음성의 감성요소 추출을 통한 감성 인식 시스템

The Emotion Recognition System through The Extraction of Emotional Components from Speech

박 창 현, 심 귀 보*
(Chang-Hyun Park and Kwee-Bo Sim)

Abstract : The important issue of emotion recognition from speech is a feature extracting and pattern classification. Features should involve essential information for classifying the emotions. Feature selection is needed to decompose the components of speech and analyze the relation between features and emotions. Specially, a pitch of speech components includes much information for emotion. Accordingly, this paper searches the relation of emotion to features such as the sound loudness, pitch, etc. and classifies the emotions by using the statistic of the collecting data. This paper deals with the method of recognizing emotion from the sound. The most important emotional component of sound is a tone. Also, the inference ability of a brain takes part in the emotion recognition. This paper finds empirically the emotional components from the speech and experiment on the emotion recognition. This paper also proposes the recognition method using these emotional components and the transition probability.

Keywords : emotion, bayesian learning, statistical method, inference, transition probability

Ⅰ. 서론

인간은 다분히 감정적인 동물이다. 말을 하지 못하는 갓난아기는 자신의 감정을 표현하는 것으로 타인과의 의사소통을 시작한다. 감정은 제 3의 언어로써 매우 중요한 역할을 수행하는 것이다. 과학 기술의 발전으로 이제 인간의 영역에 기계들의 비중이 매우 커지고 있다. 산업용으로만 사용되어지던 로봇들을 가정에서도 사용할 수 있게 되어지고 있는 것이다. 하지만, 단순히 차가운 기계에서 벗어나지 못한다면 가정용 로봇의 의미는 반감되어질 것이다. 또한 로봇의 입장에서도 감정은 생존을 위해 중요하다. 다윈이 연구한 바에 의하면 감정들은 안전한 상황과 위험한 상황, 기회를 엿볼 수 있는 상황들을 재빨리 구분할 것을 다급하게 요구하며, 긴급 상황을 대처하는데 필요한 추가의 에너지와 스태미나를 동원하게 하고 위험한 상황에 대한 공포를 느끼는 것으로 생존의 위협에서 벗어날 수 있게도 한다. 즉, 감정은 업무의 효율화와 생존을 위한 주요 기능인 것이다. 이러한 이유 때문에 감성인식의 중요성은 커지고 있다[1]. 인간이 느끼고 표현하는 감정들은 분노, 불안, 공포, 수치심, 기쁨, 사랑, 슬픔뿐만 아니라 죄책감, 선망, 질투, 긍지, 안도감, 희망, 감사, 동정심과 같은 사회적으로 미묘하다 할 수 있는 감정들이 포함되어 있다[1]. 이러한 감정들은 갑작스럽게 창조되는 것이 아니라 각 감정의 플롯에 의해 전개되는 것이다. 분노의 감정이 발생했다면, 그 감정 이전에는 어떤 상대와의 교류 속에서 분노의 요인이 되는 것을 받았기 때문이다. 그러므로 이러한 감정의 원인을 인식하고 추론할 수 있다면 현재 상대 감정의 인식을 통하여 상대의 기분이 안 좋은 경우에는 좋은 방향으로 유도할 수도 있고 우울할 때는 기분을 중화 시킬 수 있을 것이다. 하지만, 감정의 원인을 인식한다는 것은 인간이 태어나면서 학습해온 문화에 대한 수많은 요소에 대한 이해 없이는 매우 어려운 일이다. 그렇기 때문에 감정의 원인 인식 연구는 차후의 연구과제로 남겨놓고, 본 논문에서 제안하는 것은 발화로부터 감성적인 요소들과 통계적 자료를 바탕으로 추론하여 사용자의 감성을 인식할 수 있다는 것이다.

과거의 연구자들은 음성의 신호 자체로부터 감성을 인식하려고 노력하였다. 즉, 음성 신호의 주요 분석 요소인 스펙트럼과 시간 축에서의 파형 자체로부터 각 감정별 특징을 추출하여 학습하는 것이다. 본 논문도 이러한 기본 틀에서 크게 벗어나지는 않는다. 다만, 신호의 분석만으로는 불안정한 인식을 통계적인 자료를 바탕으로 추론하여 좀더 불확실한 요소를 제거하겠다는 것이다.

먼저 과거의 연구자들이 사용한 방법들을 살펴보면, Chen은 각 문장별로 피치와 RMS 에너지 윤곽선을 구하여 즐거움, 슬픔, 화, 싫음, 놀람, 두려움 6가지 감정의 특징으로 사용하였다. 하지만, 이 특징만으로는 감정의 인식 결과가 좋지 않고 표정인식을 부가하여서 각 방법에서 부족한 점을 보완해줄 수 있다고 하였다[2]. 이 논문은 오디오 정보만을 이용하여 약 70% 정도의 인식율을 보여주었지만, 실험 방법이 명확하지 않아 인식율의 의미가 부정확하고, 단지 비전이 보완 해줄 때 감성인식의 정확도가 높아진다는 것을 확인 할 수 있다. J. Nicholson은 의식적인 감정표현과 무의식적인 감정표현으로 개념을 나누어 인식하기에 더 쉬운 의식적인 감정표현에 국한하여 연구를 진행하였다. 또한, 기존의 연구자들이 분류한 감정들은

* 책임저자(Corresponding Author)
논문접수 : 2004. 6. 6., 채택확정 : 2004. 7. 18.
박창현, 심귀보 : 중앙대학교 전자전기공학부
(3rr0r@wm.cau.ac.kr/kbsim@cau.ac.kr)
※ 본 연구는 산업자원부 차세대신기술개발사업 2단계 1차년도 IIWM개발과제의 연구비 지원에 의해서 수행되었음.

- Neutrality, Joy, Boredom, Sadness, Anger, Fear, Indignation
- Anger, Fear, Sadness, Joy, Disgust
- Neutrality, Happiness, Sadness, Anger, Fear, Boredom, Disgust
- Fear, Anger, Sadness, Happiness

위의 분류와 같았고, J. Nicholson은 이 분류를 참조하여 Joy, Teasing, Fear, Sadness, Disgust, Anger, Surprise, Neutral 이 8가지 감정에 대하여 특징들을 추출하였다. 추출된 특징은 운율학적 특징(Prosodic Features)과 음성학적 특징(Phonetic Features) 크게 두 가지 속성으로 분류를 하였고, 이를 이용하여 하위 신경망들(Sub neural networks)을 거쳐 Decision logic으로부터 결과를 얻는 방법을 사용하였다. 이 논문은 학습 데이터를 이용한 Closed의 경우와 새로운 음성들에 대한 Opened인 경우로 나누어 실험을 하였는데, Closed인 경우 평균 70% 정도의 인식율을 보였으나, Opened인 경우는 평균 30% 정도의 인식율을 보였다. 이는 화자 식별이 되는 경우 각 개인의 데이터를 기반으로 감성을 인식할 경우에는 확연히 높은 인식율을 가질 수 있다는 것을 보여준다. 하지만, Opened인 경우의 인식 율은 보편적인 감성 특징을 찾는 일이 얼마나 어려운 일인가를 극명하게 보여준다[3]. 또 다른 논문에서는 17개의 특징을 이용하였다. 이 17개의 특징들은 5개의 범주로 나뉘어 지는데, 그 범주는 다음과 같다.

- Statistics related to rhythm
- Statistics on the smoothed pitch signal
- Statistics on the derivative of the smoothed pitch
- Statistics over the individual voiced parts
- Statistics over the individual slopes

위의 5 범주로부터 특징 선택과정을 통해 성능을 저하시키는 특징점들은 제외하고 성능을 향상 시킬 수 있는 특징들만을 선택한다. 너무 많은 특징들은 오히려 성능을 저하시키는 경향이 있기 때문이다. 4개의 감정이라면 적절한 2개의 특징들로 커버할 수 있다. 이 논문에서는 KNN으로 각 특징점 집합들에 대해 분류한 뒤 다수결 원칙에 의해 인식하는 방법을 이용하였고 약 80%의 인식율을 보여주었다[4]. 또한, 피치와 에너지 같은 음향적 요소 외에도 내용에 관련된 특징점을 사용하는 경우도 있다. 예를 들면, The use of swear words, Discourse information, Repetition of the same sub-dialog 같은 것들을 이용하는 것이다[5][6]. 위의 논문들을 통해 감정 특징과 분류 방법에 대한 정리를 해보았다. 이외에도 많은 연구자들이 공통적 특징으로 사용한 것은 피치이고 이 것의 패턴을 어떻게 하면 잘 반영할 수 있는가 하는 것이 특징 추출부에서의 주요 문제이다.

본 논문의 II절은 인식의 전 처리 과정에 대한 설명이다. 음성으로부터 피치를 추출하기 위해 잡음 성분을 제거해주는 Center Clipping 함수의 적용을 설명하였고, III절은 3가지 파라미터들과 4종류의 감성간의 관계를 실험을 통하여 알아보았다. IV절은 감성 인식 시스템의 구조에 대한 설명으로써 통계적인 방법을 이용한 시스템을 제안하였고 V절에서는 제안한 시스템에 대한 실험결과를 보이고 이전 연구결과와 비교를 하였다.

II. 주요 특징점 추출

가장 원시적인 형태의 감성 표현은 위급한 상황에서의 생명유지가 목적이었기 때문에 크기나 거칠기가 핵심 요소였고 인간의 문화가 발전되어 가면서 점차적으로 세밀한 변화를 거쳐 크기, 높이, 빠르기, 거칠기, 흔들림 등의 요소로 이루어진 지금의 감성 표현이 되었다.

1. 피치 추출

감정의 표현에서는 피치가 가장 많은 정보를 갖고 있기 때문에 피치를 정확히 추출하는 것이 매우 중요하다. 음성신호의 피치 추출 방법은 다음과 같이 크게 시간 축에서의 검출 방법과 주파수 축에서의 검출 방법으로 나뉘어져 있다[7].

1) 시간 축에서의 검출 방법

- PPA(Parallel Processing Approach)
- ACA(Autocorrelation Approach)
  - Center clipping 함수를 사용한 ACA
  - 3-level Center clipping 함수를 사용한 ACA

2) 주파수 축에서의 검출 방법

- CA(Cepstrum Approach)

하지만 주파수 영역에서의 피치 추출 알고리즘은 개념은 쉽지만, 계산량이 많으므로 일반적으로 사용되는 방법은 아니다. 가장 널리 사용되는 방법은 피치들이 Aautocorrelation 함수에서 잘 나타난다는 특징을 이용한 ACA이고 본 논문에서는 Center clipping 함수를 이용한 ACA를 사용하였다.

1.1. Center clipping 함수를 이용한 ACA

음성신호는 피치 이외에도 많은 정보를 함유하고 있는데, 이러한 잔여성분은 저역통과필터를 통과하더라도 존재하고 이는 곧 피치 추출의 방해요인이 된다. 그렇기 때문에 저역통과필터 외에 Center clipping 함수를 통과시키면서 이러한 방해요인을 제거하여 피치 검출의 성능을 향상시키는 방법을 이용한다.

그림 1에서와 같이 Center clipping 함수는 음성신호 (x)가 일정한 레벨(CL) 내에 있으면 그 신호를 무시하고, CL 보다 크면 원래 신호에서 CL을 뺀다. 이는 음성신호 중에서 피치에 해당하는 성분은 크기가 크게 나타나는 특징을 이용해서 잔여성분을 제거하는 방법이다. 이렇게 처리된 음성신호에 Autocorrelation 함수를 이용하여 주기성을 찾는 방법이다[7].

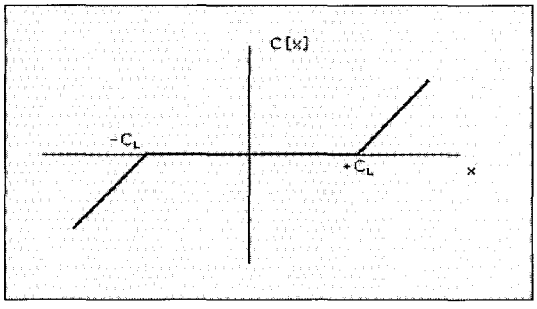


그림 1. Center Clipping 함수.
Fig. 1. Center Clipping Function.

Ⅲ. 감성과 특징점 간의 관계

감정 표현은 관습, 습관에 의해 많은 영향을 받는다. 관습, 습관의 영향을 받는 다는 말의 의미는 패턴을 갖고 있다는 말과 같다. 발화된 음성의 크기, 높이, 빠르기 등등 갖가지 작은 단위의 패턴들이 어떻게 조합되었는가에 따라서 듣는 이에게 기쁨으로 들릴 수도 있고, 화내는 것으로 들릴 수도 있는 것이다. 이 절에서는 감정 표현 시에 나타날 수 있는 요소들의 변화를 관찰한다.

1. 톤(Tone or Pitch)

본 실험은 자체적으로 만든 Analyzer/Simulator인 'NewAcous'를 이용하여 분석기의 유효성을 확인하였다. 3명의 실험자에게 톤을 높여가며 5단계로 "[아]"를 발화하도록 하였다. 이때 톤을 높여 발화하는 것은 주관적이므로 실험자 외의 5명의 사람에게 녹음한 소리를 들려주어 톤의 높낮이 에 대한 의견이 일치된 소리에 대해 분석기에서 실험한 결과는 다음의 그림과 같다.

그림 2에서 A, B, C는 발화한 실험자를 식별하고, 가로축은 좌측부터 우측으로 톤을 높이는 순서이다. 또한, 세로축은 피치를 Hz로 나타낸다. 그림에서 보는바와 같이 톤을 높일 때 피치가 선형적으로 상승하는 것을 확인 할 수 있다.

2. 크기(Loudness)

크기(Loudness)란 조용한 상태에서 큰 상태까지의 크기를 청각에 따라 순서대로 나열하는 것으로 측정한다. 이것은 주파수, 강도(Intensity)에 따라서 달라질 수 있다. 또한, 이것은 매우 주관적인 척도이므로 객관적인 양으로 측정하기가 힘들다[8]. 음향학(Acoustics)에서 Loudness는 물리적인 강도(Physical intensity)와 주관적인 크기(Judged Loudness)를 결정하여 구하나, 본 논문에서의 Loudness는 파형 의 모음 부분들에서 크기들의 평균으로 정하였다. 그림과 같이 입력되어진 샘플 음성을 절대값을 취하여 시간축에 나타내었고 에너지와 피치에 대한 기준 값에 따라 Sector 1, 2, 3으로 나누어진다. 그리고 각 Sector에서의 파형의 몇 개 샘플들의 평균값을 구하여 Sect Loud 1, 2, 3을 구한다. 본 논문에서의 Loudness는 이것들의 평균값이다.

이렇게 간단한 방식으로 Loudness를 구하는 이유는 감정 인식의 경우에 발화의 크기가 정밀하게 구해질 필요가 없고, 단지, 매우 조용, 보통, 조금 큰, 매우 큰, 이 정도의 구분만 되면 되기 때문이다. 다음은 동일한 실험자 3명으로 하여금 5단계로 소리를 크게 하도록 하였다. 아래의 그림에서 'Loudness'를 보면 청각적으로 크게 들리는 소리가 수치적으로 큰 값을 가짐을 확인 할 수 있다.

3. 분할된 구간의 개수(Sect. No.)

다음은 에너지와 피치를 이용해 구한 구간(Section)이 끊어서 천천히 발화하는 경우와 빨리 발화하는 경우, 리드미컬하게 발화하는 경우에 어떻게 나누어지는 지를 보여준다.

표 1에서 위로부터 3개의 열들은 발화된 모음의 개수가 늘어나지만 Sect. No.은 3개로 동일하고, 4번째 열에서만 Sect. No.는 4개가 된다. 5, 6, 7 열들은 [아] 음을 빨리, 개수를 늘여가면서 발화한 경우이다. 이때는 [아] 음의 개수에 따라 구간의 개수도 2, 3, 4로 증가했지만, Sect No.의 절대

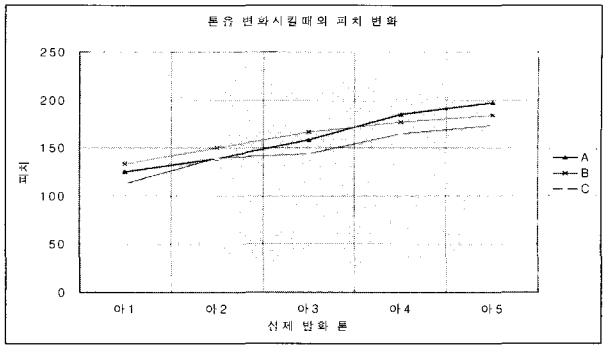


그림 2. 실제 발화에 대한 분석기의 피치 검출 실험.
Fig. 2. Pitch Extraction Experiment of 'NewAcous'.

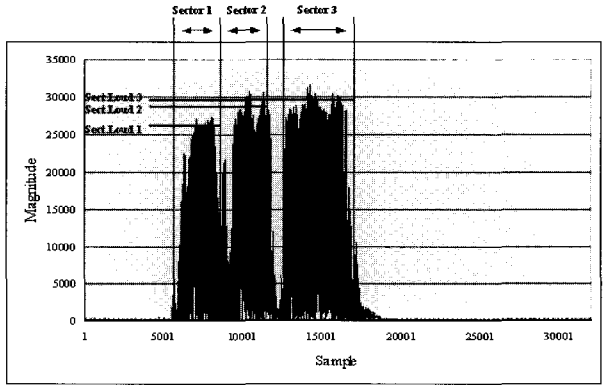


그림 3. 음성의 크기와 구간 구하는 방법.
Fig. 3. The method of getting 'Loudness' and 'Sector'.

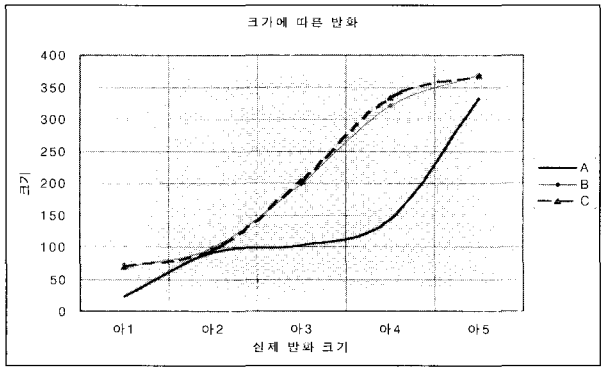


그림 4. 실제 발화 크기에 대한 Loudness의 유효성실험.
Fig. 4. The validation of 'Loudness'.

적 수치는 작다. 마지막 열은 리드미컬하게 [아]를 발화하였다. 이때는 7개의 구간으로 나뉘어졌다. 즉, 구간은 음의 발화가 불안정하게 되어졌을 때 개수가 많아지고, 음을 일정하게 발화 했을 때는 개수가 작다. 또한 동일한 음을 빨리 발화하면 모음이 뭉쳐지는 현상이 생겨서 구간의 개수가 줄어든다. 그러므로 구간의 개수로부터 음이 불안정하게 발화되는 흐느낌이나, 슬픔을 찾을 수 있고 이와 반대로 음이 안정적인 특성을 갖는 평서형을 찾을 수 있는 것이다.

표 1. 발화의 변화에 따른 구간 개수의 변화.

Table 1. The variation of Sect. No. for the variation of speaking.

발화된 소리	Sect. No.
[아아]	3
[아아아]	3
[아아아아]	3
[아아아아아]	4
[아아]:빠르게	2
[아아아]:빠르게	3
[아아아아아]:빠르게	4
[아아아아아]:리드미컬	7

표 2. Loudness에 대한 감정별 분포.

Table 2. The frequency of each emotion based on the Loudness.

감정 / 크기	평서	즐거움	화	우울
10000	0	6	6	0
8000	0	0	2	0
7000	0	2	0	0
6000	4	0	2	0
5000	4	4	4	0
4000	0	9	8	0
3000	2	4	3	0
2000	0	1	3	2
1000	8	2	5	5
800	1	0	0	0
600	2	0	0	5
400	2	0	0	9
200	12	2	2	0

4. 감정과 3가지 요소들 간의 관계 실험

감정과 앞 절들에서 살펴본 3가지 요소들과의 관계를 알아보기 위해 10명의 연기자들로 하여금 각각 4가지 감정으로 총 484개(평서형 : 35개, 즐거움 : 30개, 화 : 35개, 우울함 : 21개)의 대사를 하도록 하였다. 녹음된 파일 형태는 16bits, 11kHz, mono 이다. 표 2, 3, 4는 각각 tone, sect. no, loudness를 기준으로 감정별 분포를 나타낸다. 그러나, 먼저 주의할 점은 편의상 4개의 감정으로 나뉘어져 있지만, 실제 연기자들의 표현방식이 다소 차이가 있기 때문에 동일한 감정에서도 느낌이 다르다는 것이다. 이 사실은 요소의 분포로 확인할 수 있다. 앞서 말한 바와 같이 Loudness의 크기는 파형의 크기에 비례한 수치로써 특별하게 조정된 값이다. 예를 들면, 평서형에서는 35개의 대사를 발화하였고, 그 중 6000에 속하는 발화가 4개 5000에 속하는 것들이 4개, 3000에 2개, 1000에 8개 등이 있다. 표 2로부터 평서형의 경우에는 대부분의 데이터들이 1000 이하에 분포되어 있는 것을 확인할 수 있다. 마찬가지로 즐거움의 경우에는 1000에서 5000 사이에 많이 분포하고 또한 10000에서의 분포 확률도 높다.

표 3은 발화 문장의 구간 개수에 의한 분포를 보여준다. 이 분석에서는 문장의 길이가 구간 개수에 영향을 미칠 수 있으므로 길이가 비슷한 문장으로 녹음 되어있다. 이 부분은 3절에서 보는 바와 같이 발화시 끊김, 연음정도, 안정도가 큰 영향을 미친다. 또박또박 끊어서 얘기하면 구간의 개수가 모음의 개수만큼 생길 것이지만, 일상적인 발화에는 대체로 끊어서 얘기하는 경우가 없다고 가정한다.

표 3. Sect. No.에 대한 감정별 분포.

Table 3. The frequency of each emotion based on the Sect.No.

감정 / 크기	평서	즐거움	화	우울
10	0	2	0	0
9	0	0	2	0
8	0	0	0	4
7	0	2	0	4
6	0	4	0	10
5	4	3	4	0
4	10	7	16	1
3	12	6	6	0
2	9	6	7	2

표 4. 피치 평균에 대한 감정별 분포.

Table 4. The frequency of each emotion based on the Pitch.

감정 / 크기(Hz)	평서	즐거움	화	우울
200	0	0	4	0
190	0	1	2	0
180	4	6	6	0
170	0	9	2	0
160	1	3	4	0
150	0	3	1	0
140	6	5	2	2
130	13	2	6	1
120	1	0	4	0
110	4	1	4	8
100	6	0	0	2
90	0	0	0	6
80	0	0	0	0
70	0	0	0	2

연음이 많고 피치의 변화가 적으면서 빠르게 이어서 얘기하는 경우는 구간이 적어지므로 개수가 적어진다. 이런 식의 발화는 평서형과 '화'에서 많이 보인다.

표 3에서 평서형과 '화'의 경우는 개수가 2, 3, 4일 때 가장 많이 분포되어 있는 것을 볼 수 있다. 또한, 개수가 증가하는 경우는 음이 불안정하게 오르락내리락 하는 경우나 끊김이 많은 경우인데, 이는 우울하거나, 울먹일 때, 즐거운 상태에서도 자주 보인다. 표에서 보면 즐거운 경우보다는 우울한 경우 훨씬 개수가 많은 것을 볼 수 있다.

표 4는 피치의 평균에 의한 분포를 보여준다. 여기서 피치의 단위는 Hz다. 평서형은 100~140에서 많이 분포하는 것을 볼 수 있고, '우울'의 경우는 110 이하에서 많이 분포하는 것을 볼 수 있다. 즐거움은 대체로 140~180에 모이고, '화'는 110~200에서 골고루 분포하는 것을 볼 수 있다. 이 때 '즐거움'의 경우에는 기분이 좋아서 톤이 올라가기 때문에 피치가 높아지는 경향이 많은 반면, '화'의 경우는 목소리가 커지면서 톤이 올라가는 경향을 보이는 차이가 있다.

표 5는 위의 3개의 표들에 대한 요약이다. 표를 보면 평서형과 다른 감정들, 우울과 다른 감정들은 분류가 되는데, 즐거움과 화는 포함관계에 놓여 있는 것을 알 수 있다.

Ⅳ. 감성 인식 시스템

본 논문에서 제안하는 인식 시스템은 앞 절에서 설명한

표 5. 감정과 특징 점간의 관계 분포.

Table 5. The relation of emotions and features.

감정 \ 요소	피치평균	Sect.No.	Loudness
평서형	100~140	~4	~1000
즐거움	140~180	3~7	3000~5000
화	110~200	2~5	1000~10000
우울함	~110	6~10	~1000

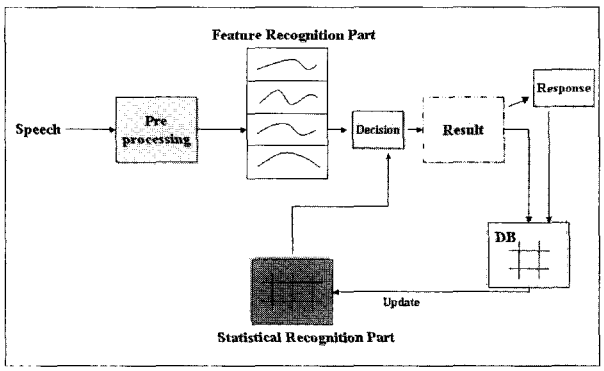


그림 5. 감성 인식 시스템 구조.

Fig. 5. The structure of an emotion recognition system.

감성 요소들에 의한 인식부(Feature Recognition Part)와 통계적 요소에 의한 인식부(Statistical Recognition Part) 두 부분으로 구성되어 있다.

그림 5에서 Feature Recognition Part는 Loudness, Pitch Mean, Sect. No.를 이용하여 감정을 분류한다. Statistical Recognition Part는 감정에 대한 통계적 자료를 이용하여 감정을 분류한다. 인간의 인지 기능들은 센서와 추론의 결합이다. 즉, 눈, 코, 혀, 귀 등의 센서로 여러 정보를 획득할 수 있지만 센서로 입력된 정보가 부족할 때 보완해 주는 것이 뇌의 추론 기능이다. 특히, 감정은 시각, 청각, 촉각, 미각, 후각 같은 인지 기능들의 상위 감각으로써 센서를 통해 입력된 정보를 종합하고 과거의 학습을 이용하여 추론하는 매우 복잡하고 민감한 감각이다. 그러므로 본 논문에서 제안한 Feature Recognition Part와 Statistical Recognition Part는 각각 센서부와 추론부에 대응될 수 있다.

1. 특징 추출 부

Feature Recognition Part는 감정들과 요소의 관계를 이용하여 가장 가능성이 높은 감정을 찾아낸다. Sect No., Loudness, Pitch Mean에 대해 각각의 감정 확률 벡터를 구성하여 4가지 감정의 확률을 계산한다.

$$E_s = \sum_{i}^{N} P(E_s | f_i) \quad (1)$$

E_s : Emotions States

f : Emotional Features

즉, 이 부분은 위의 식과 같이 어떤 감정 상태가 될 특징점들의 확률을 각각의 감정에 대해 계산하는 역할을 한다[9].

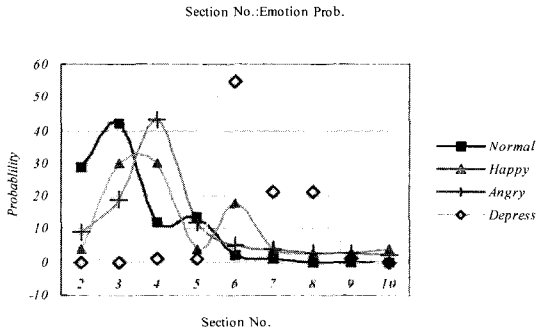


그림 6. Sect. No.에 따른 감정별 빈도수.

Fig. 6. The occurring of each emotion on the Sect. No.

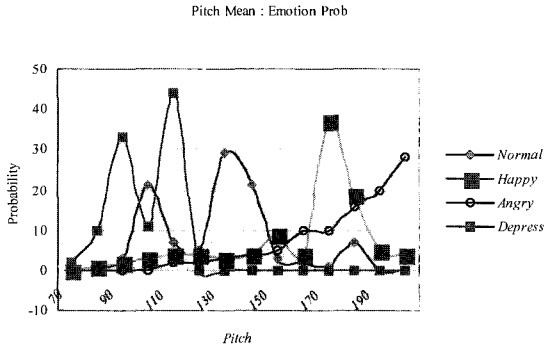


그림 7. 피치평균에 따른 감정별 빈도 수.

Fig. 7. Emotion Probability for Pitch Mean.

다음의 그림들은 각각 Sect. No., Pitch Mean, Loudness에 대해 각 감정이 발생할 확률을 그래프로 나타낸 것이다.

그림 6은 구간 개수에 따라 특정 감정으로 판단 될 확률을 나타내는 것인데 우선 가장 크게 구분되는 것은 우울함과 나머지 3개 감정이고 평서형, 즐거움, 화는 약간의 차이로 각각의 영역을 갖고 있다.

그림 7은 Pitch Mean에 의한 확률을 나타낸다. 이 그림에서는 우울함, 평서형, 즐거움이 각각 고유의 영역을 갖고 있고, 화가 3가지 감정 모두에 부분적으로 속해있는 것을 볼 수 있다.

그림 8은 Loudness에 따른 각 감정별 확률을 나타낸다. 그래프를 관찰해보면, 평서형과 우울함, 즐거움과 화가 각각 거의 유사한 확률을 갖고 있음을 알 수 있다. 그렇기 때문에 Loudness로는 두 부분으로 구분할 수밖에 없다.

2. Statistical Recognition Part

대화를 하다 보면 상대방이 한 얘기를 정확히 듣지 못하는 경우가 있다. 혹은, 누군가 소리치는데 내용을 몰라서 추측을 하는 경우가 있다. 사실, 이런 경우에는 추측을 하기위해 필요한 정보는 무수히 많다. 상대방과 자신의 관계의 정도, 둘 간의 경험들, 상대의 발화 크기, 빠르기, 톤, 바로 전의 대화 내용, 자신의 학습정도, 사회화 정도 등 매우 많은

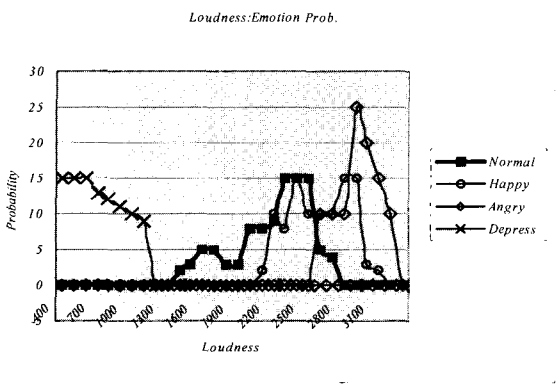


그림 8. Loudness에 따른 감정별 빈도 수.
Fig. 8. The occurring of each emotion on the loudness.

표 6. 감정의 천이 확률(단위:%).
Table 6. The Transition Prob. of Emotions(%).

현재감정 / 이전감정	평서형	즐거움	화	우울함
평서형 (63%)	62.5	12.5	20	5
즐거움 (18%)	33	65	1	1
화 (15%)	59	1	30	10
우울함 (4%)	55	5	7	33

정보들을 아주 빠른 시간동안 뇌에서 처리하고 결론을 내리는 것이다. 기계가 이렇게 많은 정보들을 처리하기 위해서는 우선 화자 인식, 음성 인식, 화상 인식, Semantic/Context Recognition 등의 기능들이 가능해야 한다.

본 논문에서는 일상 대화에서의 감정의 분포 확률과 '분위기'라는 개념을 적용한다. 일상의 대화를 살펴보면, 대부분의 경우 평서형의 대화가 많다. 특히, 이 경우 직장, 가정, 동호회 등 집단의 성격에 따라 감정의 분포가 달라 질 수 있는데, 본 논문에서는 직장, 가정, TV drama를 대상으로 조사를 하였다. 총 100개의 대화에 대해 관찰하였고, 직장에서 40, 가정에서 40, 드라마 20개에 대해서 통계를 내었다. 그래서, 대화에서 감정의 Occurring 확률은 Normal 63%, Happy 18%, Angry 15%, Depress 4%로 평서형이 가장 많이 발생한다는 걸 보여주고, 표의 이전감정과 현재감정에서의 확률을 보면 한 사람이 평서형으로 말하면 다음 말도 평서형일 확률이 높다는 걸 확인 할 수 있고, 그러다가 어떤 원인에 의해 즐거움으로 천이 되면 그 다음의 발화는 즐거움일 확률이 높다는 걸 보여준다.

감정이란 매우 변화가 다양하지만, 적당한 인과관계에 의해서 변화되기 때문에 다소 예측이 가능하다. 즉, 즐거운 상황에서는 즐거운 감정의 대답이나 적어도 평서형의 대답이 나오지 화나 우울한 상태의 발화가 나올 확률은 매우 적고, '화'난 상태였는데 다음 상태가 즐거움일 확률은 적을 것이다. 혹, 그런 적은 확률의 사건이 일어나는 경우는 코메디 같은 특별한 상황에서만 일 것이다. 이러한 예측 가능한

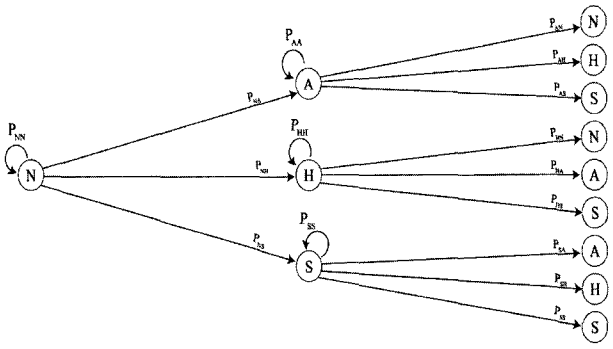


그림 9. 감정상태의 천이.
Fig. 9. The transition of emotional states.

인과관계를 이용하여, 감정 특징점들 만으로의 부족한 인식을 향상 시킬 수 있다.

위의 그림 9는 감정의 상태 천이를 보여준다. 본 시스템의 Statistical Recognition Part는 위 그림과 같은 천이구조를 갖는 표 6과 같은 벡터로 구성 되어있다. 또한, 그림 5의 DB part에서는 과거의 10개까지의 결과와 사용자의 반응을 누적 시켜, 10개마다 Statistical Recognition Part의 벡터를 갱신 시켜준다. 이는 강화학습의 간단한 개념을 이용한 것으로써 사용자로부터 좋은 평가나 벌칙을 받는 것으로부터 다음 벡터를 더욱 현재 사용자에게 적합하도록 학습해간다. 그림 5의 Decision 부분에서는 Feature Recognition Part와 Statistical Recognition Part의 간단한 계산을 통하여 적합한 결과를 결정한다.

$$E_s = W^* P(E_s | E_{s-1}) + (1 - W)^* \sum_{i=1}^{N} P(E_s | f_i) \qquad (2)$$

$$s: state(Neutral, Angry, Happy, Sorrow)$$

$$N: The number of features$$

$$w: Weight(0 \le w \le 1)$$

$$P(E|f) = \frac{P(f|E)P(E)}{P(f)} \qquad (3)$$

(2)에서 우변의 첫 번째 항이 Statistical Recognition Part에서의 결과를 나타내고 (3)은 Feature Recognition Part의 결과를 나타낸다. 그리고 W는 weight를 나타내고 이 값은 시행착오를 통해 구한다.

V. 실험 결과

특징들을 추출했던 음성데이터 녹음 환경과 동일한 환경에서 실험자들로부터 100개의 발화를 입력 받아서 인식 실험을 하였다. 발화된 문장은 학습할 때 사용되었던 문장들을 사용하였고 실험자들은 학습 때와 비슷한 감정 상태로 발화하도록 하였다. 이 때 사용된 테스트 샘플은 학습할 때 사용했던 샘플 10개와 새로운 샘플 90개로 구성되었고 실험자들의 인식률 평균을 다음의 표와 같이 정리 하였다.

다음의 그림 10의 그래프는 다른 학습 방법과 다른 특징점들을 사용하였을 때의 결과를 비교한 것이다.

표 7. 실험 결과.

Table 7. Experiment result.

감정 / 결과	평서형	즐거움	화	우울
평균	79%	70%	80%	75%

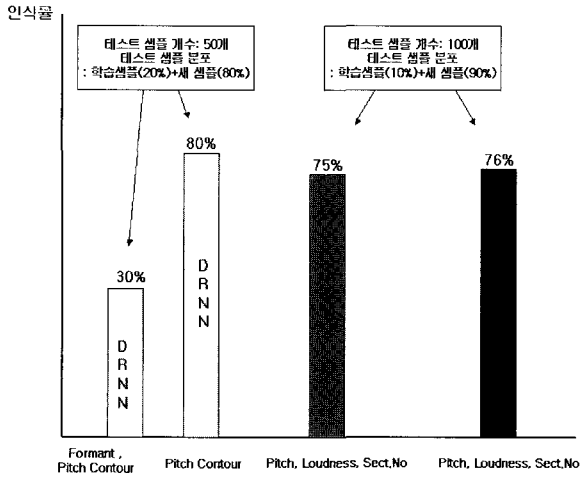


그림 10. 다른 학습 방법을 적용한 결과.

Fig. 10. The comparison with other Results.

그림 10에서 DRNN(Dynamic Recurrent Neural Network)은 시계열적인 특성을 갖는 입력값에 적합한 학습 방법으로써 그래프의 제일 왼쪽 두 개의 막대가 DRNN을 사용했을 때의 결과를 나타낸다. 포만트와 Pitch Contour를 사용했을 때는 약 30%의 결과를 보였고, 단지 Pitch Contour만 사용하였을 때는 80%의 결과를 보였다. 하지만 이 경우에는 테스트 샘플의 개수가 50개밖에 안되고 학습 샘플도 많이 포함되어 있어서 인식률의 신뢰성은 떨어진다. 3번째 막대는 ANN(Artificial Neural Network)을 사용 한 것으로써 본 논문에서 사용했던 Pitch, Loudness, Sect. No.를 사용하여 약 75%의 결과를 보였고 마지막 막대는 본 논문에서의 결과를 보여준다.

VI. 결론

본 논문은 통계적인 방법인 베이지안 학습법을 기반으로 하여 감성 인식을 하였다. 감성인식에 사용될 수 있는 다양한 파라미터들 중에서 가장 감성 정보를 많이 갖고 있는 피치, 크기, 구간 정보의 유효성을 실험을 통해 보였고 적용하였다. 또한 이 파라미터들을 기존의 연구인 신경망 적용 결과와 추론을 모방한 방법을 간단히 비교하였다. 결과적으로 두 방법간에 인식률은 우열을 가리기 힘듦을 알 수 있고 오히려 감성 인식에 대한 주요 문제는 분류 방법에 있지 않고 감성 파라미터 추출에 있다고 생각할 수 있다.

참고문헌

[1] R. S. Lazarus and B. N. Lazarus, *Passion & Reason,* Seoul, Moonye Publishing, pp. 255-256, 1997.

[2] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato and R. Nakatsu, "Emotion recognition from audiovisual information", *IEEE Second Workshop on Multimedia Signal Processing,* 1998.

[3] J. Nicholson, K. Takahashi and R. Nakatsu, "Emotion recognition in speech using neural networks," *Proc. of ICONIP '99,* vol. 2, 1999.

[4] F. Dellaert, T. Polzin and A. Waibel, "Recognizing Emotion In Speech," *Proc. of ICSLP 96 Proceedings,* vol. 3, pp. 1970 - 1973, 1996.

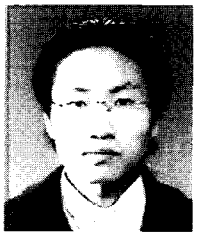
[5] S. Batliner, K. Fisher, R. Huber, J. Spilker and E. Noth, "Desperately seeking Emotions: Actors, Wizards and Human Beings", *Proc. of the ISCA Workshop on Speech and Emotion.*

[6] T. Moriyama and S. Ozawa, "Emotion recognition and synthesis system on speech", *1999 IEEE International Conf. on Multimedia Computing and Systems,* vol. 1, 1999.

[7] J. S. Han, *Speech Signal Processing,* Seoul, O-Sung-media, pp. 84-85, 2000.

[8] B. C. J. Moore, An Introduction to the psychology of hearing, Academic Press, USA, pp. 195, 2003.

[9] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, A Wiley-Interscience Publication, USA, pp. 20-26, 2001.



박 창 현

2001년 중앙대학교 전자전기공학부(공학사). 2003년 중앙대학교 전자전기공학부(공학석사). 2003년~현재 중앙대학교 대학원 전자전기공학부 박사과정 재학중. 관심분야 : 감성인식, 인공생명.

심 귀 보

1984년 중앙대학교 전자공학과(공학사). 1986년 동대학원 전자공학과(공학석사). 1990년 The University of Tokyo 전자공학과(공학박사). 1991년~현재 중앙대학교 전자전기공학부 교수. 2000년~현재 제어 · 자동화 · 시스템공학회 이사 및 지능시스템연구회 회장. 2002년~현재 중앙대학교 산학연컨소시엄센터 센터장. 2003년~현재 일본계측자동제어학회(SICE) 이사. 2003년~현재 한국퍼지 및 지능시스템학회 부회장. 관심분야 : 인공생명, 지능로봇, 지능시스템, 다개체시스템, 학습 및 적응알고리즘, 소프트 컴퓨팅(신경망, 퍼지, 진화연산), 인공면역시스템, 침입탐지시스템, 진화하드웨어, 인공두뇌, 지능형 홈 및 홈네트워킹, 유비쿼터스 컴퓨팅 등.