운율 특성 벡터와 가우시안 혼합 모델을 이용한 감정인식

Emotion Recognition using Prosodic Feature Vector and Gaussian Mixture Model

곽현석* · 김수현** · 곽윤근**

Hyun Suk Kwak, Soo Hyun Kim and Yoon Keun Kwak

Key Words : Emotion Recognition(감정인식), Pitch(피치), Energy(에너지), Characteristic Vector(특징벡터), HMM(은닉 마코프 모델).

ABSTRACT

This paper describes the emotion recognition algorithm using HMM(Hidden Markov Model) method. The relation between the mechanic system and the human has just been unilateral so far. This is the why people don't want to get familiar with multi-service robots of today. If the function of the emotion recognition is granted to the robot system, the concept of the mechanic part will be changed a lot. Pitch and Energy extracted from the human speech are good and important factors to classify the each emotion (neutral happy, sad and angry etc.), which are called prosodic features. HMM is the powerful and effective theory among several methods to construct the statistical model with characteristic vector which is made up with the mixture of prosodic features

1. 서 론

1980 년대의 대량 생산을 가능하게 한 원동력은 인간을 대신하여 일하는 생산 로봇이다. 이는 기계 산업이 새로운 경제를 이끌어 가는 중심으로 대두된 원동력이었을 뿐만 아니라, 로봇이 인간에게 실용적 의미뿐만 아니라 언젠가는 생활 속에 같이 존재할 것이라는 가능성과 기대감이 항상 존재해왔다. 그러나, 디자이너의 기본 철학이 아직 기계에서 벗어나지 못하였기 때문에, 대중적으로 필요성을 느끼지 못할 정도로 존재가 미미하였다.

이후 국내에서는 1991 년 G7 과제에서 감성 공학 관련 연구가 선정이 된 이후 인간 친화적인 접근 방법이 시도되고 있으며, 1998 년도에는 인간과 컴퓨터간의 상호작용(HCI: Human Computer Interaction)을 연구하여 최초로 상품화시킨 소니사의 아이보(AIBO)라는 유아, 성인을 위한 애완용 로봇이 등장하였다.

사람은 제공 받은 서비스나 외부 자극에 대해서 자신도 모르게 표정, 심장 박동 수, 혈압, 체온, 몸짓 그리고 목소리를 통해서 감정을 표출하기 때문에 이를 피드백 신호로 받아서, 다음 동작에 활용하여 보다 편리하고 질 좋은 서비스를 하도록 하는 것이 감정 시스템의 개념이다. 특히 음성은 공기의 매질을 통해 전달되기 때문에 화자의 위치와 거리에 자유로운 장점이 있다.

인간의 음성 신호를 이용하여 감정의 정보를 인식하기 위해서는 음운론적인 특징(phonetic feature)이 아닌 성대의 떨림 정도를 나타내는 피치(pitch)와 말하는 세기를 나타내는 에너지(energy)와 같은 운율적인 특징(prosodic feature)[1]을 정확하고 효율적으로 획득하여 화자 독립적(speaker – independent)인 특성을 나타내도록 조합해야만 한다.

Frank Dellaert[2]는 피치를 획득하여 다양한 패턴 매칭 방법을 이용하여 감정 인식률을 비교 하였으나, 단순히 피치만을 이용한 인식률은 그 결과가 저조하였다. Naoko Tosa[3]와 강봉석[4]은 음성 신호의 운율적인 특징뿐만 아니라 음운론적인 특징을 사용하였으나 훈련을 해야 할 DB 의 양이 많아지고, 그만큼 계산 시간이 많이 소요된다.

따라서 본 연구는 감정에 따른 물리적인 특징을 표현하기 위해서 운율적인 특징인 피치와 에너지 그리고 템포를 사용하였으며, 거리, 화자, 문장에 독립적인 특징을 가지도록 이들의 1 차 미분, 2 차 미분, 변화폭 및 정규화(normalizing)를 시켜 2 종류의 특징 벡터를 구성하였다. 단 발음적인 요소는 배제하였다. 패턴 매칭 방법은 음성 인식에서 가장 널리 사용하는 HMM(Hidden Markov Model)방법을 사용하였으며, 통계학적인 모델링을 할 때 사용되는 가우시안 함수의 개수를 변화시키며, 인식률을 비교하고자 한다.[5,6,7]

* 한국과학기술원 대학원
E-mail : beholder@kaist.ac.kr
Tel : (042) 869-3252, Fax : (042) 869-5201

** 한국과학기술원 기계공학과

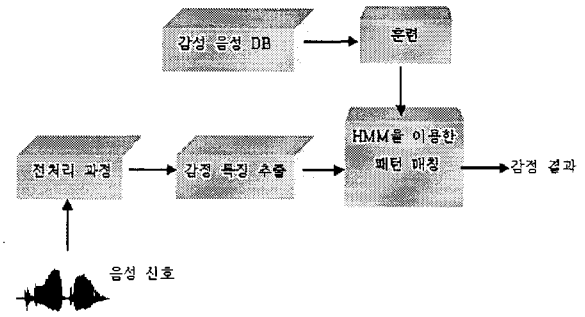

Fig. 1 감정 인식 알고리즘

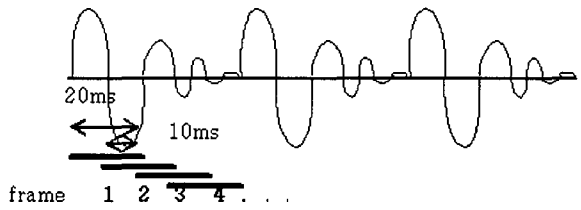

Fig. 2 음성 신호의 프레임 분석

Fig. 1 은 감정 인식을 하기 위한 전체 알고리즘을 나타낸 것이다.

2. 특징 추출 및 특징 벡터 구성

2.1 특징 추출

감정이란 주변의 특별한 자극에 의한 반응으로 일어나며, 정적인 상태의 상황이 아니라 급하게 연속되는 외부 자극으로부터 발생되는 결과의 과정이다.[1,8] 따라서 음성을 통한 감정 인식을 위해서는 각각의 감정이 음성의 운율적인 요소에 어떠한 변화를 만들어내는가 정확히 규명을 하여야 한다.

(1) 피치(Pitch) 추출

피치는 인간의 청각에 매우 민감하게 반응하는 인자로써, 허파에서 압축된 공기가 성대(vocal folds)에 진동을 일으켜 성문을 통과하는 파형으로 생기는 펄스의 주기이며, 기본 주파수(F0)라고 한다.[9,10]

사용한 음성 신호는 16kHz, 16bits 이며, Fig. 2 의 그림과 같이 데이터를 20ms(=320 samples)씩 구분하고, 10ms(=160 samples)을 중첩하여 음성 신호의 프레임을 설정하였다.

이와 같이 한 개의 감정이 담긴 문장을 로 LPC(Linear Prediction Coding)필터를 이용하여 연속적으로 고속 계산을 하기 위해 사용한 SIFT(Simple Inverse Filtering Tracking) 알고리즘[11,12]을 도시

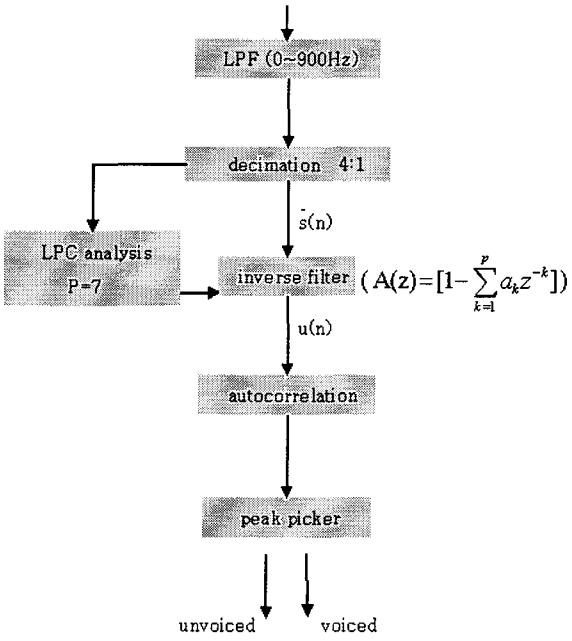

Fig. 3 SIFT 알고리즘

화하면 Fig.3 과 같다.

음성신호 s(n)을 피치가 존재하는 0~900Hz 의 저역통과필터(LPF)를 거치게 한 후, 이로 인해 줄어든 대역폭을 이용하여 4:1 로 간축(decimation)을 시켜 계산 량을 줄인다. 그리고 LPC 분석을 통해 p 차의 계수 a 값을 획득한 후 이를 이용해 만든 필터 (1)를 통과 시켜 음원의 에러신호를 획득한다.

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \qquad (1)$$

에러 신호가 성대에서 발생되는 주기적인 신호와 일치하기 때문에 정확한 값을 얻기 위하여 자기상관함수(autocorrelation)를 통해 피치를 구한다.

(2) 에너지(Energy) 추출

에너지[11]의 특성 값도 음성 신호가 단구간(10ms)인 경우에는 신호의 주기 등이 바뀌지 않는 안정적인 구간인 시불변시스템(time-invariant system)이라 가정한 후 10ms 구간을 한 프레임으로 잡고 중첩없이 구한다. 따라서 음성 신호 에너지는 (2)와 같이 정의된다.

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]w(n-m) \qquad (2)$$

$T[] = []^2$ 으로 입력 신호 x(m)의 구간별 에너지의 제곱의 합으로 정의하였으며, 시간 인덱스 n 에 따라 사각창(rectangular window)을 이용하여 획

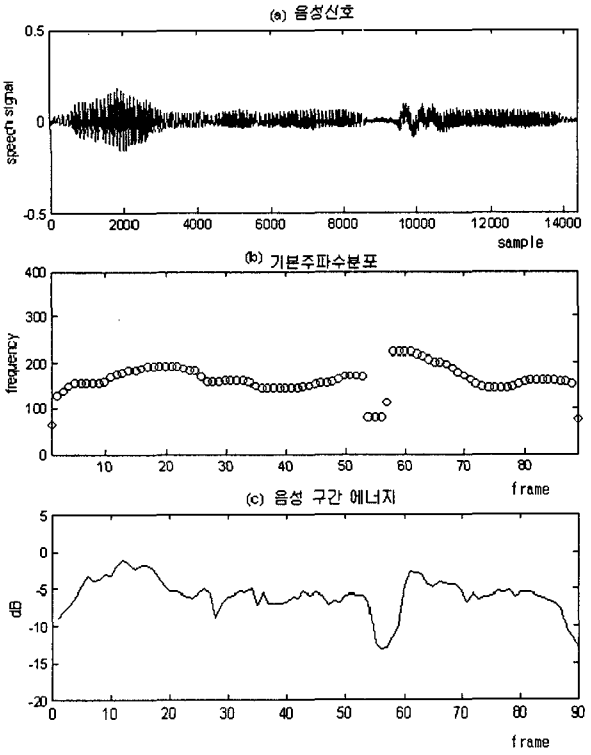

Fig.4 '마음대로 하세요' - 일반 감정 상태

득하였다.

'마음대로 하세요'라고 임의의 화자가 발성한 문장을 평서의 감정 상태에 대하여 피치와 에너지를 구하면 Fig.4 와 같다.

Fig.4 와 같이 평서, 행복함, 슬픔, 화남 총 4 개의 감정상태에 따라서 피치와 에너지의 분포를 구할 수 있다. 이 운율적인 특징은 감정에 따라 구별되는 특징을 보이며 평균 값과 변화율, 변화폭 등을 고려하여 각 프레임 별 특징 벡터를 만들 수가 있다.

2.2 특징 벡터 구성

운율적인 특성 값을 수식적으로 분석하여, 확률적인 모델을 만들기 위한 기반 뼈대를 구성하기 위해 정량적으로 분석해보았다. 우선 두 종류의 특성 벡터로 구성해 보았는데 기본적인 벡터 구성으로는 (3)과 같이 피치와 에너지의 1 차, 2 차 미분 값을 구하였다. 이러한 성분을 구성하는 이유는 화자간의 발성 차이를 보상하여 화자 독립성을 보장할 수 있다는 것이다. 또한 말의 빠르기(tempo)를 단위 시간당 변하는 음소의 개수로 정의하여 사용할 수 있으나 음소적인 DB 를 필요로 하는 제약이 생기기 때문에 단위시간당 변하는 피치와 에너지의 값을 이용한다면 말의 빠르기에 대한 정도를 표현 할 수 있다.

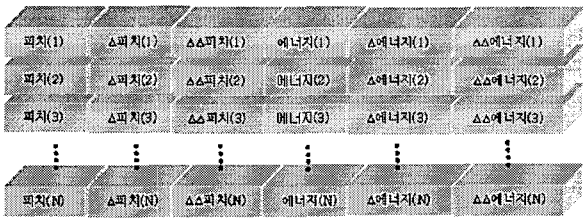

Fig. 5 6 차 특징 벡터

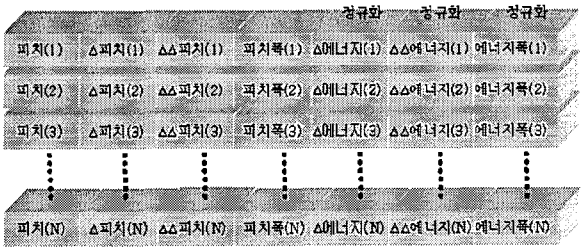

Fig. 6 7 차 정규화 특징 벡터

$$\Delta[]=[](j+2)-[](j) \qquad 0 \le j \le N-2 \qquad (3)$$

$\Delta-\Delta[]=\Delta[](j+2)-\Delta[](j)$ []: 피치 or 에너지

여기서 j 는 샘플의 인덱스이며, N 은 전체 프레임의 개수이다.

Fig. 6 의 7 차 정규화(normalizing) 특징 벡터는 피치와 에너지의 변화폭을 나타내고 발성화자의 거리에 따른 의존성을 없애기 위해서 에너지 성분은 제거하고 에너지의 1 차 미분과 2 차 미분에 에너지의 평균값을 나누어 주었으며, 각 프레임의 피치와 에너지는 평균값으로 빼주어 변화폭의 특성을 나타내었다.

3. HMM 을 이용한 통계적 모델링

3.1 감정 음성 DB

음성 DB 는 표준 연구원의 지원 아래 연세 대학교 전자 공학과 음향 연구실에서 이충용 교수 외 2 명의 연구원이 2000 년 12 월에 감정 시스템 및 테스트를 위해 제작한 데이터를 사용하였다 대상 감정은 인간의 주요 감정인 평상, 기쁨, 슬픔, 화남의 4 가지이며, 남녀 각각 5 명이 감정이 담긴 45 개의 문장을 세 번씩 발성하여 각각의 감정과 성별에 따라 675 개의 훈련 데이터를 획득할 수 있었다. 본 실험과 훈련에 쓰기 위해 제공된 파일을 PCM(Pulse Code Modulation) 필터를 사용하여 16kHz, 16bits 로 변환하였다.

DB 제작 기준은 다음의 4 가지와 같으며, 녹음한 문장의 일부를 Fig.7.에 나타내었다.

1) 3 가지 감정(기쁨, 슬픔, 화남) 상태로 발음하기에 용이한 문장.

번호	문 장	번호	문 장
1.	좋아요. 함께 가시죠.	24.	다음에 다시 얘기하자.
2.	알겠어요.	25.	아직 전화 안 왔니?
3.	모두 팔았어요.	26.	잠깐 기다려 주시겠어요?
	.....		.....
20.	마음대로 하세요.	43.	기다렸다가 같이 가죠.
21.	이리 오세요.	44.	햇볕이 쨍쨍 비친다.
22.	여보세요.	45.	있다가 다시 걸어 주시겠어요?

Fig.7 녹음 문장

2) 자연스런 감정 표현이 담긴 대화체 문장.

3) 전체적으로 우리말의 모든 음소를 고루 포함하도록 구성

4) 법, 높힘형 등 다양한 어법을 고려하도록 구성

녹음 문장을 일반인 30 명을 대상으로 실험한 주관적인 인식률은 78.2%였다.

3.2 모델 및 훈련

감정의 발성은 공통된 특성 이외에도 사람에 따라 조금씩 그 특징 값의 편차가 있기 때문에 물리적으로 하나의 모델링이 불가능하다. 따라서 음성 인식이나 영상 인식과 같은 패턴 인식 분야에서 확률적으로 모델링을 하기 위해 사용 되는 방법으로 MLB(Maximum-likelihood Bayer), KNN(K-Nearest Neighbor), KR(Kernel Regression), DTW(Dynamic Time Warping), 신경회로망(Neural Net), HMM (Hidden Markov Model)이 쓰이는데 그 중에서도 음성 인식 분야에서 가장 각광적으로 사용되고 있는 HMM 을 이용한다.

HMM[6,7]은 실제적인 관측을 통해서 변화되는 통계적인 특징을 확률적으로 모델링 하기 위하여 마코프 과정을 이용한다. 각 상태 열은 은닉되어 있고, 다른 관측 가능한 확률적인 과정들의 집합을 통하여 관측될 수 있다. 이 모델의 각 은닉 상태와 천이 상태는 이산 또는 연속 확률 밀도로 표현되는 출력 확률과 그 집합에 관련되어 있다.

HMM 은 $\lambda=(A,B,\pi)$로 간략하게 표시된다.

$$\begin{aligned}&\lambda=(A,B,\pi)\\&A=\{a_{ij}\mid a_{ij}=\Pr(s_{t+1}=j\mid s_t=i)\}\\&B=\{b_{ij}(O_t)\mid b_j(O_t)=\Pr(O_t\mid s_t=j)\}\\&\pi=\{\pi_i\mid \pi_i=\Pr(s_1=i)\}\end{aligned}\qquad(4)$$

여기에서 A 는 a_{ij}가 상태 i 에서 상태 j 로의 천이 확률을 나타내는 상태 천이 확률 분포를 나타내고 B 는 각각의 상태에 대응하는 이산 확률 분포를 나타내며, π는 초기 상태 분포를 나타낸다. 음성 인식에서 쓰이는 HMM 을 감정 인식에서 사용하기 위해서 다음과 같은 가정을 하였다. 4 개의 감정 즉, 평상, 행복, 슬픔, 화남의 감정 상태를 단어로 취급하고 문장에서 추출된 감정 특징 벡터를 각각의 상태 벡터로 취급을 하는 것이다. 즉 4 개의 단어를 인식하는 음성 인식 시스템이라고 생각할 수 있다. Fig. 8 과 같이 한 문장에서 천이할 수 있는 개수는 총 5 개로 정하였고 Left-Right HMM 구조를 택하였다.

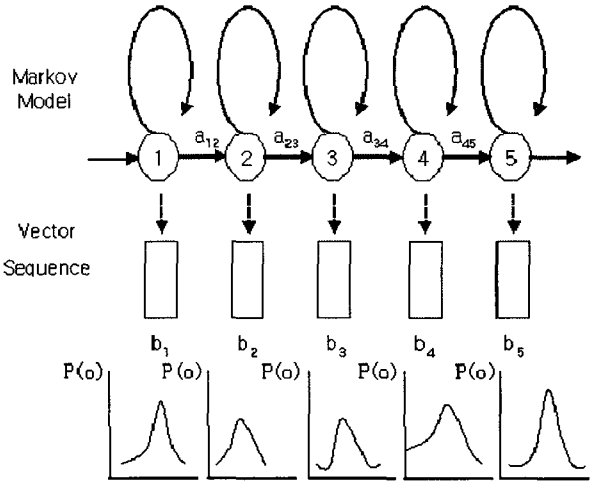

Fig.8 5 개의 상태열을 가지는 Left-right HMM 구조

특정 감정에 따른 운율 특성의 분포는 항상 균일한 패턴이 아닌 잡음이나 에러 또는 발음상의 특징 등으로 인하여 1 개의 가우시안 함수로는 정확하게 모델링 하기 힘들기 때문에 상태 열에 따라 k 개의 공간에서 각각 독립적으로 모델링 된 하위 가우시안 함수를 이용하여 조합한다.[5] 상태 1 에서 k 개의 공간 $\Omega_1,\Omega_2,,,\Omega_k$ 에서 관측 열 O 가 존재할 확률은 (5)와 같이 표현된다.

$$B(o)=\sum_{k\in\Omega_k}c_kN_k(x)=c_1N_1(x)+c_2N_2(x)+...+c_kN_k(x)\qquad(5)$$

여기에서 c 는 가중치를 의미하며, k 의 개수를 1, 2, 4, 8, 16 개로 늘여가면서 남, 녀 각각 감정이 담긴 2700 개의 특징 벡터를 훈련시켰다.

4. 결 론

음성 신호로부터 감정에 따라 정성적으로 구분될 수 있도록 운율적(prosodic)인 특징을 얻어냈으며, 이로부터 패턴 매칭을 효율적으로 할 수 있도록 6 차 및 7 차의 두 종류의 특징 벡터를 조합하였다. HMM 으로 훈련된 모델을 검증하기 위해서 남, 녀가 각각 하나의 감정당 675 개씩 발성한 문장으로 총 2700 개의 문장에 대해서 정확하게 인식된 문장의 개수를 인식률로 정의하고 Fig.

6차 특징 벡터			7차 정규화 특징 벡터		
함수개수	여 자	남 자	함수개수	여 자	남 자
1 개	56.67%	55.21%	1 개	46.15%	48.43%
2 개	63.11%	58.98%	2 개	63.19%	61.80%
4 개	67.41%	63.73%	4 개	67.89%	66.99%
8 개	68.96%	64.73%	8 개	69.22%	68.73%
16 개	51.98%	57.69%	16 개	error	65.14%

Fig. 9 가우시안 함수의 개수에 따른 감정 인식률

감정	6차 특징 벡터		7차 정규화 특징 벡터	
	여 자	남 자	여 자	남 자
Neutral	74.22%	89.48%	86.96%	80.89%
Happy	77.19%	57.48%	64.30%	69.19%
Sad	68.15%	41.19%	73.19%	71.85%
Angry	56.30%	70.77%	52.44%	52.97%

Fig. 10 감정에 따른 인식률

9 와 같은 결과를 얻었다.

모델링을 하기 위해 사용된 가우시안 함수의 개수가 증가함에 따라 인식률이 증가함을 알 수 있고 8 개로 모델링을 하였을 때가 인식률이 최고이며 16 개로 모델링을 하면 인식률 및 계산속도 모두 떨어짐을 알 수 있다. 오히려 데이터가 요구하는 이상으로 다수의 개수를 이용하여 모델링을 하면 정확성이 떨어질 뿐만 아니라 에러가 났다. 7 차 정규화 특징 벡터로 특징 벡터를 만든 것이 6 차 특징 벡터보다 인식률이 좋으며, 여자가 발성하였을 때 69.22%(1869/2700)로 나타났으며, 주관적 인식률 78.2%보다 약 10% 적게 나타났다. 남자가 여자보다 인식률이 떨어지는 것은 피치와 에너지의 변화율이 상대적으로 작은, 즉 감정의 표현이 소극적이라고 생각할 수가 있다.

다음은 감정에 따른 인식률의 변화를 Fig. 10 에 나타내었다. 이 때 HMM 으로 훈련을 하고 인식 실험을 할 때, 모델 함수의 개수를 8 개 사용하여 가장 높게 인식률이 나온 결과를 도시하였다. Fig. 10 의 결과를 보면, 4 가지의 감정에 따라 인식률이 큰 차이가 남을 알 수 있다. 6 차 특징 벡터로 패턴 매칭을 하였을 때는 평상의 감정에서 높은 인식률을 보여 주었으며, 7 차 정규화 특징 벡터로 패턴 매칭을 하였을 때는 평상과 슬픔의 감정 상태에서 높은 인식률을 보여주었다. 전자의 방법에서는 슬픔의 감정을 평상의 감정으로 잘못 인식하는 경우가 많았으며, 두 방법 모두 화났을 때의 감정을 행복함의 감정으로 잘못 인식하는 경우가 많았다. 즉 행복함과 화남의 감정을 정확하게 구분 해줄 특성 값이 필요하다.

따라서 지금의 소 용량의 데이터를 이용하여 음성 신호로부터 정확한 감정 인식을 하기 위해서는 감정간의 구별을 위한 더욱 정밀한 특성 벡터를 만들어야 한다. 화남(angry)과 행복함(happy)의 감정을 정확히 구분하는 것 이상으로 놀람(surprise)과 공포(fear)와 같은 더욱 구분하기가 모호한 감정까지 구분할 수 있도록 해야 한다.

후 기

감정 음성 DB 를 제공해준 연세대학교 전자공학과 음향 연구실 이충용 교수님 외 연구원 2 명에게 다시 한 번 감사를 드립니다.

참고문헌

(1) Iain R. Murray and John L. Arnott, 1993, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", Journal of Acoustical Society of America, V.03, No.2, p.1097~1108.

(2) Frank Dellaert, 1996, "Recognizing Emotional in Speech", ICSLP Proc. 4th Intn'l Conf., V.3, p.1970-1973.

(3) Naoko Tosa, 1996, "Life-Like Communication Agent – Emotion Sensing Character 'MIC' and Feeling Session Character 'MISE'", Proc. Of the 3rd IEEE Intn'l Conf., p.12~19

(4) 강봉석, 2000, "음성 신호를 이용한 문장독립 감정 인식 시스템", 연세대학교.

(5) Keiichi Tokuda, 1999, "Hidden Markov Model based on Multi-Space Probability Distribution for Pitch Pattern Modeling", Acoustics Speech and Signal Processing, Proc. IEEE Conf., V.2, p798~803.

(6) L. Rabiner, 1993, "Fundamentals of Speech Recognition", Prentice Hall International Inc.

(7) Steve Young, 2001, "The HTK Book (for HTK Version 3.1)", Cambridge University Engineering Department.

(8) Oatley, 1989, "The Importance of Being Emotional", New Sci. 123(Pt.1678),p33~36.

(9) 한진수, 2000, "음성 신호 처리", 오성미디어.

(10) Xueddng Huang, 2001, "Spoken Language Processing", Prentice Hall PTR.

(11) L. Rabiner, 1978, "Digital Processing of Speech Signal", Prentice Hall.

(12) John D. Markel, 1967, "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Transactions in Audio and Electroacoustics, V.AU-20, No.5, p.367~377.