말소리와 음성과학 제6권 제4호 (2014.12.31)
ISSN 2005-8063, pp. 133~139
www.speechsciences.or.kr
http://dx.doi.org/10.13064/KSSS.2014.6.4.133

포만트 공간에서의 주파수 변환을 이용한 이중 언어 음성 변환 연구

Bilingual Voice Conversion Using Frequency Warping on Formant Space

채 의 근[1] · 윤 영 선[2] · 정 진 만[3] · 은 성 배[4]
Chae, Yi-Geun · Yun, Young-Sun · Jung, Jin Man · Eun, Seongbae

ABSTRACT

This paper describes several approaches to transform a speaker's individuality to another's individuality using frequency warping between bilingual formant frequencies on different language environments. The proposed methods are simple and intuitive voice conversion algorithms that do not use training data between different languages. The approaches find the warping function from source speaker's frequency to target speaker's frequency on formant space. The formant space comprises four representative monophthongs for each language. The warping functions can be represented by piecewise linear equations, inverse matrix. The used features are pure frequency components including magnitudes, phases, and line spectral frequencies (LSF). The experiments show that the LSF-based voice conversion methods give better performance than other methods.

Keywords: Bilingual Voice Conversion, Formant Space, Frequency Warping, LSF-based Voice Conversion

1. 서론

음성 변환(voice conversion) 기법은 발성 화자의 개인성을 변환하는 것으로서, 원 화자(source speaker)의 특성을 나타내는 음성 특징을 목적 화자(target speaker)의 음성 특징으로 변환하는 것을 말한다[1]. 대부분의 음성 변환연구는 동일한 언어 사용자간의 화자 변환 등을 목적으로 한다. 즉, 동일한 언어 환경에서 남성 화자를 다른 남성 화자로 변환하거나, 남성 화자를 여성 화자, 또는 여성 화자를 남성 화자로 변환한다. 또한 음성 변환은 텍스트 입력을 음성으로 변환하는 음성 합성(TTS; Text-To-Speech) 장치에 포함되거나 음성 합성 장치의 일부 기법으로 사용된다. 본 연구에서는 일반적인 음성 변환 연구와 달리 음성 합성 장치를 사용하기 어렵거나, 음성 합성의 결과만을 이용한 환경에서 서로 다른 언어의 발성화자간의 음성 변환을 연구하였다.

음성 합성 장치에 의하여 한국어와 영어를 서비스 하는 경우, 이중 언어 환경에서 자라거나 두 개의 언어를 자연스럽게 발성하는 화자의 음성을 녹음해야 한다. 하지만, 이중 언어 화자를 구하기 어렵거나 비용이 증가하는 경우가 많아 단일 언어 화자에 의하여 음성 합성 시스템을 구축하는 경우가 많다. 이 경우, 영어 합성음이 한국어 화자에 의하여 발성된 것처럼 청자가 느끼게 하거나, 한국어 화자의 합성음과 영어 화자의 합성음이 동일한 화자에 의하여 발성되는 것처럼 느끼도록 하는 것이 연구의 동기이다.

음성 특징을 결정짓는 것은 여러 특징들이 존재하며, 어떤 한 특징에 의하여 모든 개인성 정보를 표현하기는 힘들다고 알려져 있다. 이들 특징들 중에서 포만트(formant)는 음성과 화자(발성자)의 개인성을 잘 표현할 수 있는 중요한 특징 변수중의 하나로 여겨지고 있다[1,2]. 포만트 주파수를 이용하여 음성 변환에 적용하는 연구는 많이 진행되고 있다. 대표적인 연구로는 부공간 코드북 변환[1], 신경회로망을 이용한 변환[3], 성도

1) 공주대학교, ygchae@kongju.ac.kr
2) 한남대학교, ysyun@hnu.kr, 교신저자
3) 한남대학교, jmjung@hnu.kr
4) 한남대학교, sbeun@hnu.kr

본 논문은 2014년도 한남대학교 학술연구조성비 지원에 의하여 연구되었습니다.

접수일자: 2014년 11월 15일
수정일자: 2014년 12월 9일
게재결정: 2014년 12월 13일

길이 정규화(VTLN; vocal tract length normalization)를 이용한 방법[4,5] 등이 있다.

VTLN 기법은 화자 종속적인 성도(vocal tract) 길이를 정규화하기 위하여 음성 주파수를 위상과 크기에 대하여 변환하는 기법이다. VTLN은 음성 인식에서 많이 사용되는 방법으로 화자의 개인성을 제거하여 인식 성능을 높이는 쪽으로 사용되어 왔으며 [6], 음성 변환에 적용되어 원 화자의 음성을 목적 화자에 의하여 발성된 것처럼 변환하는 연구가 소개되었다[7]. VTLN에 기초를 둔 주파수 변환 방법은 이중 선형 변환[8], 부분 선형 변환[4,5] 등과 같이 다양한 방법으로 구현되고 있다. 이들 중 몇 연구는 동일한 언어에 대하여 텍스트 종속 조건에서 실험이 진행되었으며[1,3,4,5,8], 다른 몇 연구는 서로 다른 언어에서 텍스트 독립 방법으로 연구가 진행되었다[7,9].

기존의 연구들은 음성 변환에 사용되는 다양한 매개 변수를 조정하여 진행하였으며, 대부분 음성 변환 과정을 음성 합성 시스템 속에 구현하였다. 그러나 본 연구는 음성 합성 시스템의 내부 알고리즘이나 단위 선택의 과정을 변형하지 않고, 음성 합성의 결과를 직접 주파수 변환하여 음성을 변환하고자 하였다. 음성 합성 시스템과 독립적인 음성 변환 방법은 음성 합성 시스템에서 사용되는 다양한 조절 요인들을 사용할 수 없기 때문에 문제 해결이 더욱 어려워진다. 연구의 제약조건으로 인하여 본 논문에서는 포만트 공간을 이용한 단순하면서도 직관적인 주파수 변환 방법을 이용하여 음성 변환을 시도하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구에서 사용하는 포만트 공간에 대하여 설명하고, 3장에서는 포만트 공간에서의 기존 주파수 변환 방법을 설명하고, 선형 스펙트럼 주파수를 이용한 음성 변환 방법을 제안한다. 4장에서 제안된 방법의 실험 및 결과 분석을 논의한 후, 5장에서 본 연구에 대한 요약 및 결론으로 끝을 맺는다.

2. 포만트 공간

성도는 공진(resonance)을 갖는 선형 필터로 모델링 될 수 있다. 유성음의 경우 성도의 개인 차, 혀의 위치 변화 등에 의하여 다른 공진 특성을 갖는다. 유성음에 대한 구강과 비강의 주 공진 주파수를 1차, 2차 포만트 또는 F1, F2로 표현하고 있다. 포만트의 주파수 특성은 유성음 발성 시 혀의 위치, 성도의 모양에 의하여 구분되며, 유성음의 질(quality)이나 음색(timbre) 등을 결정한다[10]. 일반적으로 F1과 F2에 의하여 대부분의 모음을 구별할 수 있으며, F3는 음소의 질 (phonemic quality), F4 이상의 포만트는 음질(voice quality)에 관여한다고 알려져 있다[17].

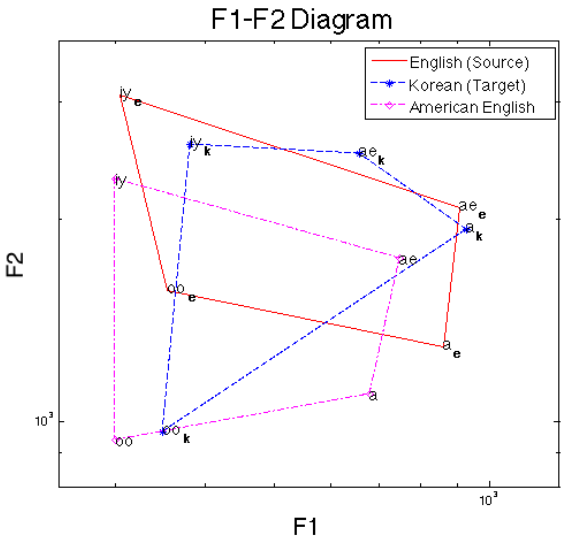

그림 1. 전형적인 미국 영어, 영어와 한국어 화자의 4개 대표 음소(/iy, oo, a, ae/)에 의한 F1-F2 다이어그램 비교

Figure 1. F1-F2 diagrams for typical American English, English and Korean speeches for four representative monophthongs (/iy, oo, a, ae/).

화자 또는 언어에 대한 유성음의 발성 분포를 살펴보기 위하여 F1-F2의 주파수 위치를 표현한 <그림 1>과 같은 F1-F2 다이어그램이 널리 이용된다. <그림 1>은 미국 영어의 전형적인 F1-F2 다이어그램[10]과 본 연구에서 사용한 한국어 화자와 영어 화자의 F1-F2 다이어그램을 표시한다. 일반적으로 포만트를 추출하는 방법은 LPC 방식이나 AR 방식을 많이 도입하나, 본 연구에서는 LPC 방식에 기초한 LSF 방식을 적용하기 때문에 LPC 방식을 적용하였다. 한국어와 영어의 음성은 모음이 포함된 단어를 생성한 후 해당 음모음을 분할한 후 LPC 방식을 적용하여 포만트를 계산하였다.

일반적으로 F1-F2 다이어그램의 경우 입 모양과 혀의 위치 등을 고려하여 9~10개의 음소에 대한 평행사변형으로 표시하나, 본 연구에서는 영어와 한국어의 경우 일대일 정합이 정확하지 않기 때문에 평행사변형의 꼭지점에 해당하는 4개의 음소(/iy, oo, a, ae/)를 지정하고, 4개의 음소에 의한 다이어그램으로 단순화시켰다.

3. 주파수 변환

본 연구는 F1-F4의 포만트 벡터를 이용하여 화자의 발성 분포를 모델링하고, 원 화자와 목적 화자의 포만트 공간을 이용한 기존의 음성 변환 방법을 개선하였다[11]. 기존의 연구 방법이 가중 선형 주파수 변환에 의하여 크기나 위상, 또는 주파수 공간에 속한 원 화자의 음성을 대상으로 변환을 하였다면, 본 연구에서는 선형 스펙트럼 주파수(LSF; Line Spectral

Frequencies)를 이용하며, 행렬 변환에 의한 직접 변환을 고려하였다.

본 장에서는 기존의 연구인 가중 선형 주파수 변환을 요약하고 선형 스펙트럼 주파수를 이용한 주파수 변환 및 역행렬을 이용한 음성 변환 방법을 제안한다. 제안된 방법은 선행 연구가 진행된 후 방법론을 개선하여 직접 주파수를 변환하는 대신, 선형 스펙트럼 주파수를 이용하여 주파수의 크기와 위상을 같이 고려한 방법이다[14].

3.1 가중 주파수 변환

<그림 1>에서 살펴본 바와 같이 원 화자(영어)와 목적 화자(한국어)의 포만트 공간은 다른 분포를 보인다. 기존의 연구[9,12,13]와 다르게 선행 연구에서는 텍스트에 독립적인 음성 변환 방법을 사용하였다. 텍스트 정보를 이용하지 않기 때문에 기준이 되는 음향학적 특징이 필요하였으며, 4개의 대표 음소(/iy, oo, a, ae/)를 선정하여 음성 변환의 지표로 삼았다. 가중 주파수 변환의 기본 알고리즘은 <표 1>에 개략적으로 정리하였다.

표 1. 가중 주파수 변환의 기본 알고리즘
Table 1. Basic algorithm of weighted frequency warping

1. 원 화자의 음성을 프레임 단위로 분할한다.
2. 유성음인 경우, 주파수 공간에서의 입력 프레임의 상대 위치를 계산한다. 상대 위치는 원 화자의 포만트 공간을 구성하는 4개의 대표 음소와의 거리 값으로부터 가중치로 계산된다.
3. 입력 프레임을 목적 프레임으로 변환하는 함수는 원 화자의 4개 대표 음소와 목적 화자의 대응되는 4개 대표 음소의 변환함수의 조합으로 계산된다. 입력 프레임을 목적 프레임으로 변환하기 위하여 2.에서 구한 가중치를 이용하여 각각의 4개 음소의 변환함수의 가중합으로 주파수 변환함수를 계산하여 적용한다.
4. 3.에서 구한 목적 프레임을 이용하여 목적 화자의 음성을 재구성한다.

음성 변환 함수를 계산하기 위하여 원 화자의 음성(영어)과 목적 화자(한국어)의 음성에서 선형 예측 분석(LPC) 방식에 의하여 포만트를 4차까지 계산하였다. 영어와 한국어의 경우 모음 음소가 정확히 대응되지 않기 때문에 포만트 공간을 잘 표현할 수 있는 4개의 음소를 선택하였으며, 그 4개의 음소들 간의 부분 선형 정합 함수는 <그림 2>와 같다.

원 화자의 음성이 입력되면 각 프레임 (원 음성 프레임)은 원 화자의 포만트 공간에서 4개의 대표 음소와의 상대적인 거리를 가중치로 상대 위치로 표현된다. 목적 화자의 음성을 구성하기 위한 목적 음성 프레임은 원 음성 프레임에 음성 변환함수를 적용하여 구할 수 있다. 음성 변환함수는 원 음성 프레임의 상대적인 위치를 나타내는 가중치를 각각 4개의 원화자의 음소로부터 목적화자의 음소로 변환하는 <그림 2>의 부분 정합 함수에 적용하여 가중 합으로 계산된다. 이 경우 원 화자나 목적 화자의 경우 4개 음소로 구성되는 포만트 공간을 벗어나는 프레임이 존재할 수 있는데 이의 처리가 음성의 질에 영향을 준다.

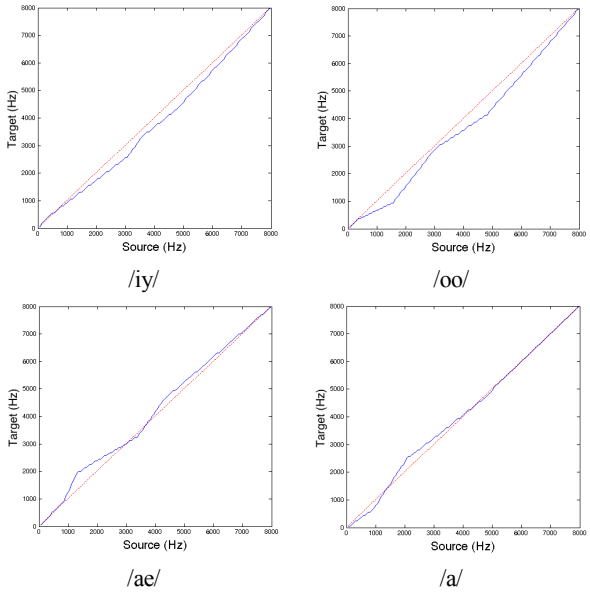

그림 2. 4개의 대표 음소에 대한 영/한 부분 정합 함수
Figure 2. Piecewise warping functions for four representative monophthongs between English and Korean languages

3.2 선형 스펙트럼 주파수 변환

선행 연구에서는 포만트 공간에서의 가중 선형 함수에 의하여 직접 주파수 변환을 시행하였다. 직접 주파수 변환의 장점은 음성 신호를 주파수로 변환한 후, 별도의 변환 과정을 거치지 않고 음성을 변환할 수 있기 때문에 변환 함수가 제대로 얻어질 수 있다면 좋은 음질을 기대할 수 있다는 점과, 원 화자의 주파수 크기 (magnitude)만을 변환하고 위상 (phase) 정보는 원 화자의 정보를 사용할 수 있다는 점이다. 그러나 변환된 주파수 크기 정보만을 변환할 경우, 목적화자의 개인성 정보를 충분히 반영하지 못한다는 단점이 존재한다.

따라서 본 연구에서는 음성 분석 및 합성, 신호처리 등에서 많이 사용하는 선형 예측 분석법(linear prediction analysis)을 이용하여 주파수의 직접 변환 대신 스펙트럼 포락(spectral envelope)를 나타내는 계수를 이동(변환)시켜 목적 화자의 음성을 얻는 방법을 제안한다. 기존 연구의 주파수 직접 변환 방법과 선형 스펙트럼 주파수 변환에 의한 음성 변환 방법을 <그림 3>과 같이 비교하였다.

기존 방법의 경우 입력 음성에 대하여 주파수로 전환한 후,

부분선형함수에 의하여 주파수 변환을 실시한 후, 다시 역변환을 실시하여 목적 화자의 음성을 생성한다. 반면 선형 스펙트럼 주파수 변환은 입력 음성에 대하여 선형예측분석에 의하여 LSF 변수를 얻는다. LSF 변수는 주파수 공간에 존재하기 때문에, 기존 연구에서 적용하였던 부분 선형 변환을 적용하여 LSF 변수를 변환한 후, 역과정을 거쳐 목적 화자의 음성을 생성한다. 이해를 돕기 위하여 각 단계의 역변환을 $^{-1}$으로 표현하였다.

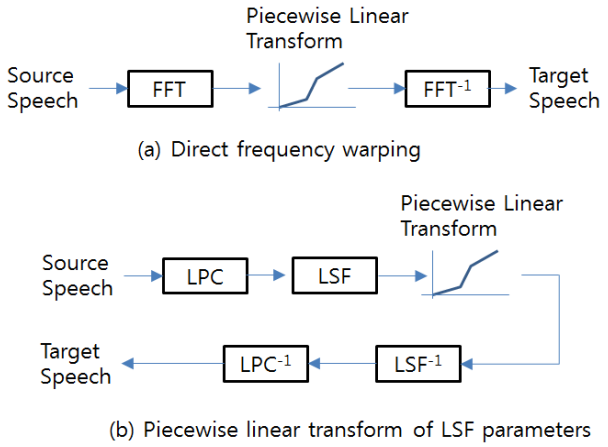

그림 3. 직접 주파수 변환과 선형 스펙트럼 주파수 변환 비교
Figure 3. Comparison between direct frequency warping and LSF parameter transform approaches.

주파수 스펙트럼과 14차 선형 예측 방법에 의한 주파수 포락 및 선형 스펙트럼 주파수의 특성을 <그림 4>에서 도식화하였다. 그림에서 살펴본 바와 같이 LPC 포락은 주파수 스펙트럼의 피크를 따라가면서 모델링하고 있고, LSF 변수는 주요 특징 위치에 존재함을 알 수 있다. 본 연구에서는 LSF 변수의 위치를 선행 연구에서 사용하였던 부분 선형 함수를 이용하여 변경한 후, LPC의 역변환을 거쳐 목적 화자의 음성을 생성한다.

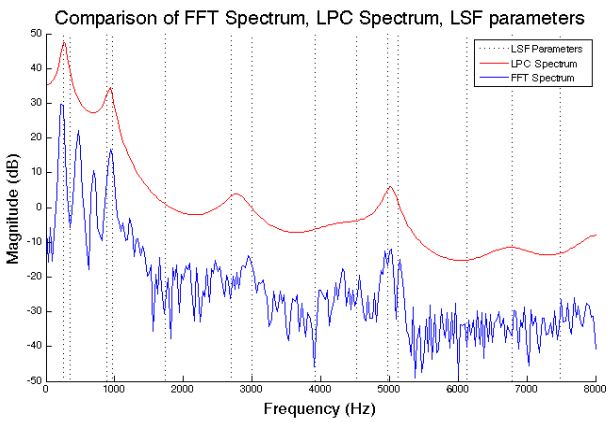

그림 4. FFT 스펙트럼, LPC 스펙트럼, LSF 변수 비교
Figure 4. Comparison of FFT spectrum, LPC spectrum, LSF parameters

3.3 행렬 연산에 의한 변환

선행 연구의 주파수 직접 변환이나 제안한 선형 스펙트럼 주파수 변환에서는 부분 선형 변환 함수를 사용하였다. <그림 1>의 F1-F2 다이어그램의 4음소에 해당하는 부분 선형 변환 함수의 가중 합으로써 입력 음성(원 화자)의 주파수 변환 함수를 결정한 후, 각 목적 화자의 포만트 공간에서 목적 음성프레임을 변환하는 방법이다. 즉, 식 (1)에 의하여 목적 화자의 음성 변환 함수 $\tau(\boldsymbol{f})$에 의하여 원 화자의 음성 프레임 $\boldsymbol{f}_s$는 목적 화자의 음성 프레임 $\boldsymbol{f}_t$로 변환한다.

$$\tau(\boldsymbol{f}) = \sum_{k=1}^{N} \boldsymbol{W}(\boldsymbol{f}, \boldsymbol{S}_k) \cdot \boldsymbol{T}(\boldsymbol{S}_k, \boldsymbol{T}_k)$$
$$\boldsymbol{f}_t = \tau(\boldsymbol{f}) \cdot \boldsymbol{f}_s \quad (1)$$

여기에서 $\boldsymbol{W}(\boldsymbol{f}, \boldsymbol{S}_k)$는 원 화자의 음성이 포만트 공간에서 차지하는 상대적 위치를 가중치로 표현한 값이다. $\boldsymbol{W}(\boldsymbol{f}, \boldsymbol{S}_k)$는 프레임 $\boldsymbol{f}$와 원화자의 대표음소 $\boldsymbol{S}_k$와의 거리$D(\boldsymbol{f}, \boldsymbol{S}_k)$를 모든 음소와의 거리 $\Sigma_1^N D(\boldsymbol{f}, \boldsymbol{S}_k)$로 정규화한 값이며, 입력 프레임 $\boldsymbol{f}$가 k개의 대표 음소로 구성된 포만트 공간을 벗어날 경우 외삽(extrapolation)으로 계산한다. $T(\boldsymbol{S}_k, \boldsymbol{T}_k)$는 F1-F2 다이어그램에 의하여 원 화자 음성의 k번째 음소의 포만트 $\boldsymbol{S}_k$를 대응되는 목적 화자의 k번째 음소의 포만트 $\boldsymbol{T}_k$로 변환하는 함수이다. 즉, 원화자의 대표음소와 목적화자의 대표음소간의 주파수 변환함수(<그림 2>의 부분 선형 함수)를 나타낸다. 본 연구에서는 원 화자 음성과 목적 화자 음성의 포만트 공간을 4개의 대표 음소로 구성하였기 때문에 사용되는 대표 음소의 수는 $N = 4$가 된다.

제안하는 방법은 음성 변환 함수 $\tau(\boldsymbol{f})$를 가중치의 합으로 계산하지 않고 대표 음소의 포만트 특성으로 표현된 행렬의 역 행렬 값으로 음성 변환 함수를 계산하는 방법이다.

$$[\boldsymbol{T}_1 \cdots \boldsymbol{T}_k]^T = [\boldsymbol{S}_1 \cdots \boldsymbol{S}_k]^T \boldsymbol{W}$$
$$\boldsymbol{W} = \left[[\boldsymbol{S}_1 \cdots \boldsymbol{S}_k] [\boldsymbol{S}_1 \cdots \boldsymbol{S}_k]^T \right]^{-1} \cdot [\boldsymbol{S}_1 \cdots \boldsymbol{S}_k] [\boldsymbol{T}_1 \cdots \boldsymbol{T}_k]^T \quad (2)$$

식 (2)와 같이 원 화자의 k개의 대표 음소와 목적 화자의 k개 대표음소가 변환함수 $\boldsymbol{W}$로서 표현된다고 하면, 음성 변환 함수는 $\tau(\boldsymbol{f}) = \boldsymbol{W}$와 같이 표현될 수 있다. 따라서 역행렬 방식에 의한 음성 변환함수 $\tau(\boldsymbol{f})$는 프레임의 특성에 상관 없이 원 화자의 음성 프레임 $\boldsymbol{f}_s$을 목적 화자의 음성 프레임 $\boldsymbol{f}_t$로 변환할 수 있게 된다.

4. 음성 변환 결과 및 분석

본 장에서는 제안된 방식의 유효성을 확인하기 위한 실험

조건과 실험 결과를 설명하고 분석한다.

4.1 음성 데이터베이스

본 연구에서는 보이스웨어의 TTS 시스템[15]에서 생성된 한국어와 영어 여자 음성을 사용하였다. 한국어와 영어는 2003과 2008년에 녹음된 음성을 사용하였으며, 한국 표준어와 미국 펜실베니아 지역성을 띄고 있다. 한국어 음성과 영어 음성은 피치 동기형 중첩 합성방식(PSOLA; Pitch Synchronous Overlap and Add)으로 합성되었으며, 16kHz로 생성되었다.

4.2 실험 및 평가

본 실험에 앞서 입력 문장에 대한 포만트 공간에서의 각 프레임 위치를 <그림 5>에 도식화하였다. <그림 5>에서 살펴본 바와 같은 많은 프레임이 포만트 공간 밖에 존재하는 것을 알 수 있다.

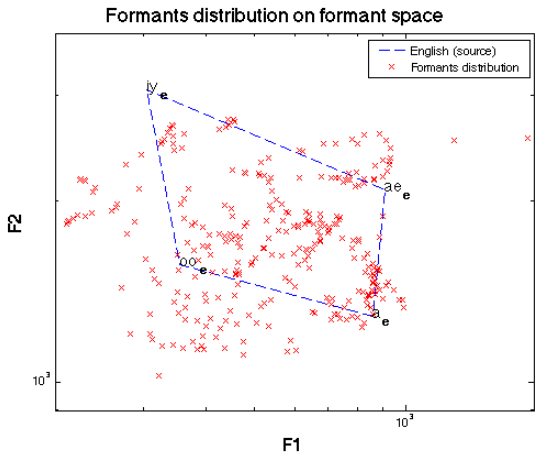

그림 5. 영어 화자 (원 화자)의 포만트 분포
Figure 5. Formant distribution of English (source) speaker

이를 고려하여, 입력 프레임이 포만트 공간 밖에 위치하면 외삽 방법에 의하여 포만트 가중치를 근사하였다. 또한, 실험 환경에서 포만트 공간 안에 프레임이 존재하는 경우에는 음성 변환을 실시하고 포만트 공간 밖에 존재하면 음성 변환을 하지 않는 조건을 추가하였다 (FS: formant space 조건).

선행 연구[11,14]의 경우에서는 성인 화자 6명이 목적 화자의 음성(한국어)과 원 화자의 음성(영어)을 모두 청취하고, 변환된 음성이 영어 화자에 의해 발성된 것처럼 느끼는지(A), 한국어 화자의 음성에 가까운지(B), 구별이 안 되는지(X)를 체크하였으며, 음질을 MOS 방법을 적용하여 5점 척도로 평가하였다. 그 평가결과는 <표 2>와 같다.

표 2. 선행연구의 예비 실험 결과 [11,14]
(A:원 화자 B:목적 화자, X:모름)
Table 2. Preliminary results of previous studies [11,14]
(A: source speaker, B: target speaker, X: none)

	ABX (target)	Voice quality (MOS)
Magnitude, Phase converted	50%	2.00
Only magnitude converted	50%	3.14
Only magnitude converted if it is in formant space	33%	4.29

본 연구에서는 각 적용 방법과 문장에 따른 효과를 분석하기 위하여 선행 연구와 달리 개별 문장에 대하여 ABX와 음질 평가를 진행하였으며, 실험 조건은 <그림 6>과 같이 총 10가지 방법으로 진행하였다. 실험 조건에서 (1)~(3) 방법은 선행 연구의 방법과 동일하다.

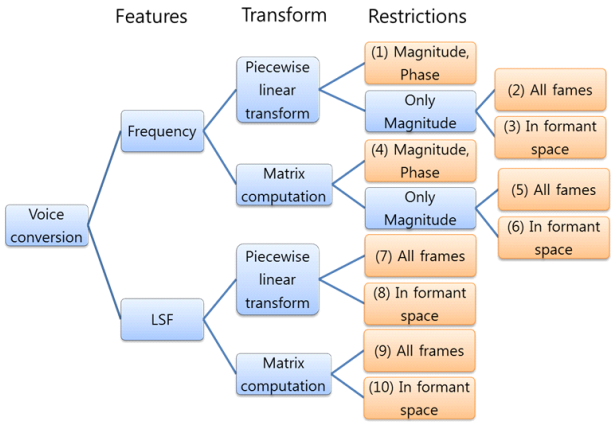

그림 6. 음성 변환 실험 조건
Figure 6. Conditions for experiments

실험 방법은 사용하는 음성 특징에 따라 주파수 직접 변환과 LSF로 구분하였으며 (Feature vs. LSF), 변환 함수의 종류에 따라 부분 선형 변환 함수와 행렬 역변환에 따라 구분하였다 (Piecewise linear transform vs. Matrix computation). 또한 입력 프레임을 목적 프레임으로 변환할 때 크기만 변환한 것인지, 위상까지 같이 고려하였는 지로 구분하였다 (Only Magnitude vs. Magnitude, Phase). 입력 프레임의 위치가 포만트 공간에 포함되었는지에 따라 다시 구분하였으며 (All frames vs. In formant space), 주파수 직접 변환 특징의 경우에는 선행 연구와의 비교를 위해 크기 정보만을 적용하여 포만트 공간의 포함 여부를 실험하였다.

실험 대상자는 목적 화자 음성(한국어), 원 화자 음성(영어) 3 문장, 그리고 이중 언어를 사용하는 화자(Jennifer Clyde)의

인터뷰 동영상[16]을 시청한 후 평가를 진행하도록 하였다. 실험에 사용된 문장은 <표 3>과 같다.

표 3. 실험에 사용한 문장
Table 3. Sentences are used for experiments.

Speaker	Sentences
Target	웹 애니웨어는 아도비 플래시가 설치된 파이어팍스 웹브라우저를 대상으로 개발이 되었습니다. (WebAnywhere has been developed targeting the Firefox web browser with Adobe Flash installed.)
Source	WebAnywhere has been developed targeting the Firefox web browser with Adobe Flash installed.
	While her husband Edward delights in her beauty, speed, and uncommon self-control, newborn Bella has never felt more alive.
	An earthquake rolled through a wide swath of Southern California late Monday morning but there were no immediate reports of damage.

대부분 20대의 남성들로 (40대 남성 1명, 20대 여성 1명, 나머지 20대 남성) 구성된 총 17명의 실험 대상자가 ABX 실험에 참여하였으며, 그중 13명이 음질 평가에 참여하였다. 실험 결과는 <표 4>와 같다. 실험에 사용된 특징에 따라 주파수(Freq.)와 14차 LPC 기반의 선형 스펙트럼 주파수(LSF), 변환 방법에 따라 부분선형변환(PLW), 행렬 연산(MAT), 특징의 제약조건에 따라 (Magnitude-Mg, Phase-Ph), 포만트 공간의 제약조건에 따라 (All, FS)로 구분하였다. LSF 특징의 경우 크기와 위상으로 분리되지 않기 때문에 모든 프레임 (All)과 포만트 공간(FS)에 포함된 프레임으로 구분하여 실험하였다.

실험결과 선행 연구와 달리 주파수 직접 변환한 경우 음성 변환이 제대로 이뤄지지 않았다고 평가하였으며 (영어 화자 음성을 영어 화자가 발성한 것으로 평가), LSF의 경우 선행 연구보다 음성 변환의 잘 이뤄졌다고 평가하였다(영어 화자 음성을 한국 화자가 영어로 발성한 것으로 평가). 또한, LSF 특징을 이용한 경우가 주파수 특징을 사용한 경우보다 음질 면에서 우수한 평가를 받았다.

주파수 직접 변환에서 우수한 음질의 경우, 원 화자의 음성이 목적 화자의 음성과 유사하게 변환되지 않았으며, LSF 특징을 이용한 경우 음질의 저하는 크지 않으면서 음성 변환이 잘 이뤄졌다고 평가되었다. 주파수 직접 변환의 경우 크기와 위상을 동시에 변환하는 경우 음질의 저하가 많이 발생하고, 위상은 원 화자의 위상을 사용하고, 크기 변환 또한 포만트 공간에 포함된 프레임만 허용하는 경우 음성 변환의 성능이 저하되어 음성 변환과 음질의 반비례관계가 형성된 것으로 판단한다. LSF 특징의 경우 선형 예측 분석에 의하여 시간 축에서 LPC 계수를 추출하고, LPC계수에서 계산된 LSF 변수의 변환/이동에 의하여 음성 변환을 실시하였기 때문에 주파수 직접변환에 의한 방법보다 음질은 유지하면서 음성 변환이 가능한 것으로 판단한다.

표 4. 실험 조건에 따른 ABX(변환 음성의 화자 선호도) 평가 결과 (A: 원 화자, B: 목적화자, X: 모름)
Table 4. ABX(converted speech preference) test along to experimental conditions
(A: source speaker, B: target speaker, X : none)

Conditions		ABX (%)			Voice quality (MOS)
		Eng	Kor	None	
(1)	Freq. PLW, Mg, Ph	41	33	25	1.85
(2)	Freq. PLW, Mg, All	**63**	33	4	2.62
(3)	Freq. PLW, Mg, FS	**61**	37	2	**3.62**
(4)	Freq. MAT, Mg, Ph	**49**	20	31	1.64
(5)	Freq. MAT, Mg, All	37	29	33	2.03
(6)	Freq. MAT, Mg, FS	**59**	37	4	**3.44**
(7)	LSF, PLW, All	39	**55**	6	**4.03**
(8)	LSF, PLW, FS	45	**51**	4	**3.49**
(9)	LSF, MAT, All	**45**	43	12	3.31
(10)	LSF, MAT, FS	**47**	45	8	**3.49**

행렬 연산에 의한 음성 변환 결과는 부분 선형 함수를 이용한 방법보다는 성능이 저하되었지만, 변환 과정이 간단하여 화자의 포만트 변이가 큰 경우에 사용할 수 있을 것이라고 판단한다. 특이할 점은 ABX 테스트에서 None의 비율이 높아졌다는 것이다. 즉, 원 화자의 음성과 목적 화자의 음성과 비교하였을 경우, 구별하기 어렵다는 평가가 증가하였다는 것이다.

5. 요약 및 결론

본 연구에서는 음성 합성기의 결과를 이용하여 다국어 음성 출력에 사용할 수 있는 음성 변환 연구를 수행하였다. 변환된 음성의 음질 및 인지도를 향상시키기 위하여 주파수 직접 변환, 행렬 역 변환에 의한 주파수 변환, LSF 특징 표현에 의한 음성 변환 등 여러 가지 방법을 사용하였다. 주파수 직접 변환의 경우 주파수 공간에서 크기와 위상의 적용 방법에 따라 음질 및 음성 변환의 성능 저하가 관측되었으며, LSF 방식의 경우 선형 예측 방법에 적용되는 계수에 직접적인 영향을 가해 음성 변환을 실시하였다. 실험 결과 LSF 특징에 기초한 변환 방법이 음질 면이나 인지도 면에서 좋은 결과를 보였다. 추후 연구로는 단일 화자인 경우에도 포만트 특성이 변화하기 때문

에, 현재의 단일 포만트 공간을 개선하여 통계적 방법을 적용한 포만트 공간을 고려한 음성 변환이 필요하다고 판단한다.

본 논문은 음성 합성기의 결과를 이용하여 연구되었으나, 화자의 포만트 정보는 동일한 음소라도 주위 음소나 감정 등에 의하여 변이가 심하기 때문에 통계적 모델링을 적용할 수 있다면 자연 음성에 적용할 수 있으며, 이 경우 몇 문장의 음성 샘플에 의하여 다중 언어 간의 음성 변환도 가능할 것으로 기대한다.

참고문헌

[1] Mizuno, H., Abe, M. (1995), Voice conversion algorithm based on peicewise linear conversion rules of formant frequency and spectrum tilt, Speech Communication, no. 16, pp. 153-164.

[2] Kuwabara H., Sagisaka Y. (1995), Acoustic characteristics of speaker individuality: Control and conversion, Speech Communication, no. 16, pp. 165-173.

[3] Narendranath M., Murthy H. A., Rajendran S., Yegnanarayna B. (1995), Transformation of foramnts for voice conversion using artificial neural networks, Speech Communication, no. 16, pp. 207-216.

[4] Sundermann D., Bonafonte A., Ney H. (2004), Time domain vocal tract length normalization, In Proc. of *IEEE In. Symposium on Signal Processing and Information Technology*, pp. 191-194.

[5] Errno D., Moreno A., Bonafonte A. (2010), Voice conversion based on weighted frequency warping, IEEE Tr. on Audio, Speech, and Language Processing, vol. 18, issue 5, pp. 922-1931.

[6] Pye D., Woodland P. C. (1997), Experiments in speaker normalization and adaptation for large vocabulary speech recognition, In Proc. of *IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pp. 1047-1050.

[7] Sundermann D., Ney H., Hoge H. (2003), VTLN-Based cross-language voice conversion, In Proc. of *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 676-681.

[8] Saheer L., Dines J., Garner P. N. (2012), Vocal tract length normalization for statistical parametric speech synthesis, IEEE Tr. on Audio, Speech, and Language Processing, vol. 20, issue 7, pp. 2134-2148.

[9] Sundermann D., Hoge H., Bonafonte A., Ney H., Black A., Narayanan S. (2006), Text-independent voice conversion based on unit selection. In Proc. of Int. Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 81-84.

[10] Huang X., Acero A., Hon H.-W. (2001), Spoken language processing – A guide to theory, algorithm, and system development, *Prentice Hall*

[11] Yun Y.-S., Ladner R. E. (2013), Bilingual voice conversion by weighted frequency warping based on formant space, LNCS 8082, pp. 137-144

[12] Sundermann D., Strecha G., Bonafonte A., Hoge H., Ney H. (2005), *Evaluation of VTLN-Based voice conversion for embedded speech synthesis*, Int. Proc. of Conference on Spoken Language Processing, pp. 3-6.

[13] Erro D., Moreno A., Bonafonte A. (2010), *Voice conversion based on weighted frequency warping*, IEEE Tr. on Audio, Speech, and Language Processing, vol 20. issue 7, pp. 2134-2148

[14] Y.-S. Yun (2013), Multilingual voice conversion using direct frequency warping, In Proc. of 2013 Korean Society of Speech Sciences Fall Conference, pp. 127-128
(윤영선 (2013), 주파수 직접 변환에 의한 다국어 음성 변환 연구, 2013 한국음성학회 가을 학술대회 발표 논문집, pp. 127-128)

[15] Voiceware Corp., VoiceTextTM, Retrieved from http://www.voiceware.co.kr/kor/product/product1.php on October 31, 2014

[16] Jennifer Clyde Interview, Retrieved from http://pann.nate.com/video/211296293 on October 31, 2014

[17] Fant G., (1970) Acoustic theory of speech production, Mouton, The Hague

• **채의근(Chae, Yi-Geun)**
공주대학교 컴퓨터공학부 컴퓨터공학전공
충남 천안시 서북구 천안대로 1223-24 (부대동)
Tel: 041-521-9233 Fax: 042-551-8104
Email: ygchae@kongju.ac.kr
관심분야: 패턴인식, 데이터통신 등
1998 ~ 현재 공주대학교 교수

• **윤영선(Yun, Young-Sun)**
한남대학교 정보통신공학과
대전시 대덕구 한남로 70 (오정동)
Tel: 042-629-7569 Fax: 042-629-7843
Email: ysyun@hnu.kr
관심분야: 음성인식, 음성처리, 웹 접근성, 내장형 시스템 등
2001 ~ 현재 한남대학교 교수

• **정진만(Chung, Jin Man)**
한남대학교 정보통신공학과
대전시 대덕구 한남로 70 (오정동)
Tel: 042-629-7574 Fax: 042-629-7843
Email: jmjung@hnu.kr
관심분야: 임베디드 시스템, 운영체제, 모바일 미들웨어 및 플랫폼 등
2014 ~ 현재 한남대학교 교수

• **은성배(Eun, Seongbae)**
한남대학교 정보통신공학과
대전시 대덕구 한남로 70 (오정동)
Tel: 042-629-7928 Fax: 042-629-7843
Email: sbeun@hnu.kr
관심분야: 센서네트워크 시스템, 임베디드 시스템, IoT 등
1995 ~ 현재 한남대학교 교수