말소리와 음성과학 제2권 제1호 (2010)
pp. 87~95

HMM 기반 혼용 언어 음성합성을 위한 모델 파라메터의 음절 경계에서의 평활화 기법

Syllable-Level Smoothing of Model Parameters for HMM-Based Mixed-Lingual Text-to-Speech

양 종 열[1] · 김 홍 국[2]
Yang, Jong-Yeol · Kim, Hong Kook

ABSTRACT

In this paper, we address issues associated with mixed-lingual text-to-speech based on context-dependent HMMs, where there are multiple sets of HMMs corresponding to each individual language. In particular, we propose smoothing techniques of synthesis parameters at the boundaries between different languages to obtain more natural quality of speech. In other words, mel-frequency cepstral coefficients (MFCCs) at the language boundaries are smoothed by applying several linear and nonlinear approximation techniques. It is shown from an informal listening test that synthesized speech smoothed by a modified version of linear least square approximation (MLLSA) and a quadratic interpolation (QI) method is preferred than that without using any smoothing technique.

Keywords: Text-to-Speech, HTS, polyglot, mixed-lingual TTS, parameter smoothing

1. 서론

컴퓨터와 인간 사이에 정보를 전달하기 위한 방법으로서 음성은 매우 편리한 매개로 사용되고 있다. 특히 음성인식과 더불어 음성합성은 컴퓨터에서 사용자에게 정보를 전달하는 편리한 방법 중 하나이다(Dutoit, 1997). 최근에는 컴퓨터뿐 아니라 PDA나 내비게이션과 같은 전자기기 등에도 음성합성 기술이 사용되는 등 그 응용범위가 점차 확대되는 추세이다.

음성합성 기술에는 여러 방법이 있다. 초기의 음성합성으로는 사람의 음성을 녹음하여 그것을 연결하여 사용하는 방법이 있다(Black & Taylor, 1994; Donovan & Woodland, 1995; Hunt & Black, 1996) 사람의 음성을 녹음하여 사용하는 연결형 방법은 좋은 음질의 합성음을 얻을 수 있다는 장점이 있으나 방대한 양의 DB를 저장해놓아야 하므로 휴대용 전자기기와 같은 적은 용량을 제공하는 장치에 적용하기 어려운 단점이 있다(Eide, Aaron, Bakis, Hamza, Picheny, & Pitrelli, 2004). 이를 보완하여 최근에는 음성으로부터 여기신호, MFCC 등 특징벡터를 추출하여 은닉 마코프 모델(Hidden Markov Model, HMM)을 학습하고 그로부터 음성을 합성하는 HMM 기반 음성 합성 방법이 많이 연구되고 있다(Black, Zen, & Tokuda, 2007; Ling, Wu, Wang, Qin, & Wang, 2006; Yoshimura, Tokuda, Masuko, Kobayashi, & Kitamura, 1999; Yu, Zhang, Tao, & Wang, 2007; Zen, Toda, Nakamura, & Tokuda, 2007). HMM 기반 음성합성 방법은 연결형 음성합성 방법에 비해 합성음의 음질은 떨어지지만 요구되는 메모리 용량이 적어 휴대용 전자기기 등에 사용하기 쉬운 장점이 있다.

음성합성의 응용분야가 점차 확대됨에 따라 요구되는 합성기술 또한 매우 다양해지고 있다. 기존 합성분야에서는 한 사람이 발화하는 한 언어에 대한 음성합성 시스템으로 충분히 그 역할을 담당했던 것에 비해 최근에는 여러 사람의 목소리를 합성해주는 화자 적응 기법을 이용한 다중 화자 음성합성시스템

1) 삼성전자 DMC jy50.yang@samsung.com
2) 광주과학기술원 hongkook@gist.ac.kr, 교신저자
이 논문은 2007년 정부의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구이며 (KRF-2007-314-D00245) 또한 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2010-C1090-1021-0007)

접수일자: 2010년 1월 27일
수정일자: 2010년 3월 9일
게재결정: 2010년 3월 14일

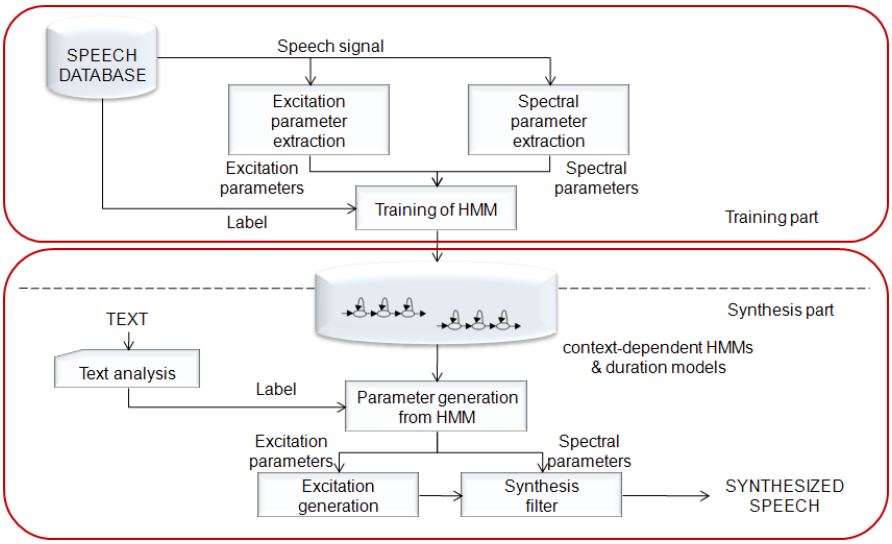

그림 1. HTS System의 개요
Figure 1. HTS System Overview

이나 여러 언어를 번갈아가면서 말해주는 다중 언어 음성합성 시스템 혹은 여러 언어가 섞인 문장을 합성해주는 혼용 언어 음성합성 시스템에 대한 요구가 확대되고 있다(Traber, Huber, Nedir, Pfister, Keller, & Zellner, 1999).

따라서 본 논문에서는 여러 언어가 섞인 혼용 언어 문장을 합성해줄 때 서로 다른 언어의 경계 간에 발생하는 문제점을 분석하고 이를 해결하기 위한 방법으로 서로 다른 언어 간의 경계를 자연스럽게 연결하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 한국어와 영어의 HMM 기반 음성합성시스템에 대해 간단히 기술하고, 3장에서는 혼용 언어 음성합성 시스템과 서로 다른 언어 간 경계를 자연스럽게 연결하는 기법에 대해 제안한다. 4장에서는 3장에서 제안된 방법을 이용해 합성한 합성음의 품질을 평가하고 마지막으로 5장에서 결론을 맺는다.

2. HMM기반 음성합성시스템

2.1 훈련 부분

HMM 기반 음성합성시스템(HMM-based Speech Synthesis System, HTS)에 대한 기본적인 구성은 <그림1>과 같다(Tokuda, Yoshimura, Masuko, Kobayashiand, & Kitamura, 2000). 시스템은 크게 문맥 종속 HMM(context-dependent HMM)과 결정 트리(Classification And Regression Tree, CART) 모델을 훈련하는 부분과 이러한 모델을 이용하여 음성을 합성하는 부분으로 나누어진다.

HTS에서 사용하는 문맥 종속 HMM은 스펙트럼과 여기신호에 대한 벡터로 구성되어 있다. 스펙트럼은 0번째 계수를 포함한 24개의 멜-켑스트럼 계수(MFCC; Mel-Frequency Cepstral Coefficient) 벡터를 추출한 후 그것을 1차 및 2차 미분한 것을 더한 것으로 구성된다. 여기신호는 음성으로부터 로그 포만트를 추출한 후 그것을 1차 및 2차 미분한 것을 더해 구성된다. 스펙트럼 신호가 연속적인 모델을 가짐으로 HMM으로 표현되는데 반하여 여기신호는 유성음과 무성음의 두 신호를 갖게 되므로 유성음/무성음을 구분하여 훈련되며 이 때 연속적인 모델과 이산적인 모델을 동시에 고려하기 위해 MSD-HMM(Hidden Markov Model based on Multi-Space probability Distribution)을 사용한다(Tokuda, Masuko, Miyazaki, & Kobayashi, 1999).

또한 HMM은 합성 시 음성의 기본 길이를 만들기 위한 길이 모델(duration model)을 가지고 있는데 결과적으로 HTS에서 문맥 종속 HMM은 MFCC뿐만 아니라 포만트와 길이모델을 포함하여 훈련된다.

2.2 결정트리 기반 문맥 클러스터링

스펙트럼, 여기신호, 길이 정보는 여러 문맥적인 요소들에 의해 영향을 받게 되는데 이러한 문맥적인 요소들을 고려하기 위해 문맥 종속적인 HMM을 훈련한다. 하지만 문맥 정보들이 증가하게 되면 그에 따른 문맥 조합은 지수적으로 증가하게 되어 제한된 음성 코퍼스로는 이들을 모두 고려한 문맥 종속 HMM을 훈련할 수 없게 된다. 이러한 문제 때문에 스펙트럼, 여기신호, 그리고 길이에 대한 결정트리 기반 문맥 클러스터링 기법을 적용한다(Odell, 1995; Shinoda & Watanabe, 2000).

2.3 합성 부분

HTS의 합성 부분에서는 가장 먼저 입력 텍스트를 문맥 종속적인 레이블로 만들어 준다. 문맥을 고려한 레이블로부터 훈련된 문맥 종속 HMM을 이용하여 하나의 HMM 문장을 만들어

준다. HMM 문장에서 각 상태의 길이는 상태길이(state duration)의 확률을 최대화하는 값으로 결정된다(Shinoda & Watanabe, 2000). 그 다음 MFCC와 피치와 같은 특징벡터의 파라메터 열을 생성하게 되는데 이 파라메터 열(식 (1))은 파라메터 발생 알고리즘(Tokuda, Yoshimura, Masuko, Kobayashiand, & Kitamura, 2000)을 이용해 식 (2)의 출력 확률을 최대화하는 값에 의해 결정된다.

$$O=[o_1^T,o_2^T,...,o_N^T]^T \quad (1)$$

$$P(O|\lambda)=\sum_{all\,Q}P(O,Q|\lambda) \quad (2)$$

여기서, o_t는 t번째 프레임에서의 특징벡터, N은 문장 전체의 프레임 수, O는 문장 전체의 파라메터 열을 의미한다. 그리고 $Q=(q_1,i_1),(q_2,i_2),...,(q_N,i_N)$ 이고 q_t와 i_t는 t번째 프레임에서의 상태와 해당 상태에서의 가우시안 확률의 i번째 혼합을 의미한다.

마지막으로 발생된 스펙트럼과 피치의 파라메터열로부터 MLSA(Mel-log Spectral Approximation) 필터를 이용해 음성 신호를 합성한다(Fukada, Tokuda, Kobayashi, & Imai, 1992).

2.4 언어의 특성

모든 언어는 언어 고유의 특성을 지니고 있다. 본 논문에서 구현된 영어와 한글 음성합성시스템을 이루는 영어와 한글은 문맥이나 음소에서 전혀 다른 언어적 특성을 지니고 있다.

결정 트리 기반 문맥 클러스터링에서는 음절 및 어절과 같은 문맥정보뿐 아니라 음소의 정보 또한 세부적으로 나누어진다. 이를 위해서는 합성하려는 언어의 특성을 파악하는 것이 중요하다. 본 논문에서는 한글에 대해서는 Kim 등(Kim, Kim & Hahn, 2006)에서 사용한 음소 특성에 따라 음소를 분류하도록 한다.

일반적으로 많은 문맥정보를 고려하면 합성음의 품질은 향상되지만(Kim, Lee, & Hirose, 2002; Kim, Lee, & Hirose, 2001) 적절히 고려되지 않은 문맥정보는 오히려 합성음의 품질을 떨어뜨릴 수 있다(Fukada, Tokuda, Kobayashi, & Imai, 1992). 따라서 본 논문에서 사용한 문맥정보는 다음과 같다.

- {선행, 현재, 후행} 음소
- 현재 음절에서 음소의 위치
- 현재 어절에서 음절의 위치
- 현재 어절에서 음절의 수

선택된 문맥정보는 기존에 한글의 음성합성 시스템에서 많이 사용되는 문맥정보를 참조하여 합성음의 품질을 높이는 문맥정보를 선택한 것이다(Kim, S. J., Kim, J. J. & Hahn, M., 2006).

3. HMM기반 다중언어 음성합성시스템

3.1 발음 구조

두 가지 이상의 언어를 혼용하여 음성합성시스템을 만들기 위해 두 언어 간 발음기호를 분류하여 사용해야 한다. <표1>과

표 1. 한글 발음의 심볼

Table 1. List of Korean phoneme symbols

발음기호	Symbol	발음기호	Symbol
ㄱ[k]	g	ㅏ[a]	a
ㄴ[n]	n	ㅓ[ə]	v
ㄷ[t]	d	ㅗ[o]	o
ㄹ[l]	l	ㅜ[u]	u
ㅁ[m]	m	ㅡ[ɨ]	U
ㅂ[p]	b	ㅣ[i]	i
ㅅ[s]	s	ㅔ[e]	e
ㅇ[ŋ]	N	ㅐ[ɛ]	E
ㅈ[tʃ]	z	ㅑ[ja]	ja
ㅊ[$tʃ^h$]	c	ㅕ[jə]	jv
ㅋ[k^h]	k	ㅛ[jo]	jo
ㅌ[t^h]	t	ㅠ[ju]	ju
ㅍ[p^h]	p	ㅒ[jɛ]	jE
ㅎ[h]	h	ㅖ[je]	je
ㄲ[k^l]	G	ㅘ[wa]	wa
ㄸ[t^l]	D	ㅝ[wə]	wv
ㅃ[p^l]	B	ㅙ[wɛ]	wE
ㅆ[s^l]	S	ㅞ[we]	we
ㅉ[$tʃ^l$]	Z	ㅚ[we]	O
silence	sil	ㅟ[wi]	wi
		ㅢ[ɯi]	xi

표 2. 영어 발음의 심볼

Table 2. List of English phoneme symbols

발음기호	Symbol	발음기호	Symbol
ɔ	AO	k	KK
ɑː	AA	ɡ	GG
iː	IY	m	MM
uː	UW	n	NN
ɛ	EH	ŋ	NG
ɪ	IH	f	FF
ʊ	UH	v	VV
ə	AH	Θ	TH
æ	AE	ð	DH
eɪ	EY	s	SS
aɪ	AY	z	ZZ
oʊ	OW	ʃ	SH
aʊ	AW	ʒ	ZH
ɔɪ	OY	h	HH
ɝː	ER	l	LL
p	PP	r	RR
b	BB	j	YY
t	TT	w	WW
d	DD	tʃ	CH
silence	sil	dʒ	JH

<표2>는 본 논문에서 분류한 영어와 한글에 대해 발음기호에 대한 심볼을 각각 보여준다. 비록 한글과 영어사이에 중복되는 발음이 있지만 같은 발음이라도 서로 다른 언어에서는 각각 사용되는 용도가 언어의 문맥 구조에 따라 다르기 때문에 각 언어의 발음 심볼은 다르게 표현해주어야 더욱 자연스러운 혼용 언어 문장을 합성할 수 있다.

3.2 기본 구조

일반적으로 혼용 언어 음성합성시스템을 만들기 위한 가장 좋은 방법은 만들려고 하는 혼용 언어에 맞는 음성 코퍼스를 이용해 혼용 언어 음성합성시스템을 개발하는 것이다. 두 언어가 섞여서 녹음된 코퍼스가 있는 경우에는 단순히 하나의 언어에 대한 음성합성시스템과 동일한 방법으로 음성합성시스템을 구축할 수 있다. 그러나 두 개 혹은 그 이상의 언어가 섞여있는 음성 코퍼스를 구축하는 것은 구축하려는 언어의 조합들을 모두 고려해야하기 때문에 음성 코퍼스의 분량이 방대질 수 있다. 또한 대부분의 음성 코퍼스가 하나의 언어에 대해 녹음되어있고 구축하려는 두 개 이상의 언어에 능숙한 성우를 구하는 일도 어려워 음성합성에 필요한 코퍼스를 구축하거나 찾는 것 자체가 어려움이 될 수 있다.

본 논문에서는 각각 별도로 녹음되어있는 한 명의 성우에 의해 발화된 한글과 영어에 대한 음성 코퍼스를 이용한 혼용 언어 음성 합성 시스템(Mixed-lingual TTS)을 고려한다.

본 논문에서 제안하는 혼용 언어 음성합성시스템의 기본 구조는 <그림2>와 같다. 혼용 언어 음성합성시스템에서는 한글과 영어에 의해 훈련된 각각의 HMM을 이용하여 입력된 혼용 언어 문장으로부터 그에 맞는 HMM을 선택해 적절한 파라메터를 생성해주게 된다.

좀 더 자세한 과정을 살펴보면, 입력된 문장을 음소정보 및 문맥을 고려한 문맥 종속 레이블로 변환한 후 문맥 종속 레이블의 현재 음소를 기준으로 현재의 음소가 한글일 경우는 한글 음성 코퍼스를 이용해 훈련된 HMM으로부터, 현재의 음소가 영어일 경우에는 영어 음성 코퍼스를 이용해 훈련된 HMM으로부터 파라메터를 생성해주게 된다.

그런데 영어에서 한글 혹은 한글에서 영어로 변하는 구간에 대해서는 현재의 음소가 한글 혹은 영어의 음소를 가진다 하더라고 선행 혹은 후행의 음소가 같은 한글과 영어의 음소를 가지지 못하는 경우가 발생하게 된다. 선행, 현재, 후행의 음소의 언어가 다른 경우에는 세 음소를 모두 고려한 파라메터를 가져오지 못하고 현재 음소에 맞는 언어에 대한 HMM으로부터 파라메터를 생성하기 때문에 언어적 특성을 충분히 반영하지 못한 파라메터를 생성하게 된다. 이로 인해 한글과 영어의 경계는 자연스럽지 못한 연결점이 발생하게 되고 이로 인해 원하지 않는 음성을 합성하게 된다.

본 논문에서는 이러한 불연속점을 파라메터 단에서 자연스럽게 연결하고 나아가 혼용 언어 음성합성시스템에서 서로 다른 언어의 경계 간에 발생할 수 있는 잡음을 최소화하는 기법을 제안한다.

3.3 선형함수를 이용한 연결 기법

HMM의 하나의 심볼은 총 5개의 state로 이루어져있다. 본 논문에서는 한글과 영어의 경계에서 파라메터를 연결하기 위한 구간으로써 이전 심볼의 마지막 두 개의 상태(state)와 다음 심볼의 처음 두 개의 상태를 정하여 총 네 개의 상태에 위치한 파라메터를 연결해주는 기법을 제안한다.

본 논문에서 제안하는 파라메터 연결기법은 크게 두 가지로 나누어진다. 하나는 선형 함수를 이용한 연결기법이고 다른 하나는 비선형 함수인 2차 함수를 이용한 연결기법이다.

선형 함수를 이용한 연결기법에서는 식 (3)과 같은 선형 함

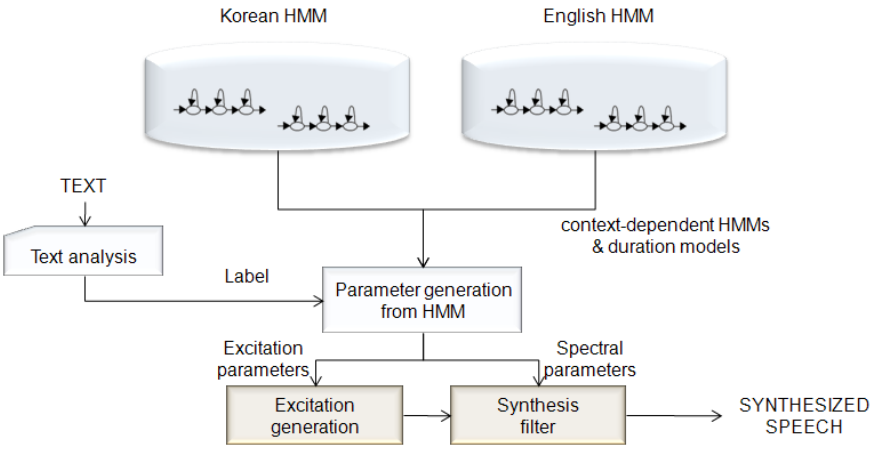

그림 2. 다중언어 음성합성시스템의 구조
Figure 2. Block diagram of a mixed-lingual TTS

수를 이용하여 경계 간 파라메터를 연결해준다. 선형 함수를 이용한 첫 번째 연결기법은 식 (4)와 같이 경계구간의 파라메터 처음 값과 마지막 값을 지나는 선형 함수를 만들어 파라메터를 보간하는 선형 보간법(Linear Interpolation, LI)이다.

$$\hat{f}(x) = mx + b \tag{3}$$

$$\hat{f}(x) = \left[\frac{y_e - y_b}{x_e - x_b}\right](x - x_b) \tag{4}$$

여기서 y는 파라메터를 나타내고 x는 인덱스를 그리고 e는 마지막 값을 b는 처음 값을 나타낸다.

선형 보간법을 이용하여 두 점을 이어주면 두 점 사이의 값들이 첫 번째 파라메터와 마지막 파라메터에 의해 이어져 불연속적인 경계를 연결할 수 있게 된다.

그러나 선형 보간법에서는 기존의 파라메터에 담겨있는 정보를 충분히 반영하지 못하고 단순히 첫 번째 값과 마지막 값에 의해 파라메터가 변형된다는 단점이 있다. 기존의 정보를 적절히 반영하면서 파라메터를 연결하기 위한 방법으로 선형 최소자승 근사법(Linear Least Square Approximation, LLSA)을 적용할 수 있다. 선형 최소자승 근사법은 식 (5) 및 (6)과 같이 기존의 값들과 만들어진 선형함수의 값과의 차이의 제곱을 최소화하는 선형함수를 만드는 것이다.

$$\hat{f}(x) = \arg min_{\hat{f}(x)} E \tag{5}$$

$$E = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{f}(x_i) - y_i\right]^2 \tag{6}$$

식 (5) 및 (6)을 만족하는 근사함수는 식 (7) 및 (8)과 같이 선형함수의 계수에 대한 편미분값을 0으로 만드는 각 계수를 구함으로써 만들 수 있다. 각 조건을 이용하여 최종적으로 식 (9)과 같이 정리할 수 있고 식 (9)의 역행렬을 구하면 근사함수를 구할 수 있다.

$$G(m,b) = \sum_{i=1}^{n}\left[mx_i + b - y_i\right]^2 \tag{7}$$

$$\frac{\partial G(m,b)}{\partial m} = 0, \frac{\partial G(m,b)}{\partial b} = 0 \tag{8}$$

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}\begin{bmatrix} b \\ m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \tag{9}$$

선형 최소자승 근사법은 원래의 정보를 최대한 반영한 근사함수를 이용하여 파라메터를 변형하므로 원래의 소리를 크게 변형하지 않으면서 파라메터를 자연스럽게 연결시켜주는 장점이 있다. 하지만 경계부분의 처음과 끝부분에서 원래의 신호와 이어지는 값이 급격히 변하여 원하지 않는 잡음이 발생하게 된다. 경계부분의 잡음을 줄이기 위하여 선형 최소자승 근사법에 의해 구해진 선형함수의 처음부분과 끝부분이 <그림3>과 같이 급격히 변화하지 않는 범위 내에 위치하는 수정된 선형 최소자승 근사법(Modified Linear Least Square Approximation, MLLSA)을 사용하면 경계부분의 노이즈를 줄일 수 있다. <그림3>의 점선은 선형 최소자승 근사법을 이용한 함수를 보여주고 실선은 수정된 선형 최소자승 근사법을 보여준다. 이 때 처음 값과 마지막 값의 제한된 범위는 그 전 신호와 그 후 신호와의 변화율을 계산하여 변화율의 크기를 넘지 않는 값으로 설정하였다.

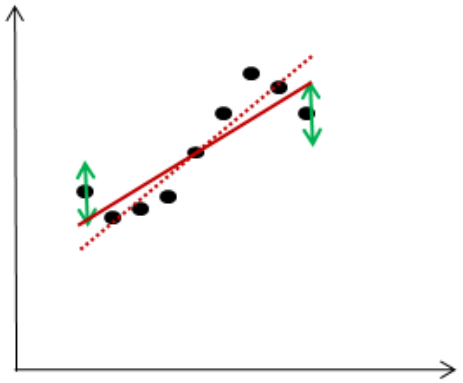

그림 3. 수정된 선형 최소자승 근사법
Figure 3. Modified linear least square approximation

<그림4>는 한영 경계구간에 대해 원래의 파라메터와 각각의 선형함수를 이용해 연결된 파라메터를 보여준다. 그림의 가로축은 인덱스열을 나타내고 세로축은 파라메터 값을 나타낸다. 원래의 파라메터가 불연속적인 파형을 그리는데 비해 선형함수에 의해 연결된 파라메터의 경우 좀 더 자연스럽게 연결되는 것을 볼 수 있다. 단, LLSA 기법의 경우 처음 값과 마지막 값

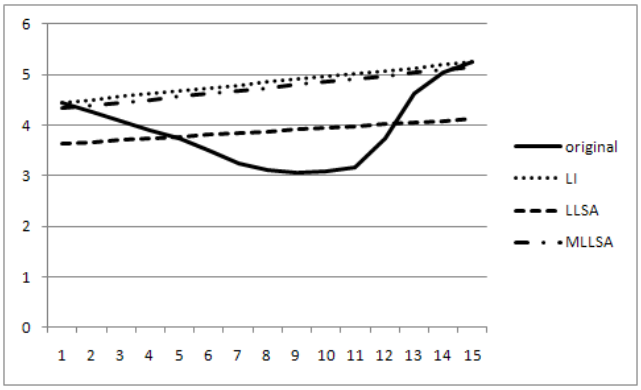

그림 4. 선형 함수를 이용해 연결된 파라메터
Figure 4. Illustration of parameters concatenated using linear functions

이 원래의 신호의 값과 많이 떨어져있어 원하지 않는 불연속점을 발생시키는 것을 볼 수 있다. 반면 LLSA 기법을 수정한 MLLSA 기법을 사용하면 선형보간법과 유사하지만 원래 신호에 가깝게 파라메터를 변화시키는 것을 볼 수 있다. 이는 LI 기법과 유사하게 파라메터 연결 부위가 크게 어긋나지 않아 자연스럽게 연결되게 하며 또한 경계부분의 다른 파라메터 값들을 고려함으로써 원래 합성하려던 음성을 더 분명하게 나타낼 수 있게 한다.

3.4 2차 함수를 이용한 연결 기법

본 논문에서는 비선형 함수로 2차 함수를 사용하여 2차 보간법과 2차 최소자승 근사법을 사용하였고 수정된 선형 최소자승 근사법과 마찬가지로 파라메터의 첫 번째 값과 마지막 값이 일정 범위 안에 들어오는 수정된 2차 최소자승 근사법을 사용하였다.

2차 보간법(Quadratic Interpolation, QI)은 경계구간의 첫 번째 파라메터와 마지막 파라메터 그리고 서로 다른 언어 간 경계가 되는 파라메터의 세 점을 지나는 2차 함수를 만들어 파라메터를 변환해주는 방법이다. 이 때 경계의 두 점이 일치하지 않기 때문에 두 점의 평균값을 세 번째 값으로 설정하였다. 식 (10)과 같은 2차 함수에 세 점을 대입한 연립방정식의 해를 찾는 것으로 원하는 2차 함수를 구할 수 있다.

$$\hat{f}(x) = a_1 + a_2 x + a_3 x^2 \tag{10}$$

2차 보간법을 이용한 연결기법은 선형 보간법이 처음과 마지막 점만을 고려해 파라메터를 연결한다는 것에 비해 두 심볼 간 경계의 값을 추가로 반영해주어 좀 더 자연스러운 합성음을 만들어주는 장점이 있다.

2차 최소자승 근사법(Quadratic Least Square Approximation, QLSA)은 선형 최소자승 근사법과 유사하게 식 (5) 및 (6)과 같이 근사함수의 값과 실제 함수의 값의 차이의 제곱이 최소가 되게 하는 함수를 찾는 것이다. 식 (11)과 같이 원래의 값과 근사 함수의 값과의 차이의 제곱에 대한 함수를 이용하여 근사 2차 함수의 각 계수에 대한 편미분 값이 0이 되도록 만드는 계수를 구함으로써 2차 근사 함수를 구할 수 있다. 이는 식 (13)과 같이 정리되고 식 (13)의 역행렬을 구함으로써 원하는 2차 함수를 구할 수 있다.

$$G(\alpha) = \sum_{i=1}^{n} \left[\hat{f}(x_i) - y_i\right]^2 \tag{11}$$

$$\frac{\partial G(\alpha)}{\partial a_1} = 0, \frac{\partial G(\alpha)}{\partial a_2} = 0, \frac{\partial G(\alpha)}{\partial a_3} = 0 \tag{12}$$

$$A\alpha = \beta \tag{13}$$

식 (13)에서 $A = \begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{bmatrix}$ 이고

$\alpha = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \beta = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \end{bmatrix}$ 이다.

2차 최소자승 근사법에서는 선형 최소자승 근사법과 마찬가지로 파라메터의 첫 번째 값과 마지막 값에서 원래의 신호와의 차이 때문에 자연스럽게 연결되지 못하는 경계점이 발생하게 된다. MLLSA 기법과 유사하게 <그림3>과 같이 연결되는 파라메터의 첫 번째 값과 마지막 값이 그 전 혹은 그 후의 값들과 큰 차이가 나지 않는 일정 범위 안에 들어오도록 값을 수정해 준 후 변경된 첫 번째 값과 마지막 값 그리고 원래의 중간 값의 세 개의 값으로부터 2차 보간법을 이용하여 2차 함수를 수정해 주는 수정된 2차 최소자승 근사법(Modified Quadratic Least Square Approximation, MQLSA)을 사용하면 원래 신호와의 경계를 더욱 자연스럽게 연결할 수 있다.

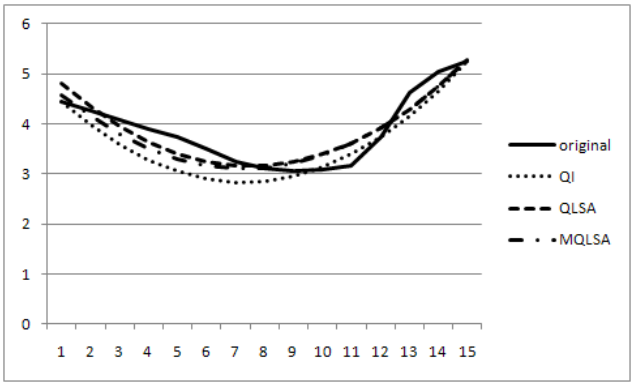

그림 5. 2차 함수를 이용해 연결된 파라메터
Figure 5. Illustration of parameters concatenated using 2th-order polynomial functions

<그림5>는 한영 경계구간에서 원래의 파라메터와 2차 함수를 이용한 각각의 연결 기법을 이용해 변화된 파라메터 값들을 보여준다. 2차 함수를 이용해 연결할 경우 선형 함수에 비해 원래의 신호와 비슷한 파형을 그리며 파라메터가 연결되는 것을 볼 수 있다. QLSA나 MQLSA의 경우 원래 신호의 정보를 기반으로 파형을 2차 함수를 만들어주기 때문에 신호의 파형에 큰 변화가 없는 것을 확인할 수 있다.

4. 실험 및 평가

4.1 실험 환경

영어 합성기를 훈련하기 위한 음성 DB는 CMU에서 제공하는 총 52분 분량의 1032문장을 사용하였고 모든 문장은 16 kHz로 샘플링되었으며 16 bit의 샘플 사이즈를 가진다. 한글 합성기를 훈련하기 위한 음성 DB는 ETRI에서 제공한 합성용 DB에서 약 100분 분량의 1000문장을 사용하였고 마찬가지로 모든 문장은 16 kHz로 샘플링되었으며 16 bit의 분해능을 가졌다. 문장 DB로부터 피치와 스펙트럼 값을 추출해내기 위한 환경설정은 다음과 같다.

- Frame length: 400 sample (25 ms)
- Frame period: 80 samples (5 ms)
- Window: Hamming window
- Pitch limit: 80~350
- MFCC order: 24

실험을 위해 사용된 프로그램은 HMM 훈련을 위해 패치된 HTK(Hunt & Black, 1996)와 SPTK(Yoshimura, Tokuda, Masuko, Kobayashi, & Kitamura, 1998)를 사용하였고 훈련된 HMM으로부터 음성을 합성하기 위해 hts_engine 2.0(Zen, Nose, Yamagishi, Sako, Masuko, Black, & Tokuda, 2006) 프로그램을 수정하여 사용하였다.

실험에 사용된 문맥종속모델은 총 5개의 state로 구성되어있고 한영 경계구간에서는 전 음소의 마지막 2개 state를 다음 음소의 처음 2개 state를 연결하여 경계구간을 정하였다.

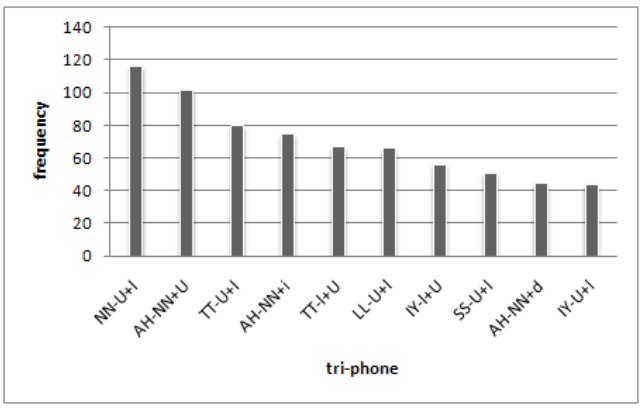

그림 6. 한영 혼용 문장의 tri-phone 빈도수
Figure 6. Histogram of tri-phone used in mixed-lingual sentences

실험에 사용한 문장은 Tabloid 한영 혼용신문에서 한글과 영어가 혼용되어 사용되는 문장 중 빈도수가 높은 것을 선택하였다. 문장 선택을 위해 Tabloid 인터넷 홈페이지(http://www.tabloid.com)에서 한글과 영어가 혼용되어 사용된 10,000개의 문장을 임의로 추출한 후 한글과 영어가 연속되어 사용되는 경우의 빈도수를 측정하여 가장 많이 사용되는 10개의 tri-phone이 포함되는 문장을 선택하였다. <그림6>은 한영 혼용 문장에서의 한글과 영어의 경계 간 tri-phone 빈도수를 나타낸 것이다. 빈도수가 가장 높은 10개의 tri-phone은 모든 경우에 한글에서 영어로 바뀌는 부분의 경계라는 것을 알 수 있고 그 중에 영어의 명사 다음 한글의 조사가 붙어있는 형식의 문장이 많이 사용된다는 것을 알 수 있다. 이러한 사실에 기초하여 영어의 명사 다음에 한글의 조사가 붙어있는 형식의 17개의 문장을 만들어 실험에 사용하였다.

4.2 평가 결과

본 논문에서는 MOS (Mean Opinion Score) 측정 방법과 선호도 테스트(Preference Test)를 진행하였다. MOS 측정방법은 모든 방법에 대해 평가를 수행하였고 선호도 테스트에서는 MOS 측정방법에서 높은 수치를 기록한 기법에 대해 추가적으로 수행하였다. 모든 성능측정은 17개의 선택된 문장을 이용하여 청각에 문제가 없는 총 8명의 2,30대 남녀 청취자들에 의해 이루어졌다.

<그림7>은 MOS 테스트에 대한 결과를 보여준다. 결과를 통해 알 수 있듯이 선형 최소자승 근사법(LLSA)을 제외한 대부분의 연결기법에서 단순 연결 기법에 비해 좋은 성능을 보여주었다. 또한 수정된 선형 최소자승 근사법(MLLSA)과 2차 보간법(QI)을 이용한 연결기법에서 가장 좋은 성능 결과를 보여주고 있다. 반면에 MQLSA는 MLLSA와 달리 QI 기법에 비해 좋지 않은 성능을 보여주었는데 이는 파라메터의 처음과 마지막 값의 불연속적인 부분 때문에 음질이 저하된 것으로 예상된다. 또한 QI 기법에서는 2차 함수를 이용하여 충분히 경계구간의 파라메터를 연결해주고 있기 때문에 QLSA 기법을 이용하여 기존의 파라메터를 고려하지 않아도 됨을 보여준다.

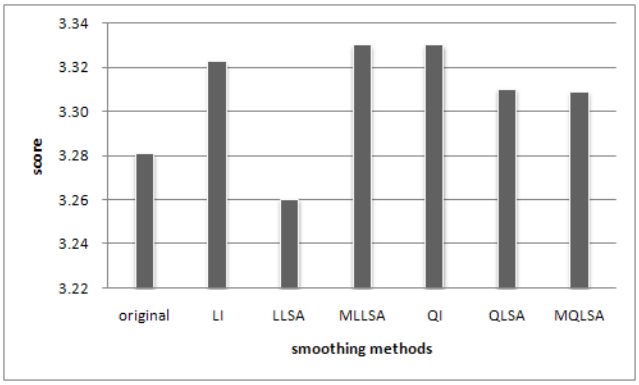

그림 7. 음질 평가 (MOS) 결과
Figure 7. Speech quality test (MOS) result

표 3. 선호도 테스트 결과
Table 3. Preference test result

Method	No Smoothing	MLLSA	No Smoothing	QI
Preference	40.0%	60.0%	34.8%	65.2%

<표3>은 가장 좋은 성능을 보인 수정된 선형 최소자승 근사법(MLLSA)과 2차 보간법(QI)에 대한 추가적인 선호도 테스트의 결과이다. 성능의 결과에 따르면 수정된 선형 최소자승 근사법으로 연결된 음성의 경우 단순히 연결된 원음(No Smoothing)에 비해 60.0%의 선호도를 가지고 2차 보간법으로 연결된 음성의 경우 단순히 연결된 원음에 비해 65.2%의 선호도를 보여준다. 이는 단순히 한글과 영어의 파라메터를 연결하는 것보다 연결되는 경계에서의 파라메터를 적절한 근사 함수를 사용하여 연결하는 것으로 불연속적인 음성을 자연스럽게 연결할 수 있음을 보여준다. 또한 선형 함수를 이용하여 파라메터를 연결하는 것보다는 비선형 함수를 이용하여 연결해주는 것이 더 좋은 음질의 음성을 합성할 수 있음을 보여준다.

5. 결론

한영 혼용 음성합성 시스템에서 한글과 영어간의 경계에 발생하는 불연속점을 단순히 연결하면 영어와 한글 각각에 의해 훈련된 HMM에 정의되지 않은 레이블 때문에 서로 다른 언어에 대한 음소를 고려해주지 못하게 된다. 이를 극복하기 위해 불연속적으로 이어지는 한영 음소에 해당하는 파라메터를 특정 연결 기법을 이용하여 연결해주게 되면 단순히 연결된 파라메터에 의해 합성된 소리보다 자연스러운 합성음을 만들어줄 수 있게 된다.

본 논문에서 사용한 파라메터의 첫 번째 값과 마지막 값의 구간을 고려한 수정된 선형 최소자승 근사법(MLLSA)과 파라메터의 첫 번째 값과 마지막 값 그리고 한영의 경계점을 잇는 2차 보간법(QI)을 사용하여 파라메터를 연결하면 한영 혼용 음성합성시 단순히 파라메터를 연결하는 것보다 자연스러운 합성음을 얻을 수 있을 것으로 기대한다.

참고문헌

Black, A. W. & Taylor, P. (1997). “CHATR: A generic speech synthesis system”, *Proc. of the 15th Conference on Computational Linguistics*, pp. 983-986.

Black, A. W., Zen, H. & Tokuda, K. (2007). “Statistical parametric speech synthesis”, *Proc. of ICASSP*, pp. 1229-1232.

Donovan, R. E. & Woodland, P. C. (1995). “Automatic speech synthesizer parameter estimation using HMMs”, *Proc. of ICASSP*, pp. 640-643.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*, Kluwer Academic Publishers.

Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M. & Pitrelli, J. (2004). “A corpus based approach to expressive speech synthesis”, *Proc. of 5th ISCA Workshop on Speech Synthesis*, pp. 79-84.

Fukada, T., Tokuda, K., Kobayashi, T. & Imai, S. (1992). “An adaptive algorithm for mel-cepstral analysis of speech”, *Proc. of ICASSP*, Vol. 1, pp. 137-140.

Hunt A. & Black, A. W. (1996). “Unit selection in a concatenative speech synthesis system using a large speech database”, *Proc. of ICASSP*, pp. 373-376.

Kim, S. J., Kim, J. J. & Hahn, M. (2006). “Implementation and evaluation of an HMM-based Korean speech synthesis system”, *IEICE Transactions on Information and Systems*, Vol. E89-D, No. 3, pp. 1116-1119.

Kim, S., Lee, Y. & Hirose, K. (2002). “A new Korean corpus-based text-to-speech system”, *International Journal of Speech Technology*, Vol. 5, No. 2, pp. 105-116.

Kim, S., Lee, Y. & Hirose, K. (2001). “Unit generation based on phrase break strength and pruning for corpus-based text-to-speech”, *ETRI Journal*, Vol. 23, No. 4, pp. 168-175.

Ling, Z. H., Wu, Y. J., Wang, Y. P., Qin, L. & Wang, R. H. (2006). “USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method”, *Blizzard Challenge Workshop*.

Odell, J. J. (1995). *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. Dissertation, Cambridge University.

Shinoda, K. & Watanabe, T. (2000). “MDL-based context-dependent subword modeling for speech recognition”, *Acoustical Science and Technology*, Vol. 21, No. 2, pp. 79-86.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashiand, T. & Kitamura, T. (2000). “Speech parameter generation algorithms for HMM-based speech synthesis”, *Proc. of ICASSP*, Vol. 3, pp. 1315-1318.

Tokuda, K., Masuko, T., Miyazaki, N. & Kobayashi, T. (1999). “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling”, *Proc. of ICASSP*, Vol. 1, pp. 229-232.

Traber, C., Huber, K., Nedir, K., Pfister, B., Keller, E. & Zellner, B. (1999). “From multilingual to polyglot speech synthesis”, *Proc. of Eurospeech*, pp. 835-838.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (1998). “Duration modeling in HMM-based speech synthesis system”, *Proc. of ICSLP*, Vol. 2, pp. 29-32.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (1999). “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis”, *Proc. of Eurospeech*, pp. 2347-2350.

Yu, J., Zhang, M., Tao, J. & Wang, X. (2007). “A novel

HMM-based TTS system using both continuous HMMs and discrete HMMs", *Proc. of ICASSP*, pp. 709-712.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. & Tokuda, K. (2006). "The HMM-based speech synthesis system (HTS) version 2.0", *Proc. of the sixth ISCA Workshop on Speech Synthesis*, pp. 294-299.

Zen, H., Toda, T., Nakamura, M. & Tokuda, K. (2007). "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005", *IEICE Transactions on Information and Systems*, Vol. E90-D, No. 1, pp. 325-333.

• **양종열 (Yang, Jong-Yeol)**
삼성전자 DMC
경기도 수원시 매탄동
Tel: 010-4475-7896
Email: jy50.yang@samsung.com
관심분야: 음성합성
2008~2010 광주과학기술원 석사과정 졸업
2010~현재 삼성전자 DMC 사원

• **김홍국 (Kim, Hong Kook)** 교신저자
광주과학기술원 정보통신공학과
광주광역시 북구 오룡동 1번지
Tel: 062-970-2228 Fax: 062-970-2204
Email: hongkook@gist.ac.kr
관심분야: 음성 및 오디오 처리, 음성인식
2003~현재 광주과학기술원 정보통신공학과 교수