정서음성 합성을 위한 예비연구

한영호* 이서배* 이정철** 김형순***
* 부산대학교 인지과학협동과정
**울산대학교 컴퓨터 · 정보통신공학부
*** 부산대학교 전자공학과

Preliminary Study on Synthesis of Emotional Speech

Youngho Han*, Sopae Yi*, Jung-Chul Lee**, Hyung Soon Kim***
*Cognitive Science, Pusan National University
**School of Computer & Information Technology, University of Ulsan
*** Dept. of Electronics Eng, Pusan National University

{woal1975, sopaeyi}@pusan.ac.kr, jungclee@ulsan.ac.kr, kimhs@pusan.ac.kr

Abstract

This paper explores the perceptual relevance of acoustical correlates of emotional speech by using formant synthesizer. The focus is on the role of mean pitch, pitch range, speed rate and phonation type when it comes to synthesizing emotional speech. The result of this research is backing up the traditional impressionistic observations. However it suggests that some phonation types should be synthesized with further refinement.

Ⅰ.서론

멀티미디어의 발전과 더불어 인간과 컴퓨터간의 보다 자연스러운 인터페이스에 대한 요구가 증대하고 있다. 이러한 자연스런 인터페이스는 다양한 형태로 발전하고 있으며 음성합성기술 역시 이러한 요구에 부응하기 위한 노력을 계속하여 왔다. 합성음에 인간의 정서를 부여하고자 하는 것은 이러한 노력의 일환이라고 할 수 있다. 그러나 인간의 정서는 실제 상황에서 아주 다양하게 나타나며 동일한 정서상태일지라도 외적 상황에 따른 차이와 개인에 따른 차이가 매우 크기 때문에 합성음제작의 기본이 되는 특정 정서상태를 포함하는 음성을 수집하는 것은 매우 어려운 작업이다[1]. 더 나아가 수집된 정서를 가지고 특정 정서를 모델링하는 것은 다루어야 하는 음향적 파라미터의 수가 매우 많고 복잡할 뿐만 아니라 사람에 따라 판단하는 정서의 유형에 크고 작은 차이가 있기 때문에 외적 타당도를 얻기가 쉽지 않다.

그럼에도 불구하고 정서음성을 수집하거나 모델링하려는 많은 시도가 이루어져왔고 상당부분 성과를 얻은 것 또한 사실이다[1]-[5]. 본 연구는 한국어 정서음성 데이터베이스 V1.0[6]에 기반하여 화냄, 슬픔, 즐거움, 공포, 지루함 등의 5가지 정서에 관한 변인들의 종류와 수준에 대해 분석하였다. 분석된 변인들은 평균피치, 피치범위, 발화속도, 발성유형등 4가지로 국한시켰으며, 포먼트 합성기인 KL_SYN88을 사용하여 정서음성을 합성하였다[7][8].

Ⅱ. 정서음성의 음향적 특징

정서에 영향을 주는 음향 파라미터는 Cahn[3]의 연구

에서 잘 나타나있다. 이중 본 연구와 크게 관련이 있는 파라미터를 정리하면 다음과 같다.

· 평균피치: 평균치는 모든 피치값을 동일한 수준으로 수정함으로써 조절할 수 있다.

· 피치범위: 피치범위는 기준이 되는 특정피치의 값을 중심으로 확장과 압축의 방식으로 조절할 수 있다.

· 피치곡선: 피치곡선은 음절과 어절을 구분하여 수평곡선과 상승 곡선, 하강 곡선으로 조절할 수 있다. 피치곡선은 정서음성 합성에 매우 큰 영향을 주는 것으로 알려져 있으나 본 연구에서는 제외하였다.

· 발화속도: 발화속도는 음소내의 안정구간에서 특정 시간비율 만큼 입력값을 삭제하는 방식으로 조절하였다.

· 발성유형: 발성유형에 따른 음향학적 특성은 Laver[9]의 연구에 잘 나타나 있다. 본 연구에서는 이를 바탕으로 가성(falsetto), 숨소리(breathy), 짜내기 소리(creaky), 긴장음(tense)을 가지고 발성유형이 정서에 미치는 영향을 연구하였다.

이외에도 강세의 빈도, 쉼구간의 연속성, 피치의 연속성, 발음의 정확성, 소리의 크기 등이 중요요인으로 작용한다[3].

정서에 관한 연구는 영어권을 중심으로 이미 많이 이루어져왔다[3][6]. 이들의 연구를 통해 정서의 음향적 특성을 알아보면 다음과 같다.

(1)즐거움

즐거움의 경우 피치와 피치의 범위를 증가시키며 크게 증가된 피치 패턴은 아주 서서히 감소하는 현상을 보인다. 조음 현상에 있어서는 때로 호기음이 섞이는 경우도 있다.

(2)화냄

화냄의 경우 일반적으로 긴장음이 나타나고 피치가 올라간다. 피치범위는 넓어지는 경향이 있고 피치의 변화가 크며 평균피치는 다른 정서에 비해 상당히 높은 편이다.

(3)슬픔

강도는 낮아지고 불규칙적인 휴지기가 발생하며 좁은 피치범위를 가지는 경우가 많다. 그리고 발성속도는 느리게 나타나며 숨소리와 가성이 많이 나타나며 피치곡선은 상승한다.

(4)공포

피치가 높고 피치범위가 넓으며 가성이 빠른 발성속도로 나타난다.

(5)지루함

낮은 평균피치를 보이며 피치범위는 좁은 편이다. 피치곡선은 수평을 이룬다.

Ⅲ. 실험방법

KL_SYN88의 입력 파라미터는 F0, 포먼트 주파수 및 대역폭등이며 사용자의 의도대로 자유롭게 파라미터를 바꿀 수 있다. KL_SYN88에 대한 자세한 내용은 [7]에 나타나 있다. 합성음 생성을 위한 원음으로 20대 중반의 남성의 화자로부터 단조로운 톤의 발화문장을 사용하였다. 사용된 문장은 한국어 정서음성 데이터베이스 v1.0에서 사용된 문장과 유사한

"저의 이름은 홍길동입니다."

이다. 이 문장은 특정 정서 상태에 편중되지 않고 다양한 정서상태에서 얼마든지 나타날 수 있다는 점과 합성된 음성문장을 실제 한국어 정서음성 데이터베이스 v1.0과 비교할 수 있다는 장점이 있다. 녹음된 문장은 ESPS/xwaes를 이용하여 F0, 5번째까지의 포먼트 주파수 크기와 대역폭을 추출하였다.

이렇게 합성된 합성음을 평균피치(3 수준), 피치범위(3 수준), 발화속도(3 수준), 발성유형(5 가지)에 변화를 주면서 총 135가지의 합성음을 제작하였다. 이를 7개의 그룹으로 무선할당하였으며, 7그룹의 순서를 달리하며 25명의 실험참가자들에게 들려주었다. 이를 통해 청취 순서에 따른 순서 효과를 없앨 수 있었다. 실험참가자들은 20대 초반의 대학생이었으며 남성 참가자 8명과 여성 참가자 17명이었다. 합성음은 1회만 듣게 하였다.

이렇게 청취된 합성음은 중립을 포함하여 화냄, 슬픔, 즐거움, 공포, 지루함 등의 6가지 문항 중에서 선택하도록 하였다.

Ⅳ.실험결과 및 분석

표 1은 평균피치와 피치범위, 발화속도에 대한 각 정서로 평가한 비율과, 파라미터와 정서간의 상관관계를 나타낸다. 표 1에서 볼 수 있는 바와 같이 평균피치는 화냄을 제외한 모든 유형의 정서음성 합성에 큰 영향을 주는 것을 확인하였다. 슬픔과 즐거움은 평균피치와 정적인(positive) 상관관계를 가지지만, 평균피치는 그대로 두는 경우가 높은 평가 수치를 보였다. 지루함은 평균피치가 내려갈수록 평가할 가능성이 높다.

표 1. 각 파라미터에 대한 정서 평가 수치(%) 및 상관계수

		화냄	슬픔	즐거움	공포	지루함
평균피치	낮게	15.5	12.1	7.0	8.2	26.0
	그대로	12.0	14.3	14.5	8.6	19.7
	높게	12.2	12.0	14.1	15.4	12.0
	상관계수	−.090	.222 *	.182 *	.228	.264 *
피치범위	좁게	22.6	15.5	4.9	13.1	21.6
	그대로	9.1	19.1	10.6	10.0	19.4
	넓게	7.7	11.4	20.4	9.1	16.8
	상관계수	−.415 **	−.114	.394 **	−.126	−.091
발화속도	느리게	4.2	22.3	4.3	13.1	40.1
	그대로	9.3	17.8	10.1	9.7	12.5
	빠르게	26.8	5.7	21.5	9.4	4.5
	상관계수	.622 **	−.462 **	.438 **	−.118	−.668 **

*. 상관계수는 0.05수준에서 유의함

**. 상관계수는 0.01수준에서 유의함

범위는 화냄과 부적인(negative) 관계가 뚜렷하며 즐거움과는 정적인 관계가 뚜렷하다. 즉 음의 변화가 단조로울수록 화냄으로 평가할 가능성이 높아지고 변화가 클수록 즐거움으로 평가할 가능성이 높아진다. 화난 정서를 분노하는 화냄과 냉소적인 화냄으로 나누었을 때 전자를 피치범위 가 높아지는 쪽으로, 후자를 피치범위가 좁아지는 쪽으로 둘 수 있으며, 피치곡선의 변화량을 따로 조작하지 않은 본 연구에서 화난 정서에 대한 높은 인식률을 보인 결과는 후자에 의한 것이라고 판단할 수 있을 것이다. 평균피치가 낮을 때, 오히려 화냄으로 평가하는 비율이 높아지는 것은 이러한 예측을 뒷받침한다. 이렇게 냉소적인 화냄에 대해서도 사람의 평가가 높게 나타나는 것은 화난 정서를 두 가지로 나누어야 한다는 Sendlmeier의 주장을 간접적으로 지지한다고 할 수 있다[4].

발화속도의 경우 공포에 질린 정서를 제외하고는 거의 대부분의 정서와 높은 상관관계를 보인다. 화냄과 즐거움의 경우 정적인 관계가 있으므로 발화의 속도가 높을수록 평가 수치가 높아지고 슬픔과 지루함의 경우는 부적인 관계가 있음으로 느릴수록 평가 수치가 높아진다. 특히 지루함은 다른 파라미터보다 발화속도와 상대적으로 높은 상관관계를 보여준다.

피치

전체적으로 평균피치, 피치범위, 발화속도는 기존의 다른 연구들에서 내린 결론과 일치한다.

그림 1은 각 정서에 대한 발성유형에 따른 차이를 보여준다. 그림에서 볼 수 있는 바와 같이 지루함의 경우 숨소리가 큰 영향을 주는 것으로 나타났다. 공포는 짜내기 소리가 강한 영향을 주며 즐거움은 긴장음이 강한 영향을 준다. 슬픔은 짜내기 소리와 숨소리가 큰 영향을 주는 것으로 나타났는데 이러한 결과는 슬픔이 숨소리에 가성을 첨가해서 의해서 만들 수 있다고 한 Sendlmerier의 연구결과와는 상이했으나[4], 발성유형만을 대상으로 정서 판단에 대한 연구를 수행한 Gobl의 결과와는 동일한 것이다[5]. 슬픈 정서 역시 울먹이는 슬픔과 조용한 슬픔으로 나눌 수 있다고 볼 때 발성유형에 있어서 나타나는 연구 결과간의 차이는 보다 세밀하게 분류된 정서에 대한 연구를 통해 밝혀질 수 있을 것이다.

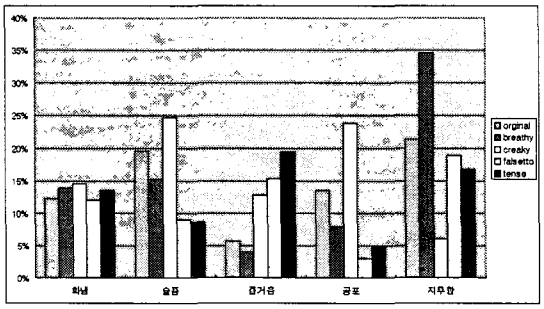

그림 1 발성유형에 따른 정서 평가 수치(%)

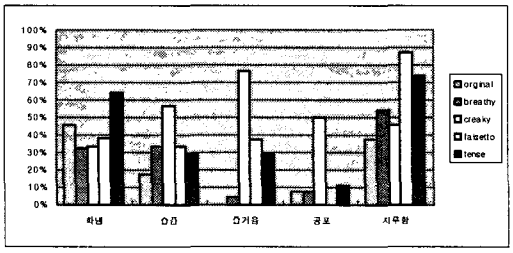

그림 2 가장 높은 평가 수치를 보이는 조건에서의 발성 양식에 따른 정서 평가 수치(%)-평균피치, 피치범위, 발화속도가 동일한 조건.

그림 2는 발성유형의 효과만을 단적으로 보여주기 위한 것이다. 각 정서에서 가장 높은 평가 수치를 보인 조건을 선택한 후, 피치와 피치범위, 발화속도 등은 동일하지만 발성유형이 다른 조건들과 비교했다. 발화속도에 크게 영향을 받는 것으로 나타난 화냄과 지루함에서의 차이가 다른 조건들에 비해 상대적으로 작아지지만, 대부분의 경우 발성유형에 따라 정서의 평가 수치가 크게 달라지는 것을 볼 수 있다. 특히 기쁨과 지루함의 경우, 평균피치, 피치범위, 발성속도가 최적화되었을 때 나타나는 발성유형이 전체평균을 통해 나타나는 유형과 다른 양상을 보이는데 이것은 발성유형과 다른 파라미터들간

의 높은 상호작용효과가 있음을 간접적으로 나타내는 것이라고 할 수 있다.

V.결론

평균피치, 피치범위, 발화속도 등의 변화에 대한 청취 평가 결과는 전체적으로 선행연구들의 결과와 일치하였다. 화냄의 경우 피치 범위는 좁아지며 발화속도는 느려지는 경향을 보인다. 이러한 결과는 냉소적인 화냄에 대한 평가가 높아졌기 때문인 것으로 보인다. 슬픔의 경우 평균피치는 높아지고, 발화속도는 느려지는 경향이 있으나 피치의 범위와는 뚜렷한 관련이 없는 것으로 나타났다. 즐거움은 피치범위는 넓어지고 발화속도는 느려지는 경향이 뚜렷했으며 평균피치는 대체로 올라가는 경향을 보였다. 공포조건은 평균피치와 피치범위, 발화속도에 의한 영향이 미미한데 상대적으로 발성유형에 많은 영향을 받는 것으로 추측된다. 지루함의 경우 평균피치는 낮아지고 발화속도는 느려지는 경향이 뚜렷하게 나타났지만, 피치범위와는 큰 관련이 없는 것으로 나타났다.

발성유형에 따른 정서효과는 전체적으로 기존의연구와 다른 결과를 나타냈다. 슬픔과 공포의 경우 짜내기 소리가 큰 영향을 미치는 것으로 나타났으며 즐거움의 경우 짜내기 소리보다는 긴장음에 더 많은 영향을 받는 것으로 나타났다. 지루함의 경우 숨소리에 의한 영향이 크게 나타났으며 화냄의 경우는 특정 발성유형에 큰 영향을 받지 않는 것으로 나타났다. 이러한 결과는 공포는 가성, 화냄은 긴장음, 지루함은 짜내기 소리에 많은 영향을 받는다는 Sendlmeier의 연구결과와 차이가 있다[4]. 이 차이는 보다 정밀하게 설계된 실험을 통해 밝혀질 수 있다. 이를 위해 우선적으로 한국어 화자에 대한 보다 다양한 정서음성의 수집과 분석이 선행되어야 할 것이다.

참고문헌

[1] 조철우, 김대현, 멀티미디어 환경을 위한 정서음성의 모델링 및 합성에 관한 연구, 한국음성과학회 Vol. 5, No.1. pp. 35-47, 1999.

[2] Breazeal, C., Emotive Qualities in Robot Speech, Proceedings of the 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems, Maui, HI. 2001.

[3] Cahn, J., Generating Expression in Synthesized Speech, M.S. thesis, MIT Media Lab, Cambridge, MA, 1990.

[4] Burkhardt, F., and Sendlmeier, W. F. Verification of Acousical Correlates of Emotional Speech using Formant-Synthesis, In SpeechEmotion-2000, pp. 151-156, 2000.

[5] Gobl, C. and Chasaide, A. N., Testing Affective Correlates of Voice Quality Through Analysis and Resynthesis, Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research. Belfast: Queen's University, pp. 178-183, 2000.

[6] 조철우, 정서정보의 변화에 따른 피치와 지속시간의 변화, 창원대학교 산기연 논문집 제11집 별쇄, 1997.

[7] Klatt, D. H., Software for a Cascade/Parallel Formant Synthesizer, Journal Acoustical Society of America, 67, pp. 971-995, 1980.

[8] Mahshie, J. and Gobl, C., Effects of Varying LF Parameters on KLSYN88 Synthesis, Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 1009-1012, 1999.

[9] Laver, J., The Phonetic Description of Voice Quality, Cambridge University Press, Cambridge, 1980.