정서정보의 변화에 따른 음성신호의 특성분석에 관한 연구

Analysis of Speech Signals According to the Various Emotional Contents

조 철 우*, 조 은 경**, 민 경 환***
(Cheol-Woo Jo*, Eun-Kyung Jo**, Kyung-Hwan Min***)

※이 연구는 1995년도 한림과학원 팀공동연구과제 연구비 지원에 의해서 이루어 졌습니다.

요 약

본 논문은 정서정보를 포함하여 수집된 음성자료를 여러 가지 신호처리 방법으로 분석한 결과에 대하여 기술하고 있다. 정서정보를 포함한 음성은 연극배우로부터 수집하였으며 분석은 주로 피치정보의 변화와 지속시간을 중심으로 행하였다. 수집된 음성에 대한 분석결과 정서정보의 변화에 따른 음성 파라미터의 변화치를 얻을 수 있었으며 이 실험은 앞으로의 정서음성정보의 분석에 필요한 기초적 실험으로 의의가 있다.

ABSTRACT

This paper describes experimental results from emotional speech materials, which is analysed by various signal processing methods. Speech materials with emotional informations are collected from actors. Analysis is focused to the variations of pitch informations and durations. From the analysed results we can observe the characteristics of emotional speech. The materials from this experiment provides valuable resources for analysing emotional speech.

Ⅰ. 서 론

최근들어 음성합성이나 인식의 분야에서는 합성된 음성의 자연성을 향상시킨다거나 인간과 자연스러운 상태의 음성을 인식하려는 시도가 진행되고 있다. 이는 기존의 음성처리 기술이 아주 제한된 범위내에서 실용화 될 정도로 발전되어 있기는 하지만 아직도 일반 사용자들의 면에서 볼 때는 사람들이 원하는 기대치와 구현된 시스템의 성능과는 많은 차이가 있는 것이 현실이다. 이러한 자연성의 구현요소로는 여러 가지가 있지만 그 중 하나가 정서(emotion)의 표현과 인식에 관한 문제이다. 인간의 정서정보는 음성에서 여러 가지 형태로 나타나고 있으며 이들 정보는 자연성에 있어서 중요한 정보의 하나이다. 그러나 이러한 정서정보를 분석하고 구현하는 데 있어서 우선적으로 해결해야할 문제는 어떻게 여러 가지 정서를 나타내는 문장을 자연스럽게 수집하는가 하는 문제이다. 낭독체 음성과는 달리 인간의 정서란 인위적으로 유발시키기가 매우 어렵기때문에 어떻게 원하는 정서를 이끌어내는가, 그리고 문장중에서 어떤 부분이 정서가 유발된 부분인가를 결정하는 것 등이 문제가 된다. 여러 가지 정서상태에서의 음성을 수집하기 위하여 다큐멘타리 녹음을 이용하든지 배우로 하여금 가상적인 정서상태에서 발성하게 하는 것등은 흔히 사용되고 있는 방법이다. 그러나 정서를 정량적으로 나타내는 것이 극히 어렵듯이 정서음성을 원하는 상태로 얻는 것 또한 극히 어려운 일이다.[1-6]

본 논문에서는 함께 수행된 정서음성 수집에 관한 연구에서 수집된 자료를 바탕으로 다양한 정서상태의 음성에 따른 음성신호 파라미터의 변환를 측정하여 통계적 특성을 구하였다. 본 연구의 목적은 주로 정서적 음성정보의 정량화에 의해 향후 합성 및 인식에 활용할 수 있는 파라미터의 추출 가능성을 연구하는 것이다. 주로 측정된 파라미터는 피치의 변화, 지속시간의 변화등이다.

Ⅱ. 정서정보 음성시료의 준비

정서정보를 포함한 음성시료는 별개로 수행된 연구에서 수집한 음성자료를 이용하였다. 우선 정서상태를 행복, 화냄, 슬픔, 두려움등 대표적인 4가지로 국한시켰다. 다른 연구에서 사용된 정서상태는 이외에도 지루함(boredom), 열정(enthusiasm), 기쁨(joy), 역겨움(indigestion), 놀

*창원대학교 제어계측공학과 부교수
**한림대학교 심리학과 조교수
***서울대학교 심리학과 교수
접수일자: 1996년 11월 13일

람(Surprise), 걱정스러움(Worry), 만족(Satisfaction)등이 있었으나 유사한 정서상태의 경우 구분이 어려워지고 자료의 양과 실험기간이 길어지는 관계로 뚜렷이 구분되는 4가지 정서상태로 한정하여 분석하였다.

음성자료의 수집을 위해서는 몇 개의 중의성 문장을 선택한 뒤 4인의 남자연극배우에 의해 4가지 정서상태로 3회씩 발성하게 하여 DAT에 기록하였다. 각 배우의 말씨는 표준말을 사용하였다. 이때 라링고그라프를 배우의 성대외벽부분에 부착하고 음성과 동시에 DAT의 2개 채널을 통해 기록하였다. 실험에 사용된 음성자료 수집을 위한 중의성 문장의 종류는 다음과 같은 5가지 종류이다. 분석에 사용된 자료는 총 240개 문장이다.

행복, 화냄, 슬픔, 두려움 등의 4가지 정서상태에 공통적으로 사용된 문장

1. 정말 그렇단 말이야
2. 난 가지말라고 하면서 문을 닫았다.
3. 야, 이제 그만하자
4. 이건 내가 원하던 게 아니야
5. 나는 ~입니다. (자신의 이름을 넣어서)

이외에도 특별한 정서상태에 대하여만 의미가 있는 몇 개의 문장이 수집되었으나 비교를 위한 대상으로는 적절하지 않기 때문에 분석에서는 제외하고 단지 참고로만 사용하였다.

정서음성의 수집에 있어서 배우를 이용할 경우 일반적으로 단순히 주어진 문장을 낭독하게 하는 방법을 많이 사용하고 있으나 이렇게할 경우 배우의 음성이 과장될 수가 있기 때문에 사전에 특정 정서를 포함한 내용을 낭독하여 충분한 정서적 분위기를 경험하게 한 후 낭독하도록 하여 정서상태를 유발시키는 방법을 사용하였다. 그러므로 하나의 정서 상태에서 하나의 문장을 읽도록 하는 데는 배경음악을 들려준다던가 자신의 과거의 경험을 회상시킨다던가 하는 과정이 들어감으로써 단순히 낭독하게 하는 것보다 훨씬 많은 시간이 소요되었다. 이와 같은 자전적 회상의 방법을 사용하는 경우는 심리학적으로 해당 정서상태를 회상하게하는데 75%정도의 성공률을 보인다는 것이 알려져 있다. 이와 같은 방법을 사용한 이유는 일반적으로 배우(연극배우나 성우들)들은 정서상태가 너무 과장되어 있어 일반적인 정서상태와는 차이가 있기 때문이다. 대개 배우를 사용한 실험의 경우 정서에 따른 특징의 차이는 명확히 나오겠지만 실제 정서상태가 어떠하다는 것과는 거리가 있는 결과를 얻기가 쉽다. 우리가 만약 정서를 인식하고자 한다면 실제 정서에 가까운 음성을 대상으로 해야 할 것이라는 것은 자명한 일이다.

Ⅲ. 일반적인 정서적음성의 특징

여러 가지 정서적 내용을 포함한 음성의 구분되는 특징은 영어의 경우 여러 가지 참고문헌에서 이미 잘 조사되어 있다. 이러한 정서적 변화들은 대개 다음과 같은 음향적 변수들을 관찰함으로써 구분될 수 있다고 Murray가 조사한 자료에서 종합적으로 언급한 바 있다.[1]

발성속도, 평균피치, 피치의 변화범위, 세기, 성질, 피치의 변화 및 조음방식등이 그것이다. 이러한 변화가 복합적으로 결합되면서 정서적 특징을 나타낸다고 하였다. 이러한 특징을 여기서 간단히 언급하면 다음과 같다.

(1)행 복

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

(2)화 냄

화냄의 경우 가장 많이 연구되어온 정서의 하나인데 이는 평균 피치가 아주 높으며 피치의 변화도 또한 크다. 발소리의 속도는 일반적으로 느려지며 긴장도가 강해진다.

(3)슬 픔

피치는 평균 피치값보다 낮아지며 속도도 느려진다. 강도는 낮아지고 불규칙적인 휴지기가 발생한다.

(4)두려움

피치의 평균치가 높아지며 피치의 범위는 커진다. 발성속도는 정상치보다 빨라진다. 발성상태는 때로 불규칙적이고 조음구조는 정상상태보다 구분되어 발성되는 경향이 있다.

이와 같은 각 정서의 특징은 주로 통계적 대표치의 결과를 바탕으로 그 경향을 분석한 것이나 Paavo Alku등은 피치패턴과 함께 역필터링 분석을 통한 성대의 움직임도 함께 분석하였다.

Ⅳ. 정서음성의 분석방법

획득된 정서음성을 분석하기 위해서는 각 정서상태에 의한 말소리를 구분시켜줄 수 있는 척도를 설정하고 규정된 척도에 의해서 음성을 분석하여 필요한 결과를 얻어내어야 하는데 본 연구에서는 정서음성을 분석하여 얻은 결과를 음성합성이나 인식등 공학적 응용에 적용할 것을 목표로 하고 있으므로 정성적 특징의 변화와 함께 정량적 특성을 측정하기로 한다.

분석에는 DAT에 녹음된 신호를 Laryngograph Processor를 거쳐 Laryngograph Ltd.의 SPG프로그램을 이용하여 분석하였다. 이 프로그램은 실시간으로 스펙트로그램, 시간축파형, 피치의 변화, Open Quotient를 구해준다.

본 연구에서 측정, 분석된 파라미터는 다음과 같다.

(1)각 발화자에 대하여 동일 문장에 대한 4가지 정서상태별 피치패턴, 발화율, 강도의 변화, 피치의 범위의 변화

(2)동일 문장, 동일 정서상태에 대한 4명의 발화자의 개인별 피치패턴, 발화율, 강도의 변화, 피치값의 범위의 변화

정서적 음성의 특징을 나타내는 가장 큰 인자가 피치

와 강도, 발화율이라는 것은 잘 알려져 있다.[14]

V. 실험결과 및 분석

앞서의 음성수집과정에서 얻어진 음성시료들을 대상으로 몇가지 파라미터들을 구하고 분석하였다. 처음 수집할 때 성대의 운동을 관찰하기 위하여 라링고그라프 신호를 동시에 같이 수집하였으나 실제 수집된 신호들을 분석해 본 결과 라링고그라프 신호의 경우 신호의 레벨이 너무 작게 기록된 경우 및 개별 화자의 감정상태의 기복에 따라서 신호가 너무 크게 변동하거나 하여 SPG에 의해서 측정이 곤란한 경우가 많았으므로 결과분석과정에서 별로 도움이 되지 않는 경우가 많았다.

정서에 관련하여 변동되는 파라미터들로는 크기의 변화, 피치의 변화, 성문파형의 변화 등을 관찰할 수 있는데 크기의 경우는 녹음레벨이 일정하지 않은 경우 피치의 추출이 어려운 관계로 수시로 레벨을 조정한 결과 수집된 음성시료에서 크기의 측정은 의미가 없고 단지 피치 주파수의 최대치, 최소치 및 범위 그리고 라링고그라프 신호로부터 측정한 성문의 여닫음의 정도를 측정하였다. 아래에 수집된 음성들중 대표적인 두가지 사례를 들고 설명한다.

표 1과 표 2는 각각 청취실험에서 가장 유도가 잘 되었다고 판정된 배우1과 배우2의 음성에 대한 분석결과이다.

표 1. 배우1의 음성측정자료(정상 피치주파수 : 135~192Hz, 폭 : 57Hz)

		행복	화냄	공포	슬픔
문장5	Pitch Max	170	274	262	350
	Pitch Min	80	160	211	162
	ΔF(Hz)	90	114	51	88
	ΔL(ms)	1100	1100	650	1160
문장3	Pitch Max	230	180	240	390
	Pitch Min	48	140	200	240
	ΔF(Hz)	82	40	40	150
	ΔL(ms)	2100	1110	780	1600

표 2. 배우2의 음성측정자료(정상 피치주파수 : 137~171Hz, 폭 : 34Hz)

		행복	화냄	공포	슬픔
문장5	Pitch Max	250	131	160	122
	Pitch Min	88	103	108	89
	ΔF(Hz)	152	28	52	33
	ΔL(ms)	1222	1325	1144	1274
문장3	Pitch Max	137	116	180	162
	Pitch Min	83	49	138	105
	ΔF(Hz)	54	67	42	57
	ΔL(ms)	1196	1894	850	1092

우선 배우1의 경우 음성의 분석결과는 다음과 같다.

먼저 행복의 경우 피치 주파수는 80~170Hz의 범위를 가지고 변화하고 있고 끝 부분이 다른 정서상태에 비해서 심하게 떨어지는 것을 볼 수 있다. 지속시간은 /~입니다/부분의 경우 1100ms를 나타내었다.

그 다음 화냄의 경우는 피치값이 160~274Hz사이에서 변화하고 있으며전체적으로 피치궤적이 올라가 있는 경향을 보여준다. 측정된 지속시간은 1100ms를 나타내었다

공포의 경우는 전체적인 피치 궤적이 가장 올라가 있으며 피치 주파수는 211-262Hz의 범위를 갖는다. 그리고 피치의 변화가 거의 없는 완만한 궤적을 그리고 있음을 알 수 있다. 지속시간은 650ms로 다른 정서상태에 비해 빨라지는 경향을 보이고 있다.

슬픔의 경우는 두 번째로 높은 평균 피치를 보이고 있으며 궤적의 기복이 다른 정서상태에 비해서 높음을 알 수 있다. 피치 주파수의 범위는 162~350Hz로 측정되었고 지속시간은 1160ms정도로 다른 정서에 비해서 길어진 것을 알 수 있다.

이상의 측정결과에서 관찰된 결과를 정리하면, 평균피치궤적은 공포>화냄>슬픔>행복의 순서로 높은 주파수를 보였으며 지속시간의 면에서는 슬픔>즐거움>화냄>공포의 순으로 길이가 긴 것을 알 수 있었다.

/나는/과 /~입니다/의 간격도 정서상태에 따라 달라지는 것이 관찰 되었는데 화냄>행복>슬픔>공포의 순으로 간격이 길어졌다.

화자2의 경우의 음성분석결과는 다음과 같다.

행복의 경우 피치 주파수는 48~230Hz였고 지속시간은 2100ms정도였다.

화냄의 경우 피치주파수는 140~180Hz, 지속시간은 1110ms였다.

슬픔의 경우 240~390Hz, 지속시간은 1600ms였다.

공포의 경우 200~240Hz, 지속시간은 780ms정도였다.

이 문장의 경우는 전체적인 피치 궤적의 높이는 슬픔>공포>행복>화냄의 순이었고 지속시간은 행복>슬픔>화냄>공포의 순이었다. 또 /야/와 /이제 그만하자/의 간격은 행복과 화냄의 경우는 음성이 서로 이어져 발성되었고 공포와 슬픔의 경우는 중간에 확실히 끊어지는 부분이 발생했다. 간격은 행복 100ms, 화냄120ms, 공포80ms, 슬픔800ms로 슬픔의 경우가 가장 간격이 길었다.

전체적으로 볼 때 신호의 강도는 화냄의 경우가 가장 컸고 그다음으로 행복, 공포, 슬픔의 순으로 크기가 작았다. 그리고 공포의 경우는 다른 정서에 비해서 기음화 현상이 강하게 나타났다.

이와 같은 결과는 기존의 영어권에서 조사된 결과와 비교적 일치하는 경향을 보였으나 일부 피치값의 범위에서는 불일치하였다. 예를 들어 행복의 경우 본 실험에서는 평균 피치값이 높아지는 경우도 있었는데 이는 배우의 연기하는 상태에 따라 다소 과장된 음성이 발성된 때문이라고도 볼 수 있으며 상황에 따라서 변화할 가능성도 있음을 말해준다. 그러나 다른 일반적인 정서상태에 따른 피치궤적이나 지속시간의 변화에 있어서는 대체로 일치하는 경향을 보여줌을 확인했다.

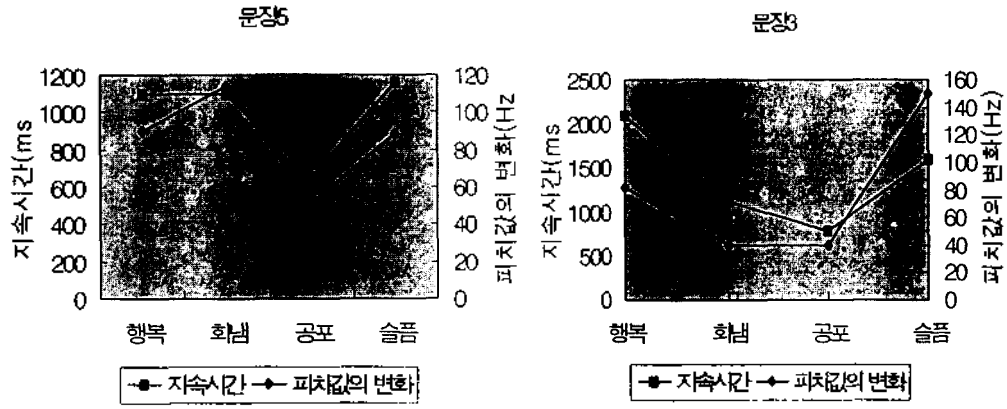

그림1. 4가지 정서에 따른 피치와 지속시간의 변화(화자1)

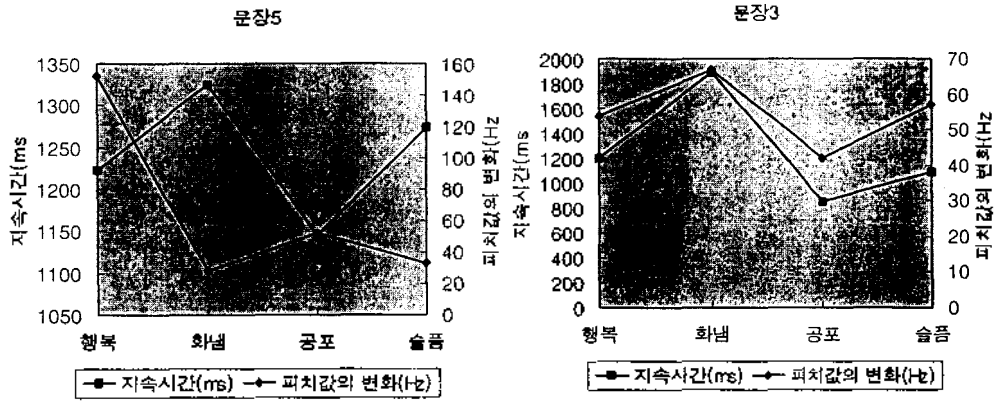

그림2. 4가지 정서에 따른 피치와 지속시간의 변화(화자2)

그림 1, 그림 2는 각각 화자1과 화자2의 각 정서상태에 따른 피치폭의 변화와 피치값의 변화를 그래프로 그린 것이다. 이 그래프들에서 주목할만한 사실은 화냄의 경우 피치값과 지속시간의 값은 같은 화자의 경우에도 경우에 따라 높아질 수도 있고 낮아질 수도 있다는 것이다. 다른 세가지 정서상태는 비교적 화자에 따라 일치하는 변화를 보이고 있다. 이러한 결과가 나온 이유는 몇가지 원인으로 생각해 볼 수 있다. 먼저 동일한 화냄상태의 경우라도 정서상태가 화자1과 화자2의 경우 또는 서로 다른 문장을 발성하면서 서로 다르게 유도된 때문으로 볼 수 있다. 이 결과가 나타내주는 사실은 정서의 유도과정에서 동일한 정서일지라도 다양한 변화속에서 유도가 가능하기 때문에 일반성을 얻기 위해서는 동일한 정서상황을 유도해 주어야 한다는 것이다. 예를 들어 두 배우에게 동일한 '화냄' 상태를 연기하라고 한 경우라도 한 배우는 열화와 같은 화낸 상태를 표시할 수 있고, 다른 배우는 신경질적인 화냄을 표시할 수가 있다. 이러한 사실은 정서정보를 공학적 응용에 이용하기 위해 수집할 경우, 또는 정서상태를 정의할 경우에도 고려해야할 사항이다.

Ⅵ. 결 론

본 연구에서는 4가지 정서상태를 잘 훈련된 배우들을 통해 유도한 다음 지시된 문장을 그 정서상태에서 읽도록 하는 방법으로 정서적 음성을 수집하고 그 수집된 음성의 피치와 지속시간 등을 측정하여 각 정서상태에서의 음성의 특성을 분석하였다.

측정결과 평균피치 궤적은 공포상태에서 가장 높게 나타났으며 지속시간은 슬픔의 경우에 가장 길게 나타났고 피치의 변동 범위는 행복의 경우 가장 크게 나타났다. 또한 4가지 정서중 '화냄'상태가 개인에 따른 편차가 크고 일관성있는 자료를 수집하기가 어려운 것으로 나타났으므로 앞으로의 연구에 주의를 기울여야할 것으로 생각된다.

본 연구에서 얻은 결과는 정서정보가 포함된 음성을 모델링하여 합성 또는 인식하거나 인간의 정서와 음성과의 관계에 대한 연구의 기초적인 실험으로 유용하다고 생각된다. 앞으로의 연구방향은 현재 수집된 자료들을 보다 구체적으로 정리하고 분석하여 일반적인 정서음성의 모델링을 구현하고자 한다.

참 고 문 헌

1. Marray, I.R. and Arnott, J.L. "Toward the Simulation of emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion", JASA, 93 (2), pp.1097-1108, 1993.
2. Chung, S.J, "An Acoustic and Perceptual Study on the Emotive Speech in Korean and French", ICPhS'95-Stockholm, Vol.1, pp.266-269, 1995.
3. Klasmeyer, G, Sendmeier, W.F, "Objective Voice Parameters to Characterize the Emotional Speech", ICPhS'95-Stockholm, Vol.1, pp.182-185, 1995.
4. Mozziconacci, S., "Pitch Variations and Emotions in Speech", ICPhS'95-Stockhom, Vol.1, pp.178-181, 1995.
5. Alku, P. et.al., "On the Perception of Emotional Content in Speech", ICPhS'95-Stockholm, Vol.1, pp.246-249.
6. M. Martin, "On the Induction of Mood", Clinical Psychology Review, pp.669-697, Vol.10, 1990.

▲조 철 우(Cheol-Woo Jo) 1961년 9월 12일생

1983년 : 고려대학교 전자공학과(공학사)
1985년 : 고려대학교 대학원 전자공학과(공학석사)
1989년 : 고려대학교 대학원 전자공학과(공학박사)
1985년~1986년 : 한국전자통신연구소 음향연구실 위촉연구원
1992년~1993년 : 영국 Keele University, Dept. of Communication and Neuroscience(Visiting Research Fellow)
1993년~1995년 : 국방과학연구소(진해) 위촉연구원
1996년 1월~2월 : 영국 Keele University, Dept. of Computer Science(Visiting Research Fellow)
1989년~현재 : 창원대학교 제어계측공학과 부교수
※주관심분야 : 음성신호처리, 음성합성, 정서음성처리, 음향신에 의한 계측

▲조 은 경(Eun-kyung Jo) 1963년 8월 22일생

1987년 : 서울대학교 심리학과(문학사)
1989년 : University of Wisconsin at Madison(심리학석사)
1993년 : University of Wisconsin at Madison(심리학박사)
1993년~1994년 : 한국형사정책연구원 선임연구원
1994년~1995년 : 한국형사정책연구원 초빙연구원
1994년~현재 : 한림대학교 심리학과 조교수
※주관심분야 : 정서와 인지, 정서조절기제, 신체변화와 정서경험, 공격성과 범죄행동

▲민 경 환(Kyung-Hwan Min) 1949년 9월 5일생
1973년 : 서울대학교 심리학과(문학사)
1975년 : 서울대학교 대학원 심리학과(문학석사)
1983년 : University of Washington(심리학박사)
1990년~1991년 : 미국 Harvard University, Yenching Institute, Visiting Research Fellow
1984년~1988년 : 서울대학교 심리학과 조교수
1988년~1994년 : 서울대학교 심리학과 부교수
1994년~현재 : 서울대학교 심리학과 교수
※주관심분야 : 성격발달, 정서발달과 정서표현