한국음향학회지 제28권 제7호 pp. 674~679 (2009)

음성정보 내용분석을 통한 골프 동영상에서의 선수별 이벤트 구간 검색

Retrieval of Player Event in Golf Videos Using Spoken Content Analysis

김 형 국*
(Hyoung-Gook Kim*)

*광운대학교
(접수일자: 2009년 8월 19일; 채택일자: 2009년 10월 12일)

본 논문은 골프 동영상에 포함된 오디오 정보로부터 검출된 이벤트 사운드 구간과 골프 선수이름이 포함된 음성구간을 결합하여 선수별 이벤트 구간을 검색하는 방식을 제안한다. 전체적인 시스템은 동영상으로부터 분할된 오디오 스트림으로부터 잡음제거, 오디오 구간분할, 음성 인식 등의 과정을 통한 자동색인 모듈과 사용자가 텍스트로 입력한 선수 이름을 발음열로 변환하고, 색인된 데이터베이스에서 질의된 선수 이름과 상응하는 음성구간과 연결되는 이벤트 구간을 찾아주는 검색 모듈로 구성된다. 선수이름 검색을 위해서 본 논문에서는 음소 기반, 단어 기반, 단어와 음소를 결합한 하이브리드 방식을 적용한 선수별 이벤트 구간 검색결과를 비교하였다.
핵심용어: 음소 기반 검색, 단어 기반 검색, 하이브리드 검색, 음성 인식, 이벤트 검색, 선수이름 검색, 음성내용 검색
투고분야: 음성처리 분야 (2.5)

This paper proposes a method of player event retrieval using combination of two functions: detection of player name in speech information and detection of sound event from audio information in golf videos. The system consists of indexing module and retrieval module. At the indexing time, audio segmentation and noise reduction are applied to audio stream demultiplexed from the golf videos. The noise-reduced speech is then fed into speech recognizer, which outputs spoken descriptors. The player name and sound event are indexed by the spoken descriptors. At search time, text query is converted into phoneme sequences. The lists of each query term are retrieved through a description matcher to identify full and partial phrase hits. For the retrieval of the player name, this paper compares the results of word-based, phoneme-based, and hybrid approach.
Keywords: Phoneme-based retrieval, Word-based retrieval, Hybrid retrieval, Speech recognition, Player event retrieval, Player name retrieval, Spoken content analysis
ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

인터넷 상에 산재해 있는 정보를 미리 수집하고 이를 체계적으로 저장한 후, 사용자가 원하는 정보를 수시로 찾을 수 있도록 해주는 일종의 데이터베이스 관리시스템인 검색엔진은 현대인의 일상생활에 있어서 절대적으로 필요한 도구 중의 하나이다.

또한, 최근 인터넷 디지털 TV와 멀티미디어 핸드폰 같은 가전제품들은 동영상이나 음악 파일 등의 대용량의 디지털 콘텐츠가 존재하는 인터넷에 접속할 수 있을 뿐만 아니라 Personal Video Recorder (PVR)를 사용하여 원하는 방송 프로그램을 녹화할 수 있게 되었다. 이러한 인터넷 상에 있는 대용량의 동영상들로부터 사용자가 원하는 내용정보를 포함하고 있는 동영상을 찾고, PVR에 저장된 동영상에서 자신이 시청하고자 하는 장면을 찾기 위한 가장 유용한 방법은 동영상 비디오 안에 포함된 음성정보 중에서 어휘들을 색인하여 이를 체계적으로 저장하는 데이터베이스 관리 시스템을 구축하는 것이다.

동영상에 포함된 오디오 스트림으로부터 음성신호를

책임저자: 김 형 국 (hkim@kw.ac.kr)
서울시 노원구 월계동 447-1 광운대학교
(전화: 02-940-5574; 팩스: 02-913-5006)

추출하고, 추출된 음성신호를 텍스트로 변환하여 검색시스템에 사용하고자 하는 Spoken Document Retrieval (SDR) 연구 [1] [2]는 지금까지 다양하게 시도되어 오고 있다.

SDR 시스템은 크게 2가지 모듈, 즉 색인과 검색 모듈로 구성된다. 현재까지 주로 사용하고 있는 SDR 방식은 대용량 연속어인식기를 이용한 단어기반 방식, 음소 열이나 음소 그래프를 색인 및 검색에 적용하는 음소기반 방식이었으며, 최근 들어 각각의 방식의 장점들을 이용하여 단어기반 방식과 음소기반 방식을 결합한 하이브리드 방식 [3]이 연구되고 있다. 이와 함께 MPEG-7과 같은 표준안에서도 단어 레벨과 음소 레벨에서 함께 색인되는 방법이 시도되고 있으며, 이런 하이브리드 방식을 통해 검색성능을 높이고 있는 추세이다.

그러나 지금까지 시도되고 있는 대부분의 연구는 연속적인 음성정보로부터 어휘를 색인하는 방법에 촛점을 맞추는 독립적인 연구가 진행되고 있을 뿐, 동영상 안에 존재하고 있는 오디오 스트림의 화자나 배경잡음을 고려한 화자 변환점 검출과 환경 사운드 식별을 함께 적용하여 분리된 음성구간으로부터 원하는 어휘를 검색하고, 그 어휘와 관련되는 특별한 구간을 찾음으로써 검색성능을 구체적으로 향상시키고자 하는 연구가 진행되지 못하고 있다.

본 논문에서는 축구나 야구경기보다 배경잡음이 상대적으로 조용한 골프 동영상에서 오디오 신호 분석을 통해 연속적인 오디오 신호로부터 이벤트사운드 구간과 음성구간을 검출하고, 이벤트사운드 구간주변의 음성정보로부터 검출된 선수이름을 통해 선수별 이벤트 구간을 검색하는 효과적인 방식을 제안한다.

본 논문의 구성은 다음과 같다. 제 II 장에서는 전체적인 시스템의 구성도와 세부 모듈의 기능과 사용된 방법을 기술한다. 제 III장에서는 선수이름을 검출하는 세 가지의 방식을 설명하고, IV장에서는 구현된 시스템의 실험결과를 분석 및 고찰하며 제 V장에서 결론과 향후 연구방향을 기술한다.

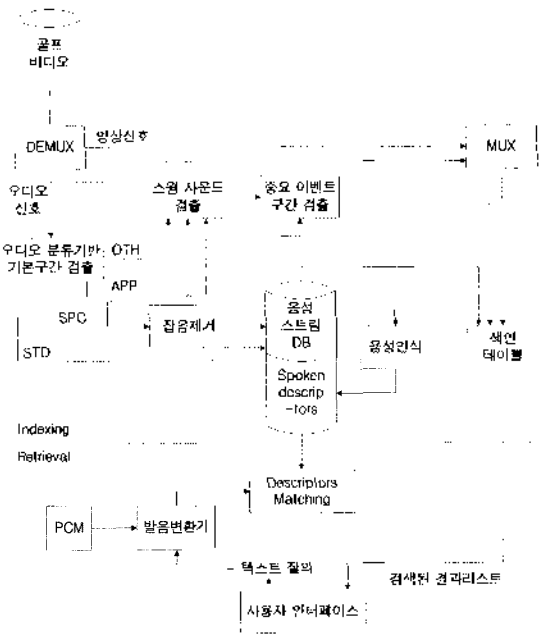
II. 시스템 구성도

그림 1은 전체적인 시스템의 개요도를 나타낸다. 제안된 시스템은 크게 색인 모듈과 검색 모듈로 구성된다.

색인 모듈의 기능은 동영상에서 분리된 오디오 스트림을 이벤트사운드구간과 음성구간으로 분할하고, 분할된 음성구간으로부터 잡음을 제거한 후에 음성특징을 추출하여 음소 및 단어 인식 등의 과정을 통해 색인된 단어 혹은 음소 그래프로 변환하여 이벤트 구간과 함께 데이터베이스화하는 역할을 수행한다.

본 논문에서의 선수별 이벤트 구간 검출을 위한 색인 모듈을 구성하는 각 세부적인 모듈의 기능과 사용된 기법은 다음과 같다.

- 오디오 분할: 연속적인 오디오 스트림을 1초 간이의 오디오 클립으로 분할하고, 분할된 오디오 클립으로부터 Mel-Frequency Cepstral Coefficient (MFCC), log energy, spectral centroid, spectral roll-off, spectral flux, zero crossing rate, SF3 등의 특징값 등을 추출하고, Support Vector Machine (SVM) [4] 기반의 Adaboost cascade [5] 분류방식을 통하여 연속적인 오디오 스트림을 스튜디오 환경에서의 아나운서의 음성구간 (STD), 선수들의 플레이에 따라 반응하는 관중들의 박수 및 환호성 소리구간 (APP), 필드에서의 레포터의 음성구간 (SPC) 및 그 외의 구간 (OTH) 등의 4가지 구간으로 분류한다. 스윙사운드는 스윙이 존재하는 APP, SPC, OTH 등의 세 구간에 대해서 impulsive onset detection과 변조스펙트럼 방법을 이용하여 골프스윙을 검출한다. 오디오 분할방식에 대해서는 [6]에 설명되어 있는 바와 같이, 각



SPC: 연속적인 오디오 스트림을 스튜디오 환경에서의 아나운서의 음성구간
APP: 선수들의 플레이에 따라 반응하는 관중들의 박수 및 환호성 소리구간
SPC: 필드에서의 레포터의 음성구간
PCM: Probability Confusion Matrix

그림 1. 전체 시스템 개요
Fig. 1. System overview.

STD, APP, SPC 구간들이 갖고 있는 구조적인 지속시간과 분류된 각 구간의 비율을 측정하여 시간에 따라 의미적인 세그먼트들의 단위로 내용구간을 조정한다. 4 개의 구간으로부터 선수 이름이 검출되는 스튜디오 환경에서의 아나운서의 음성구간, 필드에서의 레포터의 음성구간에서는 선수이름에 대한 색인시간 및 색인 프로그램의 메모리 감축을 위해서 음성길이가 15초를 초과하지 않도록 제한한다. 음성분할 길이가 지정된 지속시간을 초과할 경우에는 다음 분할구간으로 지정이 되도록 하며, 다음 분할영역이 바로 이전 영역과 1초 정도 겹치도록 하여 분할경계에서 출현한 검색어에 대해서도 검색이 가능하도록 한다.

- 잡음 제거: 골프 동영상에는 바람소리를 비롯하여 박수소리, 골프 샷 소리, 심지어 새소리 등의 다양한 잡음이 존재한다. 이러한 잡음 중에서 박수소리 및 골프 샷 소리 등은 제거하는 것이 불가능하고 따로 이벤트사운드구간으로 분할되기 때문에, 바람 소리, 파도 소리 등과 같은 잡음이 존재하는 오디오 필드에서의 레포터의 음성구간을 대상으로 잡음제거를 수행한다. 이러한 잡음을 제거하기 위해서 본 논문에서는 정적인 잡음을 제거하는 Wiener filter [7], Gaussian 분포기반의 Log Spectral Amplitude (LSA) [8] 음성추정 방식과 Gamma 분포기반의 음성추정 방식 [9]을 적용하여 그 성능을 측정하였다.
- 음성인식: 음성특징벡터로는 12차 MFCC, log-energy, 이들의 differential, acceleration 계수를 포함하여 39차 벡터를 사용하고, 변화하는 잡음환경에 강인한 특징벡터를 추출하기 위해 cepstral mean subtraction 기법을 적용한다. 음향모델에 있어서는 점프가 없는 단순한 left-to-right 모델 구조를 가지며, 각 음소들은 지속 길이의 특성에 따라 1개 또는 3개의 state를 갖는 continuous density hidden Markov model 방식을 사용한다. 그리고 음향모델의 강인성을 개선하기 위해 decision tree를 이용한 top-down 방식의 tying 기법을 적용하고, 각 state는 8개의 mixture Gaussian으로 모델링 된다. 음향모델의 훈련에 사용된 음성데이터는 약 1150명이 발화한 97,000여 발화로 구성된 음성 데이터베이스에 골프 동영상에서 추출된 배경잡음을 추가하여 사용하였다. 언어모델에서는 서로 인접해서 나타날 수 없는 음소나 트라이폰 쌍의 연결을 제약하기 위해 subword-pair 문법을 사용한다.

검색 모듈은 크게 발음 변환기, Phone Confusion Matrix (PCM)에 대한 리소스, Descriptors Matching (DM) 모듈 등의 세 가지 구성요소를 갖는다.

- 발음 변환기: 텍스트로 입력한 선수이름을 발음 열로 변환한다. 현재 한국어의 경우 99.4% 발음변환 정확률을 갖고 있다고 한다.
- PCM: 질의된 선수이름으로 표시된 descriptor symbol 열과 색인된 음소 혹은 단어 그래프에서 추출된 descriptor symbol 열사이의 유사도를 측정하는데 사용하는 PCM은 음소 인식결과를 토대로 통계적으로 혼동되는 확률 값을 추정하는 방식을 사용하였다 [10].
- DM 모듈: DM 모듈은 색인된 데이터베이스에서 질의된 선수이름의 descriptor symbol 열과 유사한 영역들을 검색한다. PCM을 이용하여 부분적인 매칭영역을 검색할 수 있다.

III. 선수 이름 및 선수별 이벤트 구간 검출

선수이름 검색 방식의 선택은 색인 단위와 밀접한 상관이 있다. 본 논문에서 사용되고 있는 방식은 크게 세 가지 형태로 분류될 수 있으며, 각 방법별로 다음과 같은 장단점을 갖고 있다.

- 단어기반 색인 및 검출: 단어기반 방식은 동영상의 오디오 내용들을 텍스트로 변환하기 위한 대용량 연속음성인식기와 텍스트 검색엔진으로 구성된다. 검색이 간편하고 검색시간이 짧다는 장점이 있는 반면에 대용량 연속어인식기에 의해 발생하는 오류로 인해 검색성능이 강인하지 못한 단점을 갖고 있다. 그리고 음성인식 엔진들이 제한된 인식대상 어휘들을 사용하여 텍스트로 변환하기 때문에 비등록 어휘에 대해서는 검색이 불가능하다는 문제가 있다.
- 음소기반 색인 및 검출 [2]: 비등록 어휘도 검색할 수 있고 학습에 큰 데이터베이스가 필요로 하지 않는 장점을 가지고 있는 방식이 음소기반 방식이다. 음소디코더는 특징 열을 입력으로 받아 각 프레임 구간마다 매칭되는 서브워드들을 탐색하여 서브워드 격자를 구성하는 음소격자 생성기, 그리고 이 격자를 그래프 형태로 변환하는 phone-graph 변환기로 구성된다. 음소격자 생성기에서는 각 프레임 구간들마다 매칭 가능한 서브워드들을 탐색하고, 각 서브워드에 대한 매칭 점수를 계산한다. 일반적으로 많이 사용되는 탐색방법이 Viterbi beam 탐색방법이다. 그러나

이 알고리즘은 maximum likelihood의 path만 유지하기 때문에 가장 좋은 path를 찾을 때는 문제가 없지만, 후보 path를 찾는 경우에는 심각한 path 손실이 존재한다. 이 path 손실을 줄이기 위한 방식으로 단어 중심, 또는 음소 중심 탐색 방식이 있는데 본 논문에서는 음소 중심 탐색 방법을 사용한다. 단어열로 색인하기 보다는 음소 열이나 음소 그래프로 색인을 하고, 사용자의 질의도 음소 열로 변환하여 음소 열의 비교를 통해 검색하는 방식으로서 식 (1)을 사용하여 매칭 confidence를 결정한다.

$$C(W, t_s, t_e|O) = \frac{\sum_{W_a, W_b} p(O, t_s, t_e|W_a, W, W_b)p(W_a, W, W_b)}{\sum_{W_s} p(O|W_s)P(W_s)} \quad (1)$$

여기서, Posteriori $P(W, t_s, t_e|O)$는 시간영역 t_s부터 t_e까지의 음소 열을 포함하는 모든 경로의 확률의 합을 나타낸다. W, W_a, 와 W_b는 각각 질의어에 따른 음소 열, 시간영역 t_s 전의 단어 시퀀스, t_e 후의 단어 시퀀스를 각각 나타내고, W_s는 단어 시퀀스를 나타낸다.

일반적으로 현재 음소 인식률은 연속어인식기의 성능에 비해 많이 낮다는 단점을 가지고 있다.

- 단어와 음소를 결합한 하이브리드 색인 및 검출 [3]: 단어기반 방식의 장점과 음소기반 방식의 장점을 결합하여 검출 성능을 향상시키기 위해 제시된 방법으로서 미등록 어휘는 사전에 첨부되어 음소 언어 모델에 의한 발화로 정의된다. 하이브리드 방식에는 posteriori 결합기법과 priori 결합기법이 있는데, 본 논문에서는 음성인식 엔진을 통해 각각 단어와 음소 열을 생성하고 그들의 시작점과 끝점을 연결해서 결합하는 posteriori 결합기법을 사용하고, 식 (2)를 통해 매칭 confidence를 결정한다.

$$C(W, t_s, t_e|O) = P_{WD} \cdot \frac{\sum_{\substack{W_a, W_b \in L_{WD} \\ t_s \in L_{WD}}} p(O, t_s, t_e, W_a, W, W_b|M_{WD})}{\sum_{W_s \in L_{WD}} p(O, W_s|M_{WD})} + P_{PH} \cdot \frac{\sum_{\substack{W_a, W_b \in L_{PH} \\ t_s \in L_{PH}}} p(O, t_s, t_e, W_a, W, W_b|M_{PH})}{\sum_{W_s \in L_{PH}} p(O, W_s|M_{PH})} \quad (2)$$

여기서, P_{WD}와 P_{PH}는 각각 가중치에 대한 priori 확률을 나타내며, L_{WD}와 L_{PH}는 각각 단어와 음소 격자를 나타내고, M_{WD}와 M_{PH}는 각각에 상응하는 사전과 언어모델을 나타낸다.

본 논문에서는 단어기반, 음소기반, 하이브리드 기반 등의 색인 및 검출 방식을 적용하여 선수이름 검출 성능을 비교한다.

골프 동영상의 이벤트구간은 선수들의 플레이에 따라 반응하는 관중들의 호응구간인 박수소리와 드라이브 샷, 아이언 샷과 퍼팅 샷 시에 발생하는 스윙구간 (SWN)을 결합하여 이벤트 구간에 포함되어 색인된다. 그러나 검출된 이벤트 구간만으로는 어떤 선수와 관련된 이벤트구간인지를 자동적으로 판별할 수가 없다. 이벤트가 발생하기 전과 후에는 아나운서나 레포터가 골프선수의 이름을 반복해서 언급하기 때문에, 검출된 이벤트 구간의 15초 전후에 존재하는 음성구간인 SPC와 STD의 음성정보로부터 색인되는 선수이름과 가장 근접한 이벤트구간의 선수이름 검출횟수를 측정하여 최종적으로 선수별 이벤트 장면을 검색한다. 즉, 반복적으로 언급되는 선수이름의 빈도수가 많을수록 이벤트구간은 검출된 선수의 이벤트로서 결정된다.

IV. 실험 결과 및 분석

본 논문에서 제안된 시스템의 성능을 평가하기 위해 스포츠 채널에서 녹화한 총 30시간 이상의 30개 골프 프로그램을 사용하였다. 그 중에서 15개 골프 동영상이 학습에 사용하였고 나머지 15개의 프로그램을 대상으로 성능을 평가하였다.

오디오 스트림은 16 kHz로 샘플링 되었으며 내용구간인 APP, SPC, STD, OTH 등의 네 개 구간과 스윙 사운드 검출구간인 SWN 그리고 하나의 이벤트구간으로 라벨화 되었다. 오디오 분할방식을 통한 기본구간 검출의 성능 평가를 위해서는 널리 알려진 precision과 recall 방식을 사용하여 검출성능을 비교하였다. 검출성능의 측정은 대부분 10초 이상 지속되는 STD, SPC는 시작과 끝점을 찾고, 매우 짧은 시간의 스윙 사운드 SWN와 특정한 규칙없이 산발적으로 분포하여 시작과 끝점이 명확하지 않은 APP는 발생횟수로 측정되었다.

표 1은 오디오 분류기반 기본구간 검출의 성능을 나타내었다.

결과를 살펴보면 STD, SPC는 매우 높은 검출 성능을 보임으로 음성정보로부터 선수이름 검출을 위한 프로그램의 구조를 뒷받침할 수 있다. 대부분의 SWN이 검출되었으나 아나운서의 감탄사나 음악에서의 강한비트, 갑자

표 1. 내용구간 검출성능 평가
Table 1. Performance of content-based segmentation.

기본구간	Recall	Precision
STD	100%	97.5%
SPC	96.2%	95.3%
OTH	93.6%	95.6%
SWN	92.7%	65.4%
APP	97.5%	96.6%

표 2. 이벤트구간 검출 성능
Table 2. Performance of event unit detection.

	Recall (%)	Precision (%)
정확률	97.5%	98.6%

표 3. 잡음제거 성능
Table 3. Performance of noise reduction.

방식	골프장 환경잡음 SNR (dB)			
	0	5	10	15
Wiener	2.2	4.9	7.4	10.7
LSA	3.3	7.3	9.4	13.3
Gamma 1	3.1	6.6	8.4	12.5
Gamma 2	2.9	6.4	8.6	12.2

기 발생한 박수소리, 클럽이 바닥에 떨어지는 소리, 골프공이 바닥에 떨어지는 소리 녹화하는 채널의 클릭 소리 등의 갑자기 발생하는 강한 사운드 등으로 인해 많은 SWN 검출오류가 발생하였으나 APP의 검출성능이 우수하여 이벤트구간 검출 성능에 큰 영향을 미치지 않았다.

표 2는 이벤트구간 검출 성능을 나타낸다.

필드환경과 같은 다양한 배경잡음에 노출된 음성정보로부터 효과적으로 선수이름을 검출하기 위해서는 잡음제거 기능이 필요하다. 본 논문에서는 골프 동영상에서 추출된 10분 길이의 STD 음성구간을 수집하고, 수집된 STD 음성에 골프 동영상에서 추출된 배경잡음을 추가한 잡음섞인 음성을 생성하여 정적인 잡음을 제거하는 Wiener filter, 음성부재확률 (speech absence probability) 기반의 LSA 음성추정 방식과 Gamma 분포기반의 음성추정방식의 Gamma1과 Gamma2를 적용하여 segmental SNR (Signal-to-Noise Ratio) 성능을 측정하였다. 표 3은 다양한 신호대잡음비율에서 네 가지 방식의 잡음제거 성능을 비교하였다.

SNR 5 dB 필드환경에서의 레포터의 음성신호에 4가지의 잡음제거 방식을 적용한 실험결과는 그림 2에 나타 있다. 그림 2에서 보는 바와 같이 LSA, Gamma 1, Gamma 2의 방식의 실험결과는 유사한 성능을 갖고 있으며, Wiener 필터를 사용한 경우는 세 방식에 비해 잡음제거 성능이

그림 2. STD 환경에서의 잡음제거 결과
Fig. 2. Results of noise reduction applied to noisy speech SPC in golf videos

표 4. 선수이름 검색 성능
Table 4. Performance of player name detection

	평균	등록어	비등록어
단어기반	32.15	64.3	0
음소기반	63.6	61.5	65.7
하이브리드	70.2	74.7	65.7

저조함을 알 수 있다.

표 3을 통해서 제안된 음성결손확률기반 LSA 음성추정방식이 다른 세 방식들보다 segmental SNR과 음성인식 정확도에 있어서 근사적으로 좋은 결과를 보였으며, Gamma분포기반 음성추정에서는 각 Gamma의 값인 κ=1과 κ=2이 비슷한 성능을 보임을 알 수 있었다. 잡음제거 실험결과를 바탕으로 제안된 전체적인 시스템의 성능평가를 위해서는 연산량이 작고 성능이 우수한 LSA기반 음성추정방식을 적용하였다.

골프동영상에서 추출된 SPC, STD 음성구간에 잡음제거처리를 통한 음성정보로부터 단어기반, 음소기반 하이브리드 방식을 적용하여 선수이름을 색인한 후, 남녀 골프 선수이름 35개를 질의어로 텍스트 입력하여 검색한 실험결과는 표 4에 나타나 있다. 실험에 있어서 검색성능의 판정은 검색결과에서 앞과 뒤 1초의 마진 내에 질의어에 상응하는 마킹된 바운더리가 포함되면 검색이 성공된 것으로 판정하고, 성능평가는 식 (3)에 나타난 전체 질의어 q에 대한 weighted average MAP (mean average Precision)인 A_{MAP}를 통해 측정하였다. A_{MAP}에서는 query 마다 relevant slot 수의 차이를 반영한다.

표 5. 선수별 이벤트 검색 성능
Table 5. Performance of player event detection

	평균	등록어	비등록어
단어기반	36.75	73.5	0
음소기반	73.25	70.2	76.3
하이브리드	79.5	82.7	76.3

$$AMAP = \frac{\sum_{q}(N_q \, maP_q)}{\sum_{q} N_q} \times 100[\%],$$

$$MAP_q = \frac{\sum_{r=1}^{N} P_r \times g_r}{S_n} \times 100[\%] \quad (3)$$

여기서, P_r는 rank r에서의 precision rate, 즉 검색된 데이터 중에서 관련된 데이터의 비율, g_r는 rank r의 검색 결과가 관련된 데이터인지를 나타내는 이진함수, N_q는 질의어 q에 대한 관련된 슬롯수, S_n는 관련된 오디오 슬롯 수를 각각 나타낸다.

표 4에 나타난 바와 같이 등록 선수이름과 비등록 선수 이름에 대한 검색성능이 하이브리드 방식을 사용할 때가 70.2%로 단어기반이나 음소기반 방식보다 성능이 우수했으며, 등록 선수이름검색에 있어서는 두 방식보다 10% 이상 우수한 성능을 나타낸다.

최종적으로 이벤트 구간 검출과 선수이름 검색을 결합한 선수별 이벤트 구간에 대한 검색성능은 표 5에 나타나 있다.

표 5에 나타난 바와 같이 검출된 이벤트 구간 전과 후에 존재하는 음성구간에서 선수이름을 검출하고, 검출된 선수이름의 횟수를 측정함으로써 선수명 이벤트 구간을 검출한 성능은 이벤트와 관계없이 음성전체 구간에 대해 선수이름을 검색한 성능보다 월등히 우수함을 알 수 있다. 즉, 관중의 박수소리와 환호성, 스윙 사운드에 의해 검출된 이벤트 구간의 전후 15초 길이의 음성구간에 존재하는 선수이름 검출과 검출된 이름의 빈도횟수 측정을 통해 선수별 이벤트 구간을 효과적으로 검색할 수 있었다.

IV. 결 론

본 논문에서는 골프동영상에서 분리된 오디오 정보를 기반으로 사운드 정보로부터는 이벤트 구간을 검출하고, 음성정보로부터는 선수이름을 검출하여 효과적인 선수별 이벤트 구간을 제공하는 시스템을 제안하였다.

음성정보로부터 선수이름을 검출하기 위해서 단어기반, 음소기반, 하이브리드 방식의 세 가지 방식을 적용한 결과 하이브리드 방식이 두 방식에 비해 등록어나 비등록어 검출에서 보다 나은 성능을 보였으며, 오디오 분류를 기반으로 분할된 APP 구간과 스윙사운드 검출을 통해 추출된 이벤트구간의 15초 전후의 음성구간으로부터 검출된 선수이름을 통해 효과적으로 선수별 이벤트 구간을 검색할 수 있었다.

차후로는 골프 동영상에 비해 잡음이 많은 축구경기와 야구경기 동영상을 대상으로 본 논문에서 구현된 시스템을 적용하여 그 성능을 측정하고, 시스템을 보완할 것이다.

참 고 문 헌

1. S. E. Johnson, P. Jourlin, J. K. Spärck and P. C. Woodland, "Spoken document retrieval for TREC-9 at cambridge university," *9th TREC9*, pp. 117-126, Mar. 2000.
2. N. Moreau, H.-G. Kim and T. Sikora, "Combination of phone n-grams for a MPEG-7-based spoken document retrieval system," *In Proc. EUSIPCO 2004*, pp. 549-552, Sep. 2004.
3. P. Yu and F. Seide, "A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech," *In Proc. ICSLP 2004*, pp. 293-296, Oct. 2004.
4. C.-C. Lin, S.-H. Chen, T.-K. Truong and Y. Chang, "Audio classification and categorization based on wavelets and support vector machine," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 644-651, 2005.
5. S. Ravindran and D.V. Anderson, "Boosting as a dimensionality reduction tool for audio classification," *In Proc. ISCAS 2004*, pp. 465-468, May 2004.
6. 김형국, "오디오 정보를 이용한 골프 동영상 자동 색인 알고리즘," *한국음향학회지*, 28권, 5호, 441-446쪽, 2009.
7. C. Jingdong J. Benesty, H. Yiteng and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218-1234, 2006.
8. I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *ScienceDirect Signal Processing*, vol. 81, no. 11, pp. 2403-2418, 2001.
9. J. S. Erkelens, R. C. Hendriks, R. Heusdens and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741-1752, 2007.
10. G. Bouselmi, D. Fohr, I. Illina and J.-P. Haton, "Fully automated non-native speech recognition using confusion-based acoustic model integration," *In Proc. Interspeech 2005*, pp. 1369-1372, Sep. 2005.

저자 약력

•김 형 국 (Hyoung-Gook Kim)