화자 인식을 통한 등장인물 기반의 비디오 요약

Character-Based Video Summarization Using Speaker Identification

이 순 탁*, 김 종 성*, 강 찬 미*, 백 중 환*

Soon-Tak Lee*, Jong-Sung Kim*, Chan-Mi Kang*, Joong-Hwan Baek*

요 약

본 논문에서는 인물 기반의 비디오 요약 방법으로써 비디오 내 음성정보를 이용하여 화자 인식 기법을 통한 등장인물 중심의 요약 기법을 제안한다. 먼저, 얼굴 영역을 포함하는 장면을 중심으로 비디오로부터 배우의 대사에 해당하는 음성 정보를 분리하고, 화자 인식 기법을 수행하여 등장인물 별로 분류하였다. 화자인식 기법은 각 화자별로 MFCC(Mel Frequency Cepstrum Coefficient) 값을 추출하고 GMM(Gaussian Mixture Model)을 이용하여 분류한다. 본 논문에서는 4명의 등장인물에 대해 GMM을 학습시키고 4명 중 1명을 검출하는 실험을 통해 학습된 GMM 분류기가 실험 비디오에 대해 0.138 정도의 오분류율을 보임을 확인하였다.

Abstract

In this paper, we propose a character-based summarization algorithm using speaker identification method from the dialog in video. First, we extract the dialog of shots containing characters' face and then, classify the scene according to actor/actress by performing speaker identification. The classifier is based on the GMM(Gaussian Mixture Model) using the 24 values of MFCC(Mel Frequency Cepstrum Coefficient). GMM is trained to recognize one actor/actress among four who are all trained by GMM. Our experiment result shows that GMM classifier obtains the error rate of 0.138 from our video data.

Key words : Content-based summarization, Speaker identification, MFCC, GMM, Minimum distance classifier.

Ⅰ. 서론

최근 멀티미디어 정보의 빠른 성장에 따라 내용 기반의 비디오 분석, 요약 및 검색에 대한 관심이 급속히 증가하고 있다. 내용 기반의 비디오 분석은 원 비디오의 구성을 파악하고 그 의미론적인 내용을 이해하는 것을 목표로 한다[1]. 비디오 요약의 대표적인 방법인 비디오 색인에 관한 연구로 장면들을 유사한 장면끼리 그룹화 하여 스토리를 형성하는 줄거리 요약 연구가 행해지고 있다[2]. 스토리 중심의 비디오 요약이 장면 중심의 비디오 요약보다 고차원의 비디오 구조를 제공하지만, 비디오 내 모든 장면이 의미를 갖는 하나의 주제를 구성하기 위해 존재하는 것은 아니다. 본 논문에서는 사용자의 요구에 맞춰 더 구체적이고 정렬된 비디오 인덱싱을 위해 등장인물 중심의 비디오 요약에 관한 연구 방법을 제안한다. 등장인물 중심의 비디오 요약에는 시각 정보와 청각 정보를 사용하는 방법이 있고. 기존의 연구는 주로 시각 정보를 이용한 비디오 분석이 주를 이루었다. 그러나 시각 정보만을 이용한 비디오 요약은 알고리즘이 복잡하고 연산이 많이 요구돼 시간적인 측면에서 효율이 떨어진다. 따라서 본 논문에서는 효율적인 비디오 요약을 위해 시각 정보만이 아닌 등장인물의 음성 정보를 이용하는 비디오 요약 방법에 대해 연구한다.

사람의 음성 신호는 성대의 주기적인 떨림에 의해서 생성되는 피치를 갖고 있으며 음성 신호의 피치는 개인별 고유한 특성을 갖고 있다. 또한, 사람마다 성도와 비도로부터 발생하는

*한국항공대학교 정보통신공학과
접수 일자 : 2005. 7. 18 수정 완료: 2005. 10. 19
논문 번호 : 2005-3-6
※본 논문은 2004년도 한국항공대학교 교비지원 연구비에 의하여 지원된 연구의 결과임.

독특한 공진 주파수(formant)를 갖고 있다. 본 논문에서는 이러한 특성을 잘 표현하는 MFCC(mel frequency cepstrum coefficient)를 이용하여 얼굴 영역을 포함하는 장면(shot)의 음성 신호로부터 화자들을 클러스터링 한다. 기존의 비디오 요약연구에서는 화자인식을 위해 전체 오디오 시퀀스에서 사람의 음성부분을 추출하는 경우가 대부분이어서 처리 속도와 정확도 측면에서 성능이 저하되었다. 본 연구에서는 우선 시각정보를 이용하여 사람이 등장하는 부분이라 판별된 샷에서의 음성신호를 사용하여 시스템을 간략화 하였다.

본 논문의 2장에서는 화자 인식을 위해 사용한 특징값인 MFCC에 대해, 3장에서는 화자 식별을 위한 분류기법인 가우시안 혼합 분포 모델(GMM)에 대해서 각각 설명한다. 4장에서는 화자 인식 성능을 평가하고 실험 결과를 살펴보며, 마지막으로 5장에서는 제안한 비디오 요약 기법에 대해 결론을 맺는다.

II. Mel Frequency Cepstral Coefficients

음성인식에 쓰이는 특징값으로 Linear Prediction Coefficients나 Linear Prediction Spectrum 등과 같은 많은 방법이 존재하지만 주파수를 피쳐로 이용하였을 때 잡음의 영향을 덜 받고 효과적인 것으로 나타났다[6]. 그 중 Mel Frequency Cepstral Coefficients (MFCC)는 음성인식, 화자 인식 등에 널리 쓰이는 유효한 특징 값 중 하나이며 음성 스펙트럼을 표현하기위해 Mel 주파수 필터로부터 계산된다[3][4]. 그림 1의 블록다이어그램은 MFCC를 추출하는 과정을 나타낸다.

식 (2)의 N는 FFT(Fast Fourier Transform)의 길이를 나타낸다. 주파수 변환된 $X(n, w_k)$의 크기(magnitude)는 필터 시퀀스의 주파수 응답에 의해 가중화되고, 이러한 필터 시퀀스는 저주파수(1000Hz이하)에서는 필터 중심주파수와 대역폭이 선형적이지만 주파수가 높아질수록 로그 스케일로 증가하는 특성을 갖고 있다. 이것은 저주파 영역의 신호에서 인간의 청각 특성이 민감한 반면 고주파 영역의 신호에서는 민감하지 않은 특성을 적용한 것이다[3].

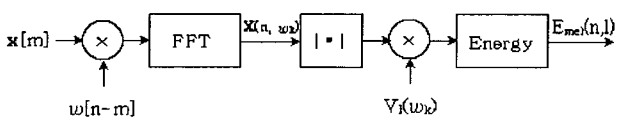

그림 1. MFCC의 추출 과정
Fig. 1. Block Diagram to Extract MFCC

$$X(n,\omega_k)=\sum_{m=-\infty}^{\infty}x[m]w[n-m]e^{-jw_k m} \quad (1)$$

$$w_k=\frac{2\pi}{N}k \quad (2)$$

그림 2는 Davies와 Mermelstein에 의해 제안된 멜 스케일(mel-scale) 필터 뱅크의 예를 보여주고 있다[5].

필터뱅크의 중심 주파수는 Mel 스케일로 존재하게 되며 수식 (3)을 이용하여 멜 스케일을 계산한다. 전체 주파수 대역을 n으로 나눈 등 간격으로 대역을 나누게 된다.

$$Mel(f)=2595\times\log_{10}(1+\frac{f}{700}) \quad (3)$$

$$BW=\begin{cases}25+75[1+1.4(\frac{f}{1000})^2]^{0.69} & f>1000\\ 1000 & f<1000\end{cases} \quad (4)$$

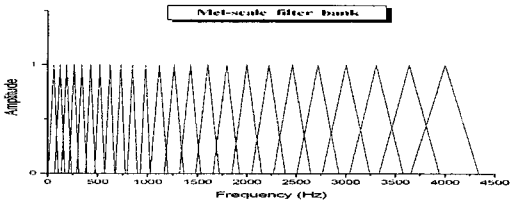

그림 2. 24차의 Mel-scale 필터 뱅크
Fig. 2. Mel-scale Filter Bank with 24 Order

n번째 필터의 중심주파수는 n번째 간격에 대응되는 주파수가 된다. 각 필터의 대역폭은 식 (4)의 critical bandwidth에 의해 결정된다. Mel 스케일 필터 뱅크의 첫 번째 필터의 주파수 응답을 $V_L(m)$라고 하면 n번째 음성 프레임에 대한 Mel 에너지는 식 5로 표현할 수 있다. L_l, U_l은 l번째 필터에서 '0'이 아닌 주파수 영역의 상한, 하한 값을 의미한다. 식 (6)은 다양한 대역폭을 갖는 필터들의 균일한 스펙트럼을 위한 정규화 과정이다.

$$E_{mel}(n,l)=\frac{1}{A_l}\sum_{k=L_l}^{U_l}|V_{l(w_k)}X(n,w_k)|^2 \quad (5)$$

$$A_l=\sum_{k=L_l}^{U_l}|V_l(w_k)|^2 \quad (6)$$

식 (7)에서처럼 Mel 에너지를 DCT(discrete cosine transform)를 취하여 멜 켑스트럼(mel cepstrum)을 구할 수 있다. DCT를 통하여 Mel 스케일 에너지를 무상관된(decorrelated) M차의 계수로 변환할 수 있다. 식 (6)을 이용하여 R개의 필터로 구성된 필터 뱅크 중 n번째 음성 프레임에 대한 m번째 계수를 계산한다.

$$C_{mel}[n,m]=\frac{1}{R}\sum_{l=0}^{R-1}\log E_{mel}(n,l)\cos\left(\frac{2\pi}{R}lm\right) \quad (7)$$

Ⅲ. 등장인물 인식

3.1 전처리

화자의 묵음 구간을 제거하기 위해서 음성 신호의 에너지를 이용하는 방법과 영교차(zero-crossing) 방법을 이용한다. 본 논문에서는 에너지를 이용하여 문턱 값 이상의 에너지를 갖는 음성 프레임에 대해 MFCC 특징 값을 사용한다.

MFCC 추출에 앞서 디지털 음성 신호는 인간의 외이 및 중이의 주파수 특성을 모델링하기 위하여 고대역 통과 특성을 갖는 디지털 프리엠퍼시스(pre-emphasis) 필터를 거친다[7]. 프리엠퍼시스 필터의 특성 함수 H(z)는 $H(z)=1-\alpha z^{-1}$와 같으며 a의 값은 0.97로 설정하였다.

필터를 통과한 음성 신호는 해밍 윈도우를 이용하여 프레임 단위로 분할된다. 프레임의 크기는 MFCC 추출에 적합한 25ms로 설정하였으며 윈도우의 이동 간격은 10ms로 설정하였다.

3.2 가우시안 혼합 모델

Gaussian Mixture Model(GMM)은 그림 3에서처럼 M개의 요소 분포를 가중치와 함께 합산한 것으로 문장 독립(text-independent) 화자 인식 시스템을 위한 화자 발성의 음향학적인 분포를 표현함에 있어서 매우 뛰어난 것으로 나타났다[4]. 다수의 화자 음성으로부터 추출된 MFCC 특징 값을 이용하여 GMM 분류기를 훈련시킨 후 실시간 화자 인식 시스템에 적용할 수 있다.

가우시안 혼합 분포는 식 (8)로 표현되며 x는 D차원의 랜덤 벡터, $b_i(x)$는 요소 분포(component density), p_i는 i번째 요소분포에 대한 가중치를 의미한다. 이때 가중치 p_i는 $\Sigma_{i=1}^{M} p_i = 1$를 만족해야 한다.

$$p(x|\lambda) = \sum_{i=1}^{M} p_i b_i(x) \tag{8}$$

$$b_i(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\}} \tag{9}$$

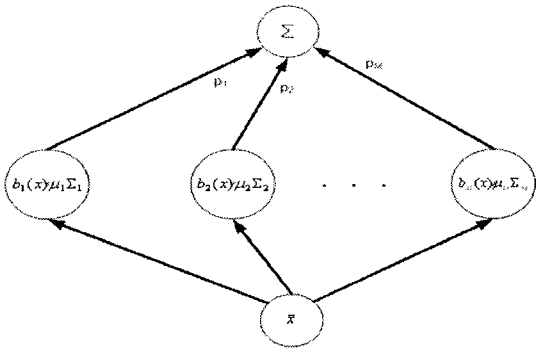

그림 3. M개의 가우시안 혼합 모델의 개념도
Fig. 3. Gaussian Mixture Model with M Components)

가우시안 혼합 모델에서 확률 밀도 함수 $p(x|\lambda)$는 각 모드(mode)에 대한 평균 벡터, 공분산 매트릭스, 가중치에 관한 함수이며 식 (10)에서처럼 3개의 매개변수를 훈련 과정에서 모델링한다. 즉, 훈련 샘플을 이용하여 각 화자에 대한 λ를 추정한다.

$$\lambda = \{p_i, \mu_i, \Sigma_i\}, \quad i = 1, \ldots, M \tag{10}$$

화자 인식을 위해 각 화자들은 하나의 GMM으로 표현되며, 이는 각 화자의 λ에 의한다.

GMM의 훈련은 Maximum Likelihood 추정 방법을 이용하여 식 (11)에 있는 GMM의 우도함수(Likelihood Function)를 최대화할 수 있는 매개변수 λ를 추정한다. 식 (10)은 훈련 샘플의 T차원의 특징 벡터 $X = x_1, x_2, \ldots x_T$에 대한 가우시안 혼합 모델에 대한 우도함수를 의미한다. 이러한 우도함수를 최대화할 수 있는 매개변수를 추정하기 위해 반복 알고리즘인 EM(Expectation Maximization)을 통해서 GMM의 매개변수를 추정한다.

$$p(x|\lambda) = \prod_{i=1}^{T} p(x_i|\lambda) \tag{11}$$

EM 알고리즘의 기본 개념은 초기 모델 λ인 혼합 모델에 $p(x|\lambda)$에 대해서 $p(x|\overline{\lambda}) \geq p(x|\lambda)$를 만족하는 새로운 모델 $\overline{\lambda}$를 추정하는 것이며 다음번의 순환 과정에서 새로운 모델은 초기 모델이 되며 특정 오차 수준에 수렴하거나 최대 순환 횟수를 만족할 때까지 반복하게 된다.

훈련된 GMM을 이용하여 새로운 음성 샘플은 식 (12)의 사후(posteriori) 확률이 최대가 되는 클래스로 분류하게 된다.

$$\hat{S} = \arg \max_{1 \le k \le S} \Pr(\lambda_k | X) = \arg \max_{1 \le k \le S} \frac{p(X|\lambda_k)\Pr(\lambda_k)}{p(X)} \quad (12)$$

임의의 음성 프레임이 입력되었을 때 학습된 GMM은 각 클래스의 사후 확률 값을 계산한다. 또한 입력된 음성 신호는 다수의 프레임을 갖고 있으므로 프레임들에 대한 사후 확률 평균값이 최대가 되는 화자 클래스로 음성 신호를 분류한다. 따라서 분류 규칙인 수식 (12)는 (13)으로 표현할 수 있다.

$$\hat{S} = \arg \max_{1 \le k \le S} \left[\frac{1}{F} \sum_{i=1}^{F} p(X_i | \lambda_k) \right] \quad (13)$$

3.3 최소 거리 분류 기법(Minimum Distance Classifier)

본 논문에서는 가우시안 혼합 모델의 성능과 최소 거리 분류 기법에 의한 화자 인식 성능을 비교한다. 최소 거리 분류 기법은 다수의 음성 프레임들에 대해서 수식 (14)를 이용하여 훈련 샘플 및 테스트 샘플들의 특징 벡터별 평균을 계산한다[3]. M은 음성 프레임의 수, L은 프레임의 길이, n은 특징 값의 차수, $-tr$은 훈련 샘플, $-ts$는 테스트 샘플을 의미한다.

$$\overline{C}_{mel}^{ts}[n] = \frac{1}{M} \sum_{m=1}^{M} C_{mel}^{ts}[mL, n],$$

$$\overline{C}_{mel}^{tr}[n] = \frac{1}{M} \sum_{m=1}^{M} C_{mel}^{tr}[mL, n] \quad (14)$$

따라서 훈련 샘플과 테스트 샘플에 대해 식 (15)의 평균 자승 오차를 이용하여 거리를 측정한다. R은 MFCC의 계수의 수, 즉 특징 값의 차수를 의미한다. 최소 거리 분류 기법에서는 평균 자승 오차 D가 문턱 값 T 이하일 때 훈련 샘플과 동일한 클래스, 즉 화자로 인식한다.

$$D = \frac{1}{R} \sum_{n=1}^{R} (\overline{C}_{mel}^{ts}[n] - \overline{C}_{mel}^{tr}[n])^2 \quad (15)$$

Ⅳ. 실험 및 고찰

실험에 사용된 음성 시퀀스는 비디오 영상에 등장하는 배경 음악, 웃음소리, 묵음 등을 제외한 등장인물의 음성만을 분리하여 GMM 분류기를 학습시켰다. 본 논문에서는 뉴스나 스포츠 비디오에 비해 그 종류가 다양하고 편집 방식이 복잡한 드라마 '대장금'을 실험 비디오로 사용하였다. 드라마는 주로 인물과 그 인물의 대사에 초점을 맞춰 사건을 전개하기 때문에, 화자인식을 통한 비디오 요약에 대한 본 논문의 실험 데이터로 적합한 조건을 갖고 있다. 선택된 실험 비디오로부터 음성 신호를 추출하여 등장인물 4명 중 1명만을 목표로 하는 두 클래스 분류기를 학습시켰다. 표 1은 화자 인식에 사용된 훈련 샘플 및 성능 분석을 위한 실험 비디오에 대해 보여준다.

표 1. 화자 인식을 위한 실험 데이터
Table 1.Experiment Data for the Speaker Identification

	클래스 A	클래스 B			
	등장인물 1 (여)	등장인물 2 (여)	등장인물 3 (남)	등장인물 4 (남)	기타
훈련 샘플(ms)	194,000	55,740	70,860	67,400	·
테스트 샘플 (장면 수)	12	5	7	5	22

본 논문에서 화자 인식을 위한 MFCC 특징 값을 표 2와 같이 12차, 24차의 MFCC 특징 값 모두와 300 ~ 3000Hz로 제한된 음성 대역에 포함되지 않은 4개의 계수를 제외한 4가지 경우에 대해 실험을 하였다.

표 2. 4가지 경우의 MFCC 특징 값
Table 2. 24 Sets of MFCC Features

	Set 1	Set 2	Set 3	Set 4
MFCC 범위	3~10	1~12	3~22	1~24
총 차수	8	12	20	24

표 3은 특징 값 차수의 변화에 따른 훈련된 가우시안 혼합 모델을 실험 영상에 적용했을 때 클래스별 오분류 확률을 보여주고 있고, MFCC 계수의 차수가 3 ~ 20인 Set 3에서 최소의 오분류 확률을 보여주고 있다.

표 3. GMM 분류기의 오분류 확률
Table 3. Misclassification Probability of the GMM Classifier

특징 값 / 오분류	Set 1	Set 2	Set 3	Set 4
$P(\varepsilon\|A)$	0.207	0.172	0.138	0.138
$P(\varepsilon\|B)$	0.0	0.034	0.0	0.034

특징 값 Set 3에 의한 실험 영상에서 등장인물 A를 검출한 결과를 그림 4에 보인다. 6332 번째 프레임은 시각 정보에서는 등장인물 C를 포함하고 있지만 청각 정보에

서는 등장인물 A의 음성 정보를 포함하고 있기 때문에 시각 정보 측면에서 오검출된 예이다. 이러한 시청각 정보가 서로 다른 등장인물을 포함하고 있을 경우는 오분류 확률에 포함하지 않았다. 이러한 오류는 시각 정보를 이용하여 입술 움직임 추적 등의 처리를 통하여 보완 되어야 한다. 그림 5는 등장인물 A의 음성 정보를 포함하지만 검출하지 못한 장면의 프레임을 보여주고 있다. 707번째 프레임이 속한 장면에서는 대부분 배경음악으로 구성되었고 나머지 세 개의 프레임에서는 등장인물의 음성 정보가 미약하거나 짧았기 때문에 검출에 실패하였다. 음성정보가 미약하면 전처리 과정에서 묵음 구간으로 간주되어 소거되고, 음성 구간이 너무 짧으면 사람 음성의 특성을 파악할 수 없기 때문에 화자 인식이 원활하게 이루어지지 않는다.

본 논문에서는 가우시안 혼합 모델(GMM)과 3.3절의 최소 거리 분류 기법에 대한 성능을 비교하였다. 식 (15)에서 거리 D가 문턱 값 T 이하를 만족할 때 테스트 샘플은 훈련 샘플의 화자와 동일한 화자로 인식하게 된다.

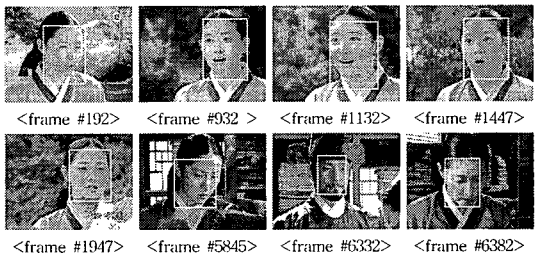

그림 4. 화자 인식을 이용한 등장인물 A를 포함하는 장면 검출

Fig. 4. The Result of Shot Detection including Actor 'A' using Speaker Identification

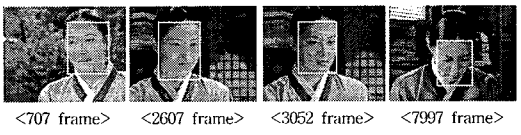

그림 5. 화자 인식에 실패한 등장인물 A의 장면

Fig. 5.The Shots failed to Detect Shots including Actor 'A'

4가지 경우의 특징 값에 대하여 T를 0.0 1~ 0.5까지 변화하면서 오분류 확률에 대한 실험을 하였다. 그림 6은 표 4에 있는 각 경우에 대한 오분류 확률을 특성 곡선(characteristic curve) 또는 ROC(receiver operating characteristic) 곡선으로 표현한 것이다. ROC 곡선은 두 클래스에 대한 오분류 확률의 보상관계(trade-off)를 시각적으로 잘 표현된다. 따라서 ROC 곡선을 이용하여 두 클래스에 대한 최적의 보상관계를 갖는 문턱 값을 결정할 수 있다. 그림 6에서 원점으로부터 가장 가까운 경우가 최적의 문턱 값이라고 할 수 있다. 따라서 표 4의 음영부분은 최소의 오분류 확률 갖는 경우라 할 수 있다. 하지만 최소 거리 분류 기법에 의한 오분류 확률은 가우시안 혼합 모델보다 훨씬 높으며 이것은 클래스, 즉 화자 인식 성능이 낮다고 할 수 있다.

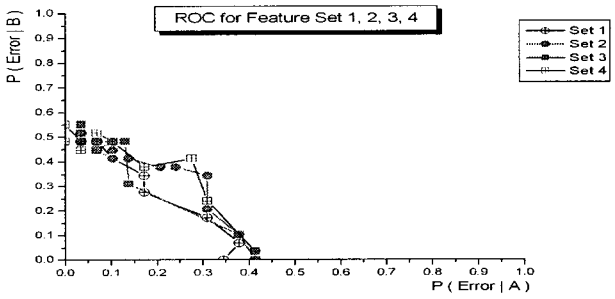

그림 6. 문턱 값(T)에 따른 ROC 곡선

Fig. 6. ROC Curve for Misclassification Probability in Table 4

표 4. 문턱 값에 따른 오분류 확률

Table 4. Misclassification Probability according to Various Threshold Value T

문턱 값 (T)	Set 1		Set 2		Set 3		Set 4	
	$P(\partial A)$	$P(\partial B)$	$P(\partial A)$	$P(\partial B)$	$P(\partial A)$	$P(\partial B)$	$P(\partial A)$	$P(\partial B)$
0.010	0.414	0.000	0.414	0.000	0.414	0.000	0.414	0.035
0.040	0.345	0.000	0.414	0.000	0.379	0.103	0.310	0.241
0.070	0.379	0.069	0.414	0.035	0.138	0.310	0.276	0.414
0.100	0.310	0.172	0.379	0.103	0.130	0.483	0.172	0.379
0.130	0.172	0.276	0.310	0.207	0.069	0.483	0.069	0.448
0.160	0.172	0.345	0.310	0.345	0.035	0.483	0.103	0.483
0.190	0.103	0.414	0.241	0.379	0.035	0.517	0.103	0.483
0.220	0.069	0.483	0.207	0.379	0.035	0.517	0.103	0.483
0.250	0.069	0.483	0.138	0.414	0.035	0.517	0.103	0.483
0.280	0.069	0.483	0.103	0.414	0.035	0.517	0.069	0.483
0.310	0.035	0.483	0.069	0.448	0.035	0.552	0.069	0.517
0.340	0.035	0.483	0.103	0.483	0.035	0.552	0.035	0.448
0.370	0.035	0.483	0.103	0.483	0.035	0.552	0.000	0.483
0.400	0.035	0.483	0.103	0.448	0.035	0.552	0.035	0.483
0.430	0.035	0.483	0.103	0.483	0.035	0.552	0.035	0.483
0.460	0.035	0.517	0.103	0.483	0.035	0.552	0.035	0.517
0.490	0.035	0.517	0.103	0.483	0.035	0.552	0.000	0.552

V. 결론

멀티미디어 정보의 빠른 성장에 따라 내용 기반의 비디오 분석, 요약 및 검색에 대한 관심이 급속히 증가하고 있다. 이러한 관심에 발 맞춰 줄거리, 사건, 인물 등에 초점을 맞춰 요약하는 연구들이 행해지고 있다. 본 논문에서는 인물 중심 요약 중에서 얼굴 영역이 포함된 장면의

배우의 대사에 해당하는 음성 정보를 비디오로부터 분리하여 화자 인식을 통해 해당 장면의 등장인물을 구별하였다. 화자 인식을 위해서는 사람 음성의 특징을 잘 나타내 화자 인식 기법에 주로 사용되는 MFCC 특징 값을 추출하여 가우시안 혼합 모델(GMM)을 통해 특정 인물을 분류하였다. 추출된 MFCC의 4가지 특징 벡터에 따른 분류 성능 실험 결과 가우시안 혼합 모델에서는 24차의 필터 뱅크로부터 추출된 24차의 MFCC 계수에서 처음과 마지막 2계수를 제외한 20차의 특징 벡터가 가장 우수한 성능 결과를 보였다. 또한, 가우시안 혼합 모델과 최소 거리 분류 기법의 성능을 비교 하였으며 최적의 가우시안 혼합 모델의 오분류(misclassification error)율이 최소 거리 분류 기법 보다 평균 0.3 정도 낮았다.

참 고 문 헌

[1] Ying Li and C.-C. Jay Kuo, "Content-Based Movie Analysis and Indexing Based on AudioVisual Cues", IEEE Transactions on circuits and systems for video technology. vol. 14, no. 8, pp 1073-1085, Aug 2004

[2] Minerva Yeung et al., "Segmentation of Video by Clustering and Graph Analysis", *Computer Vision and Image Understanding*, vol. 71, no. 1, pp. 94~109, 1998.

[3] Thomas F. Quantieri, *Discrete-Time Speech Signal Processing*, Prentice Hall, 2001

[4] Douglas A. Reynolds, Richard C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Trans. on Speech and Audio Processing*, vol 3, no. 1, pp. 72-83, Jan 1995.

[5] S. B. Davis, P. Mermelstein, "Comparison of Parametric Representations for Mono- Syllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol 28, no. 4, pp. 357-366, Aug 1980.

[6] B.Atal, "Automatic recognition of speakers from their voices" *Proc.IEEE* vol.64 pp 460-475 apr. 1976

[7] L.P. Cordella, P.Foggia, C.Sansone, M.Vento, "A Real-Time Text-Independent Speaker Identification System", *IEEE Proceeding of the 12th ICIAP*, pp. 632-637, Sept. 2003.

이 순 탁(Soon-Tak Lee)
1998년 9월 대학원 졸업(MS)
1998년 6월-2000년 8월 한국휴렛팩커드, 에질런트테크놀로지스 계측기 연구소
2000년 8월~현재 (주) 텔레칩스 미디어 연구소 선임연구원
관심분야 : 비디오 데이터베이스, 비디오 요약, 비디오 코딩, 영상 처리, 패턴 인식

김 종 성(Jong-Sung Kim)
2003년 2월 한국항공대학교 항공통신정보 공학과 졸업(공학사)
2005년 2월 한국항공대학교 대학원 정보통신공학과 졸업(공학석사)
2005년 1월~현재 LG전자 단말연구소 연구원
관심분야 : 비디오 요약 및 검색, 영상 압축, 영상 처리, 패턴 인식, 멀티미디어

강 찬 미(Chan-Mi Kang)
2004년 2월 한국항공대학교 항공통신정보공학과 졸업(공학사)
2004년 3월~현재 한국항공대학교 대학원 정보통신공학과 석사과정
2005년 1월~현재 LG전자 단말연구소 연구원
관심분야 : 비디오 요약 및 검색, 음성 신호 처리, 영상 처리, 패턴 인식

백 중 환(Joong-Hwan Baek)
1981년 2월 한국항공대학교 항공통신공학과 졸업(공학사)
1987년 7월 오클라호마주립 대학원 전기 및 컴퓨터공학과(공학석사)
1991년 7월 오클라호마주립 대학원 전기 및 컴퓨터공학과(공학박사)
1992년~현재 한국항공대학교 항공전자 및 정보통신공학부 교수
관심분야 : 영상처리, 패턴인식, 영상압축, 멀티미디어