http://dx.doi.org/10.14400/JDPM.2013.11.11.381

소리 주파수대역 기반 멀티미디어 콘텐츠의 감성 추출

권영훈*, 장재건*
한신대학교 컴퓨터공학부*

Emotion Extraction of Multimedia Contents based on Specific Sound Frequency Bands

Young-Hun Kwon*, Jae-Khun Chang*
Division of Computer Engineering, Hanshin Univ., Korea*

요 약 최근 인간의 감성에 반응하고, 감성을 유도하는 감성콘텐츠가 문화산업 분야에서 크게 주목을 받으면서 멀티미디어 콘텐츠가 유발하는 감성 추출에 초점이 모아지고 있다. 게다가 최근 멀티미디어 콘텐츠가 빠르고 방대하게 생산, 유통되는 흐름으로 볼 때 콘텐츠에서 유발하는 감성을 자동으로 추출하는 기법의 연구들이 주목받고 있다. 본 논문은 멀티미디어 콘텐츠의 소리 정보 중 특정 주파수대역의 볼륨 값을 활용하여 멀티미디어 콘텐츠 내의 감성지수를 추출하는 방법에 대해 연구하고자 한다. 이러한 연구는 동영상 콘텐츠의 감성지수를 자동으로 추출할 수 있도록 하며 추출된 정보를 활용하여 사용자의 현재 감성, 혹은 날씨 등과 같은 기타 요소에 맞추어 사용자에게 맞춤형 콘텐츠를 제공하는데 사용되어질 것이다.

주제어 : 감성 추출, 소리 주파수, 소리 감성, 멀티미디어 콘텐츠, 각성

Abstract Recently, emotional contents that induce emotions and respond to emotions are given attention in the field of cultural industries, and extracting emotion caused by multimedia contents is being noted. Furthermore, since multimedia contents have been quickly produced and distributed these days, researches automatically to extract the feeling of multimedia contents are being accelerated. In this paper, we will study the method of emotional value extraction in the multimedia contents using the volume value of the multimedia contents in a certain frequency among sound informations. This study allows to extract the emotion of multimedia contents automatically, and the extracted information will be used to provide user's current emotion, weather, etc. for the users.

Key Words : Emotion Extraction, Sound Frequency, Sound Emotion, Multimedia Contents, Emotional Arousal

1. 서론

최근 인간 감성을 자극할 수 있는 감성 서비스가 대두되면서 멀티미디어 콘텐츠가 유발할 수 있는 감성에 또한 초점이 모아지고 있다. 게다가 최근 멀티미디어 콘텐츠가 과거보다 더 빠르게 유통되고 매일 방대한 양이 새

* 본 논문은 2013년 한신대학교의 학술연구비에 의하여 지원되었음
Received 16 September 2013, Revised 9 October 2013
Accepted 20 November 2013
Corresponding Author: Jae-Khun Chang(The Society of Digital Policy)
Email: jchang@hs.ac.kr
ISSN: 1738-1916

롭게 업로드 되는 흐름으로 인해 자동으로 감성을 추출하는 기법의 연구들이 주목받게 되면서, 관련 연구가 활발하게 진행 중이다[1-3]. 소리 자극은 인간의 감성에 영향을 미칠 수 있기 때문에[4] 본 논문에서는 영상의 소리 정보에 따라 단일 감성 구간을 설정하고, 구간 내 특정 주파수 대역의 볼륨 값을 활용하여 영상의 각성지수를 추출한다.

소리의 구간을 설정함에 있어서는 영상의 볼륨정보를 활용하여 구간을 설정한다. 볼륨이란 소리의 강도를 측정하는 방법으로 사람의 청각반응이 소리 크기 자체에 비례하지 않고 대수(log)에 비례하는 현상을 반영하여 수치를 나타내는 방법이며 dB로 표현한다. 이렇게 소리의 구간을 나타내는 이유는 여러 가지 감성이 혼합된 다중 감성 영상을 하나의 단일 감성 파트로서 각성지수를 추출하기 위함이다.

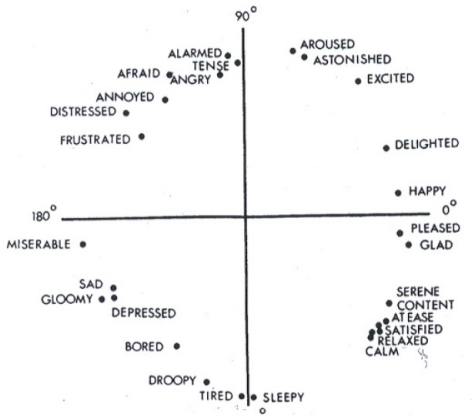
설정된 구간 내에서 각성지수를 추출하기 위하여 특정 주파수 대역의 볼륨 값을 활용한다. 주파수란 1초에 진동횟수를 나타내며 Hz로 표현하고, 소리에서는 고주파수 대역일수록 날카로운 음으로 인식된다. 일반적으로 사람은 4000Hz에서 가장 민감하게 반응하며 아기 울음소리, 여자의 비명소리, 환호소리 등이 이 4000Hz 음향에 속한다. 또한 사람의 목소리는 30Hz ~ 4000Hz 범위에 속하며 영상에서 이를 제외한 4000Hz ~ 8000Hz의 범위는 사람 목소리가 아닌 배경음으로 가정할 수 있다. 본 논문에서는 4000Hz 부근에서 소리의 최대 볼륨 값과 4000Hz ~ 8000Hz 대역의 배경음에서 평균 볼륨 값을 활용한 각성도 추출에 대하여 논의한다.

본 연구에 대한 결과는 멀티미디어의 소리로서 각성 감성지수를 추출할 수 있게 하고, 추출된 각성지수를 이용하여 사용자에게 맞춤형 콘텐츠를 제공하는데 이용될 수 있다.

2. 선행연구

2.1 감성표현 모델

러셀(Russel)이 제안한 감성모델은 정서의 개념이 쾌-불쾌, 각성-이완 차원이라는 독립적인 두 개의 차원에 따라 원형적으로 분포한다는 주장을 하였다[5]. 쾌-불쾌 각성/이완 두 차원은 각각 -5 ~ 5 사이의 분포를 따르며 각성지수가 -5에 가까울수록 이완 상태를 5에 가까울수록 각성 상태임을 나타낸다.



[Fig. 1] Mental space of emotions proposed by Russell[5]

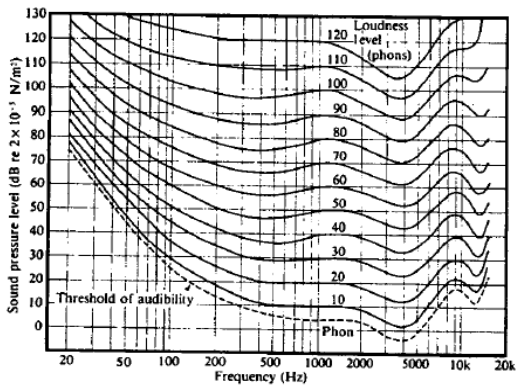
[Fig. 1]은 제임스 러셀의 감정 원형모형을 나타낸다. 직교하는 두 개의 축을 기반으로 감정의 분포를 나타내고 있으며, 가로 축은 쾌-불쾌 정도를, 세로 축은 긴장-이완 정도를 나타낸다. 가로 축에서 오른쪽으로 갈수록 쾌의 감정에 가까워지고, 왼쪽으로 갈수록 불쾌의 감정에 가까워진다. 세로 축에서는 위쪽으로 갈수록 긴장의 감정에 가까워지고, 아래쪽으로 갈수록 이완의 감정에 가까워진다. 본 논문에서는 러셀이 제안한 감정모델의 각성지수 표기를 따른다.

2.2 소리 주파수에 따른 감성

소리의 크기에 대한 귀의 민감도는 주파수에 따라 변한다. [Fig. 2]는 주파수에 따라 변화하는, 동일한 크기를 느끼게 하는 순음의 음압수준 SPL(dB)을 보여주는 등라우드니스 수준(equal loudness level) 곡선이다[6]. 청각의 민감도는 중저음 주파수 영역에서는 상대적으로 낮음을 알 수 있고, 3500Hz ~ 4000Hz에서 최대가 된다. 이는 같은 진폭의 소리 파형일지라도 주파수에 따라 사람이 느끼는 소리의 크기는 다르다는 것을 보여준다.

음악과 음향진동의 자극을 인체에 가하여 인체의 뇌파의 변화를 비교 분석 평가를 실시한 결과 강한 소음 자극에 대한 뇌파의 변화는 각성을 의미하는 베타파의 증가 및 안정을 의미하는 알파파의 감소가 나타났다[7]. 따

라서 더 큰 소리일수록 사람의 각성지수를 높인다고 알 수 있다.



[Fig. 2] Equal loudness level contours[6]

2.3 White Noise 음향

White Noise는 일반적으로 이완의 감성을 유발한다고 알려져 있다. White Noise란 사람의 가청범위 (20Hz ~ 20kHz)내의 모든 주파수를 같은 양으로 포함하고 있는 의미 없는 소리를 의미한다[8]. White Noise가 혼합된 음향에서 혼합되지 않은 음향보다 이완효과가 크게 나타났다[9]. 또한 White Noise와 원본 음향 간에 마스킹 효과가 작은 음향일수록 White Noise의 혼합으로 인한 이완효과가 큰 것을 확인할 수 있었다[10].

3. 제안하는 감성 검출방법

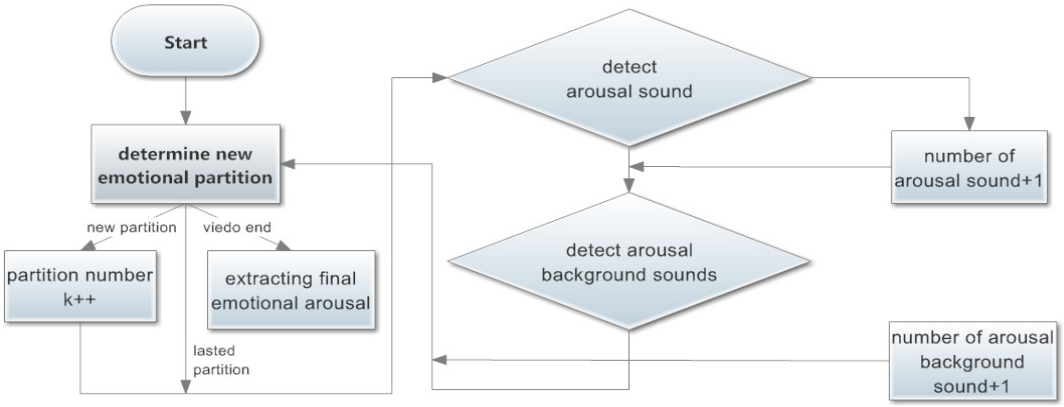
3.1 전체적 검출방법

일반적으로 소리의 크기가 강해질수록 각성을 의미하는 베타파의 증가를 보였으며[7], 사람의 청각은 4000Hz에 가장 예민하다. 따라서 본 연구에서는 사람 목소리를 제외한 영상 볼륨의 크기와 4000Hz대역에서 소리의 강도를 활용하여 각성도를 추출하는 것을 제안한다.

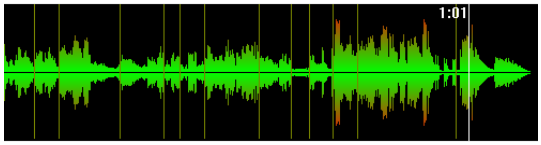
[Fig. 3]는 본 논문에서 제시된 영상 내 사운드 정보를 이용하여 각성도를 추출하는 전체 과정이다. 연산을 수행하기 위해 우선 사운드의 구간 판별을 실시한다. 새로운 구간으로 판별되지 않았을 경우에 4000Hz 부근의 최대 볼륨과 4000Hz 이상의 평균 볼륨을 구하고 각각 해당 구간에서 만족하는 프레임 수를 추출한다. 새로운 구간으로 판별되었을 경우에 구간 내 각성도가 나타난 프레임의 초당 비율을 이용하여 해당 구간에서의 각성도로 매핑을 하는 구조이다.

3.2 사운드 구간 추출

사운드 구간을 추출하는 것은 여러 감성 샷이 혼합된 영상 내에서 하나의 샷에서만 감성을 추출할 수 있게 한다. 이는 서로 다른 감성 파트는 서로 독립이기 때문이다. 그리고 최종 영상의 각성도는 전체 영상에서 각성 구간의 시간에 따른 비율로서 나타내어야 하는데, 높은 각성 파트가 오래 노출될수록 최종 영상의 각성도가 높아야 하기 때문이다.



[Fig. 3] Diagram of the proposed method



[Fig. 4] Wave form divided whole part by sound volume variation into single emotional part

[Fig. 4]은 실제 영상에서 볼륨 크기를 이용하여 동영상의 구간을 설정한 모습이다. 감성이 연속적인 부분을 새로운 구간으로서 설정하기도 하지만 서로 다른 감성 파트를 같은 구간으로 인식하지만 않는다면 동영상의 각성도를 추출할 목적으로 문제되지 않는다.

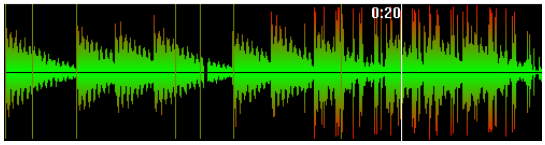
$$
\begin{aligned}
&IF\ |kVol - curVol| > \frac{kVol}{3}\ THEN\\
&\quad IF\ frame = 24\\
&\quad THEN\ NEW\ PARTITION\\
&\quad ELSE\ frame \leftarrow frame + 1\\
&ELSE\\
&\quad kVol = curVol\\
&\quad frame = 0\\
&ENDIF
\end{aligned}
$$

[Fig. 5] Pseudocode of extracting partition from sound

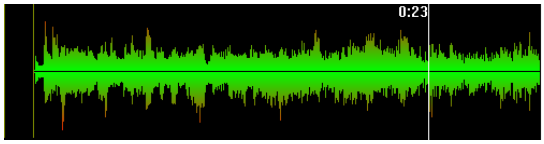
[Fig. 5]은 본 논문에서 제안하는 사운드 구간 추출 방법을 보여준다. kVol은 비교대상 볼륨 값, curVol은 현재 프레임의 볼륨 값이다. 본 논문에서는 볼륨의 1/3 크기의 갑작스런 변화 이후 그 상태(비교 볼륨 값과 현재 볼륨 값의 차이가 일정 값 이상 유지되는 상태)가 24 프레임(1초 / 24fps) 이상 지속될 때 새로운 감성 구간으로 인식한다.

3.3 각성 배경음 매핑

본 논문에서 각성도를 추출하는 첫 번째 제안하는 방법으로는 4000Hz 이상 주파수 대역에서의 평균 볼륨크기를 활용하는 것이다. 이는 주파수 대역을 활용하여 영상에서 목소리의 영향을 덜 받는 배경음을 추출하기 위함인데, 일반적으로 큰 소리일수록 각성으로 느끼기 쉽지만 단순 소리의 크기만으로는 각성도를 잘못 추출할 여지가 있다. 예를 들어, 사람의 목소리가 주를 이루는 특정 영상에서 실제 각성도가 낮음에도 각성도가 높은 수치로 나오는 경우가 있다. 이는 사람의 목소리를 강조하여 콘텐츠 이용자로부터 정보전달이 잘 이루어지게끔 영상을 제작, 편집하였기 때문으로 판단된다.



[Fig. 6] Wave form of emotional relaxes video



[Fig. 7] wave form of emotional arousal video

[Fig. 6]은 볼륨은 높지만 각성도가 낮은 A 영상의 볼륨 그래프이며, [Fig. 7]은 볼륨은 낮지만 실제 각성도가 높은 것으로 인식되는 B 영상의 볼륨 그래프이다. 두 영상을 비교했을 때, 단순 볼륨은 각성도와는 무관해 보인다. 따라서 단순 볼륨이 아닌 4000Hz 이상 주파수 대역의 평균 볼륨을 이용해야만 한다. 본 논문은 4000Hz이상 주파수 대역에서 평균 dB이 -60dB이상인 배경음을 각성 배경음으로 정의하고, 추출된 소리 구간에서 각성 배경음의 개수를 이용하여 각성도를 추출하는 것을 제안한다.

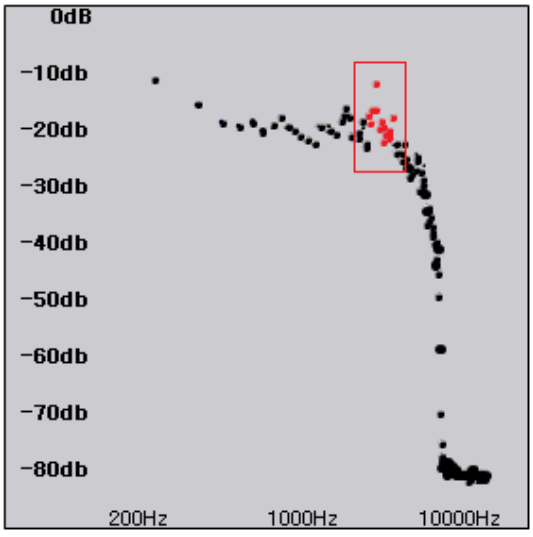
$$B_{1k} = c_1 \bullet (n_{1k} - 15)^3 \quad (n_{1k} \leqq 30) \qquad (1)$$

식 (1)은 각성 배경음을 통해 추출되어진 구간 각성도를 구하는 식이며, k는 구간번호, B_{lk}는 각성 배경음을 통해 추출되어진 구간 각성도, n_{lk}는 구간에서 추출한 각성 배경음의 초당 평균 개수, c_1은 정규화 수(5 / 153)이다. 각 구간에서 추출한 초당 각성 배경음의 평균 개수를 이용하여 구간의 각성도를 추출하고 전체 영상의 총 각성도는 구간 각성도의 평균값으로 한다.

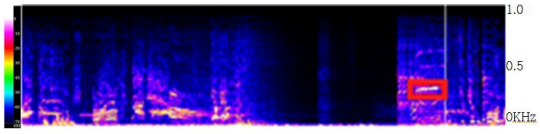
3.4 각성음 매핑

사운드에서 각성도를 추출할 때 두 번째 제안하는 방

법으로는 4000Hz 부근의 최대 볼륨 크기를 활용하는 것이다. 4000Hz 주파수 대역의 소리는 라우드니스가 가장 크기 때문에 큰 강도의 4000Hz 부근 사운드를 들은 사람은 청각적으로 예민하게 반응하고 따라서 각성도가 높아질 수 있다. 본 논문에서는 3200Hz ~ 45000Hz 범위의 주파수 대역에서 -15dB 이상이 검출될 때 각성음으로 추출한다.



[Fig. 8] Sound spectrogram of horror film



[Fig. 9] Log frequency graph of horror film

[Fig. 8]과 [Fig. 9]는 공포영화 예고편 영상에서 여자가 비명을 지르는 부분을 각각 로그 그래프와 스펙트로그램으로 나타낸 것이다. [Fig. 8]은 사운드의 로그 그래프를 나타낸 것이며 사각형 안의 빨간색 점으로 표현된 구간이 3200Hz ~ 4500Hz 구간이다. [Fig. 9]는 사운드의 스펙트로그램을 나타낸 것이며 빨간색으로 표시한 구간이 3200Hz ~ 4000Hz 주파수 대역이다. 공포영화에서 여자의 비명소리는 대표적인 각성음인데 [Fig. 8]과 [Fig. 9]에서 각각 해당 주파수 대역에 높은 반응이 있음을 확인할 수 있다. 각성 영상이 반드시 각성음이 두드러지는 것은 아니다. 하지만 각성음이 추출된 구간은 영상의 각성도가 높다고 가정을 하고 최종 구간 각성도는 구간 내에 추출된 각성음의 개수에 따라 식 (2)의 방식으로 나타낸다.

$$B_{2k} = c_2 \cdot n_{2k} \quad (2)$$

식 (2)에서 B_{2k}는 각성음을 통해 추출되어진 구간 각성도, n_{2k}는 구간 내 추출된 각성음의 개수이며, c_2는 정규화수 0.1 이다.

3.5 최종 각성도 매핑

$$B = \frac{1}{s}\sum_{k=0}^{s}(B_{1k} + B_{2k}) \quad (-5 \leqq B \leqq 5) \quad (3)$$

최종 각성도를 매핑하는 방법은 식 (3)으로 결정한다. 식 (3)에서 k는 구간번호, s는 영상 내 구간 총 개수이다. 최종 영상의 각성도는 구간 내 각성 배경음과 각성음을 통해 추출된 각성도 B_{1k}, B_{2k}의 합의 평균으로 나타낸다.

3.6 각성도 기대치

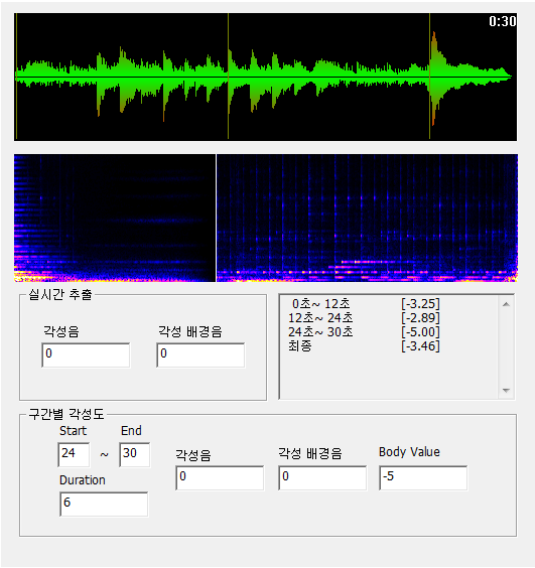
각성도는 -5 ~ +5 범위를 따르는 Russell의 감정 모델을 이용하였으며, -5에 가까울수록 이완을 +5에 가까울수록 각성을 나타낸다.

$$P = -t \cdot (t-20) \quad (4)$$

식 (4)는 -5 ~ 5 사이의 임의의 두 수가 t범위 내에 존재할 기대치 P를 나타낸 식이다. 식 (4)를 이용하여 오차범위 내에 두 수가 존재할 기대치를 구할 수 있다.

4. 실험 결과

연구의 검증성을 위하여 자체 데모를 제작하여 실험을 하였으며, 실험 환경은 VisualStudio 2010과 BASS라이브러리를 활용하여 제작하였으며, 프로그램은 영상에서 실시간으로 구간을 추출하며 구간의 각성도를 추출할 수 있게끔 설계되었다.



[Fig. 10] Proposed emotion extraction system

[Fig. 10]은 윈도우7의 샘플영상 "wildlife"를 분석한 영상 각성도 추출 시스템의 모습이다. 상단에는 볼륨 그래프와 스펙트로그램이며, 실시간 추출 탭의 각성음, 각성 배경음란은 해당 구간에서 현재까지 추출된 각성음과 각성 배경음의 개수이다. 구간별 각성도 탭은 이전 구간에서의 초당 평균 각성음, 각성 배경음을 출력하며 구간에서 최종 각성도를 표시한다. 실시간 추출 부분에서 오른쪽 탭은 추출된 구간과 구간에서의 각성도를 표시하며 영상 재생이 끝날 시 최종 각성도를 표시한다.

〈Table 1〉 Video's arousal indice

Arousal indice	Arousal degree	count
-5 ~ -3	strong relax	9
-3 ~ -1	low relax	4
-1 ~ 1	normal	16
1 ~ 3	low arousal	16
3 ~ 5	strong arousal	5

[Fig. 10]에서 "wildlife"는 볼륨 그래프에서 3개의 구간으로 추출된 모습을 확인할 수 있으며, 최종 각성도는 -3.46이 나온 것을 확인할 수 있다.

<Table 1>은 실험의 검증성을 위하여 무작위로 선정된 총 50개 영상을 100명에게 설문하여 실제 사람이 느끼는 각성도를 얻은 결과이다. 감성을 강 이완, 약 이완, 보통, 낮은 각성, 높은 각성으로 총 5가지로 분류할 때 <Table 1>과 같은 결과를 얻을 수 있었다.

<Table 2>는 설문조사 결과 값에 따라 알고리즘 일치도를 나타낸 결과이다. 오차범위는 설문한 각성지수와 데모를 통해 추출한 각성지수 두 값의 일치 범위를 나타낸 것이며 오차범위 내에 존재하였을 경우 일치하였다고 판단한다. 기대치는 식 (4)를 이용하여 추출하였다. 오차범위가 1.5일 때 일치율은 46%, 2일 때 62%, 2.5일 때 70%로 나타났다. 이때, 오차범위를 2로 설정하는 것이 오차범위를 최소화하며 60%이상의 만족할만한 일치율을 보여주므로 다른 오차범위 설정 값보다 적당하다.

〈Table 2〉 Concordant results

margin of error	1.5	2	2.5
concordance count	23	31	35
non-concordance count	27	19	15
Expectations(P)	28%	36%	44%
concordance rate	46%	62%	70%

<Table 3>는 각성도를 5가지로 분류하여 오차범위 2 이내에 설문조사 결과 값에 따라 알고리즘 일치도를 나타낸 결과이다. 그 결과, 강 각성, 강 이완 영상일수록 더 높은 일치도로 나타났음을 볼 수 있다.

〈Table 3〉 Concordant results according to emotion of media contents

Arousal indice	Arousal degree	concordance rate
-5 ~ -3	strong relax	8 / 9 (88 %)
-3 ~ -1	low relax	3 / 4 (75 %)
-1 ~ 1	normal	9 / 16 (56 %)
1 ~ 3	low arousal	7 / 16 (43 %)
3 ~ 5	strong arousal	4 / 5 (80 %)

5. 결론

본 연구에서는 멀티미디어 콘텐츠에서 소리 정보를 이용하여 각성지수를 추출하였다. 오차범위 2의 범위내에 62%의 높은 일치도를 볼 수 있고, 특히 강한 각성 영상과 강한 이완 영상에서 각각 80%, 88%의 높은 일치율

을 보여주었다.

본 논문에서는 소리 특정 주파수 대역의 볼륨값을 활용하여 각성도를 추출하였지만, 향후 소리의 날카로움이나 음색을 활용하여 각성도를 추출하는 연구가 필요할 것이며, 또한 멀티미디어 콘텐츠에서 소리를 제외한 이미지 기반에서 각성도를 추출하거나, 동영상의 영상 내부 움직임을 판별하여 각성도를 추출하는 연구 또한 필요할 것이다.

ACKNOWLEDGMENTS

This research was supported in by Hanshin University Research Grant.

REFERENCES

[1] Sang Hoon Jeong, "Development direction of emotional contents through analysis of successful cases from applying emotional technology", *KOSES, Vol. 15(1)*, pp. 121-132, March, 2012.

[2] M. W. Park, S. M. Ahn, S. D. Ha, D. U. Jeong, I. K. Lyoo, "Development of Emotion Contents Recommender System for Improvement of Sentimental Status", *KOSES, Vol. 10(1)*, pp. 1-11, March, 2007.

[3] Li, T., Ogihara, M., "Detecting emotion in music", *ISMIR, Vol. 3*, pp. 239-240, Oct, 2003.

[4] W. H. Cho, J. K. Lee, H. K. Choi, "A study on the influence of audio stimulation to human sensibility by using EEG analysis", *KSPE Autumn Conference Vol. 11*, pp. 875-876, 2011.

[5] J. Russell, "Two pancultural dimensions of emotion words,", *Journal of Personality and Social Psychology Vol.45*, pp.1285, 1983.

[6] Loudness: http://en.wikipedia.org/wiki/Loudness

[7] D. H. Moon, "The Effect on Human Body by the Stimuli of Musics and Acoustic Vibrations", *KSPE Vol 12(5)*, pp. 278-282, Nov, 2007.

[8] White Noise: http://en.wikipedia.org/wiki/White_noise

[9] J. H. Kim, M. C. Hwang, J. C. Woo, J. S. Kim, W. M. Choi, J. S. Yun, B. C. Hwang, "A Research on masking effect for mixing sound with white noise on human relaxation", *HCI 2009*, pp. 319-323, Feb, 2009

[10] J. H. Kim, M. C. Hwang, J. C. Woo, J. S. Kim, W. M. Choi, J. S. Yun, B. C. Hwang, "The effect of white noise on relaxation", *Ergonomics Society of Korea, Autumn Conference symposium 2008*, pp. 552-555, January, 2008

권 영 훈(Kwon, Young-Hun)

・2008년 3월 ~ 현재 : 한신대학교 컴퓨터공학부
・E-Mail : yhkwon44@gmail.com

장 재 건(Chang, Jae-Khun)

・1985년 2월 : 한양대학교 건축학과 (공학사)
・1989년 8월 : New Jersey Institute of Technology 컴퓨터정보학과 (MS)
・1997년 2월 : Univ. of South Carolina 컴퓨터과학과(Ph.D)
・1997년 3월 ~ 현재 : 한신대학교 컴퓨터공학부 교수
・관심분야 : 컴퓨터비전, 영상처리, ITS, 패턴인식
・E-Mail : jchang@hs.ac.kr