# 감정표현 음성합성 시스템을 위한 감정 분석

천희진, 이양희
동덕여자대학교 전자계산학과

# An Analysis on the Emotional Speech for the Speech Synthesis System with Emotion

Heejin Chun, Yanghee Lee
Dept. of Computer Science, Dongduk Women's Univ.
E-mail : heejin@cs4000.dongduk.ac.kr yhlee@www.dongduk.ac.kr

요 약

감정을 표현하는 음성 합성 시스템을 구현하기 위해서는 감정 음성에 대한 분석이 필요하다. 본 논문에서는 평상, 화남, 기쁨, 슬픔의 네 가지 감정에 대한 음성 데이터에 대해 음절 세그먼트, 라벨링을 행한 감정 음성 데이터베이스를 구축하였고, 감정 표현이 음성에 영향을 미치는 요인에 대하여, 운율, 음운적인 요소로 나누어 분석하였다. 또한 기본 주파수, 에너지, 음절지속시간에 대한 분석과 감정 음성의 기본 주파수, 에너지, 음절지속시간, 스펙트럼 포락의 인지 정도를 측정하기 위하여, 평상 음성에 감정 음성의 운율 요소를 적용하는 음성을 합성하여 ABX 방법으로 평가하였다. 그 결과, 기본 주파수의 변화가 73.3%, 음절지속시간은 43.3%로 올바른 감정으로 인지되었으며, 특히 슬픈 감정에서 음절지속시간은 76.6%가 올바르게 감정을 나타내는 것으로 인지되었다.

## I. 서 론

감정은 인간의 의사소통에 있어서 중요한 역할을 하고 있고, 감정 표현이 가능한 합성 음성이 표현되므로, 언어 장애자의 Communication Tool, 휴먼 로봇의 감정 표현의 응용도 기대된다.

감정을 표현하는 음성 합성 시스템을 구현하기 위하여, 일본어 감정 음성 코퍼스를 작성하고, 감정 음성의 코퍼스에 음향적 특징 양에 대해 분석하고, 청취 실험을 행한 연구가 보고되었다[1]. 그러나, 우리말에 있어서, 이에 대한 연구가 미흡한 실정이다.

따라서 본 논문에서는 감정을 표현하는 음성 합성 시스템을 구현하기 위해서, 감정 음성 코퍼스를 구축하고, 이 코퍼스의 음향적 특징 즉 피치, 에너지, 음절지속시간에 대한 분석을 행한다. 또한 합성 음성으로 감정 음성의 피치, 에너지, 음절지속시간의 인지 정도를 측정하는 ABX 청취 테스트를 행한다. 청취 테스트를 위한 음성 합성 방법은, 평상 음성에 각 감정 음성의 1) 피치, 2) 에너지, 3) 음절지속시간, 4) 스펙트럼 포락을 적용하여 합성한다. 이때, 각 감정 음성의 지속시간이 다르기 때문에, 지속시간을 정합시키기 위하여 음절별로 Dynamic Time Warping 을 사용한다.

## II. 감정 음성 데이터베이스 구축

감정을 표현하는 음성을 합성하기 위해 필요한

모든 운율과 음운성은 실제 인간의 감정 음성으로부터 확보되어야 한다. 그러나, 인간의 감정 음성 코퍼스를 구축하는 것은, 매우 어려운 일이다. 왜냐하면 감정을 장시간 지속하거나, 또는 인위적으로 필요에 따라 감정을 표현하기는 어렵기 때문이다.

따라서 본 연구에서는 화자의 자연스러운 감정을 이끌어내기 위하여 평상, 화남, 기쁨, 슬픔의 감정이 잘 표현될 수 있는 35개 문장을 선택하여 낭독형태로 녹음하였다. 이 35개의 문장은 특정한 감정을 일정 시간 지속시키기 위해, 대화 형식이 아니라 독백 형식의 문장을 선택하였고, 화자는 의도적으로 감정을 어느 정도 표현할 수 있는 대학 아마추어 연극 배우인 여성 화자 1인이다.

감정 표현의 음성 코퍼스는 각각 음절 단위로 세그먼트와 라벨링을 하여 감정 표현의 음성 데이터베이스를 구축하였다.

- 음절 레벨 세그먼트 (시간 정보)
- 음절 라벨링 (음운 기호)

## III. 음성 분석/ 합성계

감정 음성의 음운 및 운율의 변화를 분석하기 위해, 운율과 음운을 분리하여 다룰 수 있는 분석 합성계를 사용한다. 따라서 본 논문에서 사용되는 음성 분석/합성계의 개요는 다음과 같다.

- A/D 변환 : 8kHz, LPF
  16kHz Sampling, 16bit 양자화
- 분석 합성 파라미터 : 개량 켑스트럼[2]
- 분석
  * 프레임 주기 : 6.2msec (100 Sample)
  * 윈도우 함수
    스펙트럼 포락 : 15.5msec (256 Sample)
    Blackmann Window
    피치 검출 : 24.8msec (400 Sample)
    Blackmann Window
  * 개량 켑스트럼 차수 : 남성 - 30차
    여성 - 25차
  * 스펙트럼 포락 추출 : 개량 켑스트럼법[2]
  * 유성/무성 판별 : 스펙트럼 포락의 저주파수 부분의 에너지 평균
  * 피치 추출 : Cepstrum peak picking 법
- 합성 필터
  * LMA(Log Magnitude Aproximation) 필터[2]

## IV. 감정 음성의 음향적 특징

### 1. 기본 주파수 (f0)

감정에 의해 음성의 기본 주파수가 어떻게 변화하는가를 분석하기 위해 우선 각 감정 음성(평상, 화남, 기쁨, 슬픔)의 기본 주파수에 대한 평균치와 표준 편차는 표 1과 같다.

[표 1] 기본 주파수의 평균과 표준편차

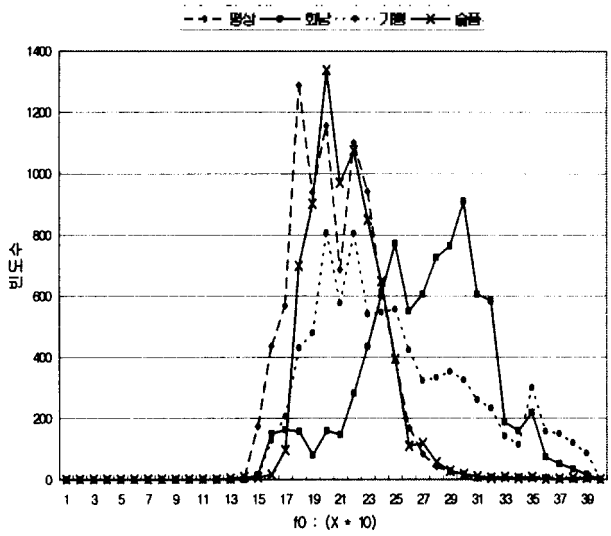
(단위: Hz)

| | 평균치 | 표준편차 |
|---|---|---|
| 평 상 | 168.2644 | 110.4880 |
| 화 남 | 247.8176 | 128.3068 |
| 기 쁨 | 199.1240 | 128.2357 |
| 슬 픔 | 151.5963 | 115.0170 |

그 결과, 화난 감정의 평균 f0는 다른 세 가지 감정의 평균 f0 보다 가장 높고, Dynamic Range도 가장 높다. 기쁜 감정이 평균 f0가 다음으로 높고, Dynamic Range도 높으며, 화난 감정과 거의 비슷하게 나타났다. 슬픈 감정의 평균 f0가 가장 낮았으며, Dynamic Range는 평상과 유사하다.

감정 음성 데이터 코퍼스의 기본 주파수의 분포는 그림 1과 같다.

그림 1에서, 화난 감정의 기본 주파수 분포는 높은 주파수가 많이 나타났고, 슬픈 감정의 기본 주파수




[그림 1] 감정 음성에 대한 기본 주파수의 분포

분포는 평상의 기본 주파수 분포와 거의 유사하다. 기쁜 감정의 기본 주파수 분포는 슬픈 감정의 기본 주파수 분포보다 높은 주파수에 많이 나타남을 알 수 있다.

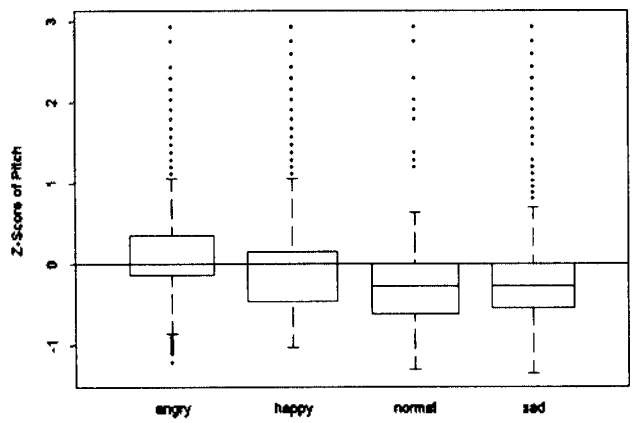
화자의 개별성을 제거하기 위해서, 식(1)과 같은 Zscore를 사용하여 기본 주파수를 정규화 하였다.

Zif = (Xif - Mf) / SDf ----- 식(1)

Zif : 감정 f의 i번째 관측치

Mf : 감정 f의 관측치에 대한 평균

SDf : 감정 f의 관측치에 대한 표준편차




[그림 2] 감정별 기본주파수의 정규화

기본 주파수에 대한 정규화의 결과는 그림 2와 같다. 이 그림의 결과, 그림 1과 비슷한 결과가 나타난다. 즉, 기본 주파수에 의해서, 감정의 표현을 구별할 수 있음을 알 수 있다.

2. 에너지

감정에 의해 음성의 에너지가 어떻게 변화하는가를 분석하기 위해 우선 각 감정 음성(평상, 화남, 기쁨, 슬픔)의 에너지에 대한 평균치와 표준 편차는 표 2과 같다.

[표 2] 에너지의 평균과 표준편차

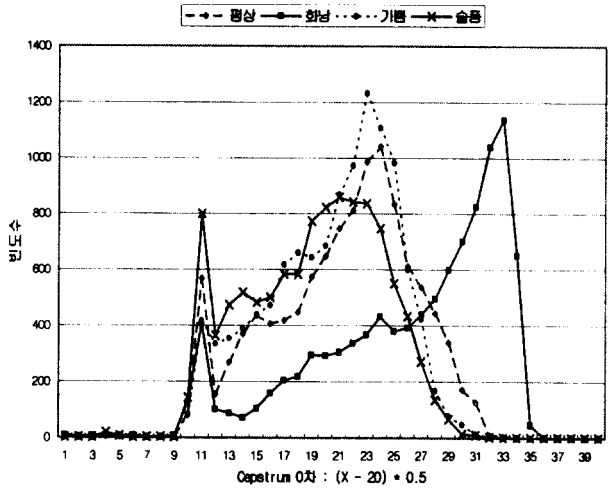
| | 평균치 | 표준편차 |
|---|---|---|
| 평 상 | 0.6018708 | 2.991825 |
| 화 남 | 3.2744480 | 3.520340 |
| 기 쁨 | 0.4364057 | 2.359750 |
| 슬 픔 | -0.2620917 | 2.572151 |

그 결과, 화난 감정의 평균 에너지는 다른 세 가지 감정의 평균 에너지 보다 높고, Dynamic Range도 가장 높다. 평상의 에너지가 다음으로 높고, Dynamic Range도 다음으로 높게 나타났다. 슬픈 감정의 평균 에너지가 가장 낮았으며, 기쁜 감정의 Dynamic Range가 가장 낮아진다.

감정 음성 데이터 코퍼스의 에너지의 분포는 그림 3과 같다.

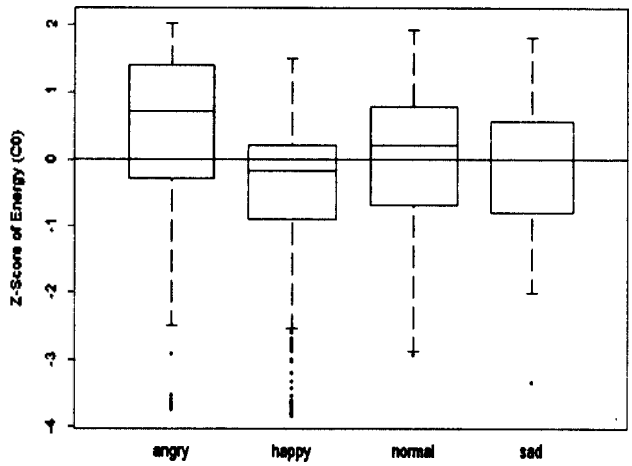
그림 3에서, 화난 감정의 에너지 분포는 높은 에너지가 많이 나타났고, 기쁜 감정의 에너지 분포는 평상의 에너지 분포와 거의 유사하다. 슬픈 감정의 에너지 분포는 다른 감정에 비해 비교적 낮은 에너지에 많이 나타남을 알 수 있다.

화자의 개별성을 제거하기 위해서, 식(1)과 같은 Zscore를 사용하여 에너지를 정규화 하였다. 에너지에 대한 정규화의 결과는 그림 4와 같다. 에너지에 대한 정규화의 결과는 그림 3와 같다.




[그림 3] 감정 음성에 대한 에너지의 분포

이 그림의 결과, 통계적으로 에너지에 의한 감정 표현의 구별이 가능하다.




[그림 4] 감정별 에너지의 정규화

## 3. 음절지속시간

감정에 의해 음성의 음절지속시간이 어떻게 변화하는가를 분석하기 위해 우선 각 감정 음성(평상, 화남, 기쁨, 슬픔)의 음절지속시간에 대한 평균치와 표준 편차는 표 3과 같다.

그 결과, 기쁜 감정의 평균 음절지속시간은 다른 세 가지 감정의 평균 음절지속시간 보다 길고,

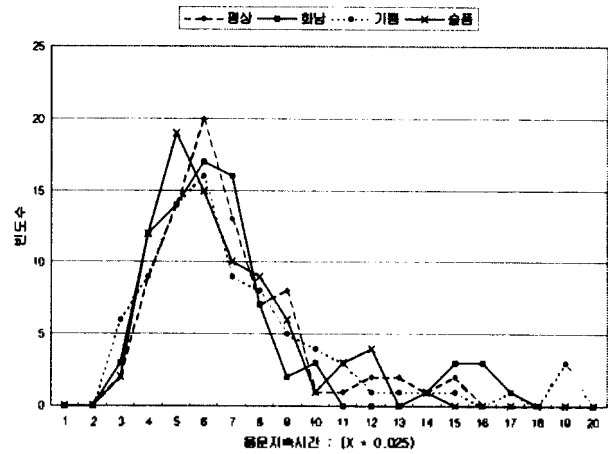
[표 3] 음절지속시간의 평균과 표준편차

(단위: msec)

| | 평균치 | 표준편차 |
|---|---|---|
| 평 상 | 159.77770 | 65.89286 |
| 화 남 | 160.66926 | 82.72853 |
| 기 쁨 | 168.24900 | 89.69467 |
| 슬 픔 | 152.09340 | 59.30874 |

Dynamic Range도 가장 높다. 화난 감정과 평상의 평균 음절지속시간은 거의 유사하지만, Dynamic Range는 화난 감정이 더 크다. 슬픈 감정의 평균 음절지속시간이 가장 낮았으며, Dynamic Range도 가장 낮음을 알 수 있다.

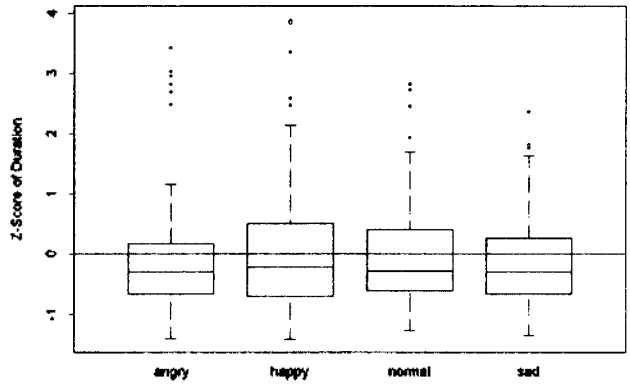
감정 음성 데이터 코퍼스의 음절지속시간의 분포는 그림 5와 같다.




[그림 5] 감정 음성에 대한 음절지속시간의 분포

그림 5에서, 감정 음성의 음절지속시간의 분포는 거의 비슷하게 나타난다. 음절지속시간을 정규화한 결과는 그림 6과 같다. 이 그림의 결과, 음절지속시간의 분포는 거의 비슷하게 나타났고, 기쁜 감정의 분포가 비교적 넓게 나타났음을 알 수 있다.

음절 지속시간의 분포 만으로는 감정 표현을 구별하기가 쉽지 않다. 여기에 문장구조, 품사, 휴지 지속시간 등의 정보에 대한 분석이 필요하다.




[그림 6] 감정별 음절지속시간의 정규화

## IV. 실험 및 평가

본 연구에서 작성한 감정 음성 코퍼스가 감정을 잘 표현하고 있는지를 평가하기 위하여, 코퍼스 내의 음성 데이터에 대해서 청취 실험을 하였다.

각 감정(평상, 화남, 기쁨, 슬픔)에 대해 10문장(총 40문장)을 대학원생 3명에게, 평상 음성과 감정 음성을 각각 비교하는 ABX 테스트(A : 평상 음성, B : 각 감정 음성, X : 각 감정 음성)로 감정 판별을 한 결과는 표 4와 같이, 슬픈 감정의 경우, 20% 정도가 오류로 나타나고, 다른 감정의 경우에는 100% 정확도를 갖는다.

[표 4] 감정 판별의 청취 테스트 결과

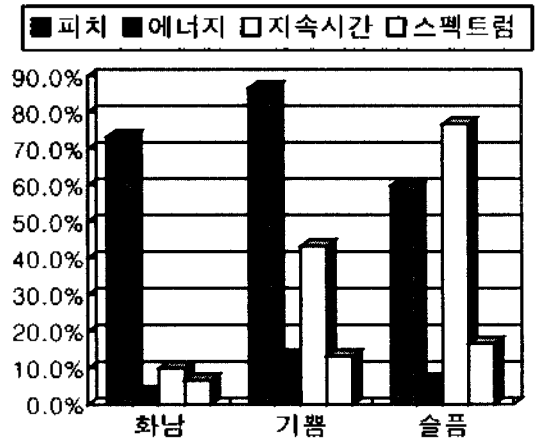
| 화남 | 기쁨 | 슬픔 |
|---|---|---|
| 100.0 % | 100.0 % | 80.0 % |

앞에서 피치, 에너지, 지속시간에 대한 분석과 감정 음성의 피치, 에너지, 지속시간의 인지 정도를 측정하기 위하여, 평상 음성을 감정 음성의 각각의 운율요소로 합성하여, 청취 테스트를 행하였다.

청취 테스트를 위한 음성 합성 방법은, 평상 음성에 각 감정 음성의 1) 피치, 2) 에너지, 3) 음절지속시간, 4) 스펙트럼 포락을 적용하여 합성하였다. 이 때, 각 감정 음성의 지속시간이 다르기 때문에, 지속시간을 정합시키기 위하여 음절별로 Dynamic Time Warping 알고리즘을 사용하였다.

테스트 방법은 청취자에게 데이터베이스에 구축된 평상 음성(A)과 감정 음성(B)을 들려준 후, 합성음(X)을 들려주어, 합성음(X)이 평상 음성(A)과 감정 음성(B) 중 어느 쪽에 가까운지를 테스트한다. 그 결과는 그림 7에 나타난 것과 같이, 감정 표현 모두가 피치의 변화에 중요한 영향을 미치고, 음절지속시간은 슬픔과 기쁨의 감정에 의해 영향을 받는다. 반면에 에너지나 스펙트럼 포락은 감정 표현에 크게 영향을 받지 않는다. 감정 음성을 합성하는 경우, 피치와 음절지속시간에 의해, 감정의 표현이 좌우될 수 있음을 알 수 있다.

따라서 규칙합성 시스템에 감정 표현을 적용하기 위해서는 이러한 요소들에 대한 감정 패턴을 생성하여야 한다.




[그림 7] 합성된 감정 음성의 청취 테스트 결과

## V. 결론

본 논문에서는 감정 음성을 분석하기 위한, 감정 음성 데이터 코퍼스를 구축하고, 코퍼스 내 음성 데이터에 대해 감정 평가하였다. 또한 감정 음성의 음향적 특징을 분석한 결과, 평균 f0의 값은 화남, 기쁨, 평상, 슬픔의 감정에 따라 구분이 되고, 평균 에

너지의 값은 화남, 평상, 기쁨, 슬픔의 감정에 따라 구분이 되지만, 음절지속시간은 감정 표현에 따라 큰 변화가 없었지만, 합성음에 의한 감정 평가를 위한 청취 테스트 결과, 기본 주파수의 변화가 73.3%, 음절지속시간은 43.3%로 올바른 감정으로 인지되었으며, 특히 슬픈 감정에서 음절지속시간은 76.6%가 올바르게 감정을 나타내는 것으로 인지되었다. 특히 에너지 7.8%, 스펙트럼 포락 12.2%로 감정에 영향을 적게 미치는 것으로 나타났다.

[참고문헌]

1. A. Iida, K.Meiseki, Nick Campbell, "Designing and Testing a corpus of emotional speech", 일본 음향학회 춘계 연구발표회 강연 논문집 pp311-312, 1997.3.

2. Y. Abe and S. Imai: "Speech Synthesis from CV-syllable cepstral parameters," Trans. IECE J64-D, pp861-868, 1981.

3. 이양희외, "한국어 음성의 규칙합성", 전자공학회지, Vol. 20, No.5, pp80-89, 1993.

4. N. Campbell, "Pragmatic Intonation: 운율 정보의 기능적 역할", ATR 음성번역 연구소 연구 발표 자료집, pp1-20, 1997.

5. W. N. Campbell, "Segmental Elasticity and Timing in Japanese Speech", in Speech Perception, Production and Linguistic Structure edited by Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Ohmsha), 1992.

* 본 연구는 1997년도 한국과학재단 특정기초 연구과제 연구지원비에 의해 연구되었습니다 *