Journal of the Korea Academia-Industrial cooperation Society
Vol. 14, No. 8 pp. 3992-3998, 2013

http://dx.doi.org/10.5762/KAIS.2013.14.8.3992

가변 운율 모델링을 이용한 고음질 감정 음성합성기 구현에 관한 연구

민소연[1*], 나덕수[2]
[1]서일대학교 정보통신과, [2](주)보이스웨어 부설연구소

A Study on Implementation of Emotional Speech Synthesis System using Variable Prosody Model

So-Yeon Min[1*] and Deok-Su Na[2]
[1]Dept. of Information and Communication, Seoil University
[2]Voiceware co. Ltd, R&D center

요 약 본 논문은 고음질의 대용량 코퍼스 기반 음성 합성기에 감정 음성 코퍼스를 추가하여 보다 다양한 합성음을 생성할 수 있는 방법에 관한 것이다. 파형 접합형 합성기에서 사용할 수 있는 형태로 감정 음성 코퍼스를 구축하여 기존의 일반 음성 코퍼스와 동일한 합성단위 선택과정을 통해 합성음을 생성할 수 있도록 구현하였다. 감정 음성 합성을 위해 태그를 사용하여 텍스트를 입력하고, 억양구 단위로 일치하는 데이터가 존재하는 경우 감정 음성으로 합성하고, 그렇지 않은 경우 일반 음성으로 합성하도록 하였다. 그리고 음성에서 운율을 구성하는 요소로 휴지기(break)가 있는데, 감정 음성의 휴지기는 일반 음성보다 불규칙한 특성이 있다. 따라서 합성기에서 생성되는 휴지기 정보를 감정 음성 합성에 그대로 사용하는 것이 어려워진다. 이 문제를 해결하기 위해 가변 휴지기(Variable break)[3] 모델링을 적용하였다. 실험은 일본어 합성기를 사용하였고, 그 결과 일반 음성의 휴지기 예측 모듈을 그대로 사용하면서 자연스러운 감정 합성음을 얻을 수 있었다.

Abstract This paper is related to the method of adding a emotional speech corpus to a high-quality large corpus based speech synthesizer, and generating various synthesized speech. We made the emotional speech corpus as a form which can be used in waveform concatenated speech synthesizer, and have implemented the speech synthesizer that can be generated various synthesized speech through the same synthetic unit selection process of normal speech synthesizer. We used a markup language for emotional input text. Emotional speech is generated when the input text is matched as much as the length of intonation phrase in emotional speech corpus, but in the other case normal speech is generated. The BIs(Break Index) of emotional speech is more irregular than normal speech. Therefore, it becomes difficult to use the BIs generated in a synthesizer as it is. In order to solve this problem we applied the Variable Break[3] modeling. We used the Japanese speech synthesizer for experiment. As a result we obtained the natural emotional synthesized speech using the break prediction module for normal speech synthesize.

Key Words : Variable Break, Emotional Speech Synthesizer

1. 서론

문자(text) 정보를 사람의 목소리로 변환 해 주는 음성 합성 기술에 있어, 이제 더 이상 좋은 음질만으로 충분치 않은 시대가 되었다. 1990년에서 2000년 초반에는 열악한 합성음 음질로 인해 정보전달의 정확성과 운율의 자

본 논문은 2012년도 서일대학교 학술연구비에 의해 연구되었습니다.
*Corresponding Author : So-Yeon Min(Seoil Univ.)
Tel: +82-2-490-7583 email: symin@seoil.ac.kr
Received June 19, 2013 Revised July 1, 2013 Accepted August 7, 2013

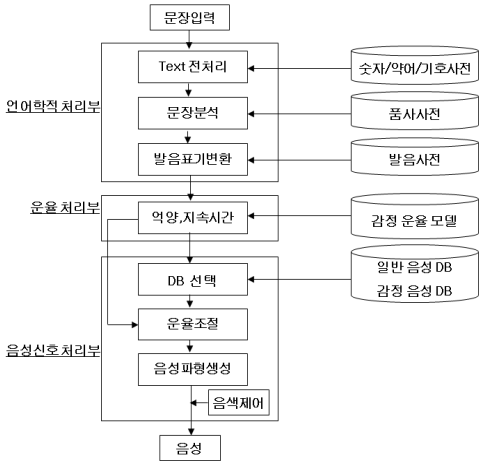
연스러움이 합성기의 평가 조건이었다. 그러나 현재는 비약적인 합성음 음질의 향상으로 인해 그 사용 범위가 확대되어, 정확한 정보전달을 목적으로 하는 네비게이션과 교통정보 안내 등에서 긴급함을 알리는 재난 재해 경고 방송, 프로그램의 흥미를 유발하는 방송 멘트 제작 및 만화 캐릭터 음성 제작까지 두루 사용되어지고 있다. 그러나 아직 까지 감정을 표현 할 수 있는 음성 합성기술의 연구가 제한적으로 이루어지고 있어 상용화되어지고 있지 않다. 기존의 감정 음성 합성기술 또한 신호처리를 기반으로 한 연구들이 대부분으로 합성음의 음질이 매우 좋지 못하였다. 따라서 현재 최고의 합성음질을 얻을 수 있는 코퍼스 기반 음성합성기술을 이용하여 다양한 감정을 표현하는 합성기술을 개발한다면 보다 다양한 응용분야에 적용할 수 있을 뿐 아니라 새로운 부가가치를 창출하는 기술로 발전할 것이다.

코퍼스 기반 음성 합성기에서 안정적인 합성음의 음질을 얻기 위해서는 하나의 발성 스타일만을 가지는 음성 코퍼스를 구축하는 것이 유리하지만 이러한 합성기는 결과적으로 단순히 정보를 전달하는 목적을 가지는 분야에만 사용될 수밖에 없다. 기쁜 뉴스나, 슬픈 뉴스 또는 정감 있고 따뜻한 소식, 흥분되는 상황, 다급하고 위험한 정보를 전달하는 분야에는 사용할 수 없고, 사용되더라도 텍스트가 가지는 전체적인 정보를 정확히 전달할 수 없게 된다. 하나의 음성 합성기가 다양한 감정을 표현할 수 있다면 위와 같은 여러 분야의 텍스트의 정보를 보다 정확히 전달할 수 있고, 미래의 인간형 로봇이나 감성지능형 컴퓨팅 등에도 적용될 수 있을 것이다.

2. 감정 음성합성기

감정 음성 음성합성을 위한 코퍼스를 구축하기 위해서는 성우가 발성한 음성신호를 언어의 운율정보를 바탕으로 재조합이 가능한 구조로 체계적으로 나누어 놓아야 한다. 세분화한 운율구 구조는 IP(Intonation Phrase), MP(Major Phrase), AP(Accentual Phrase), word, syllable, phoneme 등으로 나누어진다. 여기서 IP란 한 번의 발화(utterance)에서 나타나는 한 호흡정도의 길이(the length of breath group)를 의미한다. 그리고 하나의 IP는 하나 이상의 MP로 구성되고 MP는 다시 악센트 구(AP; Accentual Phrase)로 구성된다. 악센트 구는 하나 이상의 단어(word)로 구성되고, 단어는 하나 이상의 음절(syllable)로 구성되고, 음절은 또 하나 이상의 음소(phoneme)로 구성되어진다[1,2]. 본 연구에서 구현하는 음성 합성기는 음소를 기본 단위로 하여 음성합성이 이루어진다.

감정 음성 합성기를 구현하기 위해서는 감정 음성 코퍼스 구축뿐만 아니라 운율 모델링이 무엇보다 중요하다. 운율이란 음성의 크기, 피치, 음소 지속시간, 띄워 읽기, 악센트 등으로 감정을 표현하는 직접적인 데이터이다. 그러나 감정 음성의 운율은 변화가 불규칙하여 기존의 모델링을 이용하기가 어렵고, 모델링에 필요한 대용량의 감정 코퍼스를 구축하는 것도 힘들다. 그리고 대용량의 감정 코퍼스를 구축한다고 해도 음성파형 접합형 합성기에서 합성단위 선택을 통해 고음질의 감정 합성음을 생성하는 것은, 감정 음성의 불규칙한 운율로 인해 매우 어렵다. 따라서 본 논문에서는 제한적이지만 고음질의 감정 합성음을 얻을 수 있는 방법을 제안한다. 먼저 구축된 일반 음성 코퍼스에 감정 음성 코퍼스를 추가하여 합성단위 선택과정을 통해 억양구 단위로 일치하는 감정 음성 데이터가 존재하는 경우 감정 합성음을 생성하고 그렇지 않은 경우 일반 합성음을 생성하도록 구현하였다. 그리고 이를 위해 합성기의 휴지기 예측기에 가변 휴지기 모델[3]을 적용하였다.



[Fig. 1] Proposed Emotional TTS

3. 가변운율 모델링

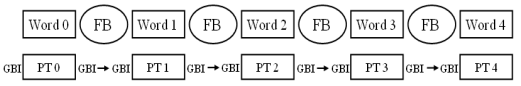
3.1 가변 휴지기

운율구의 구조를 나타내기 위해 break indices(BI)를 사용하는데, BI를 생성하는 방법은 규칙 기반 방법과 코퍼스 기반 방법이 있다. 규칙 기반 방법은 문장기호, 품사, 발음열(phoneme stream) 등의 정보와 언어학적인 정보를 이용한다. 정확한 BI를 얻기 위해서는 매우 복잡하

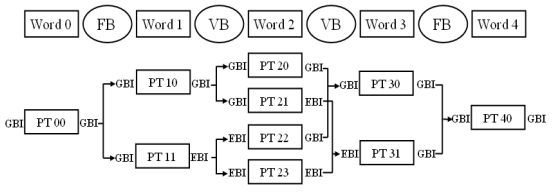
고 정교한 작업이 필요하다. 코퍼스 기반 방법에는 여러 가지 특징들을 이용하여 자동으로 decision tree를 구축하는 classification and regression trees(CART) 방법과, hidden Markov model(HMM)을 이용하는 방법이 있다[3]. 하지만 개인마다 조금씩 다른 읽기습관(reading style)에 의해 모델링의 정확도가 높지 못하고, 이것은 break index 사이의 불명확성(uncertainty, subtleness)이 그 원인이며 이것이 BI 예측을 어렵게 하는 요인이다[4].

대용량 음성 코퍼스를 사용하는 경우 합성단위 선택 과정에서 목표(target) break 뿐만 아니라 유사한 break들을 모두 이용한다면 음성 코퍼스에 포함된 다양한 break 정보와 부족한 규칙에 대한 보완을 효율적으로 수행 할 수 있고, 다양한 문맥정보를 가지는 합성단위들을 후보로 추출하여 합성음의 음질을 향상 시킬 수 있다. 감정 음성 코퍼스를 일반 음성 코퍼스와 동시에 사용하기 위해서는 이러한 유사한 break의 범위를 보다 확장하는 것이 필요하다. 이렇게 확장 가능한 break를 가변 휴지기(VB, Variable Break)로 정의할 수 있고, 그렇지 않고 확장하지 않아야 하는 것을 고정 휴지기(FB, Fixed Break)라고 할 수 있다[5]. 이러한 가변 휴지기와 고정 휴지기 정보를 이용하여 다양한 형태의 목표 운율구를 구성할 수 있다.

3.2 가변 휴지기를 이용한 Multiple Prosodic Targets



(a) Single Prosodic Target.



(b) Multiple Prosodic Targets.

[Fig. 2] Single and Multiple Prosodic Targets. (PT-Prosodic Target, GBI-Generated BI, EBI-Expanded BI)

합성기에서 합성단위 선택을 위해 BI, 음소 지속시간, 기본주파수 포락선으로 구성되는 prosodic target (PT)을 생성한다. 기존의 합성기에서는 BI가 생성되면 변하지 않기 때문에 하나의 PT(single PT)만을 사용하지만 가변 휴지기를 사용하게 되면 생성된 BI가 바뀔 수 있기 때문에 다양한 PT(multiple prosodic targets)를 나타낼 수 있다[6]. 그리고 이러한 다양한 PT를 적용하여 감정음성과 일반음성의 합성단위 선택을 수행할 수 있다.

4. 실험 및 결과

4.1 감정 음성 코퍼스

감정 음성 합성기를 구현하기 위해 먼저 감정 음성을 녹음하여 일반 음성과 동일한 코퍼스를 구축하여야 한다. 본 논문에서는 일본어 합성기를 사용하여 실험하였는데, 인사말을 Happiness, Sadness의 두 가지 Category로 분류하고, Level을 2단계로 하였다. Level 1은 감정이 적게 들어가게 낭독하고, Level 2는 Level 1 보다 감정을 더 표현하게 하였다.

[Table 1] Emotional Text Corpus (Number of Sentence)

Category	Level 1	Level2
Happiness	13	172
Sadness	53	109

[Table 2] Example of Emotional Sentence

Category	Sentence
Happiness	お電話ありがとうございます！ 毎度ありがとうございます！ お世話になります！ お世話になっております！ ありがとうございました！
Sadness	ただ今、營業時間外となっております！ 本日の營業は終了いたしました！ 本日は休業とさせていただいております！ ただ今、電話が込み合っております！ このまま、しばらくお待ちください！

4.2 휴지기 분류 (Break Index)

합성기에서 사용하는 휴지기는 Table 3과 같이 10가지 종류로 분류하여, BI 2, 3에 대해여 가변 휴지기 모델을 적용하여 합성단위 검색을 수행하였다. 감정 음성 코퍼스는 일반 음성 코퍼스와 동일한 구조의 BI 정보를 가지지만, 합성기에서 BI를 단순화 하여 합성단위를 검색하도록 하였다. 그 이유는 3장에서 언급한 것과 같이 일반 음성에 비해 휴지기 예측이 어려워 합성기에서 예측한 휴지기 정보와 코퍼스의 휴지기 정보가 일치하지 않는 경우가 많기 때문이다.

[Table 3] Break Indices used in this paper.

10 degrees of BIs	
0	No prosodic break : same as the J_ToBI usage
1	Prosodic word boundary (WB) : same as the J_ToBI usage
2	Accentual phrase or minor tone group boundary (AP) : No BPM at the end & No followed by a pause & Without pitch range resetting
3	Major phrase (intermediate phrase) boundary (MP) : BPM at the end & Followed by a pause & Without pitch range resetting
4	Intonation phrase boundary (IP) : BPM at the end & Followed by a pause & pitch range resetting
5	Sentence boundary (SB)
6	Question boundary
7	Emotional setence boundary
8	Emotional merged boundary (BI 0, 1, 2)
9	Emotional merged boundary (BI 3, 4)

실험에서 사용한 일본어는 억양구마다 피치의 변화 범위(pitch range)가 달라지고, 억양구의 끝에서는 피치가 변하는 몇 가지 패턴이 나타나는데 이러한 것을 BPMs(boundary pitch movements)[7-13]라고 한다. BI 2는 AP 경계이지만 BPM이 나타나지 않고, 바로 이어서 포즈가 오지 않으며 AP의 피치 변화 범위도 이어지는 AP에 영향을 받는 것을 의미하고, BI 3은 BPM이 나타나고, 포즈가 이어지지만 AP의 피치 변화 범위가 이어지는 AP에 영향을 받지 않는 것을 의미한다. BI 4는 BPM이 나타나고 포즈도 이어지며 AP의 피치 변화 범위도 이어지는 AP와 상관성이 없는 것을 의미하고, BI 5는 일반 문장의 끝을 나타내고, BI 6은 의문문의 끝을 나타낸다. BI 7, 8, 9는 감정 합성에 사용되는 것으로 7은 감정 문장의 경계를 나타내고, 8은 일반 문장에서 사용하는 BI 0, 1, 2를 단순화한 것이다. 즉 중간에 포즈로 분리되지 않는 BI이고, 9는 일반 문장에서 사용하는 BI 3, 4를 단순화한 것으로 포즈가 들어가는 BI이다.

Table 4는 가변 휴지기 모델을 적용했을 때, 합성기에서 예측한 BI와 합성단위 검색에서 사용되는 BI 정보이다. 합성기에서 예측한 BI가 2이고 고정 휴지기(Fixed Break)이면 일반 텍스트인 경우 BI가 2인 합성단위를 검색하고, 감정 텍스트인 경우 8인 합성단위를 코퍼스에서 검색한다. 그러나 예측한 BI가 가변 휴지기(Variable Break)인 2이면, 일반 텍스트에서는 BI가 2 또는 3인 합성단위를 함께 검색하고, 감정 텍스트인 경우 코퍼스에서 8과 9의 BI를 가지는 합성 단위를 함께 검색한다. 즉 가변 휴지기인 경우 합성단위 후보의 수가 고정 휴지기 보다 많아지게 된다.

[Table 4] Extension of Variable Break

Predicted BI		Normal Text	Emotional Text
2	Fixed	2	8
	Variable	2, 3	8, 9
3	Fixed	3	9
	Variable	2, 3, 4	8, 9

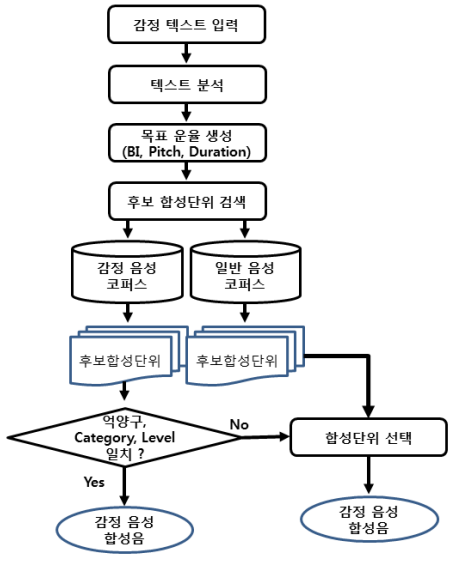
4.3 감정 텍스트 입력

본 논문에서는 감정 음성을 합성하기 위해 Markup-Language를 이용하여 사용자가 원하는 감정과 Level을 지정할 수 있도록 구현하였다.

<vtml_emotion category="happiness" level="1">お電話ありがとうございます！</vtml_emotion>
<vtml_emotion category="sadness" level="1">ただ今、営業時間外となっております！</vtml_emotion>

4.4 합성단위 검색

감정 합성음은 입력 텍스트가 위와 같이 Markup-Language를 사용하여 입력되어야 하고, Category와 Level이 일치하는 문장이 감정 코퍼스에 억양구(IP) 단위로 일치하는 경우에만 합성하도록 구현하였다. 억양구 보다 작은 악센트구(AP), 또는 단어, 음소 단위로 합성할 수도 있지만 감정 음성의 운율 변화가 심하고, 코퍼스 크기가 충분하지 못하여 고음질의 감정 합성음을 생성하기가 어렵다. 만일 억양구 단위로 일치하는 것이 없는 경우 일반 문장으로 합성하도록 하였다.



[Fig. 3] The flowchart of the proposed unit selection

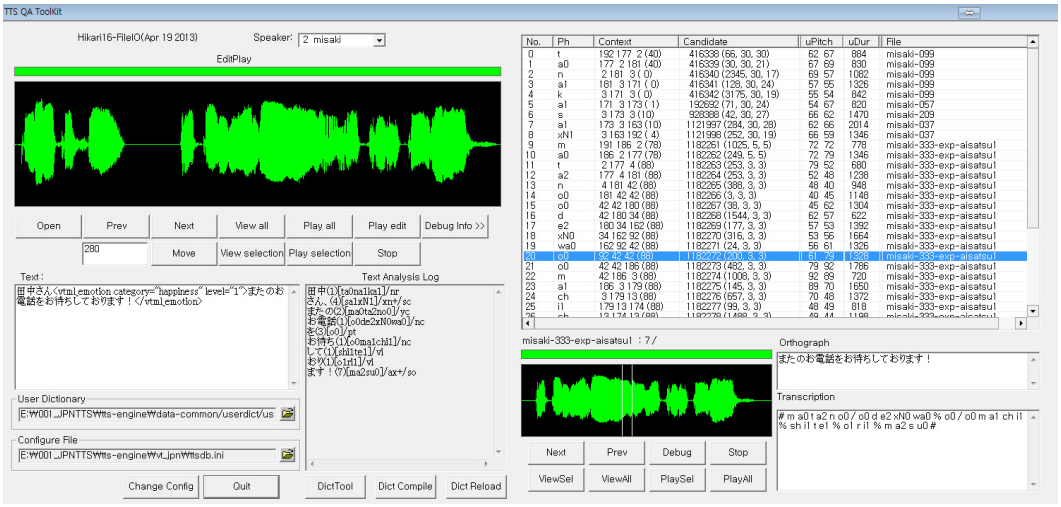
4.4 합성결과

Fig. 4와 Fig. 5는 합성 결과를 나타낸다. 그림의 왼쪽 부분은 합성음 및 텍스트 분석 결과이고, 오른쪽 상단은 합성단위 검색 및 합성단위 선택 결과이고, 오른쪽 하단은 합성에 사용된 음성 코퍼스를 보여주고 있다. 입력텍스트는 다음과 같다.

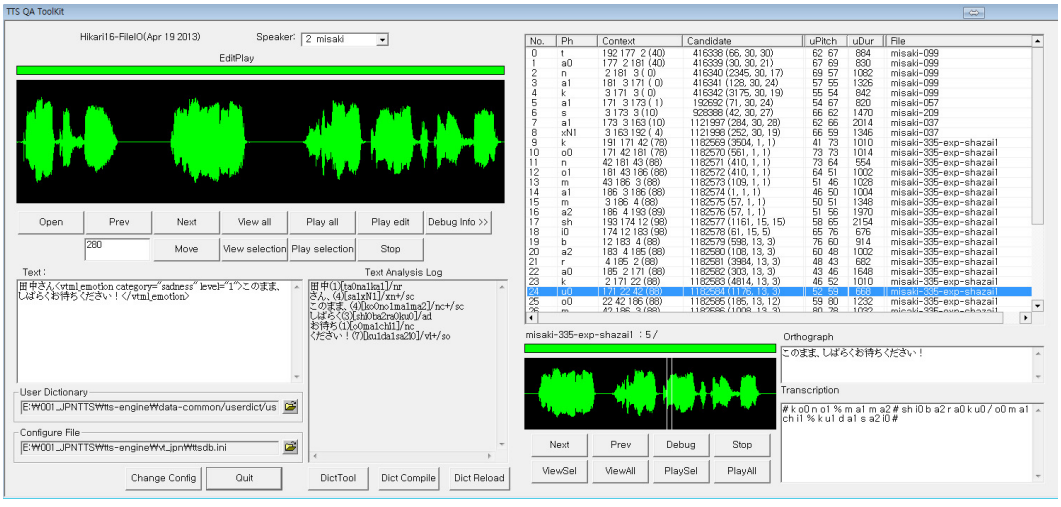
田中さん<vtml_emotion category="happiness" level="1">またのお電話をお待ちしております！</vtml_emotion>
田中さん<vtml_emotion category="sadness" level="1">このまま、しばらくお待ちください！</vtml_emotion>

"田中さん"은 일반 텍스트이고, 나머지는 감정 텍스트이다. 오른쪽 상단의 합성단위 선택 결과를 보면, "田中さん" 부분에서는 일반 텍스트의 합성과정과 같이 서로 다른 파일의 음성 코퍼스에서 후보를 가져와 연결하는 반면, 감정 텍스트 부분은 서로 연결된 후보를 가져온다. ([Candidate] column의 숫자는 후보 합성단위의 ID를 나타내는데 차이가 1이면 연결된 합성단위이다.)

그리고 텍스트 분석 결과를 보면 BI가 3인 단어가 있는데 (Fig. 4에서 を(3)[o0]/pt, Fig 5에서 しばらく(3)[shi0ba2ra0ku0]/ad), BI가 3이면 실제 포즈가 합성음에 나타나야 하지만, 가변 휴지기에 의한 합성단위 검색에



[Fig. 4] A Simulation of Emotional Speech Synthesis (Category : Happiness)



[Fig. 5] A Simulation of Emotional Speech Synthesis (Category : Sadness)

의해 포즈가 없는 BI 2의 합성단위를 선택하여 합성음을 생성한 것을 알 수 있다. 따라서 합성기의 휴지기예측 결과 고정적으로 적용하지 않고 가변적으로 적용함으로써 코퍼스에 저장된 감정 합성음을 정확히 가져올 수 있었다.

5. 결 론

음성 합성기는 대용량 코퍼스를 기반으로 하는 음성파형 접합형 합성방법의 발전으로 고음질의 합성음을 생성할 수 있게 되었다. 그러나 다양한 분야로의 사용 영역이 확대 되면서 더 이상 단조로운 낭독체의 합성음에 만족하지 못하게 되었다. 따라서 본 논문에서는 감정 음성 코퍼스와 일반 음성 코퍼스를 동시에 이용하여 입력 텍스트가 감정 음성 코퍼스에서 억양구 단위로 일치하는 경우 감정 합성음을 생성하고, 그렇지 않은 경우 일반 합성음 출력할 수 있는 합성기를 구현하였다.

References

[1] S. Kiriyama, S. Kitazawa, “Evaluation of a prosodic labeling system utilizing linguistic information,” Proc. INTERSPEECH2004, pp.2993-2996, 2004.

[2] K. Maekawa, H. Kikuchi, Y. Igarashi, J. Venditti, "X-JToBI: an extended j-toBI for spontaneous speech", Proc. ICSLP-2002, pp.1545-1548, 2002.

[3] S. H. Lee, Y. H. Oh. "The Modelling of Prosodic Phrasing and Pause Duration using CART", Proceeding of the Acoustical society of Korea, Vol. 17 No. 1, pp 81-86, 1998.

[4] Campbell, N, “Autolabeling Japanese ToBI,” Proc. ICSLP'96, vol.4, pp.2399-2402, 1996.

[5] D. S. Na, M. J. Bae, “A Variable Break Prediction Method using CART in a Japanese Text-to-Speech System," IEICE Trans. Inf. & Syst., Vol. E92-D, No.2, pp.349-352, 2009.
DOI: http://dx.doi.org/10.1587/transinf.E92.D.349

[6] D. S. Na, S. Y. Min, J. S. Lee, M. J. Bae,, “A Performance Improvement Method using Variable Break in Corpus Based Japanese Text-to-Speech System,” The Journal of the Acoustical Society of Korea, Vol. 28, No. 2, pp.155-163, 2009.

[7]. J. Venditti, J. “The J_ToBI model of Japanese intonation”, in S. A. Jun Ed., Prosodic Typology and Transcription: A Unified Approach: Oxford University Press, pp.172-200.

[8] K. Maekawa, H. Kikuchi, Y. Igarashi, J. Venditti, "X-JToBI: an extended j-toBI for spontaneous speech", Proc. ICSLP-2002, pp.1545-1548, 2002.

[9] K.-H. Kim, H.-M. Kim, K.-Y. Lee, M.-J. Lim, J.-L. Kim, "Design And Implementation of a Speech Recognition Interview Model based-on Opinion Mining Algorithm", Journal of The Institute of Webcasting, Internet and Telecommunication, Vol 12, No 1, pp. 225~230, 2012.

[10] S.-H. Kim, J.-Y. Ahn, "A Study on the Voice Interface for Mobile Environment", Journal of The Institute of Webcasting, Internet and Telecommunication, Vol 13, No 1, pp. 199~204, 2013.

[11] J. J. Im, "Development of energy expenditure measurement device based on voice and body activity", Journal of The Institute of Webcasting, Internet and Telecommunication, Vol 12, No 6, pp. 303~309, 2012.

[12] J.-Y. Ahn, S.-B. Kim, S.-H. Kim, K.-I. Hur, "A study on Voice Recognition using Model Adaptation HMM for Mobile Environment", Journal of The Institute of Webcasting, Internet and Telecommunication, Vol 11, No 3, pp. 175~180, 2011.

[13] W. Oh, E. Rhee, "Curriculum Development of Acoustics and Audio Engineering on Digital Convergence Environment", Journal of The Institute of Webcasting, Internet and Telecommunication, Vol 13, No 2, pp. 191~197, 2013.

민 소 연(So-Yeon Min) **[종신회원]**

- 1994년 2월 : 숭실대학교 전자공학과 (공학사)
- 1996년 2월 : 숭실대학교 일반대학원 전자공학과 (공학석사)
- 2003년 2월 : 숭실대학교 일반대학원 전자공학과 (공학박사)
- 2005년 3월 ~ 현재 : 서일대학교 정보통신과 부교수

<관심분야>
통신 및 신호처리, 임베디드 시스템

나 덕 수(Deok-Su Na) **[정회원]**

- 1998년 2월 : 숭실대학교 정보통신공학과 (공학사)
- 2000년 2월 : 숭실대학교 정보통신공학과 (공학석사)
- 2009년 2월 : 숭실대학교 정보통신공학과 (공학박사)
- 2001년 11월 ~ 현재 : (주)보이스웨어 연구소 책임보

<관심분야>
음성합성 시스템, 음성 신호처리