추론 능력에 기반한 음성으로부터의 감성 인식

Inference Ability Based Emotion Recognition From Speech

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

Abstract - Recently, we are getting to interest in a user friendly machine. The emotion is one of most important conditions to be familiar with people. The machine uses sound or image to express or recognize the emotion. 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

Key Words : Inference, Emotion Recognition, Pitch, Bayesian Learning

1. Introduction

인간은 다분히 감정적인 동물이다. 말을 하지 못하는 갓난아기는 자신의 감정을 표현하는 것으로 타인과의 의사소통을 시작한다. 감정은 제3의 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 파악한 바에 의하면 감정들은 안전한 상황과 위험한 상황, 기회를 엿볼 수 있는 상황들을 재빨리 구분할 것을 다급하게 요구하며, 긴급 상황을 대처하는 데 필요한 추가의 에너지와 스태미너를 동원하게 하고 위험한 상황에 대한 공포를 느끼는 것으로 생존의 위협에서 벗어날 수 있게도 한다[1]. 즉, 감정은 업무의 효율화와 생존을 위한 주요 기능인 것이다. 이러한 이유 때문에 감성인식의 중요성은 커지고 있다. 과거의 연구자들은 음성의 신호 자체로부터 감성을 인식하려고 노력하였다. 즉, 음성 신호의 주요 분석 요소인 스펙트럼 과 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, Chen and Tao 는 각 문장별로 Pitch 와 RMS energy envelope을 구하여 Happiness, Sadness, Anger, Dislike, Surprise, Fear 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 인식 율을 보여주었다. 하지만, 실험 방법이 명확하지 않아 인식 율 자체가 의미가 없고 단지, Vision이 보완 해줄 때 감성인식의 정확도가 높아진다는 걸 확인 할 수 있다. J.Nicholson은 의식적인 감정표현과 무의식적인 감정표현으로 개념을 나누어 인식하기에 더 쉬운 의식적인 감정표현에 국한하여 연구를 진행하였다. 본 논문은 발화로부터 감성적인 요소들과 통계적 자료를 바탕으로 추론하여 사용자의 감성을 인식할 수 있다는 것을 보여준다.

저자 소개
* 박 창 현 : 중앙대학 전자전기공학부
** 심 귀 보 : 중앙대학 전자전기공학부

1.1 Features

본 논문에서 사용한 특징 점들은 발화된 문장에서의 피치의 평균, 소리의 크기, Section No이다. 피치의 평균은 발화된 문장 내에서 Section 별 피치들의 평균을 구한 것이다. 소리의 크기는 입력된 음절들의 최대값들의 평균으로 구하였다. 특히 본 논문에서는 이것을 편의상 Loudness라고 부른다. Section No는 발화 문장의 에너지를 기반으로 기준치를 넘는 부분에서 Section을 나누었을 때의 개수를 구한 것이다.

표 1. 특징 점과 감성과의 관계

Components / emotions	Pitch Mean	Sect.No	Loud.
Neutral	100~140	~4	~1000
Happy	140~180	3~7	3000~5000
Angry	110~200	2~5	1000~10000
Depress	~110	6~10	~1000

표 1. 은 2명의 연기자에게 4가지 감정으로 7개의 문장을 각각 발화토록 하여 총 56개의 음성에 대해 분석 한 것이다. 특히, 4개의 감정은 범위가 큰 분류로써 각각의 분류에 대해 너무 많은 표현 방법이 있을 수 있기 때문에 분석의 간결성을 위해서 각 감정별 표현 방법을 각각 한가지로 제한 하였다. 또한, 문장의 길이를 10개 음절 내외로 정하여 Sec.No의 추출에 문장의 길이로 인한 의존성을 제거하였고 Loudness는 파형의 크기를 Scaling 했기 때문에 특정 단위를 사용하지 않았다. 위의 표를 분석하면, 흥분한 감성과 차분한 감성이 피치의 평균으로부터 구분 될 수 있는 것을 알 수 있다. 즉, Neutral, Depress는 피치평균이 140Hz 이하에 존재하고 Happy는 140~180Hz, Angry는 매우 넓은 대역을 차지하고 있고 높은 주파수 영역까지 분포하는 것을 알 수 있다. Sec.No.는 말의 빠르기와 불안정한 발화에 영향을 받는 요소이다. 즉, 말을 빨리 하는 경우에는 연음이 생기면서 Sect.No.가 작아지고 또박또박 천천히 말하거나, 흐느끼는 경우처럼 음절이 자주 끊어지는 경우에는 커지는 경향을 보인다. Loudness는 Angry인 경우 매우 커지고 폭도 큰 특징을 보이고 있고, Neutral, Depress는 1000이하인 값을 보인다. 이와 같은 분석들은 특정 패턴을 갖는 감성들이 정량적인 몇 가지 특징점으로 분류가 가능하다는 것을 보여준다.

2. Recognition System

본 논문에서 제안하는 인식 시스템은 앞 절에서 설명한 Emotional Components 들에 의한 인식부 (Feature Recognition Part) 와 Statistical Recognition Part 두 부분으로 구성되어 있다. Feature Recognition Part는 Loudness, Pitch Mean, Section Number로 Emotion을 분류한다. Statistical Recognition Part는 감정에 대한 통계적 자료를 이용하여 Emotion을 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.

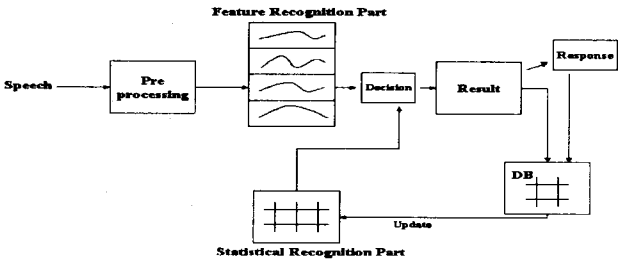

Fig 1. 감성 인식 시스템 구조

2.1 Feature Recognition Part

Feature Recognition Part는 3.3의 Emotions 과 Components의 관계를 이용하여 가장 가능성이 높은 감정을 찾아낸다. Section No., Loudness, Pitch Mean에 대해 각각의 감정 확률 벡터를 구성하여 4가지 감정의 확률을 계산한다.

$$E_s = \sum_{i=1}^{N} P(E_s \mid f_i) \qquad \text{(Eq.1)}$$

E_s : *Emotions States*

f : *Emotional Features*

즉, 이 부분은 위의 식과 같이 각 features에 대한 어떤 감정 상태가 될 확률들을 각각의 감정에 대해 계산하는 역할을 한다.다음의 그림들은 각각 Section No., Pitch mean, Loudness에 대해 각 감정이 발생할 확률을 Graph로 나타낸 것이다.

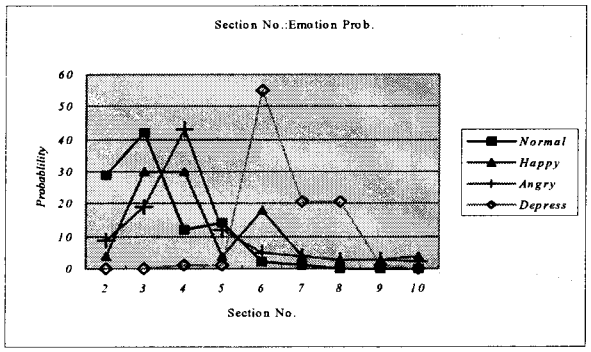

Fig.2 Sect. No. 와 감성과의 관계 분포

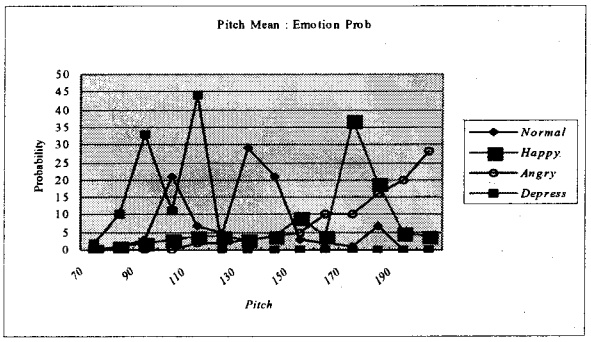

Fig.3 Pitch 와 감성과의 관계 분포

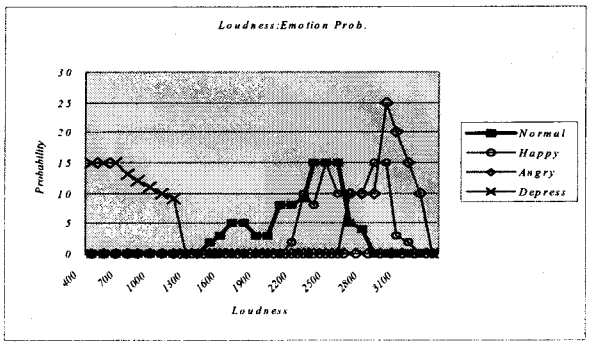

Fig.3 Loud. 와 감성과의 관계 분포

Fig 2. 은 Section 개수에 따라 특정 감정으로 판단 될 확률을 나타내는 것인데 이는 발화의 분절음 개수를 표현한 것으로써 안정성을 나타내는 것이다. 즉, 말이 떨리고 안정감이 없게 되면 이 수치가 높게 나타나고, 안정감이 있고 발화에 변화가 적으면 이 수치가 낮게 나타난다. Fig 3. 는 Pitch Mean에 의한 확률을 나타낸다. 직관적으로 알 수 있는 바와 같이 화가 났을 경우나 기뻐서 내는 소리인 경우 피치가 높은 것을 그림으로부터 확인 할 수 있다. Fig 4.는 인간의 가장 근원적인 감정 표현 형태인 소리의 크기에 대한 경우로써 화가 난 경우와 기쁜 경우의 차례로 소리의 크기가 크고 다음으로 평서형 그리고 우울한 경우의 순서이다.

2.2 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, 둘간의 경험들, 상대의 발화 크기, 빠르기, 톤, 바로 전의 대화 내용, 자신의 학습정도, 사회화 정도 등 매우 많은 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 가능 해야 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 40개의 대화에 대해 관찰하였고, 직장에서 20, 가정에서 10, 드라마 10개에 대해서 통계를 내었다. 그래서, 대화에서 감정의 Occurring 확률은 Normal 63%, Happy 18%, Angry 15%, Depress 4% 로 평서형이 가장 많이 발생한다는 걸 보여주고, Table 의 Prior state 와 Present state에서의 확률을 보면 한 사람이 평서형으로 말하면 다음 말도 평서형일 확률이 높다는 걸 확인 할 수 있고, 그러다가 어떤 원인에 의해 Happy로 천이 되면 그 다음의 발화는 Happy일 확률이 높다는 걸 보여준다.

표2. 감정 천이 분포

Prior State \ Present State	Neutral	Happy	Angry	Depress
Neutral	62.5%	12.5%	20%	5%
Happy	33%	65%	1%	1%
Angry	59%	1%	30%	10%
Depress	55%	5%	7%	33%

감정이란 매우 변화가 다양하지만, 적당한 인과관계에 의해서 변화되기 때문에 다소 예측이 가능하다. 즉, happy 한 상황에서는 happy 한 대답이나 적어도 평서형의 대답이 나오지 Angry나 depress 상태의 speech 가 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 인과관계를 이용하여, 감정 Features만으로의 부족한 인식을 향상 시킬 수 있다.

Fig 1. 의 Decision 부분에서는 Feature Recognition Part와 Statistical Recognition Part의 간단한 계산을 통하여 적합한 결과를 결정한다.

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

(Eq.2)

s: state(Neutral, Angry, Happy, Sorrow), N: The number of features, W: Weight (0=W=1)

$$P(E \mid f) = \frac{P(f \mid E)P(E)}{P(f)}$$: Bayesian rule

(Eq.3)

즉, (Eq.1) 우변의 첫번째 식이 Statistical Recognition Part에서의 결과를 나타내고 두 번째 식이 Feature Recognition Part의 결과를 나타내는 것이다. 그리고 W는 weight를 나타내고 이 값은 시행착오를 통해 구한다

3. 결 과

학습 한 음성 포함하여 200개의 샘플에 대해 실험은 75%의 결과가 나왔다. 이 결과는 기존 연구의 결과와 유사하다. 하지만 특징 점의 종류를 더욱 다양화 하면 더 나은 결과가 나올 수 있고 더욱 많은 감정의 분류에도 성공 할 수 있을 것이다.

표3. 실험 결과

Actors \ Emotions	Neutral	Happy	Angry	Depress	Average
K	57%	40%	80%	70%	62%
J	100%	73%	80%	100%	88%
Average	79%	56%	80%	85%	75%

참 고 문 헌

[1] R.S.Lazarus and B.N.Lazarus, Passion & Reason, Seoul, Moonye Publishing, pp.255-256, 19

[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, 19998