1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 1(s) 호

명료도에서 사람 목소리로 - TTS에 관하여

권철홍[1] , 최영익[1] , 이금주[2] , 심갑종[2]

[1] 대전대학교 정보통신공학과, [2] 현대자동차(주) 승용전자설계2팀

From Clarity To Human Voice

Chul-Hong Kwon[1] , Young-Ig Choi[1] , Kum-Joo Lee[2] , Kab-Jong Shim[2]

[1] Dept. of Information & Communication Eng., Taejon Univ., [2] Hyundai Motor Company

e-mail : chkwon@dragon.taejon.ac.kr

요 약

그 동안 TTS 음성합성의 평가 척도로 명료도(Clarity)와 자연성(Naturalness)을 기준으로 삼았다. 이제는 합성음의 평가 기준이 사람 목소리와 이해도가 되는 것이 좋겠다고 생각한다. 본 논문은 사람 목소리와 이해도라는 척도 중에서 사람 목소리에 관한 주제를 다루고자 한다. 이를 위하여 음성 DB의 합성 단위로 CVC type을 기본으로 하고, CV,VC type으로 보강한 단위를 선정하여 음성 DB를 구축하였다. 그리고 합성 알고리즘은 음색을 살리며 피치 변경이 용이한 PS-RELP 알고리즘을 제안하였다.

I. 서 론

그 동안 TTS 음성합성의 평가 척도로 명료도(Clarity)와 자연성(Naturalness)을 기준으로 삼았다. 이러한 기준으로 볼 때 현재 발표된 한국어 TTS 시스템은 명료도 측면에서 상당한 성과를 얻었다(여기에 언급될 수 있는 TTS 시스템은 LG전자기술원[1], 삼성종합기술원[2], 한국전자통신연구원[3] 등에서 개발된 시스템이다). 그리하여 이 시스템들은 현재 상용화되어 일부에서 이용하고 있다. 그러나 일반인들은 이 시스템들이 여전히 미흡하다고 생각하는 듯하며, 그 주요 원인은 크게 다음 두 가지로 고찰해 볼 수 있다. 먼저 합성음이 종래의 시스템에서 나오는 기계 목소리에서 크게 벗어나 있으나, 그럼에도 불구하고 사람 목소리와 거리가 있다는 점을 들 수 있다. 다시 말해서 원음의 음색을 잃어버린다는 점이다. 다른 하나는 단문을 합성하면 명료도가 우수하고 자연스러움도 어느 정도 인정해 줄 수 있으나, 여러 문장을 합성하면 자연스러움이 떨어지고 특히 이해도가 상당히 떨어진다는 점이다. 이러한 점에서 볼 때, 이제는 합성음의 평가 기준이 사람 목소리와 이해도가 되는 것이 좋겠다고 생각한다.

본 논문은 사람 목소리와 이해도라는 척도 중에서 사람 목소리에 관한 주제를 다루고자 한다. TTS의 여러 모듈 중에서 이 주제에 밀접한 관련이 있는 모듈은 음성 DB의 합성 단위 선정과 이들 DB로부터 음성 파형을 만들어 내는 합성방식이다. 이를 위해 합성 DB는 CVC(C=자음, V=모음)를 기반으로 CV, VC를 보강하여 구축하였고, 합성방식은 원음의 음색을 살릴 수 있는 PS-RELP(Pitch Synchronous Residual Excited Linear Prediction) 방식을 개발하였다. 2절에서는 제안된 음성 DB의 합성단위에 대하여 설명하고, 3절에서 PS-RELP 합성 방식을 제안하고, 4절에서 실험 데이터와 토론을 통해 제안된 방식이 유용하다는 점을 제시하고 5절에서 결론을 맺는다.

II. 음성 DB의 합성단위

TTS 음성합성 방법은 먼저 음성의 기본 단위로 DB를 구축한 후, 문장을 합성할 때 필요한 DB를 꺼내어 연결시킴으로써 합성음을 만들어 낸다. 이때 DB를 어떤 음성의 기본 단위로 구축하는가에 따라서 합성 알고리즘의 복잡성, DB 개수 및 DB가 차지하는 기억 용량 사이에 trade-off가 있을 뿐만 아니라 합성음의 음질에도 큰 영향을 준다. 그런데, 언어는 그 나름대로의 독특한 특성이 있으므로 한국어 음성합성 시스템에서는 한국어 특성을 고려하여 음성의 기본 단위를 선택하여야 할 것이다.

따라서 본 절에서는 DB를 구축하는데 이용할 수 있는 음소, diphone, 음절, 반음절 등에 대한 장·단점을 살펴보고, 한국어 음성합성 시스템에 가장 적합한 음성의 기본 단위를 제안한다.

(1) 음소

한국어의 경우 19개의 자음과 21개의 모음을 합하여 40개의 DB만 있으면 무제한 음성합성을 할 수 있다. 또한 변이음(allophone)을 고려해도 DB의 수가 많지 않으므로, 기억 용량을 적게 필요로 한다는 장점이 있다. 그러나 DB의 기본 단위로 음소를 이용하여 문장을 합성하는 경우에는 음소와 음소를 연결시켜야 하는데 이 과정에서 연결 부분에 불연속이 생겨 합성음의 명확도가 떨어진다. 또한 음성이 갖고 있는 정보 중에서 understanding에 관련된 많은 정보가 음소와 음소 사이의 천이 부분에 있으며[4], 한 음소가 인접한 음소에 영향을 미쳐 음소의 음향학적 특성을 변형시키는 상호 조음 현상도 고려하여 음소를 연결해야 하므로, 음소와 음소를 연결시키는데 필요한 규칙을 찾기 위해서는 한국어에 대한 광범위한 지식과 과학적 연구가 선행되어야 한다.

(2) Diphone

Diphone은 "한 단음(phone)의 정적인 중심 부분에서 다음 단음의 정적인 중심 부분까지의 음성 단편"이라고 정의할 수 있다[5]. 예를 들어 '가나'란 단어에서 하나의 diphone은 'ㅏ'의 정적인 중심부분에서 'ㄴ'의 정적인 중심부분까지라고 말할 수 있다. 이와 같은 diphone의 정의를 생각해 보면, 하나의 diphone는 diphone을 구성하고 있는 단음과 단음 사이의 천이 부분을 포함하게 되고, 음소인 경우에는 그들을 연결할 때 천이 부분을 규칙에 의해 연결해야 하는 것과는 달리, diphone을 연결하여 한 문장을 구성할 때는 diphone의 경계가 정적인 중심 부분이므로 diphone 끼리의 연결이 간단하다는 장점을 갖게 된다. 그러나 diphone을 연결하여 단어나 문장을 구성할 때 연결부분이 에너지가 큰 모음이 되므로 불연속성이 발생할 때 귀에 크게 감지된다는 단점을 갖는다. 본 연구자도 diphone을 DB의 기본 단위로 선정하여 TTS 개발을 한 바 있는데[5] 불연속성이 귀로 크게 감지되어 만족할 만한 합성음을 낼 수 없었다.

(3) 음절

음절은 음소의 결합으로 구성되어 있는데, 한국어의 경우 초성, 중성, 종성이 어울려 한 음절을 이루는 독특한 특징이 있다. 음절의 구성을 공식으로 표현하면 C^iVC^f(C^i=초성자음, V=중성모음, C^f=종성자음)으로 나타낼 수 있다. 한국어의 음절수는 산술적으로 약 3200개 정도가 되나, 실제 사용되고 있는 음절의 수는 음소와 음소가 연결될 때의 제약성과 현대 국어에 사용되지 않는 음절이 있으므로 1096개만 현재 사용된다고 알려져 있다[6]. 음질의 측면에서 음절을 생각해 보면, 음절은 음절을 구성하고 있는 음소들 사이에 존재하는 천이 부분의 정보를 포함하고 있으므로 단음절의 경우에는 좋은 음질의 합성음을 기대할 수 있다. 그러나 음절이 어울려 단어를 구성할 때 음절과 음절 사이도 상호 조음 현상이 존재하고, 따라서 이 음절을 주위와 독립된 단음절로 DB를 구성하면 음소인 경우와 마찬가지의 단점이 생긴다. 그러나, 주위 음소 환경을 고려한 음절을 기본 단위로 선택한다면 좋은 단위로 될 수 있다.

(4) 반음절

반음절은 음절 중 CV형, V형, VC형의 DB만 있다. CVC형의 음절은, CV형과 VC형의 음절을 모음 부분에서 연결하여 만든다. 이로 인하여 음절 DB보다 합성음의 음질이 다소 떨어지나, DB의 수가 산술적으로 계산할 때 570개 정도로 음절 DB보다 적은 기억 용량이 필요하다. 그러나 이 경우도 diphone과 마찬가지로 연결부분이 에너지가 큰 모음이 되므로 불연속성이 발생할 때 귀에 크게 감지된다는 단점을 갖는다.

(5)제안된 합성 단위

본 연구에서는 CV, VC type과 CVC type을 합성의 기본 단위로 선정하여 음성 DB를 구축하였다. 여기에서 CVC type은 단음절이 아니고 주위 음소 환경을 고려한 음절이다. 이에 관해서는 뒤에 설명이 되어 있는데, 본 연구자가 선정한 CVC type DB의 개수는 2923개이다. 이론적으로 C^iVC^f type의 DB 수를 계산해 보면, C^i에는 어절의 처음에 오는 자음 19개가 모두 올 수 있고 또한 어절 내에 오는 자음 10개(ㄱ,ㄴ,ㄷ,ㄹ,ㄴ,ㅁ,ㅇ,ㅅ,ㅈ,ㅎ)가 올 수 있다. V에는 모음 21개가 오고, C^f에는 어절 내에 오는 자음 10개(ㄱ,ㄴ,ㄷ,ㄹ,ㄴ,ㅁ,ㅇ,ㅅ,ㅈ,ㅎ)와 종성 자음 8개(ㄱ,ㄴ,ㄷ,ㄹ,ㄴ,ㅁ,ㅇ, 받침 없는 경우)가 올 수 있다. 따라서 산술적으로 CVC type DB의 수는 10,962개이다. 그러나 우리는 방대한 양의 문장으로부터 이 CVC DB에 대한 사용 빈도수를 조사하여 그 중에서 사용 빈도수가 많은 2,923개만 CVC DB로 구축하였다. 그리고 모자라는 부분은 CV, VC type으로 보충하였다. CV,VC type에는 #CV(#은 묵음), VC#, VV, VC, CV, CC 등이 있고, CVC type에는 #CVC#, CVC#, #CVC, CVC 등이 있다.

III. 합성 알고리즘

음성합성을 위해 고려할 수 있는 합성 알고리즘은 크게 세 가지 부류로 나눌 수 있다. 그 중 첫째는 파형 부호화 방식으로서, 음성 파형을 PCM 등으로 부호화

하여 컴퓨터에 저장한 뒤, 합성할 문장에 필요한 데이터베이스를 꺼내어 연결시켜 음성 파형을 만들어내는 방법이다. 이 방식은 알고리즘이 간단하며 합성음의 음질이 좋은 장점이 있으나 데이터베이스가 차지하는 기억용량이 많이 필요하고, 피치 조절이 어렵다는 단점이 있다. 이 방식의 대표적인 알고리즘에 TD-PSOLA 방식(time domain pitch synchronous overlap and add method)이 알려져 있다. 두 번째 방식은 vocoding 방식으로서, 사람의 발성기관을 수학적으로 모델링하여 음성을 합성하는 방식이다. Vocoding 방식에 의한 음성합성 방법의 대표적인 것으로 LPC(linear predictive coding) 방식을 들 수 있다. 마지막 세 번째는 Formant 합성 방식으로서, 이 방식은 음소 고유의 formant를 추출하여 데이터베이스를 구축하고, 음소와 음소가 연결할 때 이 formant의 변화를 규칙화하여 합성하는 synthesis-by-rule 방식이다.

다음에는 위에서 열거한 세 가지 합성 방식에 대하여 장·단점을 살펴보기로 한다. 먼저, TD-PSOLA 방식[7]은 프랑스에서 개발한 음성 합성 방식으로서, 이는 다음과 같은 세 단계로 나눌 수 있다. 첫째 원 음성을 피치(음성 파형의 한 주기) 단위로 분해하는 과정, 둘째 분해된 단위의 운율 처리 과정, 셋째 이로부터 합성을 하는 과정으로 나눈다[8]. 그런데 이 방식은 피치 조절을 위하여 인접한 두 피치 주기간의 음성 파형 자체를 단순히 더하기 때문에 스펙트럼의 왜곡을 피할 수 없다. 또한 음성 파형과 스펙트럼 정보와는 상관 관계가 적기 때문에 스펙트럼이 어느 방향으로 왜곡됐는지 알 수가 없어 이 왜곡을 개선시킬 방법이 없는 단점이 있다. PSOLA 방식에 대한 연구 결과를 보면, 원래 DB가 갖고 있는 피치의 작은 범위 내에서 피치를 변경시키면 스펙트럼의 왜곡이 작다는 결론을 내리고 따라서 한 DB에 다양한 피치를 갖는 여러 개의 DB를 구축하는 방법을 사용한다. 그러나 이 방법은 DB의 크기가 상당히 늘어나는 문제점을 갖게 하고, 또한 피치의 자유로운 조절에 한계를 갖는 단점을 극복할 수 없는 점이 있다.

다음에는 vocoding 방식 중 대표적인 LPC 합성 방식[9]에 대하여 살펴보자. LPC 합성 방식은 음성 발생 기관을 all-pole 필터로 modeling하고, 이 필터의 여기 신호(excitation source)로는 유성음의 경우 주기적인 펄스열을 무성음의 경우 잡음(random noise)을 사용한다. 그런데 LPC 합성 방식은 다음과 같은 문제점이 있다.

1) LPC 합성 방식은 음성 신호가 짧은 구간 동안 stationary 하다는 가정에 기초를 두기 때문에, 음성 신호의 특성이 빨리 변하는 천이(transient) 부분은 정확한 해석이 어렵다.

2) LPC 모델은 all-pole 모델로서, 음성 발생 모델의 면에서 볼 때 비음을 정확히 나타낼 수 없다. 비음은 pole-zero 모델을 이용하여 정확히 모델화할 수 있다.

3) 여기 신호가 너무 단순화되어 있고, 유성음인 경우 위와 같은 주기적인 펄스열을 사용한 결과, 합성음에 울리는(buzzy) 소리가 들린다.

LPC 방식은 이와 같은 단점으로 인해 현재 TTS 합성 방식으로 채용되지 않고 있다.

다음에는 Formant 합성 방식[10]에 대하여 설명한다. Formant란 사람의 음성 발생 기관 중 음성 스펙트럼 형성에 가장 큰 역할을 담당하는 성도의 공명 주파수를 의미하며, 음성인지(speech perception) 연구 결과에 의하면 말소리의 변별에 있어서 가장 중요한 특징이라고 알려져 있다. Formant 합성 방식은 이러한 formant의 특징을 이용하여 음성 발생 기관을 formant 주파수에 따른 일련의 공명기(resonator)들로 모델링한 것이다. 이 방식은 공명기 각각의 formant들의 주파수, 대역폭 및 크기 등을 주요 파라미터로 사용한다. 이 방식을 사용하여 미국의 Klatt가 합성기를 개발하였는데[10], 이 합성기의 성능은 현재 개발된 것들 중에서 최고를 자랑한다 하지 않을 수 없다. 그러나, 이 방식은 한 음소에서의 대표적인 formant의 주파수, 대역폭 및 크기를 찾아야 하는데 어떤 값을 대표 값으로 하는가에 대한 결정이 어렵다. 또한 음소와 음소의 연결 부분 등에서 formant의 변화 정도에 대한 데이터를 수집하여 이를 규칙화해야 하는데 이를 수행하는데 난점이 너무 많고 개발 소요 기간도 길고 개발 인력도 많아야 하는 어려운 점이 있어 Klatt 이후로는 formant 합성 방식에 대한 연구가 미비한 실정이다.

본 연구에서 제안된 음성 과형 생성 알고리즘은 PS-RELP(pitch synchronous residual excited linear prediction) 방식이다. 여기서 residual 신호는 음성 신호를 성도 필터를 반대로 통과시켜 얻은 신호를 말한다. Residual 신호를 얻는 과정을 수식으로 표현하면 다음과 같다.

$$r(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i)$$

여기에서 $s(n)$은 음성 신호이고, $r(n)$은 residual 신호이며, p는 성도 필터의 차수이고, $\{a_i\}$는 성도 필터의 coefficient이다. 이렇게 구한 residual 신호로 TD-PSOLA 방식에서와 같이 피치 조절을 한다. 두 방식의 차이는 TD-PSOLA 방식은 피치 조절을 위해 음성 신

호 자체를 이용하였고, PS-RELP 방식은 residual 신호를 이용한다는 점이다. 음성 신호를 overlap-add 방식으로 합성을 하면 음성 신호의 스펙트럼에 손상을 가져와 명료하지 못한 합성음을 도출하게 된다. 이에 반해 residual 신호는 성도 필터를 통과한 신호이므로 음성 신호가 갖는 spectrum이 제거된 신호가 된다. 이 때 음성 신호의 spectrum은 성도 필터의 coefficient에 남아 있다. 따라서 이 신호를 overlap-add 하더라도 음성 신호에는 손상이 가지 않는 이점이 생긴다.

IV. 합성 시스템의 평가

본 연구에서는 CVC type을 기본으로 CV,VC type을 보강한 합성 단위로 음성 DB를 구축하였다. CVC type에 해당하는 DB는 총 2923 개이고, CV, VC type에 해당하는 DB는 총 1333 개이며, 따라서 전체 DB 총수는 4256 개이다. 이렇게 구성된 음성 DB의 용량은 24MB이다. 이 정도 DB의 수는 전문가 1인이 4개월 정도의 시간을 투자하여 구축할 수 있었다. 또한 DB의 용량은 최근에 국내에서 발표된 TTS 시스템에 비교하여 비슷하거나 작은 수준이다[1-3]. 덧붙여, 많은 문장을 입력시켜 테스트한 결과 CVC type이 95% 이상 사용되었고 CV,VC type의 출현 빈도율은 5% 미만이었다.

본 연구에서 제안한 PS-RELP 합성 알고리즘은 피치 변경이 TD-PSOLA에 비하여 자유로와, 음성 DB의 원래 피치에 ±30% 정도의 피치 변경은 음질의 손상이나 음색의 훼손을 가져오지 않았다. 이 정도의 피치 변경 범위는 평이한 문장을 합성하는데 충분하다.

TTS 시스템의 합성음을 평가하는 객관적인 기준을 제시하기는 어렵다. 왜냐하면 서론에서도 언급했지만 TTS의 평가 기준은 명료성과 자연성으로 이를 수치화하는 것이 힘들기 때문이다. 따라서 본 논문에서는 전문가가 본 연구에서 개발한 TTS 합성음을 청취하여 평가한 주관적인 결과를 종합하여 말하겠다. CVC type을 사용한 결과 합성음이 명료하게 들렸고, CVC type과 PS-RELP 알고리즘으로 인해 합성음이 사람 목소리의 음색을 살려 주었다. 또한 긴 문장과 여러 문장을 연속하여 청취하였을 때 문장의 전체 의미를 파악할 수 있었다. 즉 이해도가 우수하다는 것을 알 수 있었다. (참고로 본 논문에서 제안된 방식으로 합성한 음성 샘플을 본 연구자의 Internet Homepage에 올려 놓았다[11].)

V. 맺음말

본 논문은 명료성을 가지며 사람 목소리의 음색을 살려 있는 합성음을 얻기 위한 주제를 다루었다. 이를 위하여 음성 DB의 합성 단위로 CVC type을 기본으로 하고, CV,VC type으로 보강한 단위를 선정하여 음성 DB를 구축하였다. 그리고 합성 알고리즘은 음색을 살리며 피치 변경이 용이한 PS-RELP 알고리즘을 제안하였다.

향후 과제로는 duration 규칙과 피치 규칙의 보완, 끊어읽기 과정의 개선 그리고 음성 DB의 보완 및 제안된 합성 알고리즘을 보완시키는 과제가 남아 있다.

참고 문헌

1. 이 준우 외, "수정된 음절을 이용한 한국어 문장-음성 변환시스템," 제13회 음성통신 및 신호처리 워크샵 논문집, pp. 237-240. 1996.

2. 김 정수, 김 상룡 외, 제품 발표회, 삼성종합기술원, 1996년 12월.

3. 김 상훈, 이 정철, ETRI 기술이전 설명회. 한국전자통신연구원, 1997년 3월.

4. 이 성주, 김 회동, 김 형순, "천이구간 정보를 이용한 음성의 가변적인 시간축 변환," 제13회 음성통신 및 신호처리 워크샵 논문집, pp. 232-236, 1996..

5. 권 철홍, 정 원국, 구 준모, 김 형순, "한국어 문자 음성 변환 시스템 : 가라사대", 한국통신학회지 제11권 9호 1994.

6. 허 웅, 국어학 -우리말의 오늘·어제-, 샘문화사, 1983.

7. F. Charpentier 외, "A diphone synthesis system based on time-domain prosodic modifications of speech", ICASSP 89, pp. 238-241, 1988.

8. 김 상훈, 지 민재, 최 운천, "한국어 문장/음성 변환에서의 TD-PSOLA 적용", 제 10회 음성통신 및 신호처리 워크샵 논문집, pp. 291-294, 1993.

9. L.R. Rabiner, R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Inc., 1978.

10. J. Allen, M. S. Hunnicutt and D. Klatt, From Text To Speech : The MITalk System, Cambridge university press, 1987.

11. 최 영익, 권 철홍, http://dragon.taejon.ac.kr/~chkwon/, 1998.