, 그 중 첫 번째는 편집 합성(concatenative speech synthesis)으로 알려진 비매개변수적(non-parametric), 예제 기반(example-based)의 방식이고, 두 번째는 통계적 매개변수 기반 음성합성(statistical parametric speech synthesis)으로 알려진 매개변수적이고 모델에 기반을 둔 방식이다. 편집 합성은 미리 녹음된 방대한 양의 음성 데이터베이스로부터 원하는 음성 유닛을 불러 모아 음성을 합성하는 방식으로 어떠한 파라미터화 과정 없이 음성 파형 단계에서 이뤄지기 때문에 음질이 우수하다는 장점을 갖고 있다.

통계적 매개변수 기반 합성 방식은 통계적 방식으로 미리 훈련된 생성모델(generative model)로부터 보코더 파라미터를 출력해서 음성을 합성한다. 텍스트 분석으로부터 출력되어진 언어학적 특징(linguistic feature)을 음향학적 특징(acoustic feature)으로 바꿔주는 과정에서 음향 모델(acoustic model)이 사용되어 음성 특징 예측을 수행하게 된다. 이때 음향 모델은 다량의 음성 데이터베이스로부터 확률 모델에 기반을 둔 통계적 기법으로 훈련되는데, 그 훈련 프레임웍은 은닉 마르코프 모델(Hidden Markov Model) 또는 깊은 신경망(Deep Neural Network)이 대표적이다. 특히 본 논문에서 사용하고 있는 'HTS'는 은닉 마르코프 모델을 기반으로 한 음성 합성을 말한다. 이 시스템은 HTK(Hidden Markov Model Toolkit)를 사용하기 때문에 합성(Synthesis)의 의미를 담아 HTS라는 이름으로 알려져 있다. 이 시스템은 기본적으로 보코더를 사용한 음성 파라미터로 음향 모델을 훈련시키는데 음성을 몇 가지의 파라미터로 나타내고 청각, 인지적으로 의미 있는 수정을 통해 새로운 음성을 출력할 수 있다는 장점이 있다. 기존에 훈련되어있는 음향 모델들 간의 보간 기법으로 목소리를 섞을 수 있으며, 적응 데이터가 주어졌을 때 적응기법을 통해 목소리를 흉내 내는 시도도 할 수 있다 (Tokuda *et al.*, 2013).

편집 합성의 음성 품질을 완전히 뛰어넘지는 못하지만, Blizzard Challenge라는 매년 개최되는 음성 합성 대회에서 통계적 매개변수 기반 음성합성 방식은 음질 측면에서 많은 발전을 보여주고 있다(Tokuda *et al.*, 2013; Zen *et al.*, 2007). 보코더와 음향 모델링에 의한 음질 저하는 극복해야 할 문제로 남아 있다. 그럼에도 불구하고 이러한 통계적 매개변수 기반의 방식이 여전히 널리 연구되고 있는 까닭은 역시 음성 특징의 유연함이라 할 수 있다. 최근 음성 합성 분야에서는 개성을 표현하고, 감정을 표현하는 음성 합성기의 필요성이 대두되고 있다. 이는 더 발전된 형태의 인간-컴퓨터 상호작용이라 할 수 있다. 가정 내에서 많이 상용화되어있는 개인 비서 시스템부터 각종 인공지능 시스템에게 목소리가 부여되면서 각 시스템의 특징을 살려내고 감정을 실어 목소리를 낼 줄 아는 시스템이 요구되고 있지만 기존에 널리 활용되고 있던 비매개변수적 방식의 음성 합성 시스템은 이러한 요구를 해결하기에는 한계가 있다. 비매개변수적 방법은 단일 화자의 음성 데이터베이스로부터 음성 합성이 이뤄지기 때문에 특정 목소리를 내기 위해서는 그 특정인의 목소리가 수 시간에서 수십 시간 수집되어야 하는데 이는 현실적으로 불가능하기 때문이다. 반면 통계적 매개변수 기반의 방법은 비록 음질 측면에서 비매개변수적 방식에 비해 열세이지만, 음향 모델 또는 음성 파라미터의 자유로운 변형을 통해 다양한 음색, 다양한 감정, 표현력 있는 음성 등을 만들어 낼 수 있는 큰 장점을 갖고 있기 때문에 앞으로도 많은 연구가 진행 될 것으로 전망된다.

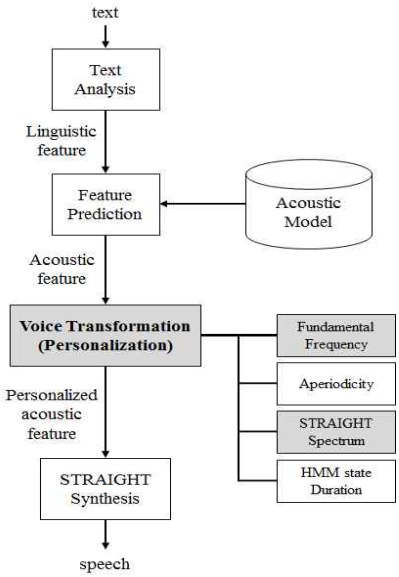

그림 1. 개인화된 HTS 개요
Figure 1. Overview of personalized HTS

본 논문에서는 앞서 설명한 통계적 매개변수 기반 음성합성 시스템의 장점을 활용한 연구를 진행하였다. 은닉 마르코프 모델에 기반을 둔 음성 합성 시스템에서 음성 파라미터의 자유로운 변환을 통해 목소리 변환(voice transformation)을 시도하였다. <그림 1>은 목소리 변환 모듈이 삽입된 개인화된 음성합성 시스템의 개요이다. 본 논문에서는 음질이 저하되지 않는 범위 내에서 두 가지 음성 파라미터를 변환하였다. 첫 번째 파라미터는 음성신호의 기본주파수로 목소리의 높낮이를 결정하는 음성 파라미터이다. 이 기본 주파수에 스케일링 인자를 두어 변환하였다. 두 번째 파라미터는 스펙트럼의 포락선이다. 특히 스펙트럼 포락선을 주파수 축에 따라 와핑시킴으로써 화자의 성도길이에 해당하는 음성 파라미터를 변환하였다. 기존의 voice

conversion 시스템과는 달리 목소리 변환(voice transformation) 시스템에서는 목표로 하는 목소리나 음색이 존재하지 않아 변환하는데 기준이 없지만(Stylianou, 2009), 본 논문에서는 위에서 언급한 두 가지 파라미터의 분석을 진행하고 이들의 상관관계를 모델링하여 목소리 변환의 기준을 제시하였다. 성별 간의 목소리 변환과 성별 내의 목소리 변환을 다양하게 시도하였으며, 본 논문의 유효성을 입증하기 위해 주관적 음질 평가와, 주관적 선호도 평가를 진행하였다.

2. 기본주파수 및 성도길이 연관성 분석

기본주파수와 성도길이의 관계는 나이 또는 성별에 따라 구분지어 그 상관관계가 설명될 수 있다. 어린이와 성인의 경우, 어린이는 신체구조가 다 자라지 않았기 때문에 성인에 비해 성대주름이 작고 성대길이 역시 짧다. 따라서 어린이의 목소리는 성인의 목소리보다 더 높은 기본주파수를 갖으며 스펙트럼 포락선을 보면 비교적으로 주파수 축에 따라 확장된 모양을 갖는다. 남성과 여성을 비교해 보면 일반적으로 남성이 여성에 비해 더 큰 신체구조를 갖는다고 할 수 있고 이는 곧 남성의 목소리는 여성에 비해 더 낮은 기본주파수를 갖고 축소된 형태의 스펙트럼 포락선을 갖는다고 할 수 있다. 본 논문에서는 다수의 화자의 목소리를 활용하여 화자별로 평균 기본주파수와 성도길이를 추정하고 이들의 상관관계를 분석한다. 분석에 사용된 음성 데이터베이스는 SNR 20dB 이상의 비교적 조용한 사무실 환경에서 16㎑ 표본화율로 수집되었다. 총 668명의 한국인 화자이며 연령대별로 정리하면 <표 1>과 같다. 화자마다 30문장을 발성하였고, 문장마다 10초 내외의 길이를 갖는다.

표 1. 분석에 사용된 음성 데이터베이스의 연령대별 화자 수
Table 1. Number of speakers per age for the analysis

연령대	20대	30대	40대	50대	계
남성	85	85	84	81	335
여성	83	86	86	78	333

2.1. 기본주파수 분석

화자의 평균 기본주파수를 추출하기 위해 STRAIGHT 보코더를 활용하였다(Kawahara, 2006). Fixed-point 분석을 통해 화자가 발성한 30문장에 대해 기본주파수를 추출하였으며, 유성음 구간에 한하여 평균을 취하였고, 이를 화자의 평균 기본주파수로 두었다.

2.2. 성도길이 분석

화자의 성도길이를 추정하기 위해 성도길이 정규화 기법(vocal tract length normalization, VTLN)을 이용하였다(Sündermann & Ney, 2003). 분석에 사용한 모든 음성 데이터로부터 일반화된 멜 스케일 켑스트럼 분석(mel-generalized cepstral analysis)을 통해 13차의 MGC(mel-generalized cepstrum) 특징벡터를 추출했다(Tokuda *et al*., 1994).

일반화된 멜 스케일 켑스트럼 분석에서 사용하는 일반화된 로그함수는 식 (1)과 같으며 본 논문에서는 γ값을 0 으로 두어 실질적으로 멜 스케일 켑스트럼을 특징벡터로 사용하였다.

$$s_\gamma(\omega)=\begin{cases}(\omega^\gamma-1)/\gamma, & 0<|\gamma|\le 1\\ \log\omega, & \gamma=0\end{cases} \tag{1}$$

멜 스케일을 위한 주파수 와핑은 식 (2), (3)과 같이 1 차 전역통과 필터의 위상 응답 특성을 갖는 쌍선형 변환 함수를 사용한다. 와핑 인자 α에 따라 사람의 청각 특성에 알맞은 주파수 스케일을 얻을 수 있고 16 ㎑ 표본화율을 갖는 음성 데이터는 α값이 0.42 일 때 멜 주파수 스케일로 근사한다고 알려져 있다.

$$\psi_\alpha(z)=\frac{z^{-1}-\alpha}{1-\alpha z^{-1}}\Big|_{z=e^{j\omega}}=e^{-j\beta_\alpha(\omega)},\quad |\alpha|<1 \tag{2}$$

$$\beta_\alpha(\omega)=\tan^{-1}\frac{1-\alpha^2\sin\omega}{(1+\alpha^2\cos\omega-2\alpha)} \tag{3}$$

성도길이 정규화 기법은 스펙트럼 최고점의 위치를 변환하기 위한 것이므로 대략적인 스펙트럼 포락선에 근거한다. 실제로 켑스트럼 차수가 9 또는 그 이상에서 처음 두 개의 포먼트를 나타낼 수 있다. 만약 켑스트럼 차수가 이 범위를 넘어서면 스펙트럼의 미세한 세부 사항을 보다 잘 표현 할 수 있지만, 스펙트럼 포락선이 13 차의 켑스트럼 차수로도 충분히 표현된다고 가정하였다. 그 후 채널 왜곡을 보상하기 위해 켑스트럼 평균, 분산 정규화 과정을 거쳤다(Saheer *et al.*, 2012).

평균 목소리 모델을 훈련시키기 위해 앞서 추출한 13 차의 MGC 특징벡터는 모두 EM 알고리즘을 통해 GMM 훈련을 진행하였다. 이때 가우시안 믹스처의 개수는 128 개를 사용하였고 diagonal 공분산 행렬을 사용하였다. 다음으로 각 화자의 MGC 특징 벡터를 grid search 를 위해 몇 가지의 와핑 인자 α 값들로 와핑시켰다. 하지만 MGC 특징 벡터는 16 ㎑ 표본화율에서 멜 주파수 스케일을 위해 이미 $\alpha=0.42$로 설정되어 있기 때문에 식 (4)로 증명되어 있는 두 개의 쌍선형 변환을 조합하는 방법을 사용하였다(Tokuda *et al*., 1994).

$$\alpha=\frac{\alpha_1+\alpha_2}{1+\alpha_1\alpha_2} \tag{4}$$

본 논문에서는 α_1은 0.42 로 두고 α_2는 -0.1 과 0.1 사이의 범위에서 0.01 만큼 변하는 값들로 설정하여 최종적으로 21 개의 와핑된 MGC 특징벡터를 구하였다. 이 후 최대 우도를 기준으로 하는 grid search를 통해 화자마다 최적의 와핑 인자를 추정하였다. 위 과정은 문장 단위로 수행되며, 각 문장별로 최대 우도를 갖는 와핑 인자를 찾고, 총 30 문장에 대해 평균을 취한 값을 해당 화자의 성도길이로 추정하였다. 즉 화자마다 하나의 와핑 인자 α를 갖는다.

<그림 2>는 추정된 성도길이의 히스토그램이다. 와핑 인자 α 값이 0 으로 추정된 화자는 평균적인 성도길이를 갖는 것으로 추정한다. 음수의 α 값을 갖는 화자는 평균보다 더 긴 성도길이를 갖고, 양수의 α 값을 갖는 화자는 더 짧은 성도길이를 갖는다. 즉 그림 2 에서 왼쪽에 분포된 화자는 남성 화자이고 오른쪽에 분포된 화자는 여성 화자를 나타낸다.

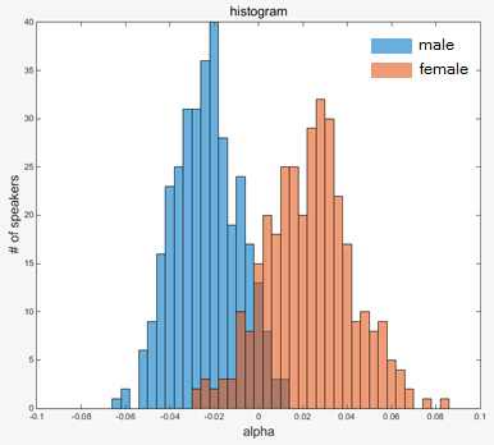

그림 2. 화자마다 추정된 성도길이의 히스토그램
Figure 2. Histogram of estimated vocal tract length

2.3. 기본주파수와 성도길이 상관관계 분석

Assmann *et al.*(2002)에 따르면 주파수축을 따라 변형된 스펙트럼 포락선과 F0의 영향에 대해 자음의 명료도를 척도로 실험을 진행하였다. 그 결과 두 인자를 개별적으로 조작하였을 때 식별 정확도가 30% 정도 떨어졌으나, 두 인자를 동시에 조작하였을 때 많은 개선을 보여줬다. 인간 청각 능력은 자연에 존재하는 음성들로부터 훈련되어지는데, 특정 기본주파수에 대해 적절한 스펙트럼 포락선 패턴이 존재하고 이러한 관계가 끊임없이 학습되고 있다고 분석하고 있다. 본 논문에서는 다량의 음성 데이터를 활용하여 이를 좀 더 세밀한 분석하고 두 인자의 상관관계를 회귀분석을 통해 모델링 하고자 한다. 앞서 구한 화자의 기본주파수와 성도길이 에 대해 scatter plot을 하면 <그림 3>과 같다.

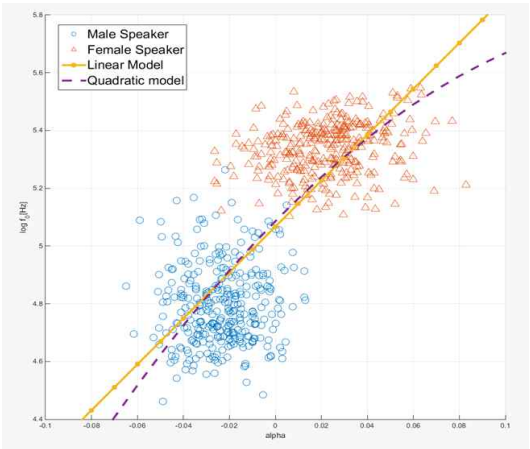

그림 3. F0와 α 의 scatter plot
Figure 3. Scatter plot of F0 and α

각각의 점은 하나의 화자를 가리키며, 성별에 따른 성도길이와 기본주파수간의 상관관계를 확인할 수 있다. 이때 기본주파수는 로그 스케일을 사용하였다. 남성 화자는 낮은 기본주파수와 긴 성도길이에 주로 분포하고 있고, 여성 화자는 높은 기본주파수와 짧은 성도길이에 주로 분포한다. 분석에 사용한 음성 데이터베이스는 20 대에서 50 대 사이의 성인 화자로만 이루어져있어 더 다양한 연령대의 목소리 특징을 반영하지 못하였다. 하지만 변성기가 오기 전의 청소년 화자 데이터가 분석되었다면, 기본주파수와 성도길이가 남성과 여성분포의 중간에 분포할 것으로 예상된다. 또 어린이의 데이터의 경우는 여성 화자보다 더 높은 기본주파수와 더 짧은 성도길이에 분포할 것으로 예상된다. 기본주파수와 성도길이의 상관관계를 회귀 분석한 결과는 다음과 같으며 <그림 3>에서 점선과 직선으로 나타내었다. 본 논문에서는 quadratic model을 사용하였다.

1) Linear model

$$f_0 = A_1\alpha + A_0 \quad ,where \;\; A_1 = 7.95, \;\; A_0 = 5.07$$

2) Quadratic model

$$f_0 = A_2\alpha^2 + A_1\alpha + A_0 \quad ,where \;\; A_2 = -22.81, \;\; A_1 = 8.11, \;\; A_0 = 5.09$$

3. 제안한 목소리변환 방법 및 실험결과

본 논문에서는 은닉 마르코프 모델 기반의 음성 합성기에서 기본주파수와 성도길이의 상관관계를 이용한 목소리 변환을 시도하였고 그 유효성을 입증하기 위해 두 가지 주관적 평가를 진행하였다. 구현한 HTS의 사양을 설명하고 실험 방법에 대한 설명, 그리고 주관적 평가 결과들에 대해 살펴보겠다.

3.1. HTS 설정

본 논문에서는 두 명의 전문 성우로부터 수집한 음성 데이터베이스로 화자 종속적 음성 합성기를 구현하였다. 성우는 30대의 남성, 여성 성우이다. 각 성우는 총 4392개의 문장을 발성하였으며, 잡음이 거의 없는 사무실 환경에서 16㎑ 표본화율로 녹음되었다. 남성 성우는 총 9.3시간, 여성 성우는 10.4시간의 데이터베이스가 구축되었다. STRAIGHT 보코더를 사용한 은닉 마르코프 모델 기반의 음성 합성기를 구현하였다(Zen *et al.*, 2007). 5-state left-to-right HMM 구조를 사용하였고 모델의 출력 벡터는 STRAIGHT 음성 파라미터로 <표 2>와 같다.

표 2. HTS의 음성 파라미터 구성
Table 2. Number of speech parameters of HTS

para.	STRAIGHT MGC	log F0	Aperiodicity	Total
static	45	1	26	216
dynamic	90	2	52	

위와 같은 사양으로 구현한 HTS의 MOS 평가는 남성 모델의 경우 3.14 점, 여성 모델의 경우 3.29 점으로 준수한 수준의 성능을 내었다. 그리고 목소리 변환 모듈을 삽입하여 개인화된 HTS를 구현하였다. 스케일링을 통한 기본주파수 변환과 쌍선형 변환을 통한 성도길이 변환을 위해 스케일링 인자와 와핑 인자를 각각 입력을 할 수 있도록 하였다.

3.2. 기본주파수와 성도길이 상관관계를 이용한 목소리 변환

앞서 구현한 목소리 변환 모듈이 내장된 음성합성기를 사용하여 2장에서 도출한 기본주파수와 성도길이의 상관관계를 이용한 목소리 변환 실험을 진행하였다. 실험에 사용된 합성문장은 훈련에서 사용되지 않은 다섯 개의 임의의 문장으로 선택되었다.

- 메이저리그 역대 삼십 번째 삼천안타 타자
- 이들 유형의 특징적인 차이는 다음에 살펴볼 의미부에서 나타난다.
- 수능을 앞두고 반짝 추위가 찾아왔습니다.
- 신용잔고 감소에 따라 증권업계에서는 증시 하락을 예상하고 있다.
- 상장기업 실적은 올해 삼 분기 들어 크게 악화된 것으로 나타났다.

본 논문에는 voice conversion 분야와 달리 목표 화자가 존재하지는 않지만, 목소리 변환에 있어서 다양한 시도를 위해 성별간의 목소리 변환과 성별 내에서의 목소리 변환을 시도함으로써 목표로 하는 성별은 존재하도록 설정하였다. 이때 음성 파라미터가 변환되는 양은 기존에 알려진 범위 내에서 그리고 실험적 경험을 통해 결정하였다.

3.3. 주관적 평가 결과

기본주파수와 성도길이의 상관관계 모델의 유효성을 입증하기 위해 두 가지 형태의 목소리 변환을 시도했다. 첫 번째 변환은 α만을 바꿔주어 오직 성도길이만 변환되도록 하는 방법이다. 두 번째 변환은 α뿐만 아니라 기본주파수도 함께 바꿔주는데 이때 본 논문에서 회귀분석을 통해 얻은 2차 회귀 모델을 활용한다. 즉, 주어진 성도길이에 해당하는 기본주파수를 구하고 이 기본주파수 대역이 되도록 기존 합성음을 변환시킨다.

평가에 참여한 실험자는 20 대에서 30 대 사이의 정상청력을 가진 남녀 16 명으로 구성되었다. 실험자들은 앞서 말한 두 가지의 합성음을 듣고 어느 합성음이 더 선호되어 듣기 거북하지 않은지 선택했다. 또한 실험의 목소리 변환이 음질 저하를 발생시켰는지 확인하기 위해 MOS 음질 평가도 진행되었다.

<그림 4>는 주관적 선호도 평가의 결과이다. 남성→남성(M→M), 여성→여성(F→F), 남성→여성(M→F), 여성→남성(F→M)에서의 선호도 결과들과 이를 종합한 결과를 구분하여 나타냈다. 여기서 VT1 은 성도길이만 변환한 것이고, VT2 는 성도길이와 기본주파수를 모두 변환한 것을 의미한다. 종합적으로 봤을 때 기본주파수는 고려하지 않고 성도길이만 변환한 목소리보다 두 파라미터의 상관관계를 고려하여 동시에 변환한 목소리가 더 선호되는 것으로 나타났다. 또 한 가지 살펴볼 점은 성별 내에서 변환이 이루어졌을 때, 어느 것도 선호하지 않는다는 응답률이 높게 나타났다. 그 이유는 성도길이의 변환은 성별간의 특징을 잘 구분하지만 동일성별 내에서는 성도길이의 변위가 많이 크지 않기 때문에 실험자들이 쉽사리 선호되는 목소리를 선택하지 못한 것이라 분석된다.

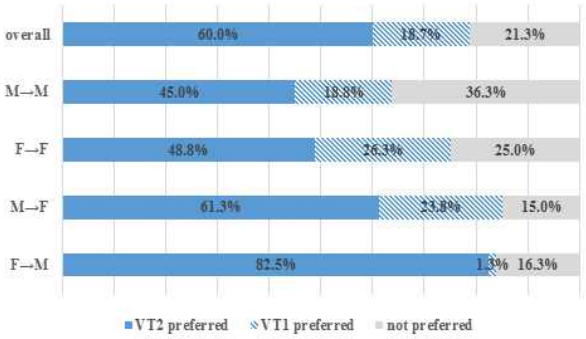

그림 4. 주관적 선호도 평가 결과
Figure 4. Subjective preference test result

<그림 5>는 주관적 음질 평가의 결과이다. Base는 목소리 변환을 하지 않은 것을 의미한다. 목소리 변환을 적용한 합성음의 음질이 미미하게 높게 평가되었다. 이는 기존에 훈련된 단일 화자의 음성 합성기가 실험자들로 하여금 덜 선호되는 목소리 특징을 갖고 있기 때문인 것으로 판단된다. 주관적 음질 평가의 결과로부터 본 논문에서 사용한 목소리 변환은 기존에 구현된 합성기의 음성 품질을 저하시키지 않는다는 것을 확인할 수 있다.

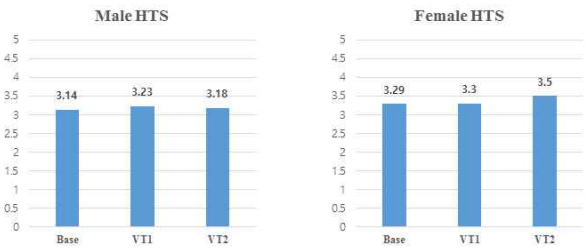

그림 5. MOS 결과
Figure 5. MOS result

4. 결론

본 논문에서는 은닉 마르코프 모델 기반의 음성 합성 시스템에서 목소리 변환을 시도하였다. 이는 통계적 매개변수 기반 음성 합성 방식이 다양한 목소리 특징을 만들 수 있다는 장점을 이용한 것으로 목소리 변환의 방법에 있어서 한 가지 기준을 제시하였다. 목소리 변환을 위해 기본주파수와 스펙트럼 포락선의 변환을 시도하였다. 이때 기본주파수는 스케일링하는 방식으로

변환하고, 스펙트럼 포락선은 일반화된 멜 켑스트럼에서 사용하는 쌍선형 변환을 활용하여 변환하였다. 여기서 스펙트럼 포락선의 변환은 화자의 성도길이의 변환을 의미한다. 본 논문에서는 이와 같은 방식의 목소리 변환은 음질 저하가 일어나지 않는다는 것을 주관적 음질 평가를 통해 확인하였다.

또한 668 명 화자의 음성 데이터베이스를 사용하여 기본주파수와 성도길이를 면밀히 분석하였다. STRAIGHT 보코더의 fixed-point analysis를 사용하여 기본주파수를 추출하고 성도길이 정규화기법을 활용하여 성도길이를 추정했다. 이를 통해 일반 화자들의 기본주파수와 성도길이가 어떻게 분포하는지 확인하였고 나아가 두 가지 음성 파라미터의 상관관계를 모델링하였다.

본 논문의 핵심은 기본주파수와 성도길이의 상관관계 모델링과 이를 활용하여 목소리를 변환하는데 있어서 기준과 방향을 제시하였다는 것이다. 이러한 목소리 변환 방법의 유효성을 확인하기 위해 다양한 주관적 평가를 실시하였고 그 결과 본 논문에서 제시한 방법이 음질을 저하시키지 않으면서 사람들로 하여금 더 선호되는 목소리를 만들 수 있다는 것을 검증할 수 있었다.

추후 연구로는 다음과 같은 것이 있다. 기본주파수와 성도길이의 상관관계가 더 정교하고 세밀한 방법으로 모델링 될 필요가 있다. 본 논문에서는 단일의 성도길이를 추정하였으나, 실제로 사람이 발성을 할 때 조음기관의 길이는 일정하지 않기 때문에 다중클래스의 성도길이를 추정하는 것이 더 정확하다(Saheer *et al.*, 2010). 또한 본 논문에서는 기본주파수와 성도길이의 상관관계 모델링을 회귀분석을 사용하여 다소 단순한 방법으로 수행했는데 통계적 모델링을 활용하여 주어지는 성도길이에 따라 추천되는 기본주파수의 범위를 제안하는 것이 더 발전된 시스템으로 나아가는 한 가지 방법이다. 본 논문과 후속연구들을 통해 향후에는 사용자가 음성 파라미터를 직관적으로 컨트롤하여 원하는 목소리를 얼마든지 생성해 낼 수 있기를 기대한다.

참고문헌

Assmann, P. F., Nearey, T. M., & Scott, J. M. (2002). Modeling the perception of frequency-shifted vowels. *Proceedings of the 7th International Conference on Spoken Language Processing* (pp. 425-428).

Kawahara, H. (2006). STRAIGHT, exploration of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Journal of Acoustical Science and Technology*, 27, 349-353.

Saheer, L., Dines, J., & Garner, P. N. (2012). Vocal tract length normalization for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7), 2134-2148.

Saheer, L., Dines, J., Garner, P. N., & Liang, H. (2010). Implementation of VTLN for Statistical Speech Synthesis. *Proceedings of the 7th ISCA Speech Synthesis Workshop* (pp. 224-229).

Stylianou, Y. (2009). Voice transformation: A survey. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (pp. 3585-3588).

Sündermann, D., & Ney, H. (2003). VTLN-based voice conversion. *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology* (pp. 556-559).

Tokuda, K., Masuko, T., Kobayashi, T., & Imai, S. (1994). Mel-generalized cepstral analysis - A unified approach to speech spectral estimation. *Proceedings of the International Conference on Spoken Language Processing* (pp. 1043-1046).

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J ., & Oura, K. (2013). Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE* (pp. 1234-1252).

Zen, H., Toda, T., Nakamura, M., & Tokuda, K. (2007). Details of the Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. *IEICE Transactions on Information and Systems*, E90-D, 325-333.

• **유효근 (Yoo, Hyogeun)**
한국과학기술원 전기및전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7617
Email: hyogeun.yoo@kaist.ac.kr
관심분야: 음성합성
현재 전기및전자공학부 석사과정 재학 중

• **김영관 (Kim, Younggwan)** 교신저자
한국과학기술원 전기및전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7617
Email: cleanthink@kaist.ac.kr
관심분야: 음성인식, 화자적응
현재 전기및전자공학부 박사과정 재학 중

• **서영주 (Suh, Youngjoo)**
한국과학기술원 전기및전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7517 Fax: 042-350-7619
Email: yjsuh@kaist.ac.kr
관심분야: 음성합성, 음성신호처리
2006~현재 전기및전자공학부 연구교수

• **김희린 (Kim, Hoirin)**
한국과학기술원 전기및전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7417 Fax: 042-350-7619

Email: hoirkim@kaist.ac.kr
관심분야: 음성인식, 화자인식, 패턴인식
2001~현재 전기및전자공학부 교수