음성합성 기술 개발 현황

오 영환
한국과학기술원 전산학과

Trend on the Technical Development of Korean Speech Synthesis

Yung-Hwan Oh
Dept. of Computer Science, KAIST

1. 서 론

음성은 인간의 가장 기본적이고 간단 명료한 정보 전달 수단으로써, 음성처리 기술들은 최근 몇 년간 급속한 발전을 이룩하였다. 음성인식, 음성합성 및 음성코딩등과 같은 음성처리 기술중에서도 특히 음성 합성 기술은 음성정보 서비스의 활성화와 함께 그 수요가 증가하여 주목받고 있는 기술로써 실용화에 가장 가까이 있는 기술의 하나이다. 최근에는 음성합성 기술의 발전에 힘입어 오디오텍스(Audio-tex), 자동 음성 응답 시스템(ARS), 음성 우편 시스템(VMS)등이 속속 개발 실용화되고 있으며 음성정보 서비스의 대중화를 가속화시키고 있다.

이러한 발전의 원인은 크게 세가지로 요약되는데 첫째는 컴퓨터 성능의 향상이고 둘째는 디지탈 신호처리에 적합한 음성분석 기술의 개발이며 세째는 반도체 기술의 발전에 따른 고속, 대용량의 특성을 갖는 디지탈 신호처리 소자 및 기억소자의 개발이라 볼 수 있다.

최근의 음성합성 기술 연구방향은 시시각각으로 변하는 정보에 능동적으로 대처할 수 있으며 인간의 편리를 추구하여 보다 자동화된 방법에 대하여 진행되고 있다. 이를 위해서는 단순히 음성을 디지탈화하여 저장한후, 편집하여 재생해주는 수준을 넘어서 매체변환 즉, 임의의 문장을 음성으로 변환하여 주는 무제한 어휘 음성합성 기술이 필수적이다.

본 고에서는 최근 음성합성 기술의 주연구대상이 되고 있는 무제한 어휘의 음성합성 즉, 문서-음성 변환에서 합성음질을 향상시키기 위한 기존의 연구와 최근 국내의 연구에 대하여 소개하고 앞으로의 연구 방향에 대하여 논하고자 한다.

2. 음성합성 기술의 개요

음성합성 기술은 합성 대상 어휘에 따라 크게 제한 어휘 합성 방식과 무제한 어휘 합성 방식으로 나눌 수 있다.

제한 어휘 합성 방식은 합성하고자 하는 어휘들을 미리 분석하였다가 이들의 조합에 의해 말을 합성하는 방법으로써 구조가 간단하고 미리 사람이 발음한 내용을 편집하는 것으로 자연스러운 음질을 갖는 장점이 있다. 그러나 출력하고자 하는 문장에서 저장된 단어의 위치, 억양에 따라 합성가능한 어휘수에 제약이 따르게 되며 미리 저장된 음성만을 출력할 수 있는 단점이 있다. 따라서 이 방식은 주로 단어 또는 문장 단위의 음편들을 연결한 몇가지의 합성음성만으로도 사용가능한 지하철 안내방송 또는 ARS등에 이용되고 있다.

무제한 어휘 합성은 언어의 기본 단위인 음소 또는 음절등을 저장시킨 후 합성하고자 하는 문장을 분석하여 저장된 합성단위 음성들로부터 합성음을 생성해내는 방식으로 문서-음성 변환(text-to-speech conversion) 이라고도 한다. 이 방식은 어떠한 형태의 문장이라도 출력시킬 수 있으며 제한적 음성합성 방식보다 훨씬 더 높은 메모리 효율을 갖는다.

그러나 현재까지의 연구결과로는 제한어휘 합성 방식보다 음성의 명료성 및 자연성이 떨어지고 있다. 이때 명료성은 듣는이에게 합성음의 내용을 정확하게 전달하는 정도를 말하며, 자연성은 합성음과 사람이 발성한 음성과의 유사도를 말한다.

무제한 어휘 합성에서 합성음의 명료성과 자연성을 향상시키기 위해서는 적절한 코딩방법 및 합성단위의 선택, 그리고 언어처리에 기반한 운율제어 규칙 등을 고려해야 한다. 이중 합성음의 명료성은 현재 일정 수준에 이르러 단위음 접속을 통한 무제한 어휘 합성이 가능한 시스템이 개발되고 있는 실정이다. 명료성이외에 합성음의 자연성을 위해서는 자연음에서 나타나는 억양, 강세, 길이, 휴지기등의 운율이 합성음에 반영되어야 한다. 운율정보가 합성음에 반영되면 합성음이 자연스러워지는 것은 물론, 운율정보가 갖는 의미적 요인으로 인하여 보다 명확한 의미 전달이 가능하다.

3. 문서-음성 변환 시스템

문서-음성 변환 시스템은 입력된 문장에 대해 음운변환 및 운율제어 규칙등을 적용하여 음성신호로 변환하는 장치로, 구성은 처리과정에 따라 크게 언어처리부, 운율제어부, 음성생성부의 3부분으로 구성된다.

언어처리부에서는 입력된 문장에 포함된 기호나 약어등을 발음에 알맞은 구술적인 표현으로 변환시킨 후 구문론과 관련된 사전 정보에 의해 문장의 구조를 분석하고 의미를 파악한다. 운율제어부에서는 언어처리부에서 추출해낸 정보에 의해 끊어읽기, 강세, 억양등을 지정하는 운율기호를 결정한다. 마지막으로 음성생성부에서는 전단에서 입력된 음운기호열과 운율기호열을 조합하여 음성출력을 내보낸다. 문서-음성 변환 시스템은 위와 같은 3가지 요소들로 구성되며 합성을 위한 기본단위, 코딩방식, 운율정보 추출 과정등에 따라 여러가지 형태의 시스템으로 구현될 수 있다.

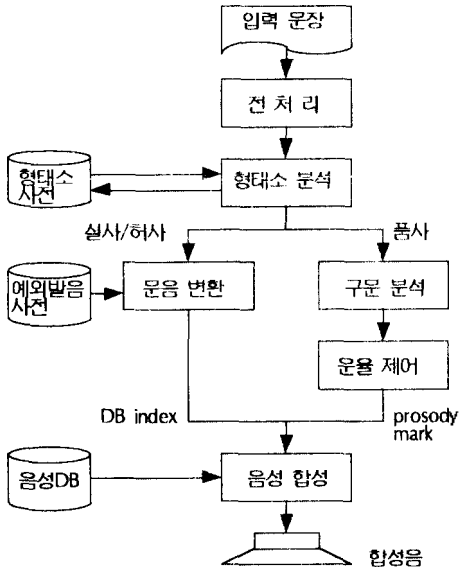

그림 1. 문서-음성 변환 시스템의 구성

3.1 언어 처리부

인간이 문장을 읽을 경우 문장 전체의 의미나 내용을 파악하여야 하는 것처럼 음성합성에 있어서도 높은 품질을 얻기 위해서는 문장의 언어적 이해가 필수적이다. 언어 처리부에서는 입력된 문장에 포함된 기호나 약어등을 발음에 알맞은 구술적인 표현으로 변환시킨후 문자의 구조를 분석하고 의미를 파악한다. 문장의 구조를 이해하기 위해서는 형태소 분석, 구문 분석등이 필요하며 운율 제어부에서는 이러한 정보를 이용하여 운율 파라미터를 추출한다.

1) 형태소 분석

형태소(morpheme)란 일정한 뜻이 있는 언어의 가장 작은 단위, 즉 최소의 유의적 단위로 정의된다. 형태소 분석은 자연언어의 처리에 있어서 가장 기본적인 분석과정으로 문장의 구조를 이해하기 위한 구문분석 과정은 물론 음운변환(grapheme-to-phoneme) 과정에서도 형태소 분석 결과를 이용하여 문장을 처리하므로 문서-음성 변환에 있어서 필수적인 단계이다.

한국어에 있어서 형태소 분석이란 띄어쓰기의 단위인 어절의 구성요소 즉, 형태소를 밝히는 것으로 정의된다. 문장을 구성하는 최소단위인 어절은 형태소가 모여 이루어진 것으로, 형태소 분석시 그 어절을 구성할 수 있는 모든 가능한 형태소 조합을 찾게 된다.

한국어에는 약 50여개의 형태소 분류가 있으며 분류별로 사전에 저장되는데, 형태소 분석시 주어진 어절에 대하여 가능한 형태소를 찾기 위하여 빈번한 사전 검색이 요구된다. 그러나, 이때 입력된 문장의 모든 어휘가 사전에 등록되어 있다고는 할 수 없으므로 이에 대한 대처가 필요하다. 예를 들어 일반적인 문장에서는 복합어, 새로운 지명 및 인명등 미지의 어구가 수시로 나타나는데, 보통의 지능을 가진 인간은 이러한 미지어(unknown-word)에 대해서도 문맥 및 과거의 지식에 의해 문법적 역할이나 의미를 추측하여 전체의 대의를 파악할 뿐만 아니라 이를 적절하게 발음할 수 있다.

따라서 문서-음성 변환에서도 형태소 분석시 미지어에 대한 처리를 적절히 해주어야 할 필요가 있다.

2) 구문 분석

인간이 긴 문장을 적절히 끊어 읽는 것은 문장의 구문구조를 분석해 낼 수 있기 때문이다. 문서-음성 변환 시스템에서 구문 분석기는 자연스러운 운율정보를 생성하기 위하여 문장의 구문정보를 추출해 내는 역할을 한다.

우리말에서의 구의 기능, 상호 결합관계 정보는 구의 마지막에 고정되어 있는 조사, 어미등의 기능어 문법정보로부터 구할 수 있다. 이와 같은 점에 착안하여 복잡한 처리과정을 생략하고 조사와 어미등의 형식형태소와 구문요소간의 결합규칙만을 이용하여 구문분석을 한 문서-음성 변환 시스템의 예도 있으나, 임의의 문장을 분석하는데는 미흡한 점이 있다.

자연언어 처리분야에서 연구되고 있는 구문분석 방법으로는 기존에 많이 쓰여왔던 구-구조문법(phrase structure grammar)에 의한 분석 방법과 최근 한국어 분석에 이용되고 있는 의존문법(dependency grammar)에 의한 분석 방법이 있다. 이중 의존문법에 의한 분석이 많은 주목을 받고 있는데, 그 이유는 첫째 한국어에 있어서 어순의 자

유성에 의한 문제점을 쉽게 해결할 수 있으며, 둘째 구성요소의 불연속성이나 생략등과 같은 현상에 큰 영향을 받지 않아 매우 견고하다는 점이다. 따라서 구문분석에 의존문법을 이용한다면 한국어의 이해에 매우 효과적일 것으로 생각된다.

그러나 자연언어 처리분야에서 사용했던 알고리즘을 그대로 음성합성에 적용하는데는 해결해야할 점들이 있다. 먼저, 자연언어 처리분야에서 구문분석기의 목적은 하나의 문장을 정확히 분석하는데 있다. 따라서 처리과정이 복잡하고 시간이 늦어지고, 어떤 문장은 하나 이상의 모호한 결과가 나오게 된다. 반면, 문서-음성 변환시스템의 파서는 주목적이 운율정보를 추출하는데 있고, 잘못된 문장이라도 실시간에 문서-음성 변환하여야 하므로 분석된 결과는 항상 하나가 나와야 하며, 잘못된 문장이라도 그 자체를 읽어주어야 한다.

따라서 문서-음성 변환을 위해서는 견고하고 실시간 처리가 가능한 구문분석 알고리즘의 개발이 시급한 과제라 할 수 있다.

3.2 운율 제어부

운율이란 발성시 나타나는 억양, 강세, 리듬등의 특성을 말하는데 이는 기본주파수, 음소길이, 음량, 휴지기 길이등에 의해 결정된다. 운율은 합성음의 이해도와 자연성에 중요한 요소로 작용하며 정보 전달에 큰 영향을 끼치는데, 운율 제어부에서는 언어 처리부에서 분석된 결과들을 이용하여 운율 파라미터들을 제어한다.

사람이 한번 숨을 쉬어 발성하는 말의 단위를 발화단위라 하는데 발화 단위내에서는 기본 주파수가 점차 낮아지는 경향을 갖는다. 이를 억양의 '기본 기울기'라 하며, 억양의 기본 기울기에 단어, 음절의 강세 및 문 구조에 따른 억양 패턴이 더해져서 전체 억양 패턴이 구성된다.

음소의 길이 및 휴지길이는 억양과 함께 합성음의 자연도를 결정하는 중요한 요소이다. 음소의 길이는 음소 자체의 성질 뿐만 아니라 주변의 음소환경, 한 단어내의 음소 갯수, 단어내에서의 음소의 위치, 강세여부등 다양한 요소에 의해 영향을 받는다.

휴지기 길이도 음소길이와 마찬가지로 전후의 음소환경에 의해 영향을 받게 되는데, 그 이외에 발화단위 사이에서 긴 휴지기가 존재한다. 발화단위는 하나의 발화단위 내의 어절갯수, 음절갯수 뿐만아니라 문장의 구조 및 의미에 의해 결정된다.

일본의 경우 이미 1960년대에 Fujisaki에 의하여 기본 주파수 윤곽을 만들기 위한 모델이 제안되어 널리 쓰이고 있으나, 국내에서는 이러한 획기적인 모델의 제안은 아직 없으며 계속 연구중에 있다.

3.3 음성 생성부

음성생성부에서는 언어 처리부에서 변환된 합성단위음 열과 운율제어부에서 생성된 운율기호로부터 실제 음성신호를 합성하게 되는데, 본 절에서는 음성생성과 관련하여 코딩 방법과 합성단위에 대하여 살펴보기로 한다.

1) 코딩 방법

무제한 어휘 합성을 위한 대표적인 코딩방법으로는 포만트 합성방식과 LPC 계열의 분석합성방식이 있는데 2가지 모두 음원 코딩법이라는 공통점을 가지고 있다. 1960년 Fant에 의해 음성생성의 디지탈 모델이 발표된 이후 음성합성에 큰 발전을 가져온 음원 코딩법은 인간의 성도 특성을 모델링하여 특징 파라미터의 시간적 변화 정보에 의해 음성을 합성한다. 파형 코딩법에 비해 연산량이 많고 음질도 떨어지나, 데이타 압축률이 높고 특히, 특징 파라미터의 변환에 따라 말의 속도, 음높이, 스펙트럼 변환등이 용이하여 대부분의 무제한 어휘 합성에 적용되어 왔다.

먼저 포만트 합성방식은 순수 규칙합성 방식으로서 성도의 변화/필터특성은 각 음소 그리고 음소간의 포만트 변화를 나타내는 규칙을 사용하여 기술한다. 영어권에서는 이미 오래전부터 Klatt형 합성기가 상용화되어 쓰이고 있다. 국내에서도 몇몇 대학 및 연구소를 중심으로 포만트 방식의 음성합성에 대한 연구가 이루어져 왔으나, 음성분석에 대한 많은 자료를 필요로 하기 때문에 음질개선을 위해서는 앞으로도 보다 많은 연구가 필요할 것으로 생각된다.

LPC 계열의 합성방식은 음성신호의 예측 계수를 이용하여 all-pole 성도모델 필터를 구성하고 음원 신호를 필터링하면 쉽게 음성신호를 합성해낼 수 있다는 장점이 있다. 국내의 경우 초창기에는 대부분의 연구기관에서 비교적 시스템와 구현이 간단하고 쉽다는 장점때문에 이 방식을 택하였으나, 합성음질에 한계를 느끼고 지금은 별로 사용하지 않고 있다.

최근에는 음성신호를 파라미터화하지 않으면서 피치를 변화시킬 수 있는 PSOLA(Pitch Synchronous Overlap & Add) 방법이 개발되어 많은 연구기관에서 이를 문서-음성 변환에 이용하고 있다. PSOLA 방법은 음성신호의 각 피치 단위의 신호를 피크값을 중심으로 분석창을 이용해 음성소편(short term signal)을 만든 후, 합성시 각각의 음성소편을 중첩시켜 연속된 음성신호를 만드는데, 중첩 길이를 변화시킴으로써 피치 주기를 조절할 수 있다.

PSOLA 방법은 음성신호를 시간 영역에서 부호화하므로 음원 코딩법에 의한 합성음에 비하여 음질이 좋고 합성시 연산량이 많지 않아 실시간 처리가 용이하다. 반면 합성시 필요한 데이타량이 많으며 음성소편을 자동적으로 분류해내기 어려우므로 음성소편 사전 구성시 수작업이 많이 필요한 단점이 있다.

2) 음성합성 단위

음성합성 단위는 연속음을 내기 위한 합성의 기본 단위로써, 기존의 음성합성 단위에는 음소, diphone, 반음절, 음절등이 있으며 이때 알고리즘의 복잡도, 데이타의 갯수, 음질등이 크게 영향을 받는다. 음소를 기본단위로 하여 음성을 합성할 경우 매우 적은 갯수의 데이타로 무제한 음성합성을 할 수 있는 장점이 있으나 음소와 음소를 연결시키는 과정에서 일어나는 조음결합 현상을 나타내기 어려우므로 불연속이 생겨 합성음의 명료성이 떨어진다.

diphone은 서로 연결할 경우 음소처럼 불연속이 생기지는 않으나 diphone 자체의 수가 상당히 많아 많은 기억 용량을 필요로 하게 된다. 음절을 기본 단위로 할 경우 비교적 좋은 결과를 얻을 수 있으나 음절과 음절 사이에 일어나는 조음결합 현상과 천이 구간상의 문제가 발생할 수 있고 데이타의 수가 많아지는 단점이 있다.

최근의 합성단위 선택의 동향은 음성의 전후환경을 고려하여 합성단위를 정의하여 합성음의 음질을 향상시키려는 시도들이 나타나고 있다. 일본 ATR에서 제안한 COC (context-oriented-clustering), 전자통신연구소에서 제안한 CDU(context dependent unit), 삼성종합기술원에서 제안한 modified diphone등은 모두 이러한 흐름에서 제안된 단위들이다.

이러한 방법들은 최근 하드웨어의 발달로 기억장치의 집적도 및 가격면에서 유리해진 것을 이용하여 조음결합 법칙의 추출에서 발생하는 노력 및 비용을 절감하고 양질의 합성음을 얻을 수 있다는 장점이 있다.

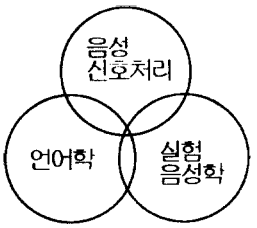

그림 2. T-T-S 관련 연구

4. 맺음말

앞에서 살펴본 바를 정리하면 최근 연구 동향은 연구소 및 기업체 중심의 실용적인 연구가 두드러짐을 알 수 있다. PSOLA 방식과 같은 시간축상의 코딩방법을 문서-음성변환 시스템에 적용하여 합성음질의 명료도 향상을 꾀하였고, 합성단위에 있어서도 음소의 환경을 충분히 흡수하고자 많은 수의 합성단위를 음성합성에 이용하는 추세이다.

이러한 합성음의 명료성 확보를 바탕으로 자연언어처리 기법을 적용하여 합성음의 자연성을 향상시키고자 하였으나, 아직 한국어 처리 기술 자체가 완벽한 단계에 이르지 못했고 이를 문서-음성 변환에 이용하려는 시도가 오래되지 않아 이에 대한 기술은 초보적인 수준을 벗어나지 못하고 있다.

매체 변환 기술의 하나인 문서-음성 변환 기술은 신호처리 분야 뿐만아니라 언어학 및 실험음성학적인 연구를 포함한 종합적인 기술이라 할 수 있다. 따라서, 문서-음성 변환 기술의 발전을 위해서는 실험 음성학 및 자연어 처리 분야와의 연계가 필수적이다. 이를 통하여 현재 기술에서 가장 시급히 개선되어야 할 점으로 지적되고 있는 자연성을 향상시킬 수 있으며 실용화 시스템에의 적용이 앞당겨질 수 있을 것으로 예상된다.

한편, 현재까지의 음성합성 기술을 통한 합성음은 개인적인 정보가 모두 제거된 방법이었다. 그러나, 앞으로는 점점 더 다양화되고 고급화되어가는 사용자들의 욕구를 충족시키고 보다 양질의 음성서비스를 제공하기 위하여 꼭 개발해야 할 기술이 음색변환 기술일 것이다.

음색변환 기술과 문서-음성 변환 기술이 결합되면 임의의 화자가 발성한 얼마간의 음성데이타를 이용하여 그 화자의 음성을 무한적으로 합성할 수 있어서 자동통역 전화에서도 통역된 내용들이 본래 화자의 음색을 잃지 않고 전달이 가능해지며, 다양한 음색의 합성음으로 사용자들의 욕구를 만족시킬 수 있을 것으로 예상된다.

참고 문헌

1. 진용옥; "음성 정보처리 기술 및 음성 정보서비스의 발전과 전망", 음성통신 및 신호처리 워크샵 논문집, pp.12-26, 1992
2. 김상룡, 김정수; "형태소 해석을 이용한 합성 음성의 음운 및 운율 처리", 전자 공학회지 20권 5호, pp.508-515, 1993
3. 이윤근, 안승권; "음성 합성 기술 분야", 전자 공학회지 20권 5호 pp.523-532
4. 김상훈, 지민제, 최운천; "한국어 문장/음성 변환에서의 TD-PSOLA 적용", 음성통신 및 신호처리 워크샵 논문집, pp.291-294, 1993
5. 정인종, 경연정, 이양희; "한국어 음성의 규칙 합성", 전자 공학회지 20권 5호 pp.587-596
6. 공병구, 김상룡, 김정수; "이질음 접속에 의한 음질 저하 및 극복 대책 연구", 음성통신 및 신호처리 워크샵 논문집, pp.279-284, 1993
7. 구준모 외; "한국어 무제한 음성합성 시스템 ; 가라사대", 음성통신 및 신호처리 워크샵 논문집, pp201, 1992