FFT와 MFB Spectral Entropy를 이용한 GMM 기반의 감정인식

이우석, 노용완, 홍광석
성균관대학교 정보통신공학부

Speech Emotion Recognition Based on GMM Using FFT and MFB Spectral Entropy

Woo-Seok Lee, Yong-Wan Roh and Kwang-Seok Hong
School of Information and Communication Engineering, Sungkyunkwan University

Abstract - This paper proposes a Gaussian Mixture Model (GMM)-based speech emotion recognition methods using four feature parameters; 1) Fast Fourier Transform(FFT) spectral entropy, 2) delta FFT spectral entropy, 3) Mel-frequency Filter Bank (MFB) spectral entropy, and 4) delta MFB spectral entropy. In addition, we use four emotions in a speech database including anger, sadness, happiness, and neutrality. We perform speech emotion recognition experiments using each pre-defined emotion and gender. The experimental results show that the proposed emotion recognition using FFT spectral-based entropy and MFB spectral-based entropy performs better than existing emotion recognition based on GMM using energy, Zero Crossing Rate (ZCR), Linear Prediction Coefficient (LPC), and pitch parameters. In experimental Results, we attained a maximum recognition rate of 75.1% when we used MFB spectral entropy and delta MFB spectral entropy.

1. 서 론

인간의 감정을 인식하고 해석하며 표현하는 감성적 지능은 인간의 의사 교환 및 결정에 있어 매우 중요한 역할을 하고 있다 [1]. 컴퓨터가 감정을 인식할 수 있다면 이는 인간과 컴퓨터간의 상호작용에 있어서 보다 개선된 인터페이스를 제공해 줄 수 있을 것이다 [2]. 이를 반영하듯, 최근에는 음성, 얼굴표정, 몸짓 등을 이용하여 인간의 감정 상태를 인지하고 분석하기 위한 다양한 학제적 연구 또한 활발히 추진 중에 있다. 특히 인간의 음성으로부터 추출할 수 있는 에너지, 포만트, 템포, 지속시간, 주파수 변이, 진폭 변이, Mel Frequency Cepstral Coefficient(MFCC), Teager 에너지 등과 같은 특징들은 인간의 감정을 인식하고 이해함에 있어 매우 중요한 요소로 사용되어 지며, 기존 연구에서는 일반적으로 에너지와 피치 특징을 이용하여 인간의 감정을 인식하는 것에 그 중점을 두고 있다 [3]-[4], 따라서 본 논문에서는 음성신호로부터 인간의 감정을 인지하고 표현하기 위한 새로운 방법으로, FFT 스펙트럴 엔트로피와 델타 FFT 스펙트럴 엔트로피, MFB 스펙트럴 엔트로피 그리고 델타 MFB 스펙트럴 엔트로피 특징 파라미터를 이용한 GMM 기반의 감정인식 방법을 제안하고 그 성능을 평가한다.

2. 특징 추출

엔트로피는 Shannon의 정보 이론에서 데이터에 포함되어 있는 정보의 양을 나타낸다[5]. 이를 음성정보처리의 관점에서 나타내면, 엔트로피는 음성특징에 대한 정보량이 된다.

2.1 FFT 스펙트럴 엔트로피와 델타 FFT 스펙트럴 엔트로피

FFT 스펙트럼 엔트로피를 구하는 과정은 다음과 같다. 음성은 먼저 프리엠퍼시스 필터를 사용하여 주파수의 진폭 특성을 평탄하게 유지 한다. 프리엠퍼시스 필터를 통과한 음성 신호는 단구간 신호로 분할되며 이를 프레임이라고 한다[6]. 프레임 길이는 32ms(512 sample data)로 하였으며 두 인접 프레임 사이의 중첩은 16ms(256 sample data)로 하였다. 그후 각 프레임마다 해밍 윈도우를 사용하였다. 시간 영역의 음성 신호를 주파수 영역으로 변환하기 위해 FFT를 사용하였으며, 식(1)에서 $X(i,n)$으로 정의하였다.

$$X(i,n)=\sum_{m=1}^{M} x(m,n)e^{-j\frac{2\pi}{N}im} \tag{1}$$

여기서, $X(i,n)$은 n번째 프레임의 i번째 주파수 성분을 나타내며 M은 FFT 포인트의 개수를 나타낸다. $x(n,m)$은 시간 영역의 음성신호 n번째 프레임의 m번째 샘플데이터를 의미한다. FFT를 취한 후 각 프레임의 파워스펙트럼 $S(i,m)$는 식 (2)에 의해 계산되어 진다.

$$S(i,n)=|X(i,n)|^2 \tag{2}$$

스펙트럼의 확률밀도는 주파수 성분의 정규화 방법을 사용하여 얻을 수 있다. 파워 스펙트럼 정규화는 $P[S(i,n)]$로 정의되며 식 (3)으로 표현하였다.

$$P[S(i,n)]=\frac{S(i,n)}{\sum_{m=1}^{M/2} S(m,n)} \tag{3}$$

마지막 단계는 엔트로피를 계산하는 단계로서, 엔트로피 $H(n)$은 식 (4)로 표현할 수 있다.

$$H(n)=-\sum_{i=1}^{M/2} P[S(i,n)]\log P[(S(i,n)] \tag{4}$$

델타 FFT 스펙트럴 엔트로피는 주파수의 시간적인 변화에 대한 인접 프레임간의 차에 대한 엔트로피로 정의된다. 델타 FFT 스펙트럴 엔트로피는 $n+1$번째 프레임의 i번째 주파수 성분으로부터 n번째 프레임의 i번째 주파수 성분을 차감 후 정규화 및 스펙트럴 엔트로피를 구한다. 델타 FFT는 $S^d(i,n)$으로 정의되며 식 (5)에 나타내었다.

$$S^d(i,n)=S(i,n+1)-S(i,n) \tag{5}$$

델타 FFT에 절대값을 취해 파워 스펙트럼을 구한 후 정규화 를 위한 수식은 식 (6)과 같이 표현할 수 있으며, 델타 FFT 스펙트럴 엔트로피 $H^d(n)$은 식 (7)에 주어진다.

$$P[S^d(i,n)]=\frac{|S^d(i,n)|}{\sum_{m=1}^{M/2}|S^d(m,n)|} \tag{6}$$

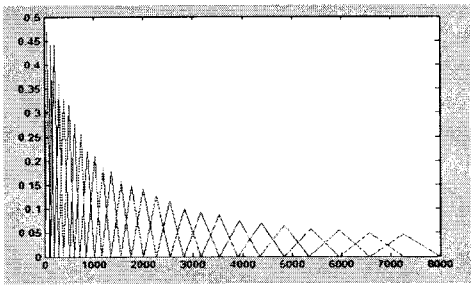
$$H^d(n)=-\sum_{i=1}^{M/2} P[S^d(i,n)]\log P[S^d(i,n)] \tag{7}$$

2.2 MFB 스펙트럴 엔트로피와 델타 MFB 스펙트럴 엔트로피

심리 음향 연구에 따르면 인간의 소리를 인지할 때 주파수에 대해 비선형적인 스케일을 가지며 이를 멜 스케일이라 한며 [7]-[8], 멜 스케일은 식 (8)로 정의되어 진다.

$$f_{mel}=2595\log_{10}(1+\frac{f}{700}) \tag{8}$$

MFB는 27개의 서브밴드 필터로 구성되어 있으며 FFT에 멜-스케일 필터를 곱한 것으로 그림 1로 도식화 하였다.



<그림 1> 멜-주파수 필터 뱅크의 스케일

멜-스케일은 저주파 대해 민감한 인간의 청각 특성을 반영 한 것으로 [6], MFB 스펙트럴 엔트로피는 주파수 정규화와 스펙트럴 엔트로피를 구하기 전에 멜-스케일 필터를 곱한 것이다. 그림 1에서와 같이 음성은 프리엠퍼시스 필터 $1-\alpha Z^{-1}$을 사용하며 음성 입력 신호의 고주파 영역을 강조하기 위해 사용된다. 여기서 "α"의 범위는 0.9에서 1사이의 값을 가지며 기본 값으로 α는 0.97를 사용하였다. 프리엠퍼시스된 음성 신호는 단구간으로 분할되며 분할된 프레임마다 해밍윈도우 처리를 하게 된다. 해밍윈도우 처리 방법은 입력 신호를 일정한 프레임으로 분할함으로써 생기는 불연속을 보완할 수 있는 방법이다. 윈도우 처리에 의하여 분할된 프레임에 존재하는 각각의 입력 신호를 주파수 영역의 신호로 변환하며 FFT(Fast Fourier Transform)를 이용하여 시간 영역의 입력 신호를 주파수 영역의 신호로 변환한다. 변환된 주파수 영역의 신호를 멜-스케일 필터와 곱하여 MFB 신호를 얻게 된다. MFB는 $M(b,n)$으로 정의 되며 식 (9)와 같이 표현된다.

$$M(b,n)=\frac{\sum_{i=L_b}^{U_b} V_b(i)S(i,n)}{\sum_{i=L_b}^{U_b} V_b(i)} \quad (9)$$

여기서 $M(b,n)$는 $S(i,m))$에 b번째 MFB의 i번째 멜-스케일을 곱하여 얻을 수 있다. L_b와 U_b는 i번째 멜 필터의 시작 주파수와 끝 주파수를 나타내며 멜-주파수 스펙트럴 에너지를 $M(b,n)$로 정의 한다. $M(b,n)$는 n번째 프레임의 b번째 MFB 에너지이다. MFB 스펙트럴 에너지에 절대값 및 로그 연산을 취한 후 정규화를 한다. 정규화는 $P[M(b,n)]$으로 정의되며 식 (10)으로 표현하였다.

$$P[M(b,n)]=\frac{M(b,n)}{\sum_{m=1}^{B} M(m,n)} \quad (10)$$

여기서 B는 멜-필터의 개수를 나타내며 본 논문에서는 27개의 필터를 사용하였다. 제안된 MFB 스펙트럴 엔트로피는 식 (11)에 나타내었다.

$$H_{MFB}(n)=-\sum_{b=1}^{B} P[M(b,n)]\log P[M(b,n)] \quad (11)$$

델타 MFB 스펙트럴 엔트로피는 MFB의 시간적인 변화에 대한 인접 프레임간의 차로 정의된다. 델타 MFB 스펙트럴 엔트로피 $M^d(b,n)$은 $n+1$번째 프레임의 b번째 필터 에너지로부터 n번째 프레임의 b번째 필터 에너지를 차감 후 절대값 및 로그를 취한다. 델타 MFB 에너지의 정규화 $P[M^d(b,n)]$는 식 (12)에 주어진다.

$$P[M^d(b,n)]=\frac{|M^d(b,n)|}{\sum_{m=1}^{B} |M^d(m,n)|} \quad (12)$$

여기서 B는 멜 필터의 개수를 나타내었다. 제안한 델타 MFB 스펙트럴 엔트로피는 식 (13)에 나타내었다.

$$H^d_{MFB}(n)=-\sum_{b=1}^{n} P[M^d(b,n)]\log P[M^d(b,n)] \quad (13$$

3. GMM을 이용한 패턴인식 알고리즘

GMM은 입력된 음성신호를 M개의 성분 분포들의 선형 조합으로 근사화 할 수 있으며, 이는 식 (14)로 정의되어진다.

$$P(x|\lambda)=\sum_{i=1}^{M} p_i b_i(x) \quad (14)$$

b_i는 데이터 x에 대한 가우시안 확률 밀도 함수를 의미하며, p_i는 혼합 가중치(mixture weight)로 각 확률밀도함수 또는 각 성분의 상대적인 중요도를 의미한다. GMM을 이루는 파라미터는 식 (15)에 주어진다.

$$\lambda=\{p_i, u_i, \Sigma_i\} \quad i=1,2,3,\ldots,M \quad (15)$$

p_i는 혼합 가중치(mixture weight)이며, u_i는 평균 벡터(mean vector)이다. Σ_i는 공분산 행렬이며, 이들 세 가지 파라미터의 집합이 어떤 화자나 감정의 가우시안 혼합 분포를 표현할 수 있는 모델이 된다. GMM을 이용한 훈련 과정에서는 감정별 음성 데이터마다 MLE(Maximum Likelihood Estimation)와 EM(Expectation Maximization) 알고리즘을 이용하여 최대 가우시안 혼합 분포 값을 갖는 GMM 파라미터를 추정한다. 또한 GMM 기반의 감정인식 과정에서는 추정된 감정별 GMM 파라미터를 이용하여, 입력된 음성 데이터의 특징 벡터에 대한 각각의 가우시안 혼합 분포를 구하고 그 중 가장 큰 확률 값을 가지는 GMM 파라미터를 인식 결과로 선택한다.

4. 실험 및 결과

4.1 데이터베이스

본 논문에서는 인간의 주요 감정인 화남, 기쁨, 슬픔, 평소의 4가지 감정을 인식 후보로 결정하였고, 감정음성 데이터는 감정 표현을 훈련하는 아마추어 연극단원 남/여 각 15명을 대상으로 45개의 문장을 3번씩 발성 및 녹음한 연세대학교 멀티미디어 통신 연구실의 데이터베이스를 이용하였다.

4.2 스펙트럴 엔트로피를 이용한 실험

감정음성 데이터베이스의 45개의 문장 중 35개의 문장은 훈련 데이터로 사용하고, 나머지 10개의 문장을 실험 데이터로 사용하였다. 훈련에 사용된 문장은 총 12,600개이다. 실험의 진행은 훈련 데이터와 실험 데이터에서 200개의 문장을 임의적으로 선택하여 실험을 진행하였다.

<표 1> 감정인식 결과 : (a) FFT 스펙트럴 엔트로피; (b) 델타 FFT 스펙트럴 엔트로피

	훈련 데이터		실험 데이터	
	남성	여성	남성	여성
화남	78%	69%	72%	66%
슬픔	80%	74%	86%	78%
기쁨	52%	58%	46%	66%
평소	60%	62%	52%	54%
평균	67.5%	65.5%	64%	66%

(a)

	훈련 데이터		실험 데이터	
	남성	여성	남성	여성
화남	82%	84%	84%	78%
슬픔	78%	72%	44%	76%
기쁨	64%	52%	48%	54%
평소	62%	58%	66%	50%
평균	71.5%	66.5%	60.5%	64.5%

(b)

<표 2> 감정인식 결과 : (a) MFB 스펙트럴 엔트로피; (b) 델타 MFB 스펙트럴 엔트로피

	훈련 데이터		실험 데이터	
	남성	여성	남성	여성
화남	78%	66%	74%	62%
슬픔	76%	84%	76%	86%
기쁨	54%	58%	42%	54%
평소	68%	72%	72%	70%
평균	69%	70%	66%	68%

(a)

	훈련 데이터		실험 데이터	
	남성	여성	남성	여성
화남	92%	80%	84%	78%
슬픔	72%	92%	72%	96%
기쁨	56%	58%	54%	52%
평소	86%	74%	82%	74%
평균	76.5%	76%	73%	75%

(b)

표 1과 2에서는 개별 특징들을 이용한 GMM 기반의 감정인식 실험결과를 보여준다. 실험 결과, FFT 및 델타 FFT 스펙트럴 엔트로피를 이용한 실험의 경우 동일하게 65.75%의 인식률을 보였다. 반면, MFB 스펙트럴 엔트로피를 이용한 실험의 경우는 68.25%의 인식률을 보였으며, 델타 MFB 스펙트럴 엔트로피를 이용한 실험의 경우 75.13%의 인식률을 보였다.

5. 결 론

본 논문에서는 FFT 스펙트럴 엔트로피, 델타 FFT 스펙트럴 엔트로피, MFB 스펙트럴 엔트로피, 그리고 델타 MFB 스펙트럴 엔트로피 특징 파라미터를 이용한 GMM 기반의 새로운 감정인식 방법에 대하여 기술하였으며, 4가지 감정인식 후보에 대하여 개별 특징들을 이용한 성능평가 실험을 수행하였다. 그 결과 FFT 스펙트럴 엔트로피 보다 MFB 스펙트럴 엔트로피의 성능이 보다 우수하였으며 MFB 스펙트럴 엔트로피 특징을 이용했을 경우 최대 인식률을 보임에 따라 향후 감정인식 시스템에서 보다 적합한 특징으로 사용되어질 수 있음을 알 수 있었다. HMM과 같은 기존의 패턴인식 알고리즘과 더불어 델타-델타 FFT / MFB 스펙트럴 엔트로피 특징을 선별적으로 선택 및 적용함으로써 보다 진보된 감정인식기의 구현을 향후 과제로 남겨둔다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT 연구센터 지원 사업의 연구결과로 수행되었음 (IITA-2008-(C1090-0801-0046))

[참 고 문 헌]

[1]D. Goleman,: "Emotional Intelligence", Bantam Books, New York, 1995
[2]Borchert, M.; Dusterhoft, A.: "Emotion in Speech-Experiments with Prosody and Quality Features in Speech for Use in Categorical and Dimensional Emotion Recognition Environments", Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on, 30 Oct.-1 Nov. 2005
[3]Sung-ill kim, Sang-hoon Lee, Wee-jae Shin and Nam-chun Park,: "Recognition of Emotional states in Speech using Hidden Markov Model", Proceeding of KFIS Fall Conference, Volume 14, Number 2, 2004
[4]Li Zhao, Yujia Cao, Zhiping Wang and Cairong Zou,: "Speech Emotional Recognition Using Global and Time Sequence Structure Features with MMD", Lecture Notes Computer Science, Vol.3784.Springer-Verlge, 2005
[5]정미옥, 김현숙, 송점동, 이정현, "음성인식 시스템에서 엔트로피를 이용한 거절", 한국정보과학회 가을 학술발표논문집, Vol. 26. No. 2, 1999
[6]Young-Wan Roh and Kwang-Seok Hong,: "Delta FBLC based Speech/Non-Speech Frame Decision in Real Car Environment", The 4th Conference on New Exploratory Technologies(Next 2007)
[7]Chakroborty S., Saha G., "Improved Text-Independent Speaker Identification using Fused MFCC & IMFCC Feature Sets based on Gaussian Filter", International Journal of Signal Processing Vol 5, Number 1, 2008
[8]Ben Gold and Nelson Morgan, "Speech and Audio Signal Processing", Part-IV, Chap.14, pp 189~203, John Willy&Sons, 2002