Journal of the Korea Academia-Industrial cooperation Society
Vol. 15, No. 9 pp. 5763-5768, 2014

http://dx.doi.org/10.5762/KAIS.2014.15.9.5763
ISSN 1975-4701 / eISSN 2288-4688

HMM 기반 감정 음성 합성기 개발을 위한 감정 음성 데이터의 음색 유사도 분석

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

Analysis of Voice Color Similarity for the development of HMM Based Emotional Text to Speech Synthesis

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

요 약 하나의 합성기에서 감정이 표현되지 않는 기본 음성과 여러 감정 음성을 함께 합성하는 경우 음색을 유지하는 것이 중요해 진다. 감정이 과도하게 표현된 녹음 음성을 사용하여 합성기를 구현하는 경우 음색이 유지되지 못해 각 합성음이 서로 다른 화자의 음성처럼 들릴 수 있다. 본 논문에서는 감정 레벨을 조절하는 HMM 기반 음성 합성기를 구현하기 위해 구축한 음성데이터의 음색 변화를 분석하였다. 음성 합성기를 구현하기 위해서는 음성을 녹음하여 데이터베이스를 구축하게 되는데, 감정 음성 합성기를 구현하기 위해서는 특히 녹음 과정이 매우 중요하다. 감정을 정의하고 레벨을 유지하는 것은 매우 어렵기 때문에 모니터링이 잘 이루어져야 한다. 음성 데이터베이스는 일반 음성과 기쁨(Happiness), 슬픔(Sadness), 화남(Anger)의 감정 음성으로 구성하였고, 각 감정은 High/Low의 2가지 레벨로 구별하여 녹음하였다. 기본음성과 감정 음성의 음색 유사도 측정을 위해 대표 모음들의 각각의 스펙트럼을 누적하여 평균 스펙트럼을 구하고, 평균 스펙트럼에서 F1(제1포만트)을 측정하였다. 감정 음성과 일반 음성의 음색 유사도는 Low-level의 감정 데이터가 High-level의 데이터 보다 우수하였고, 제안한 방법이 이러한 감정 음성의 음색 변화를 모니터링 할 수 있는 방법이 될 수 있음을 확인할 수 있었다.

Abstract Maintaining a voice color is important when compounding both the normal voice because an emotion is not expressed with various emotional voices in a single synthesizer. When a synthesizer is developed using the recording data of too many expressed emotions, a voice color cannot be maintained and each synthetic speech is can be heard like the voice of different speakers. In this paper, the speech data was recorded and the change in the voice color was analyzed to develop an emotional HMM-based speech synthesizer. To realize a speech synthesizer, a voice was recorded, and a database was built. On the other hand, a recording process is very important, particularly when realizing an emotional speech synthesizer. Monitoring is needed because it is quite difficult to define emotion and maintain a particular level. In the realized synthesizer, a normal voice and three emotional voice (Happiness, Sadness, Anger) were used, and each emotional voice consists of two levels, High/Low. To analyze the voice color of the normal voice and emotional voice, the average spectrum, which was the measured accumulated spectrum of vowels, was used and the F1(first formant) calculated by the average spectrum was compared. The voice similarity of Low-level emotional data was higher than High-level emotional data, and the proposed method can be monitored by the change in voice similarity.

Key Words : Emotional Speech Synthesis, Voice Color Similarity

본 논문은 2013년도 서일대학교 학술연구비에 의해 연구되었음.
*Corresponding Author : So-Yeon Min(Seoil Univ.)
Tel: +82-2-490-7583 email: symin@seoil.ac.kr
Received May 27, 2014 Revised June 16, 2014 Accepted September 11, 2014

1. 서론

음성 합성기는 대용량 코퍼스 기반의 음성파형 접속형 합성기와 HMM(Hidden Morkov Model) 기반 음성합성기로 활발히 연구되고 상용화 되고 있다.

대용량 코퍼스를 이용한 음성합성기의 경우 고음질 합성이 가능하여 웹사이트 뿐 아니라 휴대형 단말기인 내비게이션과 스마트폰 등에 널리 사용되고 있다.

HMM 기반 음성합성기도 이전에는 학교나 연구소에서 주로 연구용 합성기로 많이 사용되었지만 최근에는 음질의 향상으로 상용 합성기형태로 점차 보급이 늘어나는 추세이다. HMM 기반 음성 합성기는 대용량 코퍼스를 이용한 음성파형 접속형 합성기에 비해 적은 녹음 음성으로도 합성기 구현이 가능하고, 화자 적응 기술을 이용한 음색 변환이 가능하여 다양한 언어와 여러 화자의 합성기를 함께 사용하기를 원하는 분야에 접목되고 있다.

음성 합성기의 사용 분야가 늘어남으로써, 사용자들은 기존의 고음질의 합성음과 더불어 보다 재미있고 다양한 응용을 할 수 있는 합성기를 원하는 방향으로 바뀌고 있다. 즉 합성음을 이전의 정보 전달 목적 뿐 아니라 의도나 감정까지 전달하는데 사용할 수 있기를 원하고 있다. 그러나 일반적으로 감정을 표현한 음성은 피치나 지속시간 등의 변화 범위가 커서 음성파형 접속형 합성기를 이용하여 감정 음성 합성기를 구현하기 힘들어진다. 현재 이러한 합성기에서 감정을 구현하는 방법으로는 특정 문장이나 단어 등에 해당하는 녹음된 감정 음성을 이용하여 일반 합성음 사이에 삽입하여 감정 합성음의 효과를 주는 방법이 있다. 즉 이전의 음향 데이터를 합성음 중간에 삽입하는 형태와 같이 감정 음성 데이터를 삽입하여 보다 다양한 합성음을 출력할 수 있게 한 것이다. 그리고 HMM 기반 음성합성 방법을 이용한 감정 음성 합성도 활발히 연구되어지고 있는데, HMM 기반 합성기는 합성 방법에 따른 단점인 보코더 음질과 운율의 평탄화로 인해 성능 측면에서 상용화에 제한적이었다. 그러나 최근에 다양한 모델링 기법이 도입되어 많은 음질 개선이 이루어지고 있다[1-4]. 이러한 추세로 캐릭터 제품이나 스마트폰과 같은 다양한 단말기에서 HMM 기반 음성 합성기를 탑재하고 있다.

본 논문에서는 HMM 기반 음성 합성기를 이용하여 감정합성기를 구현하고, 감정 음성 데이터의 감정 레벨에 따른 음색 변화를 분석하였다. HMM 기반 음성 합성기는 기본적으로 운율이 평탄화 되는데, 이것은 운율뿐만 아니라 음색에도 영향을 미치게 된다. 녹음과정에서 감정 표현의 정도가 클수록 그 영향은 커지게 되는데, 하나의 합성기에 기본 음성과 여러 감정 음성을 합성하는 경우 기본 음성과의 음색 차이로 인해 성능에 문제가 될 수 있다. 따라서 녹음 음성의 음색을 미리 분석하고 감정을 모니터링하면서 녹음을 진행하는 방법이 필요하게 된다. 본 논문에서는 대표 모음의 평균 스펙트럼과 F1(제1포만트)을 이용하여 감정 음성 합성기 구현 및 평가에 필요한 기본 음성과 감정 음성의 음색을 비교할 수 있는 방법을 제안한다.

2. 감정 음성 데이터

감정 음성 데이터를 구축하기 위해서는 우선 녹음 시나리오를 작성해야 한다. 일반적으로 음성 녹음에 필요한 시나리오는 다량의 텍스트 코퍼스를 구축하여 음소의 균형을 고려한 문장을 추출하게 되는데[5], 감정 음성 합성기에 필요한 시나리오도 비슷한 과정을 통해 얻을 수 있다. 그러나 감정을 표현한 문장은 제한적이어서 다량의 텍스트 코퍼스 구축이 어렵고 감정에 따른 분류가 선행되어야 하므로 시나리오 작성이 용이하지 않다.

본 논문에서는 감정을 기쁨, 슬픔, 화남의 3가지로 분류하여 해당 감정에 대한 텍스트 코퍼스를 구축하고, 각각 30분의 녹음 시간에 필요한 음소 균형이 고려된 시나리오를 작성하였다. 그런데, 감정 시나리오는 감정을 표현하는 문장이나 단어가 제한적이라는 특징이 있어 모든 음소 환경을 포함하기 어려워서, 감정이 표현되지 않은 일반 문장의 시나리오로 보완하는 것이 필요하다. 각 감정의 시나리오와 일반 문장의 시나리오를 이용하여 각각 1시간의 감정 음성 데이터를 녹음하였다. 우선 감정 시나리오를 이용하여 30분의 음성을 녹음하고, 일반 문장 시나리오도 동일한 감정으로 읽게 하여 30분을 추가한 감정 음성 데이터를 구성하였다.

음성에서 감정은 매우 주관적인 요소이기 때문에 합성기에 필요한 감정 표현의 강도를 정하는 것이 어렵고, 정해진 강도를 유지하면서 장시간 녹음을 진행하는 것 또한 쉽지 않다. 따라서 초기의 테스트 녹음과 정확한 모니터링이 중요하다. 따라서 감정 표현의 강도를 시각화할 수 있는 방법이 동원된다면 감정 음성 녹음의 시행착

오 및 품질 유지에 도움이 될 수 있을 것이다.

감정 음성을 녹음하는데 있어 중요한 것이 화자의 음색을 유지하는 것이다. 감정을 과도하게 표현한 음성에서는 종종 화자의 음색이 유지되지 않는 특징이 있는데, 이러한 데이터로 합성기를 구현하면 합성음이 원하는 화자의 음색을 나타내지 못하는 결과를 얻을 수 있다. 그러나 음색에 초점을 맞추어 감정의 표현을 너무 제한하면 합성음에서 감정이 제대로 나타나지 않는 단점이 있을 수 있다. 따라서 감정의 표현과 음색 유지의 두 가지 측면을 고려하며 녹음이 진행되어야 한다. 본 논문에서는 감정이 과하지 않게 표현되면서 음색도 유지되는 정도의 Low level과 음색 유지 측면보다 감정 표현에 초점을 맞춘 High level의 두 가지 녹음을 진행하여 비교하였다. 단 High level의 경우에도 과도한 음색 변화는 방지하였다.

3. 감정 음성의 음색 비교 방법

기존의 감정 음성에 대한 연구는 입력 음성신호에서 감정의 분류를 위해 감정을 인식하거나, 일반 음성 인식의 성능을 높이기 위해 감정으로 인한 왜곡을 처리하기 위한 방향으로 진행되어져 왔다. 그러나 감정 음성 합성기에서는 감정을 분류하여 모니터링 된 감정 음성을 녹음하고 이를 가공하여 합성음을 생성하는 것으로, 감정을 인식해야 하는 처리와 다를 수 있다. 하지만 기존의 연구결과인 감정을 나타내는 특징과 감정으로 인해 음성신호의 변화 등을 이용할 수 있다. 감정 음성 신호를 분석하는데 사용되는 파라미터는 기본 주파수와 에너지 등의 변화율과 모음의 포만트 정보 등이다. 그리고 대용량 감정 음성 코퍼스를 이용하여 대표 모음인 /A/, /E/, /I/, /O/, U/ 의 포만트 변화를 측정해 보면 감정의 변화에 따라 조금씩 변화되는 것을 알 수 있다. 즉 모음의 포만트 변화를 이용하면 감정을 분류하거나 인식도 가능 할 수 있는 것이다[6]. 또, 화자 식별과 같은 분야에서 화자의 특징을 분석하기 위해 주로 멜켑스트럼과 같은 파라미터를 사용하는데, 이것은 음소의 포만트 정보가 화자의 음색을 나타내기 때문이다.

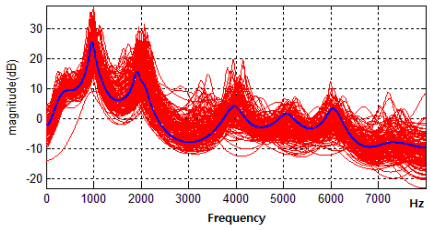
본 논문에서는 모음의 포만트 변화를 동일한 감정에 적용하여 음색의 변화를 관찰 할 수 있는지 테스트 하였다. 음색은 감정 음성과 일반 음성을 동시에 합성하는 합성기 측면에서는 합성음의 품질을 결정하는 중요한 요소가 될 수 있다. 즉 감정 음성의 녹음 데이터에서 과도한 음색의 변화는 하나의 합성기에서 서로 다른 음색의 합성음을 생성하는 결과를 초래할 수 있기 때문이다.

화자의 음색과 감정을 모니터링하면서 녹음 한 데이터의 음색 정보를 분석하기 위해 /A/, /E/, /I/, /O/, U/ 와 같은 대표 모음의 LPC 스펙트럼을 추출하고, 이 스펙트럼을 누적시켜 평균 포만트 스펙트럼을 생성하였다.

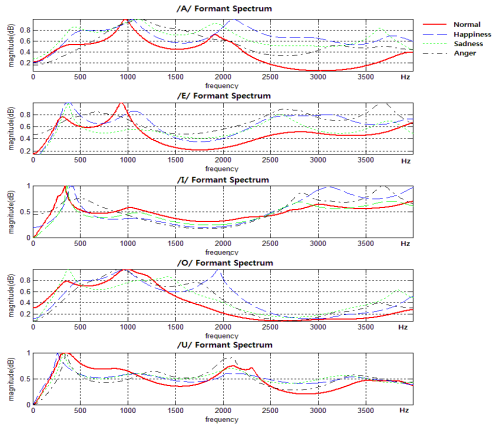
$$S(w) = \frac{1}{N}\sum_{n}^{N} s_n(w) \tag{1}$$

$$\tilde{S}(w) = \frac{S(w) - \min(S(w))}{\max(S(w))} \tag{2}$$

식 (1)은 포만트 스펙트럼을 누적하여 평균 포만트 스펙트럼을 구한 것이고, 식 (2)는 평균 포만트 스펙트럼의 정규화 식이다. 감정 음성과 일반음성의 음색변화를 관찰하기 위해 스펙트럼의 크기를 정규화 하였는데, 이것은 포락선의 모양 변화를 통해 음색 변화를 유추하기 위한 것이다. Fig. 1은 모음 /A/의 평균 포만트 스펙트럼을 나타낸 것이고, Fig. 2는 일반 음성과 감정 음성에 대해 각 대표 모음별 정규화 된 스펙트럼의 변화를 나타낸 것이다.

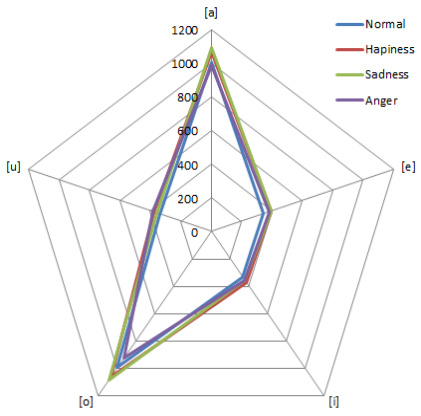


[Fig. 1] Average Formant Spectrum of /A/

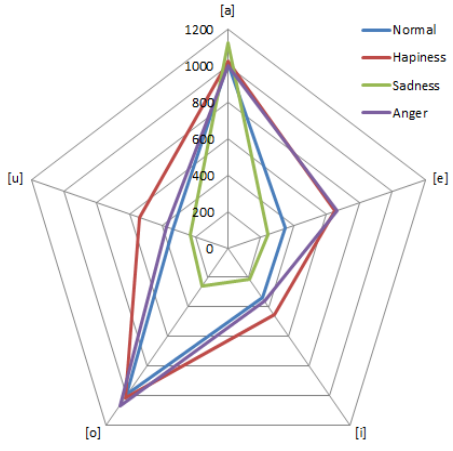


[Fig. 2] Average Formant Spectrum of Low Level Emotional Voice

특히 Fig. 2는 녹음 시 성우에게 Low level의 감정 표현을 요구하여 수집한 음성 데이터를 이용하였다. 전체적인 스펙트럼 형태는 비슷하나 감정 마다 스펙트럼의 피크인 포만트 정보가 차이나는 것을 알 수 있다. 주로 200Hz~2000Hz 대역에 분포하는 제 1포만트(F1)는 감정 및 음색에 따라 변화하는 것으로 연구되어 지고 있다[6]. 따라서 본 논문에서는 평균 포만트 스펙트럼에서 F1을 측정하고, 이것을 도식화하여 음색 변화를 추정할 수 있는 Fig. 3, Fig. 4와 같이 모음도를 구성하였다. Fig. 3은 감정 표현을 Low level로 유지하면서 녹음한 데이터이고, Fig. 4는 감정 표현을 High level로 유지하면서 녹음한 데이터를 분석한 모음도이다.



[Fig. 3] Vowel Chart using Low Level Emotional Voice



[Fig. 4] Vowel Chart using High Level Emotional Voice

감정 표현의 강도가 높은 데이터에서 차이가 두드러지는 것을 알 수 있다. 즉 감정 표현이 크면 클수록 동일한 모음의 F1의 차이가 커지는 것을 알 수 있다. 실제 High level 감정데이터를 청취하면 특정 문장에서는 음색이 유지 되지 않는 경우도 발견할 수 있었다. 특히 Sadness의 경우 일반 음성과의 음색 유사도가 낮아지는 경향이 나타났는데, Fig. 4에서도 이러한 특징이 잘 나타난다. 따라서 F1의 모음도를 녹음과 함께 실시간으로 분석한다면, 녹음 시 음색에 대한 모니터링을 실행하면서 녹음하는 것도 가능해 질 것으로 생각된다.

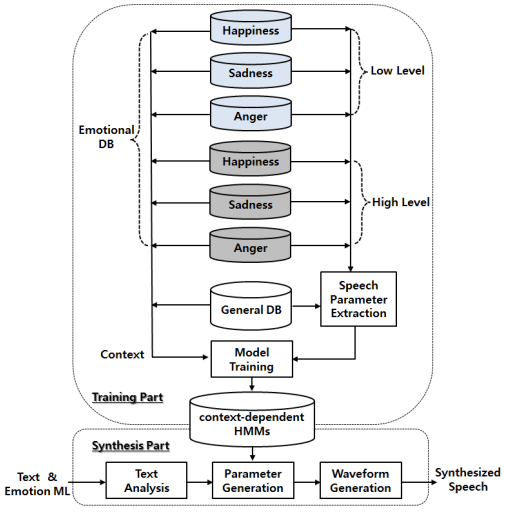
4. HMM 기반 감정 음성합성

HMM 기반 음성 합성기는 통계적 모델링을 이용하는 합성 방법으로 안정성은 우수하나 운율이나 음색이 평탄화 되는 단점이 있다. 감정 음성은 일반음성에 비하여 운율의 변화가 심하고 대용량 코퍼스를 구축하기가 힘들어 파형 연결 합성방식의 합성기로 구현하기가 매우 어렵다. 따라서 비교적 적은 데이터로 안정적인 성능을 낼 수 있는 HMM 기반 음성 합성 방식으로 구현하는 것이 효율적이다. 그런데 HMM 기반 음성합성 방식은 합성음의 운율이 평탄화 되는 단점이 있어 감정을 합성하는 경우 녹음 데이터의 감정 표현의 강도를 어느 정도로 유지할 것인지를 결정하는 것이 매우 중요하다. 감정 표현의 강도가 약할 경우 평탄화로 인해 감정 합성음과 일반 합성음의 차별성이 없어지게 되고, 강도를 강하게 녹음할 경우에는 음색 및 안정성을 유지하면서 녹음하는 것이 어려워지기 때문이다. 본 논문에서 사용한 감정 음성은 3번의 테스트 녹음을 통해 두 가지 감정 level에 적합한 감정 표현의 강도를 결정하고 지속적인 모니터링을 통해 녹음하였다. 그러나 다양한 텍스트에 대하여 일정한 감정 표현의 강도를 유지하는 것과 그것을 모니터링 한다는 것은 쉽지 않은 일이고, 경험 및 주관적 판단에 따라 녹음 데이터의 품질이 달라 질 수 있다.

Fig. 5는 본 논문에서 구현한 감정 음성 TTS의 구성도이다.

전체 구성은 훈련부(Training Part)와 합성부(Synthesis Part)로 이루어져 있고, 훈련부에서는 2가지 level의 감정 음성 데이터와 일반 음성 데이터의 context 정보와 음성 특징 파라미터를 추출하고 훈련을 통해 HMMs를 구축하는 것이다. 3가지 감정(Happiness, Sadness, Anger)과 2가지 Level(Low Level, High Level)

을 context로 처리하여 훈련하였다. 합성부는 일반적인 HTS의 구성도와 유사하고, 단지 TTS의 입력에서 감정 음성을 합성하기 위한 Emotion ML(Markup Language)을 설계하여 사용하였다. Category는 3가지 감정 중 하나를 지정해야하고, level은 Low Level을 1, High Level을 2로 정의하고 2가지 중 하나를 지정하도록 하였다. Emotion ML을 사용하지 않는 경우에는 일반 음성으로 합성하도록 하였다.



[Fig. 5] The block diagram of the emotional TTS using in this paper

```
<emotion
   category = " happiness | sadness | anger"
   level = " 1 | 2 " >

   text

</emotion>
```

[Fig. 6] The Emotion ML(Markup Language) using in this paper

5. 실험 및 결과

High level과 Low level로 감정을 조절하여 녹음 된 데이터를 이용하여 HMM 기반 음성 합성기를 구현하여 일반 음성의 합성음과 비교하여 음색이 어떻게 변화되는지 청취테스트를 진행하였다. 그리고 합성음의 음색은 음질의 영향도 있다고 생각하여 청취 테스트에 사용된 합성음의 음질에 대해서도 테스트를 진행하였다. 음질 테스트는 MOS(Mean Opinion Score)[7]를 사용하였다.

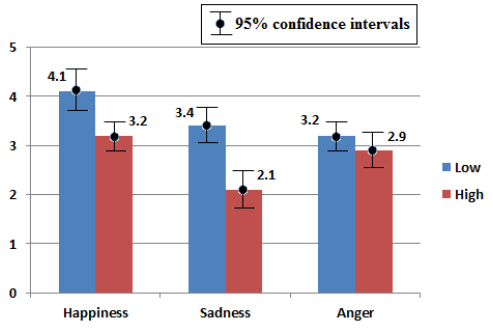
음성 DB는 일본인 여성화자 데이터로 일반 음성 8.3시간(약 2800문장)과 감정 음성 6시간 분량을 사용하였다. 감정 음성은 세부적으로 각 level에 3시간씩, 하나의 level에서는 각 감정에 1시간씩의 데이터를 녹음하여 사용하였다. 한 감정에 사용된 시나리오는 700~800문장 정도이다. 훈련 및 합성 시스템은 (주)보이스웨어에서 개발한 Japanese VoiceText Micro[8]를 사용하였다.

청취테스트에 사용할 합성음은 임의의 텍스트 10문장을 사용하여 일반음성 및 각 level의 3가지 감정음성을 합성하였다. 임의의 텍스트는 훈련에 포함되지 않은 문장으로 150자 이하의 대화체 문장을 사용하였다.

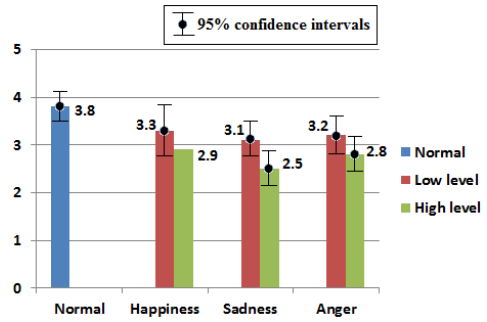
청취 테스트는 일본인 5명이 음질과 음색 유사도를 평가하였고, 음색 유사도 평가는 먼저 감정이 없는 일반 합성음(Normal)을 들려준 후 테스트 음성을 들려주고 MOS와 비슷하게 1~5의 점수를 주도록 하였다(1:다른 사람의 목소리, 2:비슷하지 않음, 3:약간 비슷함, 4:매우 비슷함, 5: 같은 사람의 목소리).

테스트 결과 실제 합성음에서도 Low level의 감정 합성음이 High level의 감정 합성음 보다 음색 유사도 및 음질이 높게 나타났다. 즉 감정의 강도가 강한 합성음에서 음색 및 음질이 나쁘게 나타날 수 있음을 보여주는 결과이다.

실험 결과를 3장의 녹음 데이터 모음도 분석 결과와 비교 했을 때 모음도의 거리가 크게 나타나는 High level의 감정 데이터를 이용하여 합성음을 생성하는 경우 합성음에서도 음색의 차이가 크게 발생 할 수 있음을 확인할 수 있었다.



[Fig. 7] Test of Voice Color Similarity



[Fig. 8] Test of Voice Quality(MOS)

6. 결론

본 논문은 HMM 기반 음성 합성기를 이용하여 감정 합성기를 구현하는 경우 발생할 수 있는 음색 변화에 대한 연구이다. 감정 음성의 녹음 시 표현 강도에 따른 합성음에서의 음색 변화를 살펴보고 적절한 감정의 강도를 설정함에 있어 시행착오를 줄이는 것이 목적이다. 감정 음성 녹음 데이터에서 음색의 변화를 유추할 수 있는 파라미터를 추출하고 이것을 효율적으로 보여줌으로써 적절한 감정의 표현 강도를 전체 녹음과정에 유지할 수 있도록 도와 줄 수 있는 방법을 제안하는 것이다. 또한 앞으로는 감정의 강도 및 음색 변화를 수치로 표현하는 방법을 연구해야 할 계획이다.

References

[1] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," IEICE Transactions, vol. E90-D, no.5, 816-824(2007)
DOI: http://dx.doi.org/10.1093/ietisy/e90-d.5.816

[2] Z-.H. Ling, Y. Hu, and L. Dai, "Global variance modeling on the log power spectrum of LSPs for HMM-based speech synthesis," Proc. INTERSPEECH, 825-828(2010)

[3] Z. Yan, Q. Yao, S.K. Frank, "Rich Context Modeling for High Quality HMM-Based TTS," INTERSPEECH 2009, 1755-1758(2009)

[4] J. Yamagishi, K. Onishi, T. Masuko, T. Kobayashi, " Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. on Inf. & Syst., vol.E88-D, no.3, 503-509(2005)
DOI: http://dx.doi.org/10.1093/ietisy/e88-d.3.502

[5] M. Isogai et al., "Recording script design for corpus-based TTS system based on coverage of various phonetic elements," Proc. ICASSP, vol. I, 301-304(2005)

[6] Seo-Bae Lee, "An Analysis of Formants Extracted from Emotional Speech and Acoustical Implications for the Emotion Recognition System and Speech Recognition System," Journal of the Korean society of speech sciences , No.3 Vol1, 45-50(2011)

[7] D. S. Na and 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, 349-352(2009)

[8] http://voicetext.jp/news/archives/2570

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