특집

음성 합성 및 발성 변환 기술

김종국, 이기영(관동대학교 정보통신공학과), 배명진(숭실대학교 정보통신공학과)

I. 서 론

음성은 인간과 인간의 의사소통 수단으로 가장 편리하게 사용되는 매체이다. 음성 중에는 여러 가지 정보가 포함되어 있지만 가장 기본적이고 중요한 것이 의미정보 즉 언어적 정보이다. 또한 음성에는 누가 말하고 있는가를 나타내는 개인성 정보, 말하는 사람의 감정을 전해주는 정서 정보 등이 있다. 말하고 있는 모습을 보지 않고도 지금 말하는 사람이 누구인지를 알 수 있는 능력은 누구라도 가지고 있어서 상대의 감정을 음성으로부터 알아내는 능력과 함께 일상생활에서 중요한 역할을 담당하고 있다.

따라서 최근에는 인간과 컴퓨터와의 자연스러운 통신을 위해 음성 언어를 이용한 human interface 기술의 관심이 점점 커지고 있다. 음성 언어를 이용한 인터페이스를 실현하기 위해서는 기계가 음성을 이해하고, 기계가 음성을 생성하는 음성 인식, 음성 합성 기술이 필요하다. 특히 음성 합성 기술은 현재 상용화되어 서비스에 응용될 정도로 많은 기술적 발전이 있었다. 음성 합성 기술은 기계가 인간의 음성을 합성하여 내도록 하는 기술이며, 기계에 의한 합성 음성은 명료하고 자연스러워야 한다.

최근 음성신호처리 기술의 발달과 함께 man machine interface에서 처리결과를 무제한 음성 메시지로 비교적 명확하고 자연스러운 목소리로 들려주기 위한 합성방식으로 활발한 연구 및 개발이 진행되고 있다. 그러나 아직까지 규칙합성에 의한 합성음의 음질 수준이 자연음성에 비하여 매우 큰 열세를 보이고 있는 것은 자연스러운 음성의 합성을 위한 기본적 기술인 언어처리 기술, 운율 제어기술 및 음성 생성기술 모두가 현재 충분한 수준에 수준에 도달하지 못하였기 때문이다. 이에 대해 음성 변환기술은 현재의 음성 합성 방법만으로 해결하지 못한 음성 생성기술이나 운율 제어기술 등의 한계를 극복하기 위해 연구되고 있다. 음성 변환은 합성음에 감정정보를 포함시키거나 정감도를 높이거나 특정 화자의 음색으로 변환하는 기술이다[13].

본 고에서는 음성 합성 기술을 소개하고 음성 변환 기술들의 전반적인 현황에 대하여 기술하고자 한다. 서론에 이어 Ⅱ장에서는 음성 합성 기술에 대하여 기술하고 Ⅲ장에서는 음성 합성 시 고려사항을, Ⅳ장에서는 음성 변환 기술에 대하여 설명하고 마지막으로 결론을 맺는다.

II. 음성 합성 기술

음성합성(speech synthesis)은 인간의 발성모델을 토대로 연구되고 있으며 성도의 전달함수와 성대의 진동특성을 모델링 하여 구현되고 있다. 이러한 모델링에 의한 합성방식의 대표적인 것으로 LPC 계열의 파라미터 합성방식이 주류를 이루어 오고 있으나 모델 자체의 한계와 합성단위 연결시 스펙트럼 왜곡이 발생하며 운율조절에 의해 합성음성을 떨어뜨린다.

이에 프랑스의 CNET에서는 non-parametric 합성방식으로 pitch-synchronous 하게 분석하여 운율조절이 용이한 PSOLA 합성방식을 개발하여 합성음성의 명료도와 자연성을 대폭 향상시켰다. 응용할 수 있는 음성합성의 종류는 그 원리에 따라 세 가지로 나뉘고 있다[13]. 음성 파형을 그대로 이용하는 파형부호화-합성 방법(waveform coding synthesis), 통신량을 감소하기 위하여 음성을 분석한 파라미터를 구한 후 이를 이용하여 다시 음성을 합성하는 분석-합성 방법(analysis-synthesis) 및 음성학과 언어학의 규칙을 기본으로 문자와 운율 특징으로부터 음성을 합성하는 규칙-합성 방법(synthesis by rule)이 있다. 음성 합성 연구의 궁극적인 목표는 아직까지 인간의 발성기관을 모델링 하는 것이지만, 컴퓨터의 연산 속도 및 기억용량이 급속히 발전하면서 음성합성에 대한 연구는 단순히 인간의 발성기관 모델링에 그치지 않고 문서 처리 기술을 포함한 문서-음성 변환(TTS)시스템 기술로 확장되었으며 더불어 음성 변환 기술을 발전을 가져왔다[6]. 다음은 현재의 많이 사용되고 응용되는 음성 합성 기술을 소개 한다.

1. 파형부호화(Waveform coding) 합성

파형부호화(waveform coding)합성은 미리 녹음해 둔 음성 파형을 적당한 단위로 연결하여 합성하는 방법이다. 여기서 사용하는 단위란 합성단위를 말하며 그 종류로는 문장 단위, 어절 단위, 단어 단위, 음절 단위 및 그 이하의 단위(복합 음소:triphone, 반음절, 음소, 반음소 등)가 있다. 어절 이상의 단위를 이용한 합성음은 명료성이나 자연성이 매우 뛰어나지만 실제적인 합성을 위한 기억 용량이나 무제한의 연속 음성의 합성에서 자연성 등이 떨어지는 등 제한점이 많다.

대표적인 방법에는 PCM, ADPCM등이 있다. 그러나 최근 실제적으로 접해볼 수 있는 합성된 음성으로 전화 번호 안내나 ARS 응답의 경우 모두 파형부호화-합성을 이용하고 있으며 제한된 단어와 문장이긴 하지만 합성 음성의 명료성이나 자연성이 뛰어난 것을 알 수 있다.

2. 분석 합성(Synthesis by analysis)

분석-합성에 의한 합성에서는 단어나 어절을 음성 생성 모델에 의해 분석하여 특징 파라미터 시계열을 추출한 후 다시 이들을 이용하여 합성하는 방법이다. 이 방법에서는 파형부호화-합성에 의한 방법보다 적은 정보량을 요하므로 기억용량이나 통신량을 감소시킬 수 있다. 또한 발화속도나 피치, 스펙트럼 등의 급격한 변화를 감쇄시킬 수 있는 장점이 있다. 합성법에는 LPC, PARCOR, LSP등이 있다[14].

3. 규칙 합성(Synthesis by rule)

규칙-합성(synthesis by rule)은 음성학과 언어학의 규칙을 기본으로 문자와 합성될 음성 파형의 크기, 억양, 장단 등의 운율 특징과 함께 음성을 합성하는 방법이며, 음성학과 언어학적 음성의 지식을 기초로 한다. 예를 들어 반음절 단위를 이용한 규칙-합성에서 한국어 문장 "학교에 가요"를 합성한다고 하자. 첫 문자 '학'이라는 문자로부터 '하-'라는 전반부 반음절과 '-악'이라는 후반부 반음절을 합성 단위 데이터베이스로부터 취합하여 '학'이라는 음절의 합성 파형을 구성한다. 이 같은 방법으로 마지막 음절 '요'까지 합성 파형을 구성한다. 이와 동시에 취해 줘야 할 것은 이 문장 전체의 운율(prosody)이다[1].

4. PSOLA 합성

프랑스의 CNET에서는 non-parametric 합성방식으로 피치동기(pitch-synchronous) 하여 분석하여 운율조절이 용이한 PSOLA (pitch-synchronous overlapped adding) 합성방식을 개발하여 합성음성의 명료도와 자연성을 대폭 향상시켰다. 1990년 Charpentier는 CNET 의 PSOLA를 TD-PSOLA(time dependent PSOLA) 와 LP-PSOLA(linear predicting PSOLA) 로 나누어 개발하였다. 음성 합성 방법을 합성 단위에 따라 분류하면 문장, 어절, 단어, 음절, 복합 음소, 반음절, 음소, 반음소 합성 단위 등으로 나뉘어 질 수 있다. 그러나 문장에서 단어 단위까지를 파형부호화-합성 방법으로, 그 이하 단위는 분석-합성이나 규칙-합성으로 하였을 때 그 합성 음성으로부터 명료성과 자연성이 청취하기에 판별이 가능한 범위이다[12].

III. 음성 합성시 고려사항

1. 단위음 결합

단위음 결합 합성방식에서는 합성에 사용되는 음성의 합성단위가 저장되어 있어야 한다. 합성단위를 길게 할 것인가 짧게 할 것인가에는 각각의 장단점이 있다. 긴 합성 단위를 이용할 경우에는 긴 시간 간격동안 합성음의 자연성을 그대로 유지할 수 있으며 한 문장을 합성하는데 결합되는 부분이 적어지므로 깨끗한 음성을 생성할 수 있다. 그러나 긴 합성 단위를 사용할 경우 임의의 문장을 음성으로 생성하기 위해서는 많은 양의 합성 단위가 필요하게 되므로 보다 짧은 길이의 합성 단위를 이용하는 것도 유효하다.

모든 종류의 신호처리는 음질저하를 초래한다. 따라서 이상적인 단위음 결합 합성 시스템은 신호처리 없이 스펙트럼과 운율이 합성하고자 하는 문장에 가장 적절한 합성 단위를 선택하여 합성할 수 있어야 한다. 이러한 단위음을 선택하려면 엄청난 양의 합성 단위가 필요하다. 일반적인 단위음 결합 합성 시스템은 수동으로 분할된 같은 종류의 합성단위를 이용하여 합성에 사용한다[15].

2. 언어 처리

모든 문서-음성변환(TTS)시스템에서 임의의 문서를 입력받아 합성 단위로 바꾸어주는 부분은 필수적이다. 대부분의 TTS 시스템은 음소 또는 음소군을 합성 단위로 이용하기 때문에 임의의 문서를 음소단위로 변환하여야 하며 이때 입력된 문서의 문맥정보, 형태소 정보 등은 그대로 유지되어야 한다. 입력 문서가 문법에 벗어나지 않는 문자의 열로만 이루어져 있더라도 문

서-발음 변환은 상당히 어려운 작업이며 이를 해결하기 위하여 여러 단계의 처리 과정을 거치고 있지만 아직 완벽한 문서-발음 변환 시스템은 개발되어 있지 않다[4].

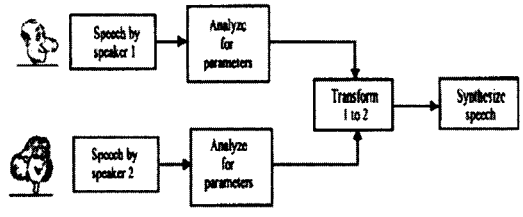
3. 운율 조절

음성에 있어서 운율(prosody)이란 억양, 강세 그리고 리듬으로 지각된다. 운율의 물리적인 의미는 기본주파수(F0), 분절음의 길이, 에너지 등과 관련이 있다. 그밖에도 발음상 또는 음향학적인 요인에 의해 영향을 받는다. 인간의 음성에 있어서 운율은 단어 그 자체는 물론 화자(speaker)의 의도, 발성당시의 청취자, 그리고 화자의 감정적 신체적 상태 등 여러 가지 환경에 따라 달라질 수 있다. 일반적으로 이러한 현상은 대화 시에는 물론 문서를 낭독할 때에도 나타나는데, 그 이유는 인간은 문서를 읽을 때 읽고 해석한 후 발성을 하기 때문이다. 그러므로 문서-음성 변환을 위해서는 인간의 지능과 유사한 형태를 이용하여 음성 합성을 해야겠지만, 현재까지 대부분의 문서-음성 변환(TTS) 시스템에서는 적용되고 있지 않아서 의미 또는 감정이 전혀 개입되지 않은 평서체의 음성 합성만이 이루어지고 있는 형편이다[2]. 입력된 문서로부터 운율조절 파라미터를 추출하기 위해서는 운율 경계결정, 기본 주파수 궤적 생성, 그리고 합성 단위음 지속시간 조절 등이 필수적이다. 그밖에 에너지 궤적 생성도 운율조절에 있어 중요한 부분이지만, 기본 주파수 궤적 생성 결과를 스케일(scaling)만 바꾸어서 이용하는 것으로 충분하다고 알려져 있다. 일반적으로 이에 단어상 또는 의미적 강세를 표현하기 위해, 강세를 나타낼 음절의 에너지를 증가시키는 규칙을 포함한다[9].

Ⅳ. 음성 변환 기술

1. 개요

음성변환(voice conversion)이란 화자의 개인성 정보를 수정하거나 치환하는 기술로, 일반적으로 입력 음성을 목적화자가 발성하는 것처럼 들리도록 변환하는 것을 뜻한다[6]. 다음은 일반적인 음성 변환 기술의 개념을 그림 1에서 보여주고 있다.



〈그림 1〉 음성 변환 기술의 개념도

이러한 음성 변환은 최근 문서-음성 변환(TTS) 시스템의 급증하는 수요로 인하여 그 중요성이 커지게 되었다. 일반적인 TTS 시스템은 한 화자의 음성 데이터베이스를 구축하고 이를 접속하는 형식으로 무제한 합성을 수행한다. 그러나 합성을 위한 음성 데이터베이스의 구축은 매우 많은 노력이 들어가는 작업이므로 하나 이상의 데이터 베이스를 작성한다는 것은 불가능한 일이다. 따라서 문서-음성 변환(TTS) 시스템이 대화 시스템으로 발전하고자 하거나, 구성된 데이터베이스의 음성외 음성을 출력하고자 할 때 음성 변환 기술은 필수적인 조건이 된다. 또한 음성 변환은 보안이나 신분 보호를 위한 음성변조나 게임, 애니메이션 캐릭터를 위해 음성을 변환하는 데도 활용될 수 있을 것이다.

음성 변환을 위해 고려되어야할 화자의 개인성 요소는 크게 음향학적 요소와 운율적 요소로 나눌 수 있다. 음향학적 화자의 개인성 요소는 발성기관의 해부학적 구조의 차이, 발성기관을 이용한 조음 방법의 차이, 성대에서의 여기 신호의 특성 등에 의해 나타나는 포만트 주파수, 포만트 대역폭, 스펙트럼 경사와 성문 파형(glottal waveform) 등이 있으며 운율적 개인성 요소에는 기본 주파수 궤적, 음소별 지속시간, 휴지기, 에너지 등이 있다.

완전한 음성 변환을 위해서는 이러한 요소들의 변환이 모두 이루어져야 한다. 그러나 운율요소의 변환은 화자의 발성습관을 모델링 하여야 한다는 점에서 매우 어려운 작업이며, 현재의 음성 변환 기술들도 음향학적 요소의 변환에 주력하고 있는 실정이다. 일반적으로 음성 변환을 위해 여러 음향학적 요소를 포함하는 스펙트럼 포락(spectrum envelope)의 변환과, 개인성 요소에 가장 큰 영향을 미치는 운율 요소인 피치 주기 값만을 변환시킨다[13].

2. 음성 변환의 발전과정

음성의 음성 변환이란 원시 화자의 음성(a source speaker's speech)을 목적 화자의 음성(a target speaker's speech)으로 변환하는 기술이다[1]. 각 화자 목소리의 개성을 음질(voice quality, timber) 또는 음성의 개인성(voice individuality)이라 표현한다. 따라서 음성 변환 기술을 좀더 구체적으로 음성의 언어적인 정보보다는 비언어 정보에 속하는 각 화자 목소리의 개성인 음질을 변환하는 기술이라 할 수 있다. 이는 음질로 인해 서로 다른 각 화자의 음성을 듣고 누구인지 판별할 수 있는 음성의 개인성을 들려주기 때문이다. 음성의 개인성을 나타내는 음질은 그의 스펙트럼 포락의 모양에서 나타나며, 피치 주파수, 에너지 및 지속시간 또는 발성속도로 나타나는 운율특성에 의해 좌우된다. 스펙트럼의 포락 특성과 운율특성은 음질을 구별하는 특징 파라미터인 동시에 발성기관을 움직이는 주요 합성요소이기도 하다.

음성의 음질 변환을 위한 연구는 이미 오래 전부터 진행되어 왔다. 발성 또는 지속시간의 스케일링 변환(time scaling modification)은 청취하여 이해하는 시간을 주기위하여 발성속도를 늘이거나 제한된 시간 안에 녹음된 음성자료를 스캔해 내기 위해 발성속도를 줄이기 위해 출발된 기술이다[2]. 음성의 주파수 영역에서 압축 및 신장하는 기술은 대역이 제한된 공간에서 음성통신을 하기 위해 적용된 기술이었다[7].

1986년 Charpentier[4]등에 의해 피치동기 중복가산(pitch synchronous overlap add, PSOLA)이라는 합성기술을 제안하면서 피치변환을 통해 여성의 음성을 남성의 음성 또는 어린이의 음성으로 변동시키는 음질변환에 대하여 기술하였다. 음성의 스펙트럼특성의 중요성을 인식한 Abe[5]등은 특히 합성된 전화음성으로 하여금 화자의 개인성을 주기 위하여 벡터양자화를 이용한 음질의 스펙트럼 변환기술을 제안하였다[6-9].

그 이후 TD-PSOLA[10] 또는 LP-PSOLA[11]를 이용한 음질변환 기술 및 대응 스펙트럼 벡터의 정합을 얻기 위한 방법으로 DTW 방법에서 HMM에 이르기까지 보다 완벽한 변환음성을 얻기 위해 연구가 진행되어 왔으며 통계적인 선형다변회귀방법[6]도 연구되고 있다. 또한 피치 궤적으로 나타나는 억양을 변환하기 위한 방법으로 통계적인 방법을 기초로 운율구를 이용한 방법 등이 연구되고 있다[16].

3. 음성 변환 기술의 문제

음성 변환은 음성 처리에 거의 모든 주요한 주제가 음성 변환의 문제와 관련 있는 것처럼 음성 처리를 연구하기 위해서 무궁무진한 분야이다. 첫째, 음성 변환의 분석분야는 화자의 특성 정보를 파악하는 것과 음향 모델링, 음성 코딩, 그리고 심리음향을 위하여 밀접하게 관련된 모델 파라미터들을 추정하는 적합한 모델들을 개발하는 것에 관련이 있다[11]. 둘째, 원래의 모델과 목표 모델들 사이의 관계가 결정되어지고 그리고 관찰되지 않은 데이터에 일반화되어야 한다. 학습과 일반화 처리는 음성 인식, 패턴 인식 그리고 기계 학습과 함께 음성 변환과 관련이 있다. 음성 변환 기술에서 주요한 학습 방법을 목적으로 하는 방법들은 Vector Quantization, Hidden Markov Models, Gaussian Mixture Models, Artificial Neural Networks, 그리고 Radial Basis Function Networks등이 있다.

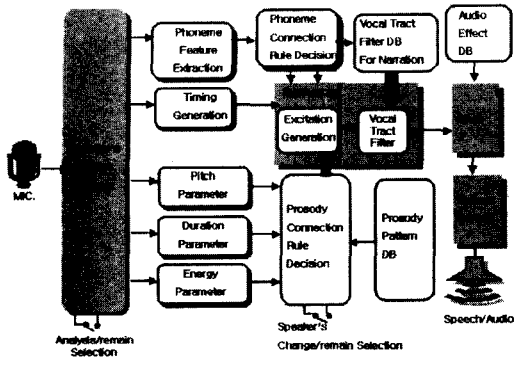
또한 훈련의 중요한 단계인 전처리와 원시 화자와 목표 화자의 학습 데이터의 적절한 할당, 음성 모델의 선택에서 화자들의 중요한 특징들을 추출하고, 음성 모델의 파라미터들을 추정하기 위해 학습 데이터들을 분석해야 한다. 더불어 원시 화자와 목표 화자 사이의 매핑(mapping)을 추정하기 위해 자동적으로 학습 방법들을 수용해야 한다. 마지막으로, 정확한 방법들은 최소화된 왜곡과 목표한 화자의 합성음의 유사도 극대화와 함께 원래의 음성을 처리하는데 수용되어야만 한다. 이러한 방법들도 또한 음성 합성과 음성 코딩의 응용 등에 적용되어진다[15].

음성 변환의 강인성은 아마 그 목적이 원시 화자와 목표 화자의 폭넓은 변화를 위해 잘 수행할 방법들을 개발해야 하는 것이 음성 변환에 있어서 가장 중요한 점이다. 음성 변환 기술들은 위에서 설명된 모든 3가지 방법과 연관된 새로운 방법을 알아보는 것 즉, 분석, 학습과 합성, 사용자 인터페이스의 필요를 최소화 하는 강인하고 자동화한 시스템을 개발하는 것이목적이다. 합성음의 성능 테스트의 하나인 주관적인 평가는 청각 시스템과 음성 변환 시스템들의 성능 평가 특성을 조사하기 위해 가장 현실성 있는 방법이다. 음성 변환의 기술에 관련한 특징들인 성도, 피치 궤적, 음소의 지속 기간, 에너지 궤적을 연구해 야 한다. 특히 음성 변환에 가장 중요한 화자의 정확한 특성중의 하나인 성도 모델링의 문제점을 조사하고 정확한 추정과 성도 스펙트럼의 변경에 대하여 새로운 방법들을 연구해야 한다. 음성의 운율 정보 변경은 또 다른 관점이며 화자 식별의 인식을 위해 중요한 단서인 억양 특성들을 추출하는 것이다. 더불어 우리들은 자세한 피치 궤적 변환에 대하여 새로운 방법을 연구해야 한다[15].

4. 음성 변환 시스템의 구성

일반적인 음성 변환 시스템은 크게 분석부(analysis part), 변환부(transformation part), 합성부(synthesis part)로 나눈다. 분석부에서는 입력음성을 매 분석 구간마다 분석하여 변환을 수행할 특징 파라미터를 추출한다. 추출되는 특징 파라미터는 변환부와 합성부에 종속되지만, 크게 스펙트럼 포락 변환을 위한 파라미터(스펙트럼 파라미터)와 운율 요소 변환을 위한 파라미터로 나뉘어져 분석된다[13].

일반적으로 성도특성을 반영하는 스펙트럼 포락정보의 변환을 위해 포만트 주파수(formant frequency)나 켑스트럼 계수(cepstrum



〈그림 2〉 음성의 발성 변환 시스템

coefficient)가 사용되며, 간단한 운율 요소의 변환을 위해 피치 주기 값을 추출하여 피치 변환을 수행한다. 변환부에서는 분석부에서 넘어오는 스펙트럼 파라미터와 피치 주기가 각각의 학습된 변환 방법에 의해 목적 화자의 특징 파라미터로 변환된다. 변환 방법의 학습에 있어서 고려되어야 할 사항은 화자의 음성공간을 표현하기 위한 화자 모델링의 방법과 그에 기반한 사상관계 학습이다.

일반적으로 화자의 모델링을 위해서 벡터 양자화 기반의 코드북(code book)이 주로 사용되어져 왔다. 이 방법에서 전체 음성공간은 코드북의 크기로 군집화(cluster)되고 각 클래스의 대표값을 코드워드(code word)에 저장해 화자의 전체 음성공간을 표현하게 된다. 합성부는 변환부에서 반환된 파라미터들을 음성으로 재합성하는 일을 수행한다. 일반적으로 LPC 보코더 계열의 합성기를 이용하여 변환된 스펙트럼 파라미터와 피치 주기를 입력으로 변환 음성을 출력한다. 그림 2는 운율을 고려한 발성 변환을 위한 시스템의 구성도를 보이고 있다[3].

5. 여러가지 음성 변환 방법들

가) 스펙트럼 특징의 변환

스펙트럼의 변환기술은 학습과정과 변환과정으로 나눌 수 있다. 스펙트럼 변환을 위한 기술로는 벡터양자화와 DTW 또는 HMM을 적용하여 대응 코드벡터를 이용하는 방법과 스펙트럼 포락비율을 이용한 방법 및 포만트 이동과 스펙트럼 기울기의 변환을 이용한 방법 등이 있다.

A. 대응 코드벡터를 이용한 방법

학습과정에서는 원시 화자와 목적 화자 사이의 스텍트럼 대응 쌍을 구하기 위한 절차는 다음과 같다.

① 이미 원시 화자A에 의해 녹음되어 있는 학습용 동일 단어나 문장을 목적 화자B가 발성하여 벡터 양자화에 의한 코드북(code book)을 구축하고 음성의 프레임별 코드벡터를 작성한다.

② 원시 화자의 코드북에 의해 작성된 학습용 동일 음성의 프레임별 코드벡터와 DTW를 수행하여 대응되는 원시 화자A의 코드벡터에 대한 목적화자B의 코드벡터들을 따로 모아 히스토그램을 작성한다.

③ 이 과정에서 얻은 히스토그램을 가중치로 하여 목적화자B의 코드북을 선형조합에 의해 매핑 코드북으로 갱신한다.

④ DTW에 의한 코드벡터 사이의 정규화거리가 수렴할 때까지 과정②와 ③을 반복하여 원시 화자A의 코드벡터가 목적 화자B의 코드벡터로 보다 가까워지도록 한다. 변환과정에서는 학습과정에서 얻은 최종적으로 선형조합에 의해 얻은 매핑 코드북에 의해 목적화자B 음성의 코드벡터를 프레임별로 재작

성하며 LPC 보코더에 의해 재합성한다[13].

B. 스펙트럼 포락비율을 이용한 방법

학습과정에서도 대응 코드벡터를 이용한 방법에서와 같은 학습과정이 필요하다. 즉, 동일한 학습단어를 원시 화자와 목적 화자가 발성해야 하며 DTW 또는 HMM을 이용하여 먼저 대응되는 스펙트럼 쌍을 구한다. 다시 말해 원시 화자와 목적화자의 대응하는 스펙트럼 포락의 비율을 이용한다.

C. 포만트 이동과 스펙트럼의 기울기 변환

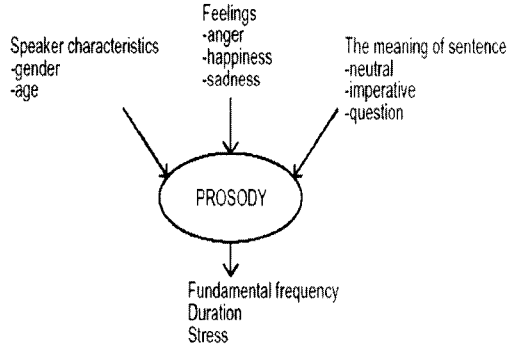
원시 화자와 목적 화자 사이의 대응하는 스펙트럼을 이용한다. 여기서는 대응하는 스펙트럼에서 포만트 주파수를 추출하여 원시화자의 포만트 주파수를 목적 화자의 포만트 주파수로 이동하고 스펙트럼의 기울기도 변환하여 준다. 이를 다시 푸리에 역변환 하여 목적 화자로 변환된 음성을 얻는다[8].

D. 선형다변회귀 모델

선형다변회귀 모델은 DTW(Dynamic Time Wapping)로 시간 정렬을 수행한 후 벡터 양자화(VQ)와 선형다변회귀를 이용해 음성 변환을 수행한다. 회귀분석이란 독립변수(independent variable)로부터 종속변수(dependent variable)를 예측하기 위하여 사용되는 회귀방정식(regression equation)이라는 두 변수 사이에 구체적인 함수 관계를 규명하는 데 이용되는 통계적 분석 방법을 말한다. 이러한 선형다변회귀를 이용하여 LPC 켑스트럼을 원시 화자에서 목적 화자로 변환시키게 된다. 변환된 LPC 켑스트럼과 잔차신호(residual)를 컨벌루션(convolution)하여 신호를 합성할 때 두 파라미터간의 왜곡된 관계 때문에 음질의 저하가 발생한다[8].

나) 운율 특징의 변환

일반적으로 글을 읽을 때나 자연스러운 대화를 할 때의 음성에는 각 화자마다 서로 다른 특성의 운율정보를 포함하고 있다. 운율정보는 단음절로 구성된 단어음성으로부터 시작하여 구나 절 단위의 음성에 이르기까지 발성된 음성신호에 포함되는 언어적 정보이면서 화자고유의 특성이기도 하다. 대화를 하거나 낭독을 하는 화자는 이 운율을 이용하여 의미, 감정, 의도 및 마음가짐 등을 전하며, 청취자는 화자가 발성한 음성으로부터 운율을 이용해서 전체적인 의미를 파악한다. 아래의 그림 3은 운율 정보(prosodic information)를 보여주고 있다[13].



〈그림 3〉 운율 정보

따라서 문장을 합성하는 경우 운율 정보를 합성음에 반영하면 보다 명확한 의미 전달이 가능해 지며 화자의 음색을 잘 나타낼 수 있다. 목소리는 지문이나 신체적인 특징 등과 함께 개인 정보를 제공할 수 있는 화자 고유의 특성을 가지고 있기 때문이다. 이러한 각 화자 고유의 목소리의 특징 중에서 운율 정보를 분석하여 데이터베이스를 구축하고 합성하는 음성 변환처리하면 음성 변환시스템, 각종 제품의 자기 목소리 입력,

발성 장애인의 목소리 재현 등 다양한 분야에 응용이 가능하다.

운율 특징은 피치 궤적으로 나타나는 억양, 에너지 및 지속시간을 말한다. 에너지나 지속시간의 변환기술은 PSOLA 방법에 의해 변환이 가능한 기술이므로 본 고에서는 피치 궤적(pitch contour)을 변환하는 기술을 중심으로 기술한다[10].

A. 통계적인 피치궤적의 변환, 가우스 정규화

피치궤적의 변환은 원시 화자에서 목표 화자로 향한 피치의 평균과 변동 범위를 일치시킬 필요가 있다. 이 모델링은 원시화자의 피치 값들을 원하는 피치 주파수로 매핑(mapping)하는데 사용된다. 임의 화자들의 각 피치 값들의 평균과 표준편차 를 이용하여 원시 화자의 피치 주파수(F0)를 목표 화자의 피치 주파수로 변환시켜준다. 가우스 정규화(Gauss normalization)를 위한 과정은 첫째 원시화자와 목표 화자의 음성으로부터 피치 궤적을 추출한다. 둘째 각 화자의 피치 값에서 각각의 평균과 표준 편차를 구한다. 최종적으로 원시 화자의 피치 궤적을 목표 화자의 피치 궤적으로 변환시킨다[16].

B. 변형된 가우스 모델링

가우스 정규화(Gauss normalization)는 문장 음성의 시작부분이나 마지막 부분의 피치 궤적이 항상 가우스적으로 일정하게 분포되어 있다는 가정에서 시작되었기 때문에 발성된 문장 음성의 피치궤적에 적용하기에는 한계가 있다. 따라서 변형된 기울어진 가우스 모델링을 제시하였으며 이 모델링은 발성 문장 음성의 피치 궤적이 자연적으로 기울어져 있다는 가정에서 시작하였으며 이것은 기울어진 피치궤적의 기저선(base line)에 적용한 가우스 정규화 모델링이다.

억양구(intonational phrase)를 형성하는 단위인 강세구(accentual phrase)는 주로 피치 궤적(F0 contour)에 의해 특징지어진다. 하나의 강세구가 세 음절 이하로 구성된 어절일 경우 주로 L H (low-high)의 피치 궤적으로 나타나며, 네 음절 이상일 경우 L H L H의 피치 궤적으로 나타난다. 또 억양구의 마지막 강세구에는 경계톤이 기본 강세구의 피치 궤적에 덥혀짐으로써 기본 L (H L) H 피치 궤적이 아닌 경계톤의 피치 궤적으로 나타나게 된다. 따라서 강세구를 이용한 가우스 모델링은 문장음성 전체의 가우스 분포를 이용하는 것이 아니라 각 강세구의 가우스 분포를 이용하여 피치 궤적을 변환하는 것이며 이에 따라 문장 전체에 해당하는 이전 방법들보다 목표 화자의 피치 궤적 특성을 이식하기에 용이하다[1].

V. 결 론

음성합성 시스템의 이상적인 목표는 자연성과 명료성이다. 명료성을 위해서 음운 변동의 규칙, 합성방식, 합성단위의 접속 기술이 있으며 자연성 면에서는 한국어 언어 처리에 기반을 둔 운율정보의 생성 및 제어기술의 개발이 필요하다. 이를 위해서 많은 연구자들의 부단한 노력에도 불구하고 시료음성의 개성을 그대로 재현하거나 한 화자의 음색을 다른 화자의 음색으로 변환하여 합성하는 기술에 그치고 있다.

각 개인의 고유한 음성 특징 파라미터를 추출 분석하여 화자특성에 의해 발성하는 처리기술 외에 합성 음성의 음색을 상황에 따라 변화시킬 수 있는 방법으로 휴지기, 피치, 에너지 및 지속시간 등의 운율 정보를 운율패턴으로 작성하여 데이터베이스화하고 이를 통하여 음색을 변환

처리하는 기술을 개발해야 한다. 더불어 음성 변환 시스템은 음성 합성분야의 응용시스템으로써 고음질의 합성음 및 음색변환을 위해서는 음성 화자간의 운율 변경을 수행해야 하며 운율 변경을 위해서는 한국어의 운율구 패턴을 기반으로 지속시간, 피치 및 에너지의 변경이 수반된 음성 변환 시스템이 필요하다.

참고문헌

[1] K. Lee, "Statistical Approaches to Convert Pitch Contours Based on Korean Prosodic Phrases", Proc. The Journal of The Acoustic Society of Korea, Vol. 23, No. 1E, 2004.

[2] J.K. Kim, W.R, Jo, M.J Bae, "A Study on Real Time Prosody Control of Speech", CCCT2003, Vol. 5, ,pp. 195-198, 2003.

[3] M.J. Bae, "On a Voice Color Change in the Fairy Tale Narration System with Parent' s Voice Color", J., Acoust., Society, Korea, Vol.16, No.8, pp.131-135, November 1997.

[4] F. J. Charpentier, M. G. Stella, "Diphone Synthesis Using an Overlap-add Technique for Speech Waveform Concatenatio", Proc. ICASSP' 86, 1986.

[5] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice Conversion Through Vector Quantization", Proc. ICASSP' 88, 1988.

[6] Abe, Masanobu, Shikano, Kiyohiro, and Kuwabara, Hisao,"Cross-Language Voice Conversion", Proc. ICASSP' 90, 1990.

[7] Abe, Masanobu, "A Segment-Based Approach to Voice Conversion", Proc. ICASSP' 91 , 1991.

[8] H. Mizuno, M. Abe, "Voice Conversion Algorithm Based on Piecewise Linear Conversion Rule of Formant Frequency and Spectrum Tilt", Speech Communication, Vol. 16, 1995.

[9] Levent M. Arslan and David Talkin, "Voice Conversion by Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum", Proc. EuroSpeech97, 1997.

[10] Valbret, H., Moulines, E. and Tubach, J.P., "Voice transformation using PSOLA technique", Speech Communication 11, 1992.

[11] Xuejing Sun, "Voice Quality Conversion in TD-PSOLA Speech Synthesis", Proc. ICASSP' 2000, 2000.

[12] F. M. Gimenez de los Galanes, M. H. Savoji, J. M. Pardo, "New Algorithm for Spectral Smoothing and Envelope Modification for LP-PSOLA Synthesis", Proc. ICASSP' 94, Vol. 1, 1994.

[13] M.J. Bae, "The TTS Speech Synthesis Techniques", Proceedings of Korea Inst. Commun. Sciences, Vol.11, No.9, pp.67-78, Sept. 1994.

[14] H. Kwon, K. Bae, "Voice Conversion Using Linear Multivariate Regression Model and LP-PSOLA Syntjesis Method", Proc. The Journal of The Acoustic Society of Korea, Vol. 20, No. 3, 2001.

[15] D. T. Chappel, J. H. L, "Speaker-Specific Pitch Contour Modeling and Modification", Proc. ICASSP' 98, Vol. 1, pp. 885-888, 1998

[16] K. Lee, C. Choi, K. Choi, H. Lee, "Intonation Conversion Using the Other Spreaker' s Excitation Signal", Proc. The Journal of The Acoustic Society of Korea, Vol. 14, No. 4, 1995.

Acknowledgements : This work was partially supported by interdisciplinary research grants of the KOSEF. (Subject Number:R01-2002-000-00278-0)

저자소개

김 종 국

현 재 숭실대학교 대학원 정보통신공학과 박사과정
주관심분야 음성인식, 음성합성, 음성코딩, 통신 및 신호처리 등

이 기 영

현 재 관동대학교 교수
주관심분야 음성인식, 음성합성, 언어처리, 신호처리 등

배 명 진

1986년 - 1992년 호서대학교 전자공학과 조교수
1992년 - 현 재 숭실대학교 정보통신공학과 교수
주관심분야 음성인식, 음성합성, 음성코딩, 통신 및 신호처리 등