한국음향학회지 제27권 제5호 pp. 244~251 (2008)

비교사 토론 인덱싱을 위한 시청각 콘텐츠 분석 기반 클러스터링

Audio-Visual Content Analysis Based Clustering for Unsupervised Debate Indexing

금 지 수*, 이 현 수*
(Ji-Soo Keum*, Hyon-Soo Lee*)

*경희대학교 컴퓨터공학과
(접수일자: 2008년 2월 4일; 수정일자: 2008년 5월 21일; 채택일자: 2008년 6월 20일)

본 연구에서는 시청각 정보를 이용한 비교사 토론 인덱싱 방법을 제안한다. 제안하는 방법은 BIC (Bayesian Information Criterion)에 의한 음성 클러스터링 결과와 거리기반 함수에 의한 영상 클러스터링 결과를 결합한다. 시청각 정보의 결합은 음성 또는 영상 정보를 개별적으로 사용하여 클러스터링할 때 나타나는 문제점을 줄일 수 있고, 토론 데이터의 효과적인 내용 기반의 분석이 가능하다. 제안하는 방법의 성능 평가를 위해 서로 다른 5종류의 토론 데이터에 대해 음성, 영상 정보를 개별적으로 사용할 때와 두 가지 정보를 동시에 사용할 때의 성능 평가를 수행하였다. 실험 결과 음성과 영상 정보를 결합한 방법이 음성, 영상 정보를 개별적으로 사용할 때 보다 토론 인덱싱에 효과적임을 확인하였다.
핵심용어: 시청각 콘텐츠 분석, 비교사 토론 인덱싱, 화자 인덱싱
투고분야: 음성처리분야 (2.4)

In this research, we propose an unsupervised debate indexing method using audio and visual information. The proposed method combines clustering results of speech by BIC and visual by distance function. The combination of audio-visual information reduces the problem of individual use of speech and visual information. Also, an effective content based analysis is possible. We have performed various experiments to evaluate the proposed method according to use of audio-visual information for five types of debate data. From experimental results, we found that the effect of audio-visual integration outperforms individual use of speech and visual information for debate indexing.
Keywords: Audio-Visual Content Analysis, Unsupervised Debate Indexing, Speaker Indexing
ASK subject classification: Speech Signal Processing (2.4)

I. 서 론

최근 방송 뉴스와 영화, 스포츠 그리고 토론에 대해 시청자의 관심이 높아지고 있는 가운데 이러한 멀티미디어 데이터를 효과적으로 검색 (retrieval)하거나 인덱싱 (indexing) 또는 내용을 이해 (understanding)하려는 연구가 활발하게 진행되고 있다 [1][2].

특히 정치, 경제, 사회, 교육 등 다양한 주제의 토론에 대해 시청자의 관심이 높아지고 있으며, 각 방송사에서는 토론을 정규방송으로 편성하여 심도 있게 방송하고 있다. 따라서 토론 데이터에 대한 연구도 기존의 뉴스나 영화, 스포츠에 대한 연구만큼 중요하다고 할 수 있다. 토론 데이터를 분석하면 토론자의 의견을 자동으로 요약할 수 있을 뿐만 아니라 토론의 내용 기반 구성이 가능하며, 시청자의 요구에 따른 검색을 효과적으로 수행할 수 있다 [3][4].

내용 기반 분석을 위한 접근 방법으로는 크게 오디오 분석을 기반으로 하는 연구와 비디오 분석을 기반으로 하는 연구, 그리고 두 가지 방법을 결합한 연구로 나눌 수 있다 [2][5][6].

오디오 처리에 기반을 둔 대표적인 연구로는 화자 인덱싱 (speaker indexing)이 있다 [7][8]. 화자 인덱싱은 발성 화자의 수와 특성 등의 사전 (prior) 지식이 없는 상태

책임저자: 이 현 수 (leehs@khu.ac.kr)
경기도 용인시 기흥구 서천동 1
경희대학교 컴퓨터공학과
(전화: 031-201-2947; 팩스: 031-202-1723)

에서 화자 변화 구간을 찾고, 동일 화자 구간을 묶는 기술로 방송 뉴스, 토론, 드라마 등에 효과적으로 적용할 수 있다. 하지만 이를 위해서는 화자의 발성 특성을 잘 표현할 수 있는 특징 파라미터의 선별과 적은 데이터로 부터의 화자 모델 구성, 그리고 환경 변화에 강인한 화자 적응과 동일한 화자에 대해 여러 개의 화자 모델이 구성될 때 하나의 화자 모델로 효과적으로 묶을 수 있는 연구가 수행되어야 한다.

그리고 비디오 분석을 기반으로 하는 연구에서는 주로 동영상에서 샷 (shot) 경계를 검출하고, 검출된 샷들을 유사한 내용으로 클러스터링 (clustering) 하거나, 동일한 사람의 데이터를 검색하기 위해 얼굴검출 및 얼굴인식에 관한 연구가 다양하게 진행되고 있다 [9][10].

또한 오디오와 비디오 정보를 결합한 방법은 음성 정보만 사용하여 인덱싱을 수행할 때 화자를 잘 표현할 수 있는 특징 파라미터와 학습 데이터 부족의 문제를 화자의 얼굴이나 외형적인 영상 정보를 이용하여 해결하고, 영상 정보만 사용할 때 여러 명의 사람이 화면에 나타나거나 한 명도 나타나지 않을 때는 음성 정보를 이용하여 두 가지 방법을 개별적으로 사용할 때의 단점을 줄이기 위한 방법이다 [2][9].

현재 방송되고 있는 토론을 살펴보면 토론에 참여하는 사람들 중에서 진행자를 제외한 토론자는 매번 바뀌기 때문에 화자 모델의 구성과 학습에 어려움이 많다. 특히 토론자가 사투리를 사용하거나 액센트와 억양 패턴 등 개인적 발성 특성의 변화가 큰 경우에는 그렇지 않은 토론자 보다 개인의 특성을 잘 나타낼 수 있지만, 오히려 화자 인덱싱에서 변화 구간 검출과 클러스터링 과정에서 어려움이 될 수 있다. 그리고 영상 정보를 사용할 때 토론자들은 지정된 위치에 앉아서 토론에 참가하고 있지만 여러 방향으로 고개를 돌리면서 발성하고 있기 때문에 방향과 표정, 크기 변화에 강인한 얼굴인식 알고리즘이 요구된다. 뿐만 아니라 토론자가 발성하고 있는 동안에 화면에 나타난 토론자가 발성하고 있는지 확인하기 위해서는 입술의 움직임을 확인해야 하는데, 토론자 전체를 카메라로 잡거나 보조 자료가 나타나는 화면에서는 여러 명의 토론자가 나타나고 영상에서의 입술 크기가 작기 때문에 움직임을 확인하기가 대단히 어렵다.

본 연구에서는 비교사 (unsupervised) 방법으로 토론 데이터 인덱싱을 수행하기 위해서 BIC (Bayesian Information Criterion)를 사용할 때, 화자의 피치 (pitch) 변화가 큰 경우 클러스터링 과정에서 동일 토론자에 대해서 다수개의 클러스터 (cluster)가 생성되는 문제점을 영상 정보를 결합하여 줄이고자 한다. 이를 위해서 화면의 밝기 정보와 토론자와 주변의 외형적 특징들을 이용하여 샷들을 클러스터링한 후 음성 정보에 의한 BIC 클러스터링 결과와 결합하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 제 II장에서는 음성 정보를 이용한 화자 변화 검출 및 클러스터링 방법을 기술하고, 제 III장에서는 영상 정보를 이용한 클러스터링 방법을 기술한다. 그리고 제 IV장에서는 음성과 영상 정보를 결합하여 인덱싱 하는 방법을 제안한다. 제 V장에서는 음성과 영상의 개별적인 사용 방법과 음성과 영상 정보를 결합한 방법을 실험하고 결과를 분석 고찰하며, 마지막으로 VI장에서는 결론과 향후 연구 방향을 기술한다.

II. 음성 정보를 이용한 화자 클러스터링

2.1. 화자 변화 검출

음성 정보를 이용하여 동일한 화자의 음성 구간을 클러스터링하기 위해서는 우선 오디오 데이터가 음성인지 아닌지를 판단하고, 음성 구간에 대해 화자 변화 검출을 수행한 후 동일 화자 구간을 묶어야 한다 [11].

기존의 화자 변화 검출 방법으로는 BIC 방법과 GLR (Generalized Likelihood Ratio) 방법, 그리고 이 두 가지를 결합한 방법 등 다양한 방법들이 있다 [12][13]. BIC 방법은 거리 기반의 방법이 임계값에 의한 영향을 크게 받는데 비하여 임계값의 영향을 적게 받고 화자 변화 검출을 할 수 있으며, 화자 모델링을 수행하지 않고 화자 변화 검출 결과에 대해 비교사적인 방법으로 화자 클러스터링이 가능하다. 따라서 본 연구에서는 [14]의 연구에서 BIC 방법이 2초 보다 짧은 구간에 대해 화자 변화 검출이 어렵다는 단점을 지적하였으나, 토론자간의 의견 교환이 짧고 빈번하게 발생하는 경우를 제외하면 비교적 오랜 시간 의견을 제시하기 때문에 화자 변화 검출과 클러스터링을 모두 수행할 수 있는 BIC 방법을 사용하였다.

BIC를 이용한 화자 변화 검출은 크게 세 단계로 구성되어 있다 [12]. 첫 번째 단계에서는 큰 분석 윈도우를 이동하면서 화자 변화를 검출하고, 두 번째 단계에서는 첫 번째 단계에서 검출된 후보 지점을 중심으로 보다 작은 분석 윈도우를 사용하여 정확한 화자 변화 지점을 검출한다. 그리고 마지막 단계에서는 검출된 변화 지점을 기준으로 인접한 세그먼트들 간의 BIC 차이를 계산하여 결과를 검증한다.

2.2. 화자 클러스터링

BIC를 이용한 화자 클러스터링은 화자 변화 검출 단계에서 찾아진 모든 세그먼트들을 클러스터로 설정하고 수행되는데, i번째 클러스터와 j번째 클러스터에 대해 ΔBIC를 계산하여 두 클러스터를 하나의 클러스터로 묶는다. 이러한 클러스터링 과정은 모든 클러스터들에 대해 각 클러스터들이 ΔBIC 조건을 만족할 때까지 반복한다.

$$R(C_i, C_j) = \frac{N_{i+j}}{2}\log|\Sigma_{C_{i+j}}| - \frac{N_i}{2}\log|\Sigma_{C_i}| - \frac{N_j}{2}\log|\Sigma_{C_j}| \quad (1)$$

$$\Delta BIC(C_i, C_j) = R(C_i, C_j) + \frac{\lambda}{2}\left(p + \frac{p(p+1)}{2}\right)\log(N_{i+j}) \quad (2)$$

식 (1)은 클러스터 C_i와 C_j간의 우도비 (likelihood ratio)를 나타내고, 식 (2)는 모델의 복잡도를 고려한 ΔBIC를 나타낸다. 식 (1)에서 N_{i+j}와 N_i, N_j는 클러스터를 구성하는 특징 벡터의 개수로서 각 클러스터의 전체 프레임 개수를 나타내고, Σ_{i+j}와 Σ_i, Σ_j는 각 클러스터에 대한 공분산 행렬이며 $|\cdot|$은 행렬식을 나타낸다. 그리고 식 (2)에서 p는 특징 벡터의 차원이고, λ는 penalty 계수이다.

III. 영상 정보를 이용한 샷 클러스터링

샷 클러스터링은 그림 1과 같이 크게 3개의 블록으로 구성되어 있다. 첫 번째 블록에서는 토론 데이터에서 샷 경계를 검출하고 키 프레임 (key frame)을 저장한다. 두 번째 블록에서는 키 프레임에서 토론자와 방청객의 위치 정보를 추출하고, 세 번째 블록에서는 두 번째 블록에서 추출한 정보를 이용하여 샷 클러스터링을 수행한다.

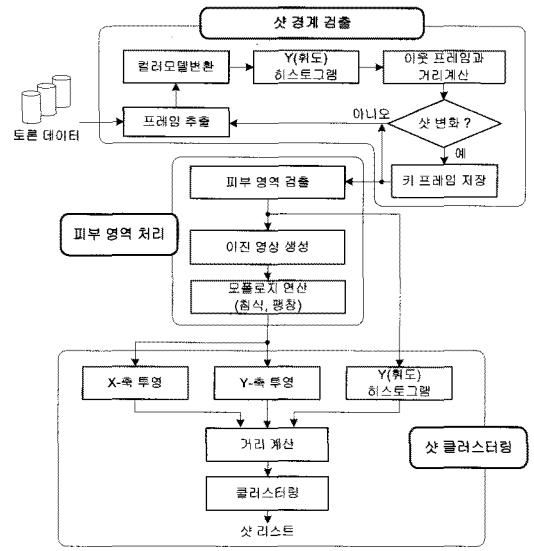


그림 1. 영상 정보를 이용한 샷 경계 검출 및 클러스터링
Fig. 1. Flow diagram of shot change detection and clustering using visual information.

3.1. 샷 경계 검출

토론 데이터에 대한 샷 경계 검출은 다음과 같이 수행한다. 먼저 동영상 파일의 각 프레임을 순차적으로 읽은 후 RGB 컬러 모델을 YC_bC_r 모델로 변환한다. 그리고 이웃하는 프레임과의 Y 히스토그램 차이를 계산하여 임계값 이상이 되면 샷 경계로 결정하고 샷 경계를 기준으로 중간 프레임을 키 프레임으로 저장한다. 이때 임계값은 10개의 프레임에 대한 Y 히스토그램의 평균을 구한 후 이 값보다 3배 이상 크면 샷 경계로 하였다. 이러한 샷 경계 검출은 토론자의 고개 돌림이나 손동작, 카메라의 이동에 영향을 적게 받는다.

3.2. 샷 클러스터링

토론 데이터의 화면 구성을 살펴보면 토론자의 배경이 단조로운 색으로 구성된 경우와 방청객이 토론자들의 뒤에 착석하여 토론을 방청하며 토론에 참여하는 경우가 있다. 이러한 토론장의 구성은 토론자의 샷 클러스터링에 효과적으로 사용될 수 있다.

제안하는 샷 클러스터링 방법은 다음과 같다. 먼저, 샷 경계 검출에서 저장한 키 프레임의 YC_bC_r 컬러 모델에서 C_bC_r에 대해 피부색의 범위를 지정하여 토론자와 방청객의 대략적인 얼굴 위치를 검출한다. 이때 C_bC_r의 범위는 실험을 통해 C_b = [95, 130], C_r = [130, 160]으로 정했다. 이러한 얼굴 위치는 이진 영상으로 저장되며 침식 (erosion)과 영역 팽창 (area opening) 같은 모폴로지 (morphology) 연산을 적용하여 작은 객체들을 제거한다. 다음으로 토론자와 방청객의 얼굴 및 위치 정보를 나타내는 이진 영상에 대해 X축과 Y축에 대해 투영한 히스토그램과 YC_bC_r 컬러 모델의 Y 값에 대한 히스토그램을 계산한다.

샷 클러스터링을 위해 정의한 거리함수는 식 (3)과 같다. 식 (3)에서 X_{hi}와 Y_{hi}는 이진 영상에 대해 각각 X축과 Y축으로 투영한 히스토그램을 나타내고, Y_i은 Y값의 히스토그램을 나타낸다. 그리고 α와 β는 히스토그램들의 차이에 대한 가중치 파라미터이며, M과 N은 영상의 X와 Y축의 크기를 나타낸다.

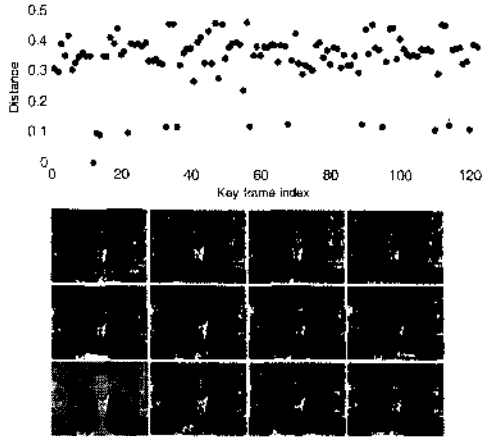


그림 2. 키 프레임 12에 대한 거리와 클러스터링 결과
Fig. 2. An example of distance and clustering results for key frame 12.

$$D(i,j) = \frac{\alpha}{M \cdot N}\sum_{x=1}^{M}(|X_{pri\_i}(x) - X_{pri\_j}(x)|) + \frac{\beta}{M \cdot N}\sum_{y=1}^{N}(|Y_{pri\_i}(y) - Y_{pri\_j}(y)|) + \frac{(1-\alpha-\beta)}{M \cdot N \cdot 255}\sum_{r=0}^{255}(|Y_{i\_r}(x) - Y_{j\_r}(x)|) \quad (3)$$

샷 클러스터링을 하는데 있어서 배경이 단조로운 경우에는 토론자의 배경과 외형적 특성을 중심적으로 활용하는 것이 효과적이고, 방청객이 참석한 경우에는 토론자의 외형적 특성과 방청객의 배석 정보를 활용하는 것이 클러스터링에 적합하다. 식 (3)에서 정의한 거리 함수는 토론자의 배경이 단조로운 경우에는 Y 값이 중요하게 작용하고, 방청객이 참여하는 토론의 경우에는 샷의 밝기인 Y값과 방청객의 위치 정보를 활용하여 효과적으로 클러스터링 할 수 있다.

그림 2는 방송된 토론에 대해 식 (3)에 정의된 거리함수를 적용한 예를 나타낸다. 그림은 키 프레임 12에 대해 검출된 다른 키 프레임들과의 거리를 구하고 임계값을 적용하여 동일한 클러스터로 묶인 샷들을 나타낸다.

IV. 음성/영상 정보를 이용한 클러스터링

음성 정보만을 이용한 화자 인덱싱에서는 적은양의 데이터에서 화자의 특성을 추출하여 화자 변화를 검출하고 동일 화자를 묶어야 하기 때문에 효과적인 특징 파라미터의 선별이 중요하다. 현재 사용되고 있는 특징 파라미터로는 음성인식과 화자인식에서 주로 사용되는 것들로 음성 정보와 화자 정보를 모두 포함하고 있기 때문에 화자의 정보만 추출하기가 어렵다. 또한 화자 모델링에 있어서 부족한 화자의 데이터에 따른 문제를 줄이기 위해 사전에 참조 모델 (reference model)을 구성한 후 모델을 적응시키는 방법들도 연구되고 있으나 화자수의 증가에 따라 성능의 영향을 받고 있다 [7].

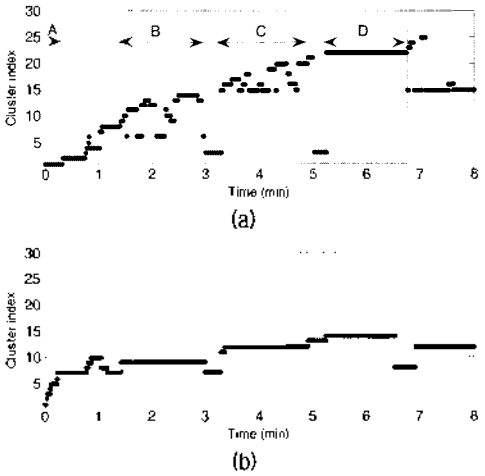


그림 3. 음성과 영상 정보에 의한 클러스터링 결과 (a) 음성 (b) 영상
Fig. 3. Clustering results by audio and visual information (a) Speech, (b) Visual.

그리고 영상 정보를 이용한 클러스터링 방법에서는 입술 움직임을 검출하여 발언자를 찾은 후 얼굴인식을 거쳐 동일한 사람의 샷들을 묶거나, 발언자를 잘 표현할 수 있는 외형적 특성을 추출하여 클러스터링 해야 한다. 그러나 토론과 같이 참여자가 빈번하게 변경되는 경우에는 참여하는 사람마다 얼굴인식을 위한 등록 과정이 필요하고 항상 정면을 향한 얼굴이 나타나지 않기 때문에 적용에 어려움이 있다.

본 연구에서는 음성 정보만 이용하여 화자 인덱싱을 수행하는데 있어서 나타나는 개인적 발성 특성에 의한 화자 클러스터링의 문제점을 줄이고자 한다. 개인의 발성특성은 다른 화자들과의 차이를 크게 나타낼 수 있어서 화자간의 분별력은 높일 수 있으나, 동일 화자의 음성이 BIC 클러스터링 과정에서 다수개로 나뉠 수 있기 때문에 화자 인덱싱을 위해서는 해결해야할 사항이다.

그림 3은 음성과 영상 정보에 의한 클러스터링 결과를 나타낸다. 그림 3.(a)에서 A구간은 음악이고 B~D 구간은 서로 다른 화자의 음성 구간을 나타낸다. 그리고 그림 3.(b)는 영상 정보에 의해 클러스터링된 결과를 나타낸다. 음성 정보에 의한 BIC 클러스터링 결과를 살펴보면 화자 B와 C는 다수개의 클러스터가 생성된 것을 볼 수 있고 D는 하나의 클러스터로 묶인 것을 알 수 있다. 그리고 영상 정보에 의한 클러스터링 결과인 A구간은 음악이 나오는 토론의 시작부분으로 샷의 변화가 많기 때문에

다수개의 클러스터를 생성한 것이다.

그림 4는 음성 정보만 이용한 BIC 클러스터링 결과 중에서 그림 3.(a)의 B, C, D 구간에 대해 분석 세그먼트의 길이를 2초로하고 1초씩 중첩하면서 [14]의 연구에서 사용한 [15]의 방법으로 피치를 찾은 후 구한 평균 피치와 표준 편차를 나타낸다.

그림 3과 4를 살펴보면 B와 C구간에 대한 피치의 변화가 D구간과 비교해 큰 것을 확인할 수 있다. 그리고 피치의 변화가 큰 경우에는 BIC에 의한 클러스터링 결과에서 다수개의 클러스터가 생성된 것을 확인할 수 있다. 반면에 D구간에 대해서는 피치의 변화가 작은 것을 볼 수 있고, 이에 대한 BIC의 결과는 하나의 클러스터로 묶인 것을 알 수 있다.

위와 같은 실험 결과를 바탕으로 BIC에 의한 클러스터링 결과에서 동일 화자의 피치 변화가 큰 경우 발생할 수 있는 다수개의 클러스터를 영상 정보에 의한 클러스터링 결과와 결합하여 하나의 클러스터로 묶고자 한다. 하나의 세그먼트에 대해 나타날 수 있는 BIC 클러스터링 결과와 영상 정보에 의한 클러스터링 결과는 크게 4가지 패턴으로 그림 5와 같다.

그림 5에서 (a)는 음성과 영상 모두 하나의 클러스터로 묶인 경우이고, (b)는 동일한 토론자가 의견을 제시하는 동안 자료화면 등이 연속하여 나타나면서 샷의 변화가 많은 경우이다. 그리고 (c)의 경우에는 피치의 변화가 큰 한명의 토론자가 의견을 제시하거나 여러 명의 토론자가 의견을 제시하는 동안 샷의 변화가 없는 경우이고, (d)는 피치 변화가 큰 한명의 토론자가 의견을 제시하거나 여러 토론자 음성구간일 수 있고, 자료 화면이 연속하여 나타나는 경우를 나타낸다.

음성 정보만 이용할 때 (a)와 (b)는 가장 이상적인 클러스터링 결과로 한명의 음성 구간이 하나로 묶인 경우이다. 그리고 (c)와 (d)는 한명의 토론자 음성인지 아니면 여러 명의 토론자 음성구간인지 판단하기 어려운 경우이다. 이러한 경우에는 영상을 분석하여 음성이 존재하는 구간에서 입의 움직임을 정확히 찾아야 한다. 하지만 토론장 전체를 화면으로 잡거나 자료 화면이 나타나는 경우에는 얼굴이 없어 입술 움직임 검출을 할 수 없기 때문에 본 연구에서는 BIC 클러스터링 결과와 샷 클러스터링 결과를 이용한다. 특히 음성과 영상 클러스터링 결과를 결합하여 음성, 영상 정보를 개별적으로 사용할 때 주로 나타나는 그림 5.(b)와 5.(c) 같은 패턴의 개수를 줄였다.

음성과 영상 정보를 이용한 클러스터링은 다음과 같이 수행된다. 먼저 영상 정보에 의한 샷 클러스터링 결과를 바탕으로, 음성 정보를 이용한 클러스터링 과정에서 동일 화자에 대해 생성된 다수개의 BIC 클러스터를 하나로 묶는다. 이때 중요한 토론자의 샷은 10초 이상 계속되고, 그 구간에서는 적어도 2초 이상 의견을 제시한다고 가정하여 영상 클러스터링 결과에서 샷의 지속 시간이 10초 이상이고, 그 샷에 나타나는 음성 세그먼트의 길이가 2초 이상인 세그먼트들만 하나로 묶는다. 이런 방법으로 토론 데이터 전체 구간에서 동일한 샷이 나타나는 구간의 음성 클러스터를 하나로 묶는다. 그리고 새롭게 만들어

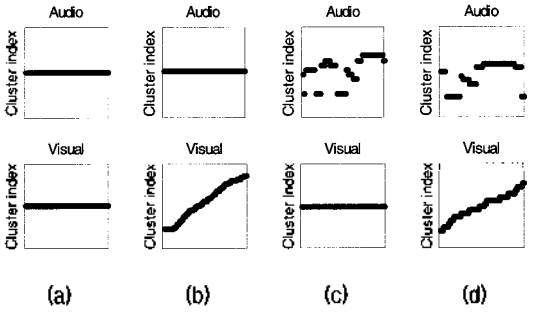


그림 5. 음성/영상 정보에 의한 클러스터 패턴
Fig. 5. Cluster patterns by audio-visual information.

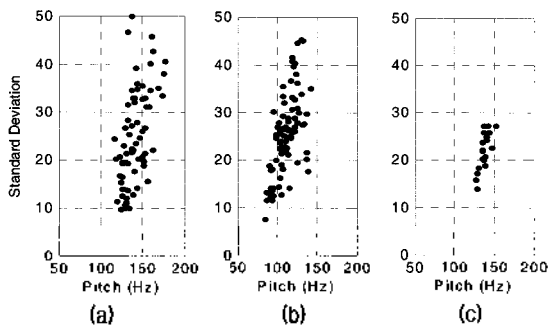


그림 4. 3명의 화자에 대한 평균 피치 및 표준 편차
Fig. 4. Mean and standard deviation of pitch for three speakers.

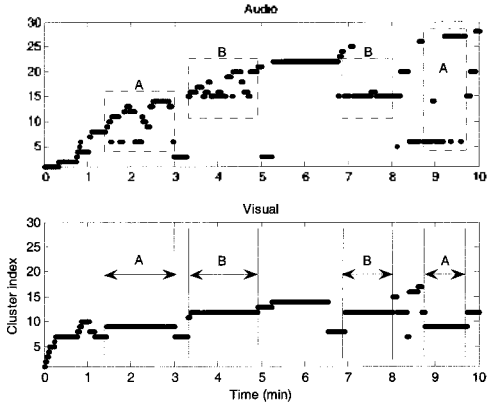


그림 6. 영상 정보에 의한 음성 클러스터링
Fig. 6. Audio clustering by visual information.

진 클러스터를 이용하여 클러스터링 되지 않은 구간을 인덱싱 한다. 음성과 영상 정보 결합에 의한 클러스터링 과정은 다음과 같다.

단계1 : BIC를 이용한 음성 클러스터링
단계2 : 거리 함수에 의한 샷 클러스터링
단계3 : 샷 클러스터링 결과를 바탕으로 클러스터링 조건을 만족하는 음성 클러스터링
클러스터링 조건: 중요 샷은 10초 이상 지속되며, 샷 구간에서의 음성 세그먼트 길이는 2초 이상
단계4 : 위의 단계 3에서 클러스터링된 음성 클러스터를 바탕으로 샷 클러스터링

클러스터링 과정에서 단계1과 단계2는 개별 정보에 사용에 의한 과정을 나타내고, 단계3은 영상 정보를 바탕으로 그림 5.(a)와 5.(c)의 음성 정보를 클러스터링하는 과정이다. 그리고 단계4에서는 단계3에 의한 클러스터링 결과를 바탕으로, 음성 정보를 이용하여 그림 5.(b)와 5.(d)를 클러스터링 한다.

그림 6은 단계3에 의해 영상 정보를 바탕으로 그림 5.(c)와 같은 패턴을 하나로 묶은 것이다. 그림에서 영상의 A와 B구간은 샷 클러스터링에 의해 동일한 클러스터로 묶인 구간이며, 이를 바탕으로 음성 구간의 클러스터들을 묶은 것이다. 이와 같은 방법으로 음성과 영상 정보를 결합하여 토론 데이터를 클러스터링하면 음성과 영상 정보를 개별적으로 사용할 때 생성되는 클러스터의 개수를 줄일 수 있다.

V. 실험 및 결과 분석

5.1. 실험 데이터의 구성

토론 데이터의 인덱싱 실험에는 서로 다른 5종류의 토론 데이터를 사용하였다. 실험 데이터는 TV 수신 카드를 사용하여 비디오에 대해서는 비압축 방식으로 초당 30프레임씩 저장하였으며, 오디오 데이터는 16 kHz로 샘플링하였고 16 bit로 양자화 하여 저장하였다. 구성한 실험데이터의 분량과 토론에 참가한 토론자의 수는 표 1과 같다.

음성 정보를 이용한 화자 변화 검출과 클러스터링에서 사용한 특징 파라미터는 32 ms의 프레임에서 12차의 MFCC를 사용하였다. 그리고 화자 변화 검출 1단계에서의 분석 윈도우는 3초로 하였고, 2단계에서는 2초로 하였으며 0.2초씩 이동하면서 BIC를 이용한 변화 검출을 수행하였다.

표 1. 토론 데이터의 분량 및 토론자의 수
Table 1. The length and number of speakers of debate data.

토론	1	2	3	4	5
분량 (min:sec)	49:06	39:44	49:30	36:31	44:09
토론자 수	4	5	15	5	6

5.2. 음성 정보에 의한 실험 결과

음성 정보만 이용하여 BIC 방법으로 화자 변화 검출과 클러스터링을 수행한 실험 결과는 그림 7과 같다. 그림 7.(a)를 살펴보면 토론 2와 3, 4는 1과 5의 데이터와 비교하여 두 사람 이상의 음성이 섞인 세그먼트 (mixed segment)가 많은 것을 볼 수 있는데 이것은 두 화자의 특성이 유사하여 화자 변화 검출을 정확히 하지 못한 결과이기도 하지만 토론의 성격과 관련이 크다고 할 수 있다. 실질적으로 토론 2와 3, 4는 다른 토론자의 의견 제시가 다 끝나지 않았는데도 중간에 의견을 제시하는 경우가 많았으며, 토론 1과 5는 한 토론자의 의견 제시가 끝난 후 다른 토론자가 의견을 제시하는 형식으로 진행되고 있어 두 사람의 음성이 섞인 세그먼트수가 적다.

그리고 그림 7.(b)는 그림 7.(a)의 잘못된 클러스터링 (missed clustering) 결과에 대해 두 사람 이상의 음성이 섞이지 않은 세그먼트에 대한 BIC 클러스터링 결과로 잘못된 세그먼트의 길이에 의한 영향을 나타낸다. 결과를 보면 기존 연구에서 보고된 바와 같이 BIC 방법이 2초보다 짧은 세그먼트에 대해 많은 에러가 발생한 것을 확인할 수 있다.

음성 정보만 이용하여 생성된 클러스터에서 세그먼트 길이가 2초 이상인 세그먼트들의 BIC 클러스터링 결과는 표 2와 같다. 표에서 화자의 수는 토론에 직접 참여한 토

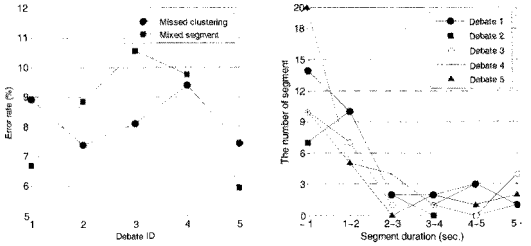


그림 7. BIC에 의한 클러스터링 결과 (a) 에러율, (b) 잘못된 클러스터링에 대한 세그먼트 길이
Fig. 7. A clustering results using BIC (a) Error rate, (b) Segment duration of missed clustering.

표 2. 음성 정보를 이용한 BIC 클러스터링 결과
Table 2. Results of BIC clustering using audio information.

토론	1	2	3	4	5
토론자 수	4	5	15	5	6
음성 세그먼트	359	271	284	277	404
클러스터 I	43	46	72	44	55

론자와 방청객 중에서 발언 시간을 가지고 토론에 참여한 사람의 수를 나타낸다. 표에서 음성 세그먼트 (speech segments)는 BIC에 의해 화자 변화 검출된 세그먼트의 수를 의미한다. 그리고 생성된 클러스터의 수는 BIC 방법이 2초보다 짧은 음성 세그먼트에 대해서 화자 변화 검출과 클러스터링의 단점을 보이기 때문에 2초보다 짧은 세그먼트로 구성된 클러스터를 제외한 클러스터의 수를 나타낸다. 가장 이상적인 화자 인덱싱 결과는 생성된 클러스터의 개수가 토론에 나타난 화자의 수와 같은 경우지만 표에 나타난 것과 같이 화자 수보다 많은 클러스터가 생성되었고, 이와 같은 결과는 토론에 참석한 사람들이 문장 독립의 발성을 하고 있기 때문에 사용한 특징 파라미터에 의한 영향과 개인 특성의 변이 때문으로 분석된다.

5.3. 영상 정보에 의한 샷 클러스터링 결과

거리 함수에 의한 샷 클러스터링 결과는 표 3과 같다. 표에서 비디오 샷 (video shots)은 밝기 정보를 이용하여 검출한 샷의 개수이다. 그리고 클러스터 II는 검출된 모든 샷에 대해 클러스터링한 결과이며, 클러스터 III는 토론장 전체 또는 여러 명의 토론자를 보여주는 샷 그리고 자료 화면을 보여주는 샷을 제외한 단일 토론자에 대한 클러스터링 결과이다.

표 3에서 클러스터 III의 결과를 분석해 보면 토론 1과 5는 음성 정보에 의한 실험 결과에서 기술한 바와 같이 토론자의 의견 제시가 끝난 후 다른 토론자가 의견을 제시하는 형식으로 구성되어 있고, 대부분 토론자가 바뀔 때 카메라가 같이 바뀌기 때문에 토론자 수와 비슷한 클러스터를 생성하였다. 하지만 토론자의 프로필을 소개하기 위해 자막이 나타난 경우와 고개를 숙인 경우가 포함되어 토론자 수보다는 많은 클러스터가 생성된 것이다.

그리고 토론 2와 4는 토론자의 수보다 훨씬 많은 클러스터가 생성되었는데 토론자가 의견을 제시하면서 손동작을 자주하는 경우 같은 토론자에 대해서 다수의 클러스터가 만들어졌다. 또한 토론자 뒤에 앉아 있던 방청객이 중간에 좌석 위치를 변경한 경우가 있었으며, 토론자의 프로필 자막으로 다수의 클러스터가 생성되었다.

표 3. 영상 정보를 이용한 샷 클러스터링 결과
Table 3. Results of shot clustering using visual information.

토론	1	2	3	4	5
토론자 수	4	5	15	5	6
비디오 샷	123	121	264	109	87
클러스터 II	38	53	183	61	32
클러스터 III	6	20	81	22	8

토론 3은 다른 데이터와 비교하여 카메라의 이동이 많은 것을 확인하였다. 특히 동일한 토론자에 대해서도 다양한 각도와 줌-인/아웃 (zoom-in/out)이 되어있는 샷들이 많아 비교적 많은 클러스터들이 생성되었다. 따라서 샷의 각도와 크기 변화에 강인한 처리를 위해서는 [16]의 연구와 같은 방법을 적용하여 먼저 카메라 샷의 종류를 분류한 후 특정 샷에 대해서 토론자의 영상 정보를 추출하는 것이 효과적인 방법일 수 있다.

5.4. 음성/영상 정보를 이용한 실험 결과

표 4는 음성과 영상 정보를 이용한 실험 결과를 나타낸다. 표에서 클러스터 IV는 BIC에 의해 화자 변화 검출이 정확히 이루어진 2초 이상의 세그먼트들에 대한 클러스터링 결과이다.

생성된 클러스터의 개수를 보면 음성과 영상 정보를 개별적으로 사용한 결과와 비교하여 높은 성능 개선을 확인할 수 있다. 특히 토론 1과 5는 토론자를 잡는 카메라의 각도 변화가 적고, 토론자가 의견을 제시한 후 다른 토론자가 의견을 제시하는 형식으로 구성되어 있어서 실제 토론자와 비슷한 수의 클러스터가 생성되었다.

그리고 토론 2와 3, 4에 대해서도 높은 성능 개선이 있었지만 카메라의 변화가 커서 영상 정보에 의해 동일 토론자에 다수개의 클러스터가 생성되었고, 또한 음성에 의한 클러스터링도 다수개의 클러스터가 생성되어 실질 화자수보다 많은 클러스터가 생성되었다.

본 연구에서 제안한 방법으로 토론 데이터에 대해 화자 인덱싱을 수행한 결과 음성, 영상 정보를 결합한 방법이 정보를 개별적으로 사용하는 방법보다 높은 클러스터링 결과를 얻을 수 있었다. 그리고 BIC에 의한 음성 클러스

표 4. 음성/영상 정보를 이용한 토론 인덱싱 결과
Table 4. Results of debate indexing using audio-visual information.

토론	1	2	3	4	5
토론자 수	4	5	15	5	6
클러스터 IV	4	11	34	11	8

터링 방법의 단점은 줄일 수 있었으나, 화자 특성 변화에 정확한 변화 검출과 클러스터링이 되지 않아 이에 대한 연구가 필요하다. 또한 영상에 의한 거리 기반의 방법은 전체적으로 토론자에 대해 강인한 클러스터링 결과를 보였으나, 동일한 토론자에 대해 여러 각도와 화면의 크기 변화에 대해서는 블록 매칭과 같은 알고리즘을 부가적으로 사용하면 성능을 더욱 개선할 수 있을 것이다.

VI. 결 론

본 연구에서는 토론 데이터를 대상으로 화자 인덱싱을 수행하는데 있어서 음성이나 영상 정보를 각각 사용하는 경우에 발생하는 문제점을 발견하고 이 문제점을 두 가지 정보를 이용하여 성능을 향상 시키는 방법을 제안하였다. 음성 정보만 사용한 클러스터링에서 동일 화자의 음성이 다수개의 클러스터로 나뉘는 문제점을 영상 정보를 결합하여 클러스터의 개수를 줄이는 방법을 제안하였고, 영상 정보만 이용한 클러스터링 과정에서 어느 화자의 클러스터로도 포함시킬 수 없는 샷들을 음성 정보를 이용하여 특정 화자의 클러스터로 묶을 수 있었다.

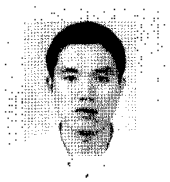
향후 연구 방향으로는 본 연구에서 제안한 방법을 개선하여 음성/영상 정보를 이용하면서 온라인으로 비교사 화자 인덱싱을 수행할 수 있는 방법을 연구할 계획이다.

참 고 문 헌

1. Xinbo Gao, Xiaoou Tang, "Unsupervised Video-shot Segmentation and Model-free Anchorperson Detection for News Video Story Parsing," IEEE Trans. Circuits and Systems for Video Technology, 12(4), 765-776, 2002.
2. Alberto Albiol, Luis Torres, Edward J. Delp, "The Indexing of Persons in News Sequences using Audio-visual Data," in Proc. International Conference on Acoustics, Speech, and Signal Processing, 3, 137-140, 2003.
3. Yuya Akita, Masahiro Hasegawa, Tatsuya Kawahara, "Automatic Audio Archiving System for Panel Discussions," in Proc. International Conference on Multimedia & Expo, 3, 1895-1862, 2004.
4. Alfred Dielmann, Steve Renals, "Automatic Meeting Segmentation Using Dynamic Bayesian Networks," IEEE Trans. Multimedia, 9(1), 25-36, 2007.
5. 한학용, 허강인, 김수훈, "오디오 데이터의 특징 파라메터 구성에 따른 내용기반 분석," 한국음향학회지 21(2), 182-189, 2002.
6. 손종목, 배건성, 강경옥, 김재곤, "내용기반 비디오 색인 및 검색을 위한 음성인식기술 이용에 관한 연구," 한국음향학회지, 20(2), 16-20, 2001.
7. Soonil Kwon, Shrikanth Narayanan, "Unsupervised Speaker Indexing Using Generic Models," IEEE Trans. Speech and Audio Proc., 13(5), 1004-1013, 2005.
8. Sue E. Tranter, Douglas A. Reynolds, "An Overview of Automatic Speaker Diarization Systems," IEEE Trans. Audio, Speech and Language Proc., 14(5), 1557-1565, 2006.
9. Ying Li, Shrikanth Narayanan, C.-C. Jay Kuo, "Audiovisual-based Adaptive Speaker Identification," in Proc. International Conference on Acoustics, Speech, and Signal Processing, 5, 812-815, 2003.
10. Ki Tae Park, Doo Sun Hwang, Young Shik Moon, "Anchor Frame Detection in News Video Using Anchor Object Extraction," IEICE Trans. Fund., E88-A(6), 1525-1528, 2005.
11. 금지수, 임성길, 이현수, "스펙트럼 분석과 신경망을 이용한 음성/음악 분류," 한국음향학회지, 26(5), 207-213, 2007.
12. Scott Shaobing Chen, P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via The Bayesian Information Criterion," DARPA Broadcast News Transcription & Understanding Workshop, 1998.
13. P. Delacourt, C. J. Wellekens, "DISTBIC: A Speaker Based Segmentation for Audio Data Indexing," Speech Communication, 32, 111-126, 2000.
14. Min Yang, Yingchun Yang, Zhaohui Wu, "A Pitch-based Rapid Speech Segmentation for Speaker Indexing," in Proc. IEEE International Symposium on Multimedia, 2005.
15. Xuejing, "Pitch Determination and Voice Quality Analysis using Subharmonic-to-Harmonic Ratio," in Proc. International Conference on Acoustics, Speech, and Signal Processing, 1, 333-336, 2002.
16. Maria Zapata Ferrer, Mauro Barbieri, Hans Weda, "Automatic Classification of Field of View in Video," in Proc. International Conference on Multimedia & Expo, 1609-1612, 2006.

저자 약력

•금 지 수 (Ji-Soo Keum)



1998년 2월: 강남대학교 전자계산학과 (공학사)
2000년 2월: 경희대학교 전자계산공학과 (공학석사)
2000년 3월~ 현재: 경희대학교 컴퓨터공학과 (박사과정)
※주관심분야: 멀티모달 화자인덱싱, 패턴인식, 신경망

•이 현 수 (Hyon-Soo Lee)

1979년 2월: 경희대학교 전자공학과 (공학사)
1982년 4월: 일본 게이오대학원 전기공학과 (공학석사)
1985년 4월: 일본 게이오대학원 전기공학과 (공학박사)
1999년 9월~2000년 8월: 미국 오레곤 주립대학교 전기 및 컴퓨터공학과 방문연구원, 미국 캘리포니아대학교(U.C.I) 전기 및 컴퓨터공학과 방문연구원
1985년~ 현재: 경희대학교 컴퓨터공학과 교수
2005년~ 현재: 경희대학교 전자정보대학 학장 및 정보통신대학원 원장
※주관심분야: 컴퓨터구조 및 VLSI, 병렬처리, 패턴인식, 신경망, 음성처리