음성합성시 에너지 정규화가 음질에 미치는 영향
Effect of Energy Normalization on the Quality of Synthetic Speech

정은석, 최의선, 이철희
연세대학교 전자공학과
Eunsuk Jung, Euisun Choi and Chulhee Lee
Dept. of Electronic Engineering, Yonsei University

요 약

본 논문에서는 코퍼스 기반 음성합성시 각 음성 세그먼트의 에너지 정규화가 합성된 음성의 음질에 미치는 영향에 대하여 연구한다. 음성합성에 사용되는 음성 세그먼트들은 실제 자연 음성 데이터로부터 추출된 것으로 다양한 발음세기를 가진다. 따라서 이들을 조합하여 만든 합성음성의 음질은 일반적으로 음량이 고르지 못하고 듣기에 부자연스럽다. 이러한 문제를 해결하기 위해 음성합성시 음성 세그먼트의 에너지를 정규화하는 방법을 제안하고 정규화방법으로 최대진폭 정규화방식을 사용하였다. 녹음환경이 비교적 일정한 코퍼스와 그렇지 않은 환경에서 녹음된 코퍼스를 사용하여 정규화 없이 합성한 음성의 음질과 정규화를 거쳐서 합성한 음성의 음질을 비교한다. 실험결과 음성 세그먼트의 에너지를 정규화한 경우 합성음성의 음질이 개선되었다.

Ⅰ. 서 론

음성합성 특히 문서음성변환(TTS) 시스템은 입력된 문서의 내용을 특정한 알고리즘에 의해 음성으로 출력시키는 시스템으로 자동응답기, 언어교육, 장애보조기구, 말하는 장난감, 기계-인간 대화 등 다양한 분야에 응용할 수 있는 시스템이다 [1]. 이러한 음성합성 시스템은 문서입력을 음성출력으로 전환하는 방법에 따라서 크게 두 부류로 나뉘어진다.

첫째는 포만트(formant) 합성기와 같은 규칙합성기(synthesis by rule)이다. 규칙합성기는 사람의 음성으로부터 운율조절규칙과 합성규칙들을 미리 찾아낸 후 이 규칙을 토대로 음성을 합성한다. 이 방법으로 합성된 음성은 의사전달이 가능할 정도로 명료하지만 자연스럽지 못한 단점이 있다 [1].

두번째 부류는 연결합성기이다. 연결합성기는 실제 자연 음성으로부터 추출한 음성세그먼트들을 연결시켜 합성음성을 만드는 방법이다. 이 방법으로 합성된 음성은 충분한 명료성을 가지며 규칙합성기에 비해서 더 자연스러운 장점이 있다. 그러나 연결합성기는 규칙합성기에 비해서 많은 양의 데이터가 필요하다. 대표적인 연결합성기로 PSOLA(Pitch Synchronous OverLap and Add)합성기와 코퍼스 기반의 음성합성기 등을 들 수 있다. 전자는 후자에 비해서 단위음 연결시 음질저하가 심하다 [2-3]. 코퍼스 기반의 음성합성기는 PSOLA합성기의 합성방법이 지닌 문제점을 개선하기 위하여 제안되었으며 방대한 양의 자연음성 데이터를 녹음하고 이로부터 최적의 음성세그먼트를 검색하여 연결시키는 방법으로 음성을 합성한다 [2]. 이 방법으로 합성된 음성은 충분한 명료성을 지닐 뿐 아니라 규칙합성기에 비해 자연스럽고 단위음 연결시 PSOLA합성기에 비해 개선된 음질을 보여준다 [3]. 그러나 코퍼스의 음성데이터들간에 발음세기의 차이가 존재할 경우 합성된 전체음성의 음질은 부자연스러울 수 있다.

본 논문에서는 이러한 문제점을 해결하기 위해 각 음성 세그먼트의 에너지를 정규화하여 음질의 개선 가능성에 대해 연구한다.

Ⅱ. 코퍼스내 음성세그먼트의 음량의 차이로 인한 코퍼스 기반 음성합성시스템의 문제점

대량의 음성을 저장하여 만든 코퍼스를 기반으로 하는 음성합성시스템은 음성을 합성하기 위해 코퍼스를 먼저 분석하여 검색을 위한 사전정보를 획득한다. 그 후에는 코퍼스를 합성단위별로 분할하여 원하는 후보음성세그먼트들을 찾아내고 이들 중에서 문맥요소와 운율요소가 최적인 음성세그먼트를 선택한다. 이렇게 선택된 음성세그먼트들을 연결시켜서 원하는 음성이 합성된다. 후보음성세그먼트에서 최적 음성세그먼트 선택시 고려되는 운율요소는 피치주기, 지속시간, 음량과 같은 세부요소를 가지며 문맥요소

는 전후에 존재하는 음소이다 [3]. 결과적으로 코퍼스 기반의 음성합성시스템은 같은 연결합성 시스템인 PSOLA방식에 비해 인위적인 조작이 적으며 따라서 단위음 연결시 음질저하도 적다.

그러나 이상적인 조건의 후보음성세그먼트가 항상 있는 것은 아니므로 최적 음성세그먼트 선택시 가장 유사한 후보음성세그먼트가 선택된다. 따라서 음량이나 피치, 지속시간, 문맥요소의 불일치가 발생하며 이로 인해 합성된 음성의 음질은 저하될 수 있다. 예를 들어, 같은 음소지만 문맥요소나 전체 문맥의 운율에 따라 다양한 발음세기를 가지므로 음질저하가 일어날 수 있다(그림 4.(a)참조). 특히 음량 즉, 발음세기의 차이는 문장의 운율이나 문맥요소에 따라서 변할 뿐 아니라 코퍼스의 녹음환경에 따라서 변하게 된다. 녹음환경은 발화자의 건강상태, 피로도, 감정뿐 아니라 녹음기기의 상태 등 다양한 변수를 가지고 있다.

그림 1과 그림 2는 동일인을 화자로 하여 생성된 코퍼스내에서 17번 발음되는 "여러분 안녕하십니까~"라는 동일한 문장이 가지는 최대진폭의 다양성을 보여준다. 실험에 사용된 코퍼스의 녹음환경에 따라서 샘플문장내의 최대진폭이 최고 15dB까지 차이가 나며, 샘플문장의 평균에너지는 최고 20dB까지 차이를 나타내고 있다. 그림 3.(a)는 17개의 샘플들 중에서 발음세기가 매우 크게 차이나는 샘플문장 3개의 파형을 보여준다. 이와 같은 음성세그먼트의 발음세기의 차이는 합성음성은 부자연스러운 발음세기의 변화를 가져와 음질을 저하시킬 수 있다.

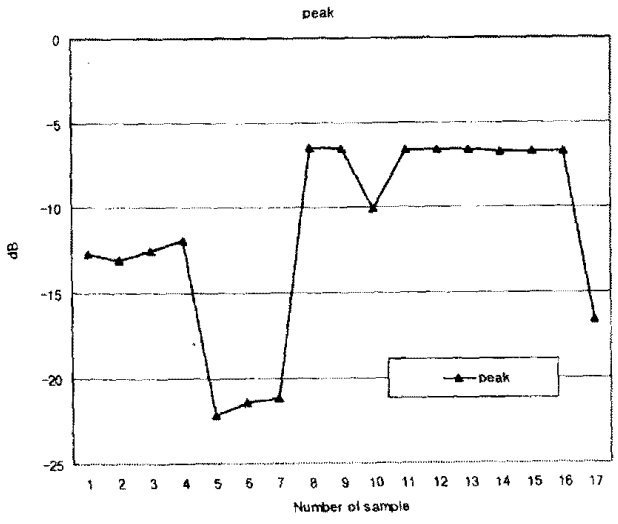


그림 1. 17개 샘플문장 각각의 최대진폭.

이러한 문제를 해결하기 위해 발음세기 정규화를 제안하고 앞서 사용했던 "여러분 안녕하십니까 ~"의 각 샘플들의 최대진폭을 정규화하였다. 그림 2에서 최대진폭 정규화를 함으로 샘플들의 평균에너지도 정규화되었음을 알 수 있다. 그림 3.(b)는 그림 3.(a)에 보여진 샘플문장들을 최대진폭 정규화방식으로 정규화한 후의 파형들이며 정규화를 통해서 샘플문장들이 매우 유사해졌음을 보여준다. 그리고 정규화가 이루어진 음성 청취시에도 서로 다른 세기를 가졌던 17개의 샘플문장이 아주 유사해졌음을 알 수 있다. 이로부터 최대진폭 정규화를 통하여 발음세기 정규화가 가능함을 알 수 있다.

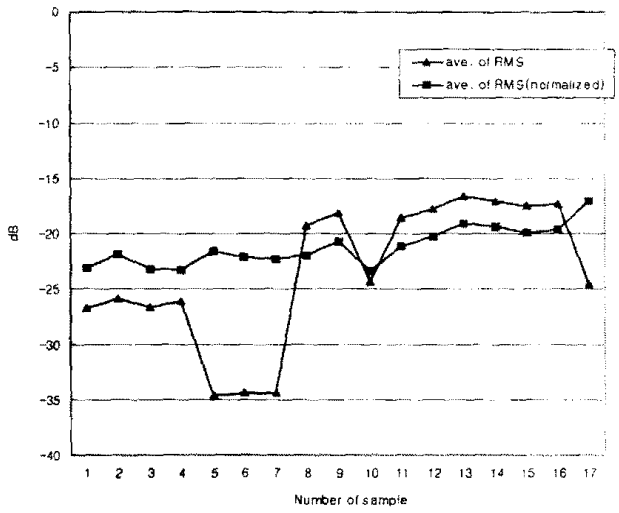


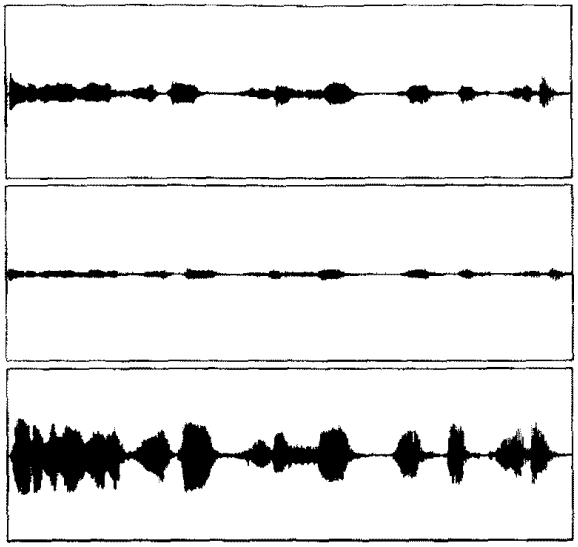
그림 2. 17개 샘플문장 각각의 평균에너지.

따라서, 이와 같은 방법으로 합성단위만큼의 크기를 가지는 각 음성세그먼트의 에너지를 정규화하면 각 음성세그먼트마다 비슷한 발음세기를 가져 합성된 음성의 부자연스러운 발음세기 변화가 개선될 것으로 예측된다.

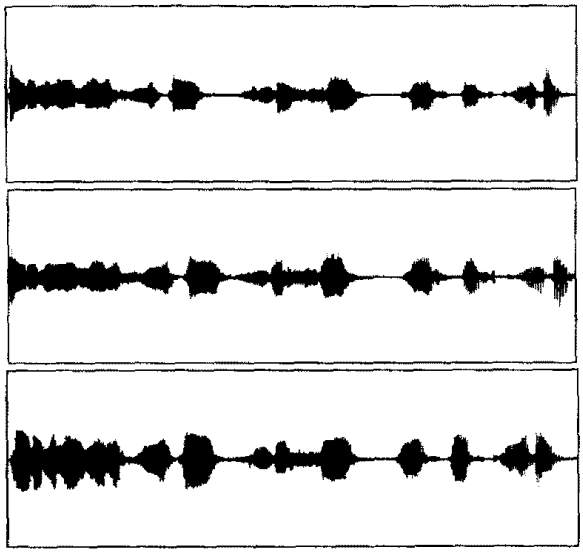
Ⅲ. 실험 및 고찰

3.1 실험방법제시 및 실험과정서술

제안한 기법의 성능을 실험하기 위하여 다음과 같은 실험을 수행하였다. 녹음환경이 비교적 일정한 코퍼스와 녹음환경의 변화가 심한 코퍼스를 사용하여 음절크기 합성단위로 5개의 문장을 합성하였다. 녹음환경이 일정한 코퍼스에서 2문장, 녹음환경의 변화가 심한 코퍼스에서 3문장을 합성하되 각각의 문장에 대해서 음성세그먼트별로 최대진폭 정규화를 한 경우와 정규화를 하지 않은 경우로 나누어 2개의 합성음성을 만들어 서로 비교하였다. 공정한 평가를 위하여 평가 이전에 두 합성음성이 가지는 평균에너지를 균일하게 맞추었다.



(a) 정규화 이전 파형.



(b) 최대진폭 정규화 이후 파형.

그림 3. "여러분 ~"의 대표적 3개 파형.

합성단위 선택은 음성세그먼트 정규화를 염두에 두고 선택하였다. 음성의 에너지는 자음보다는 모음에 집중되어 있으므로 정규화하고자 하는 진폭최대치를 항상 모음에서 찾을 수 있도록 모음 하나를 반드시 포함시키는 합성단위인 음절을 선택하였다.

합성한 문장 5개는 다음과 같다.

(1)우리는 민족중흥의 역사적 사명을 띠고 이 땅에 태어났다.

(2)조상의 빛난 얼을 오늘에 되살려 안으로 자주 독립의 자세를 확립하고, 밖으로 인류 공영에 이바지할 때다.

(3)이에, 우리의 나아갈 바를 밝혀 교육의 지표로 삼는다.

(4)성실한 마음과 튼튼함 몸으로, 학문과 기술을 배우고 익히면서

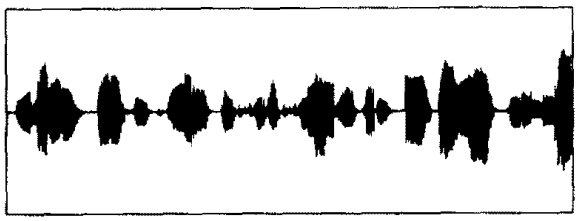
(5)민족의 슬기를 모아 줄기찬 노력으로 새 역사를 창조하자.

(1),(4),(5) 문장은 녹음환경의 변화가 심한 코퍼스를 사용하여 합성하였고 (2),(3) 문장은 비교적 일정한 녹음환경의 코퍼스를 사용하여 합성하였다.

3.2 실험결과 및 고찰

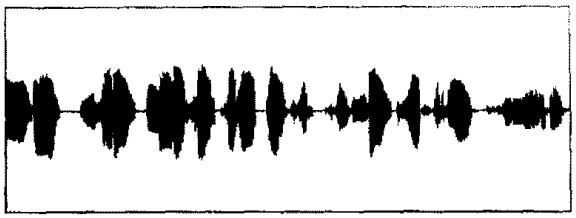
그림 4의 (a),(b),(c),(d),(e)는 (1)~(5) 각각의 문장을 음성세그먼트별 정규화 없이 합성한 음성의 파형

(a)



(b)

(c)



(d)

(e)

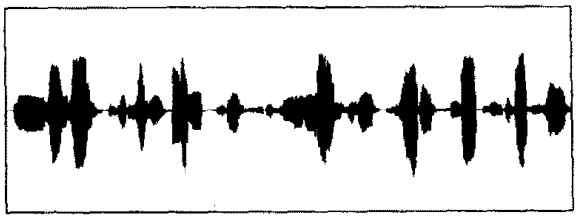


그림 4. 비정규화 합성음성파형.

이다. 그림 5의 (a),(b),(c),(d),(e)는 각 문장을 음성세

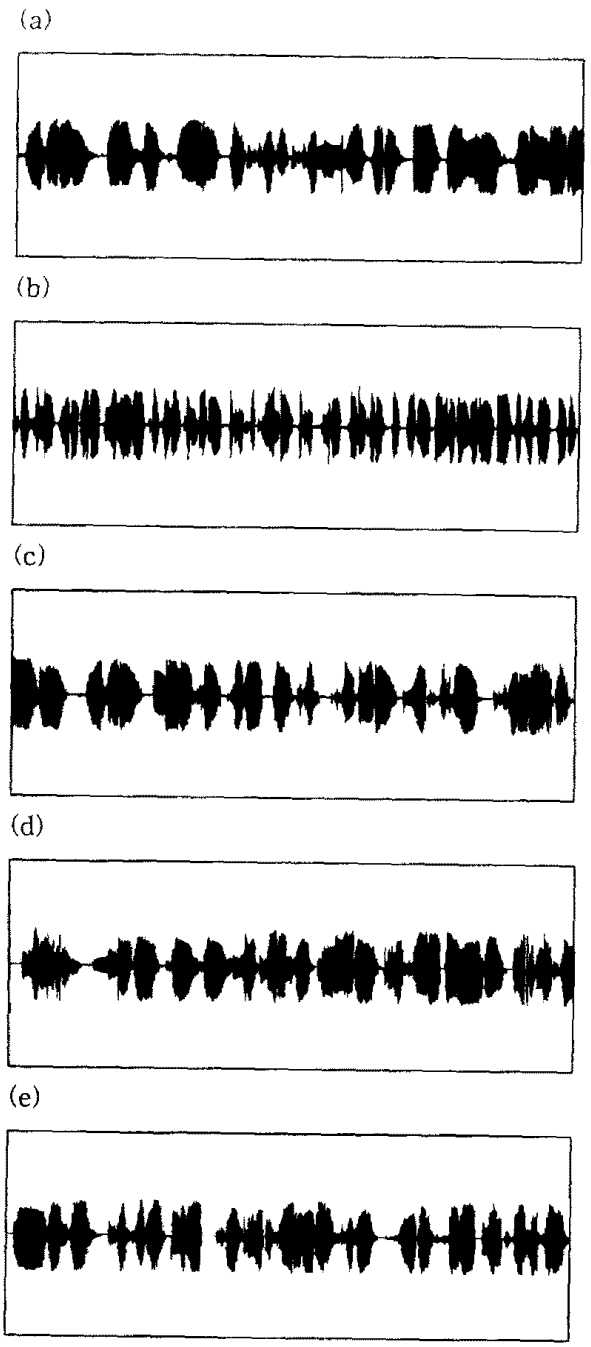


그림 5. 정규화 합성음성파형.

그먼트별 에너지 정규화를 한 후 합성한 음성의 파형이다.

그림 4의 합성음성 파형에서 (a),(d),(e)는 각 모음의 진폭의 변화가 크며 (b),(c)는 진폭의 변화가 비교적 적음을 볼 수 있다. 파형에서 관찰되는 이러한 특징은 실제음성 청취시 더욱 명확하게 된다. 그림 4의 (a),(d),(e)에 해당하는 음성에서는 청취시 앞서 논의한 발음세기의 차이로 인한 음질저하가 쉽게 확인되었지만 (b),(c)에 해당하는 음성은 발음세기 차이로 인한 음질저하정도가 미미하였다. 이는 문맥요소나 운율요소로 인한 발음세기의 차이가 녹음환경의 변화로 인한 발음세기의 차이에 비하여 음질저하에 미치는 영향이 적음을 알 수 있다.

음성세그먼트별 에너지 정규화를 거쳐서 합성된 음성의 파형은 그림 5에 보여지고 있으며 그림 4와 비교하여 (a),(d),(e)의 경우 진폭에 뚜렷한 차이가 관찰된다. 그림 4의 파형에서는 모음의 진폭변화가 불규칙적이고 크다면, 그림 5의 파형에서는 모음의 진폭변화가 적음을 알 수 있다. 따라서 음성세그먼트별 에너지 정규화를 통한 음질개선을 기대할 수 있으며 실제로 청취해본 결과 뚜렷한 음질개선이 있었다. 그러나 그림4와 5의 (b),(c) 파형에서는 큰 차이가 관찰되지 않았다. 실제 음성을 청취시에도 음성세그먼트별 정규화한 경우의 합성음성과 정규화 없이 바로 합성된 음성의 음질은 비슷하였다.

Ⅳ. 결 론

본 논문에서는 코퍼스 기반 음성합성시 각 음성세그먼트의 에너지 정규화가 음질에 미치는 영향과 음질의 개선가능성에 대하여 연구하였다. 코퍼스의 녹음환경이 일정하지 않은 경우 음성세그먼트의 에너지 정규화를 통해 음성을 합성한 경우가 정규화를 하지 않은 경우보다 음질이 개선되었다.

참고문헌

[1] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 1997.

[2] F. Chou, C. Tseng, "Corpus-based Mandarin Speech Synthesis with Contextual Syllable Units Based on Phonetic Properties," *ICASSP*, 1997.

[3] 김재홍, 조관선, 이철희, "코퍼스에 기반한 반음절 단위의 한국어 음성합성 시스템," 1998 한국통신학회 추계학술대회.