

헬스케어 로봇으로의 응용을 위한 음성기반의 감정인식 알고리즘 구현

Implementation of the Timbre-based Emotion Recognition Algorithm for a Healthcare Robot Application

Jung-Shik Kong^{**}, Oh-Sang Kwon^{**}, Eung-Hyuk Lee^{***}
공 정 식^{**}, 권 오 상^{**}, 이 응 혁^{***}

Abstract

This paper deals with feeling recognition from people’s voice to fine feature vectors. Voice signals include the people’s own information and but also people’s feelings and fatigues. So, many researches are being progressed to fine the feelings from people’s voice. In this paper, We analysis Selectable Mode Vocoder(SMV) that is one of the standard 3GPP2 codecs of ETSI. From the analyzed result, we propose voices features for recognizing feelings. And then, feeling recognition algorithm based on gaussian mixture model(GMM) is proposed. It uses feature vectors is suggested. We verify the performance of this algorithm from changing the mixture component.

요 약

음성신호는 화자에 대한 고유한 정보와 주변의 음향환경에 대한 정보는 물론 감정과 피로도 등 다양한 정보가 포함되어 있다. 이에 음성신호를 이용한 연구분야에서 감정 상태를 파악하기 위한 연구가 지속되어 왔다. 이에 본 논문에서는 화자의 감정을 인식하기 위해 ETSI의 3GPP2 표준코덱인 Selectable Mode Vocoder(SMV)를 분석한다. 이를 기반으로 감정 인식에 효과적인 특징들을 제안한다. 이후 선정된 특징 벡터를 이용하여 Gaussian Mixture Model(GMM) 기반의 감정 인식 알고리즘을 개발하고 Mixture component 개수를 변화시키면서 성능을 검증한다.

Key words : Feeling Recognition, Selectable Mode Vocoder, Gaussian Mixture Component

1. 서론

IT 기술은 과거에 기술 및 시설 인프라 구축 중심에서 인간을 중심으로 진보되어 왔다. 인간중심의 발전 방향은 지속될 것이고 그에 따른 서비스의 중요성도 더욱 부각될 것이다. 감정인식 분야는 현재 많은 연구가 진행되고 있고 미래기술로 주목받고 있다. 최

근에는 휴대용 기기와 로봇 등 감성 인터페이스에 대한 관심이 고조되면서 국내는 물론 해외에서도 중요한 연구주제로 떠오르고 있다. 인간과 다른 매체간의 감성 인터페이스를 위해서는 단순한 외적요소의 인지를 넘어 감정 상태와 선호 경향까지 파악할 수 있는 기술이 요구된다. 최근의 연구들은 화남, 슬픔, 즐거움, 중립 감정 등의 기본 감정을 기반으로 그 범위를 확대해 나가고 있다 [1-3].

음성신호에는 화자에 대한 고유한 정보와 주변의 음향환경에 대한 정보는 물론 감정과 피로도 등 다양한 정보가 포함되어 있다. 음성신호를 이용한 연구 분야에서 감정 상태를 반영하는 효과적인 특징들을 추출하는 것이 성능을 결정하는 데 가장 중요한 요소이다. 그동안의 연구에서 피치와 에너지가 감정인식에 있어서 매우 효과적인 특징임은 밝혀졌지만 [4], [5], 그 자체로 완벽한 분류가 어렵고 다양한 감정에 대한 분류에서는 한계가 있으므로 각 감정 상태를 잘 반영할 수 있는 특징에 대한 연구가 요구된다.

패턴인식 기법은 HMM (Hidden Markov Model),

* 仁德大學 機械設計科
(Department of Mechanical Design, Induk University)
** 京畿工業大學 自動化로봇科
(Department of Automation and Robot Kyonggi Institute of Techonology)
*** 韓國産業技術大學校 電子工學科
(Department of Electronical Engineering, Korea Polynomic University)
★ 교신저자 (Corresponding author)
接受日:2009年 12月 12日, 修正完了日: 2009年 12月 27日

1) LPC prediction coefficients :

$$a_m(i) = \alpha_i^{(10)} \quad ; 1 \leq i \leq 10 \quad (1)$$

Reflection coefficients, prediction coefficients, LP prediction error는 다음의 Levinson-Durbin algorithm을 통해서 계산된다.

$$\begin{aligned} E^{(0)} &= R_m(0) \\ i &= 1 \\ \text{while } (i < 10) \{ \\ & k_m(i) = -\frac{R_m(i) + \sum_{j=1}^{i-1} \alpha_j^{(i-1)} \cdot R_m(i-j)}{E^{(i-1)}} \\ & \alpha_i^{(i)} = k_m(i) \\ & j = 1 \\ & \text{while } (j < i-1) \{ \\ & \quad \alpha_j^{(i)} = \alpha_j^{(i-1)} + k_m(i) \cdot \alpha_{i-j}^{i-1} \\ & \quad j = j+1 \\ & \} \\ & E^{(i)} = (1 - k_m^2(i)) E^{(i-1)} \\ & i = i+1 \\ & \} \end{aligned}$$

여기서 $E_{LPC} = \sum_{n=0}^{L_{LPC}-1} s^2[n]$ 로써 segment energy가 계산되고, 신호는 LPC window에 의해 곱해지며 10차의 특징벡터를 추출하여 사용한다.

2.2.2 음악 분류 알고리즘 특징 벡터

1) Running mean of the first LSF coefficient :

$$\overline{lsf_1}(1) = 0.75 \cdot \overline{lsf_1}(1) + 0.25 \cdot lsf_1(1) \quad (2)$$

여기서 $lsf_1(1)$ 는 식 $A(z) = 1 + \sum_{i=1}^{10} a(i) \cdot z^{-i}$ 을 통해서 구해진 10차의 lsf 값 중 첫 번째 값을 나타낸다.

2) Running mean of energy :

$$\overline{E} = 0.75 \cdot \overline{E} + 0.25 \cdot E \quad (3)$$

식 $E = \max\left(10, 10 \cdot \log_{10}\left(\frac{R_1(0)}{L_{LPC}}\right)\right)$ 에 의해서 구해진 에너지 값의 running mean 이다.

3) Spectral difference :

$$SD_4 = \sum_{i=1}^{10} (k_1(i) - \overline{k_N}(i))^2 \quad (4)$$

$\overline{k_N}$ 은 잡음/무음 구간의 reflection coefficients의 running mean 이다.

4) Running mean of the partial residual energy :

$$\overline{E_N^{res}} = 0.9 \cdot \overline{E_N^{res}} + 0.1 \cdot E^{res} \quad (5)$$

VAD가 동작하지 않을 때 $\overline{k_N}$ 에 따라 업데이트 된다.

5) running mean reflection coefficients of noise/silence :

$$\overline{k_N}(i) = 0.75 \cdot \overline{k_N}(i) + 0.25 \cdot k_1(i) \quad ; i = 1, \dots, 10 \quad (6)$$

6) Running mean of the normalized pitch correlation :

$$\overline{corr_P} = 0.8 \cdot \overline{corr_P} + 0.2 \cdot \left(\frac{1}{5} \cdot \sum_{i=1}^5 corr_P^B(i) \right) \quad (7)$$

$\overline{corr_P^B}(i)$ 는 5개의 normalized pitch correlation이다.

7) Running mean of the periodicity counter :

(8) ~ (13)의 특징 벡터와 설정된 문턱 값의 비교를 통해서 증가하거나 감소하고, $\overline{c_{pr}} \geq 18$ 이면 프레임은 음악으로 분류한다.

$$\overline{c_{pr}} = \alpha \cdot \overline{c_{pr}} + (1 - \alpha) \cdot c_{pr} \quad (8)$$

여기서 사용되는 가중치 α 값은

$$\alpha = \begin{cases} 0.98 & c_{pr} > 12 \\ 0.95 & c_{pr} > 10 \\ 0.90 & c_{pr} > \text{otherwise} \end{cases} \quad (9)$$

8) Running mean of the normalized pitch correlation :

$$\overline{corr_P^N} = 0.8 \cdot \overline{corr_P^N} + 0.2 \cdot \left(\frac{1}{5} \cdot \sum_{i=1}^5 corr_P^B(i) \right) \quad (10)$$

매우 낮은 주파수 잡음 플래그 F_f 가 설정되는 것은 VAD가 동작하지 않거나 $lsf_1(1) < 110$ Hz 또는 $lsf_1(1) < 150$ Hz일 때이다. $\overline{corr_P^N}$ 은 정정된 VAD가 동작하지 않거나 F_f 가 설정될 때 업데이트된다.

9) Running mean of the music continuity counter :

(2)-(8), (10)의 특징 벡터와 설정된 문턱 값의 비교를 통해서 증가하거나 감소하고, $\overline{c_M} \geq 200$ 이면 프레임은 음악으로 분류한다.

$$\overline{c_M} = 0.9 \cdot \overline{c_M} + 0.1 \cdot c_M \quad (11)$$

3. GMM 기반의 감정 인식 시스템

본 논문에서는 별도의 처리과정 없이 SMV의 인코딩 과정에서 추출되는 파라미터들을 특징 벡터로 이용하되 효과적인 GMM을 구성하기 위해 SMV 파라미터를 선택적으로 사용하여 계산량 감소와 성능 향상을 도모하였다. 추출한 중요 특징벡터를 상태 열 N개의 특징 벡터 $X = \{x_1, x_2, \dots, x_N\}, x_t \in R^D$ 라 할 때,

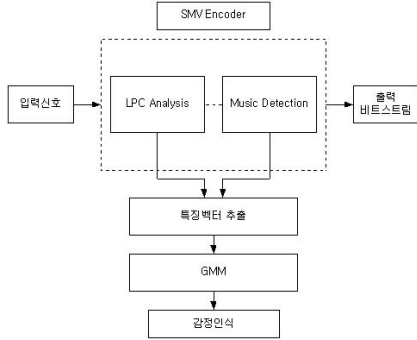


Fig. 2 Emotion Recognition with GMM from SMV parameter

그림 2. SMV 파라미터를 이용한 GMM 기반 감정 인식

1) M개의 혼합성분 (Mixture Component)을 가지는 가우시안 확률밀도함수의 우도 (Likelihood)는 다음과 같다.

$$p(x_t|\lambda) = \sum_{i=1}^M p_i b_i(x_t) \quad (11)$$

$$b_i(x_t) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x_t - \mu_i)^T (\Sigma_i)^{-1} (x_t - \mu_i)} \quad (12)$$

$$\sum_{i=1}^M p_i = 1, 0 \leq p_i \leq 1 \quad (13)$$

GMM모델의 번째 성분 파라미터 λ 는

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, i = 1, \dots, M \quad (14)$$

가우시안 혼합 성분 밀도의 가중치 (Mixture Weight : p_i), 평균 벡터 (Mean Vector : μ_i), 공분산 행렬 (Covariance Matrix : Σ_i)로 구성되고 각각의 식은 아래와 같다.

$$p_i = \frac{1}{T} \sum_{t=1}^N p(i|x_t, \lambda) \quad (15)$$

$$\mu_i = \frac{\sum_{t=1}^N p(i|x_t, \lambda) \cdot x_t}{\sum_{t=1}^N p(i|x_t, \lambda)} \quad (16)$$

$$\Sigma_i = \frac{\sum_{t=1}^N p(i|x_t, \lambda) \cdot x_t^2}{\sum_{t=1}^N p(i|x_t, \lambda)} - \mu_i^2 \quad (17)$$

이를 통해 i번째 성분 사후확률 (A Posteriori

Probability)은 다음과 같이 주어진다.

$$p(i|x_t, \lambda) = \frac{p_i b_i(x_t)}{\sum_{k=1}^M p_k b_k(x_t)} \quad (18)$$

본 논문에서는 Expectation Maximization (EM)을 사용해 최적 모델 λ 을 추정하며 $p(x|\lambda') \geq p(x|\lambda)$ 가 되는 새로운 모델 λ' 이 정해진 문턱 값 (Threshold)에 도달할 때까지 반복하여 모델을 설정한다. 실제로 구성된 각 클래스별 모델의 실제 데이터의 특징벡터를 입력 받아 감정플래그 (Emotion Flag)를 선택한다.

$$EF = \arg \max_{1 \leq s \leq k} \sum_{t=1}^N \log p(x_t | \lambda_s) \quad (19)$$

$k = \text{umber of emotion}$

IV. 성능평가

특징 선택에 따른 성능의 변화를 살펴보기 위해 LPC prediction coefficients와 pitch는 공통적으로 포함되고 9개의 Music detection 파라미터를 모두 사용한 20차의 특징벡터 결과와 그 중 6개를 선택한 17차의 특징벡터 결과를 비교하여 보았다. 동일한 조건에서의 인식성능을 알아보기 위해 GMM의 mixture component 개수는 16으로 맞추었다. 실험을 통해 확인된 인식 성능은 20차의 특징벡터를 사용하였을 때 54.73%였고, 17차의 특징벡터의 경우가 65.01%였다. 결과는 위의 표와 같이 혼동행렬 (confusion matrix)로 나타내었다.

Table 3. Confusion matrix of 20th characteristic vector
표 3. 20차의 특징벡터를 이용한 경우 혼동행렬

	angry	fast	neutral	question	slow
angry	0.73	0.14	0.03	0.04	0.06
fast	0.20	0.72	0.02	0.05	0.01
neutral	0.22	0.24	0.18	0.31	0.05
question	0.27	0.15	0.09	0.47	0.02
slow	0.15	0.01	0.09	0.11	0.64

Table 4. Confusion matrix of 17th characteristic vector
표 4. 17차의 특징벡터를 이용한 경우 혼동행렬

	angry	fast	neutral	question	slow
angry	0.84	0.04	0.01	0.07	0.04
fast	0.16	0.67	0.11	0.05	0.00
neutral	0.15	0.22	0.48	0.13	0.02
question	0.22	0.04	0.10	0.63	0.01
slow	0.19	0.01	0.11	0.06	0.62

다수의 확률밀도함수로 이루어진 GMM의 경우 mixture component 개수의 변화에 따라 계산량이 달라지며 대체적으로 mixture component 개수가 증가함에 따라 성능 역시 향상됨을 확인할 수 있다. Mixture component 개수 변화에 따른 결과를 살펴보면 16차일 때 65.01%, 32차일 때 67.74%, 64차일 때 73.84%, 128차일 때 70.73%, 256차일 때 75.68%, 512차일 때 76.69%의 정확도를 나타내었다. mixture component의 개수를 증가시킬수록 높은 성능을 확인

할 수 있지만 계산량의 증가로 실시간 적용에는 제약이 따르게 된다. 64차를 사용하였을 때 결과가 512차의 결과와 비교하여 근소한 성능차이를 나타내는 반면 계산소요 시간은 현저히 감소하므로 64차의 경우가 최적화된 알고리즘이라고 할 수 있다. 각각의 결과를 표로 나타내어 보면 아래와 같다.

Table 5. Confusion matrix from 32th mixture order
표 5. Mixture order가 32차일 경우의 혼동행렬

	angry	fast	neutral	question	slow
angry	0.88	0.06	0.02	0.03	0.01
fast	0.02	0.82	0.15	0.01	0.00
neutral	0.10	0.40	0.43	0.03	0.04
question	0.26	0.05	0.08	0.60	0.01
slow	0.15	0.03	0.14	0.03	0.65

Table 6. Confusion matrix from 64th mixture order
표 6. Mixture order가 64차일 경우의 혼동행렬

	angry	fast	neutral	question	slow
angry	0.79	0.09	0.00	0.11	0.01
fast	0.03	0.84	0.00	0.13	0.00
neutral	0.10	0.40	0.43	0.02	0.04
question	0.02	0.02	0.06	0.90	0.00
slow	0.17	0.03	0.00	0.13	0.67

Table 7. Confusion matrix from 128th mixture order
표 7. Mixture order가 128차일 경우의 혼동행렬

	angry	fast	neutral	question	slow
angry	0.71	0.03	0.08	0.16	0.02
fast	0.00	0.72	0.21	0.07	0.00
neutral	0.02	0.12	0.65	0.20	0.01
question	0.01	0.03	0.10	0.86	0.00
slow	0.11	0.01	0.24	0.05	0.59

Table 8. Confusion matrix from 256th mixture order
표 8. Mixture order가 256차일 경우의 혼동행렬

	angry	fast	neutral	question	slow
angry	0.80	0.02	0.05	0.10	0.02
fast	0.01	0.80	0.13	0.06	0.00
neutral	0.05	0.12	0.58	0.18	0.07
question	0.02	0.08	0.04	0.85	0.01
slow	0.09	0.01	0.13	0.02	0.75

Table 9. Confusion matrix from 512th mixture order
표 9. Mixture order가 512차일 경우의 혼동행렬

	angry	fast	neutral	question	slow
angry	0.82	0.08	0.01	0.05	0.04
fast	0.01	0.80	0.14	0.04	0.01
neutral	0.04	0.09	0.57	0.16	0.14
question	0.05	0.06	0.09	0.77	0.03
slow	0.04	0.01	0.07	0.01	0.87

V. 결론

본 논문에서는 휴먼 인터페이스의 지능형 로봇에 적용하기 위해 임베디드 시스템에 적합한 감정 인식 알고리즘을 구현하였다. 제한된 시스템 리소스 환경에 적용하기 위해 현재 모바일 휴대폰에 사용되고 있는 SMV의 파라미터를 특징벡터로 사용하였고, 다양한 감정 패턴에 보다 효과적인 성능을 보일 수 있도록 음성 인식과 음악 인식에서 뛰어난 성능을 보이는

Expectation-Maximization (EM) 알고리즘 기반의 패턴인식기법인 GMM 알고리즘을 사용하였다. 최종적으로 GMM의 Mixture component 개수를 변화시켜 감정 인식에 적합한 최적의 Mixture component 개수를 적용하였다.

참고문헌

- [1] Q. Ji, P. Lan, C. Looney, "A Probabilistic Framework for Modeling and Real-Time Monitoring Human Fatigue," *IEEE Transaction on systems, man, and cybernetics Part A : Systems and humans*, vol. 36, no. 5, Sep. 2006.
- [2] S. Casale, A. Russo, S. Serrano, "Multi-Style Classification of Speech Under Stress Using Feature Subset Selection Based on Genetic Algorithms," *Speech Communication*, 2007.
- [3] R. Faltlhauser, T. Pfau, G. Ruske, "On-line Speaking Rate Estimation Using Gaussian Mixture Models," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2000.
- [4] O. W. Kwon, K. Chan, J. Hao, T. W. Lee, "Emotion Recognition by Speech Signals," *Proc. Eurospeech*, 125-128, 2003.
- [5] S. Ramamohan, S. Dandapat, "Sinusoidal Model-Based Analysis and Classification of Stressed Speech," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 3, May 2006.
- [6] R. O. Duda, P. E. Hart and D. G. Stork, Pattern classification, *John Wiley & Sons, INC.*, 2001.
- [7] S. Craig Greer, and A. Dejaco, "Standardization of the selectable mode vocoder," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 953-956, May 2001.
- [8] G. Yang, E. Shlomot, A. Benyassine, J. Thyssen, S. Huan-yu and C. Murgia, "The SMV algorithm selected by TIA and 3GPP2 for CDMA applications," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 709-712, May 2001.
- [9] 3GPP2 Spec., "Selectable Mode Vocoder (SMV) Service Option for Wideband Spread Spectrum Communications Systems," *3GPP2-C.S0030-0*, v3.0, Jan. 2004.

공 정 식 (정회원)

1998년 : 인하대학교 자동화공학과 졸업 (공학사)
 2006년 : 인하대학교 대학원 자동화공학과 (공학박사)
 2007년 3월~2009년 2월 : 대덕대학 마이크로 로봇과 전임강사
 2009년 3월~현재 : 인덕대학 기계설계과 전임강사

<주관심분야> 지능 제어, 로봇 제어, 재활 보조 시스템, 헬스케어

권 오 상 (비회원)

1990년 : 인하대학교 전자공학과 졸업 (공학사)
 1992년 : 인하대학교 대학원 전자공학과 (공학석사)
 1999년 : 인하대학교 대학원 전자공학과 (공학박사)
 2007년 3월~현재 : 경기공업대학 자동화로봇과 조교수

<주관심분야> 지능형 로봇 시스템, 신호처리

이 응 혁 (정회원)

1985년 : 인하대학교 전자공학과 졸업 (공학사)
 1987년 : 인하대학교 대학원 전자공학과 (공학석사)
 1997년 : 인하대학교 대학원 전자공학과 (공학박사)
 2000년 3월~현재 : 인한국산업기술대학교 전자공학과 교수

<주관심분야> 지능형 서비스 로봇 제어, 재활 보조 시스템, 의용전자기기 및 신호처리, 임베디드 시스템