2000年度 第13回 信號處理合同學術大會 論文集 第13卷 1號

멀티미디어 정보의 색인 및 검색을 위한 핵심 사운드 검출

이 용 주, 배 건 성
경북대학교 전자·전기공학부

Detection of Keysound for Indexing and Retrieval of Multimedia information

Yong Ju Lee, Keun Sung Bae
School of the Electronic & Electrical Engineering, Kyungpook National University

draball@mmirl1.knu.ac.kr, ksbae@ee.knu.ac.kr

요 약

멀티미디어 정보의 보다 효율적인 검색을 위해서는 비디오 요약정보의 생성 및 색인 작업이 필요하며, 이러한 요약정보를 만들기 위해서는 많은 시간과 비용이 소요된다. 스포츠 비디오 프로그램의 요약정보를 만들 때 오디오 신호를 이용하여 주요 장면을 검출할 경우 이러한 시간과 비용을 줄일 수 있다. 본 연구에서는 축구경기 비디오에서 주요장면을 나타내는 핵심 사운드로 주심의 호르라기 소리 및 아나운서의 "슛" 음성을 정의하고 이를 오디오 신호에서 검출하는 방법에 대해 연구하였다.

Ⅰ. 서 론

멀티미디어 정보의 홍수 속에서 비디오 프로그램의 주요 장면에 대한 요약정보를 생성하고 색인함으로써 사용자는 보다 빠르고 효율적으로 멀티미디어 정보에 접근할 수 있게 된다. 그러나, 비디오 프로그램에서 주요 내용에 대한 요약정보를 생성하는 작업은 많은 시간과 비용이 요구되므로 이러한 시간과 비용을 줄일 수 있는 방법에 대한 기술 개발이 필요하다.

비디오 프로그램에서 오디오 신호는 정보 내용의 의미 파악을 위한 주요 key가 될 수 있다. 중요한 의미를 가지는 오디오 신호로 사람의 음성, 박수 소리 등 다양한 형태로 존재할 수 있으며, 뉴스, 스포츠, 뮤직, 드라마 등 멀티미디어 정보의 종류에 따라 중요시되는 오디오 신호의 형태는 다양하게 나타난다. 따라서, 멀티미디어 정보의 종류에 따라 오디오 신호에서 내용 및 특징을 잘 나타낼 수 있는 핵심사운드를 정의하고 이를 검출함으로써 비디오 요약정보 생성 작업을 보다 효율적으로 수행할 수 있다.

본 논문에서는 축구경기 비디오 프로그램에서 경기의 내용 및 특징을 잘 나타낼 수 있는 오디오 신호로 주심의 호루라기 소리와 아나운서의 "슛" 음성을 핵심 사운드로 정의하고, 음성신호 처리 기법과 핵심어 검출 기법을 이용하여 주어진 오디오 신호로부터 이러한 핵심사운드를 검출해내는 방법을 연구하였으며, 이에 따른 실험 결과를 제시한다.

Ⅱ. 호루라기 소리의 특징 및 검출

호루라기 소리의 시간축상의 파형과 스펙트럼을 살펴보면, 호루라기 소리는 주기신호에 가까우며 16kHz로 샘플링된 신호에서 4kHz 부근에서 몇 개의 톤(tone) 성분을 가지는 것을 볼 수가 있다. 그러나 관중의 함성이나 아나운서의 음성이 섞일 경우 파형과 스펙트럼이 달라지게 되는데 그림 1에서와 같이 전체적인 스펙트럼의 모양은 배경잡음에 의존하지만 4kHz 부근의 톤성분

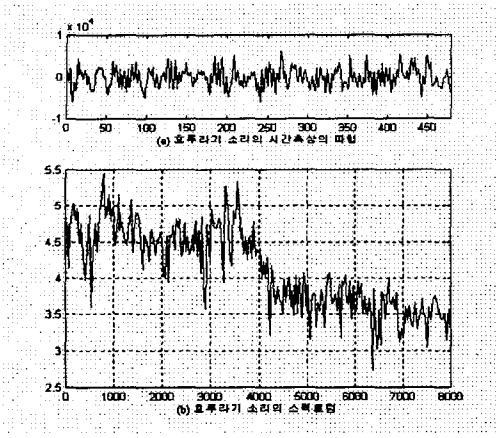

그림 1. 관중함성이 포함된 호루라기 소리의 파형 및 스펙트럼

은 상대적으로 큰 에너지를 가지므로 아나운서의 발음이나 관중 함성 등의 배경잡음에 덜 민감한 특성을 가진다. 음성신호처리 기법을 이용하여 이러한 특징을 추출하였는데, 1시간 30분 정도의 축구경기 비디오 프로그램에서 검출실험을 한 결과 2.5배의 오검출 비율을 허용할 경우 78% 정도의 검출율을 얻었다[1].

Ⅲ. 핵심어 검출기법을 이용한 "슛" 검출

축구경기 비디오에서 "슛", "골인"과 같은 아나운서의 음성은 경기의 주요장면에 해당되며, 이를 검출할 경우 요약정보 생성에 효율적으로 이용할 수 있다. 본 연구에서는 "슛" 음성을 핵심사운드로 정의하여 검출하고자 하였는데, 검색방법으로 핵심어 검출 기법[2,3]을 주로 사용하였으며, 단어의 개수가 적은 점을 감안하여 filler model을 사용하지 않는 핵심어 검출기법[4]을 사용하였다.

축구경기에서 아나운서의 음성신호는 관중의 함성에 의해 많이 손상되는데, 특히 "슛", "골인" 등의 주요한 이벤트가 발생하는 부분에서는 관중의 함성이 더욱 커지는 특성을 나타낸다. 그림 2는 축구경기에서 "슛"이 발생하는 상황에서의 오디오 신호를 나타낸 그림이다. 일반적으로 음성인식 또는 핵심어 검출 시스템은 배경잡음에 의해 그 성능이 현저하게 감소되는 경향이 있으므로, 관중함성이 포함된 음성신호를 그대로 사용할 경우 인식이 아주 어렵게 된다. 그러므로 음성인식의 전단계로서 배경잡음 제거를 위한 전처리 과정이 필요한데, MMSE(Mininum Mean Square Error)를 이용한

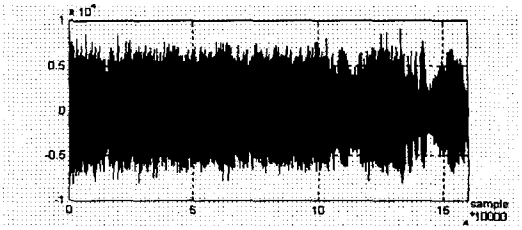

그림 2. "슛"이 포함된 오디오신호

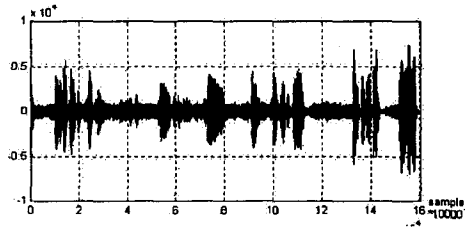

그림 3. MMSE STSA기법을 이용하여 잡음을 제거한 후의 음성

STSA(Short Time Spectral Amplitude estimator) 방식[5]을 이용하여 관중함성 등의 배경잡음을 제거하여 주었다. 그림 3은 MMSE STSA 방식을 이용하여 배경잡음이 제거된 오디오 신호를 나타낸 것이다.

MMSE 기법을 이용하여 배경잡음이 제거된 음성신호에서 핵심사운드를 검출하기 위해 "슛" 음성에 대한 HMM(Hidden Markov Model)을 작성하였다. 축구경기에서 발생하는 "슛" 음성은 일상대화에서 발성되는 것과는 달리 화자가 흥분된 상태에서 발음되므로 단어단위로 HMM 모델을 형성하는 것이 좋은 성능을 나타낼 것으로 생각되어 실험에서는 축구경기 중에 발음 된 "슛" 음성의 6부분을 발췌하여 모델 작성에 사용하였다. 특징 파라미터로는 13차의 MFCC(Mel-Frequency Cepstral Coefficient)를 사용하였으며, HMM 모델의 state수는 3으로 하였다.

본 실험에서는 핵심어 검색의 계산량을 줄이기 위해 프레임단위의 관측확률을 계산하여 핵심어 후보구간을 정하고 이 부분에 대해서만 핵심어 검출을 수행하는 방법을 사용하였다. 즉, 작성된 HMM을 이용하여 상태천이에 무관하게 프레임단위로 관측확률을 측정하였다. 다시 설명하면, 이전 상태에 무관하게 현재 프레임이 각 상태에서 관측될 확률을 구하고 최대 관측확률을 현재프레임의 관측확률로 결정하였다.

그림 4는 MMSE STSA 기법을 이용하여 잡음이 제거된 음성신호에 대해 프레임단위로 구해진 관측확률

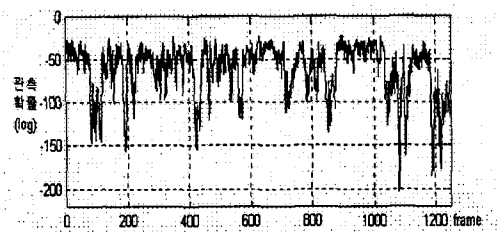

그림 4. 프레임단위 관측확률

을 나타낸다. 그림에서 보면, 아나운서의 음성이 포함되지 않은 관중함성 구간(이하, 비음성 구간이라 한다)에서 관측확률 값이 높게 나타났으며 "슛" 음성에 해당되는 구간은 다른 음성들의 경우보다는 관측확률이 높지만 비음성 구간보다는 조금 낮게 나타났다. 비음성 구간에서의 높은 관측확률은 핵심어 검출에서 오검출의 원인이 되므로 미리 제거해 줄 필요가 있으므로, 실험에서는 에너지를 이용한 음성/비음성 프레임의 판단을 통해 음성프레임에 대해서만 관측확률을 측정하였다. 그림 5는 음성/비음성 프레임의 판단을 통해 전체신호에서 음성부분 만을 나타낸 그림이고, 그림 6은 음성프레임으로 판단된 구간에 대한 프레임단위 관측확률을 나타낸 것이다. 아나운서의 음성중 유성음 부분이 주로 음성으로 판단되었으며, 다른 음성에 비해 핵심어 "슛"이 존재하는 구간에서 관측확률이 높게 나타났었다.

주어진 오디오신호에서 핵심어가 포함된 구간은 관측확률이 높게 나타나므로 프레임단위 관측확률을 이용하여 핵심어 후보구간을 정할 수 있다. 프레임단위 관측확률에서 핵심어 후보로 채택된 구간에 대해서 핵심어의 시작프레임과 마지막프레임을 이동시키면서 핵심어 관측확률을 구한다. 구해진 최대 관측확률을 미리 정의된 기준값과 비교하여 핵심어의 유무를 판단하는 방식을 이용하였다.

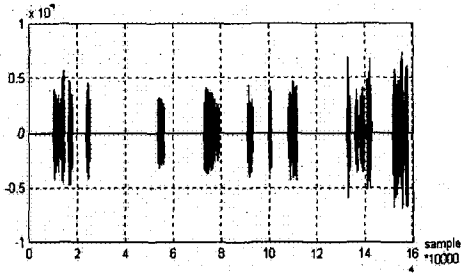

그림 5. 에너지를 이용한 음성 비음성 판단

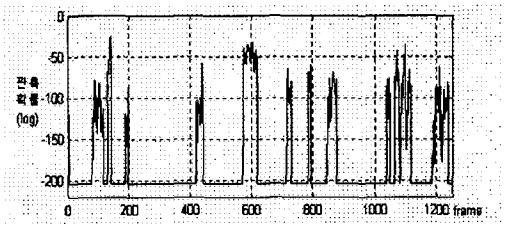

그림 6. 음성부분에 해당하는 프레임별 관측 확률

호루라기 소리와 "슛" 음성 검출 실험을 위해 실제 축구경기 비디오 프로그램에서 오디오 신호를 수집하였다. 서로 다른 3개의 축구경기에 대해서 각각 12분, 28분, 46분 가량의 오디오 신호를 녹음하여 실험에 사용하였고, 두 번째 경기와 세 번째 경기에서 "슛" 단어가 발생하는 부분의 음성신호를 발췌하여 핵심어 검출에 사용하였다 이때, 오디오신호는 16kHz로 샘플링 하였고, 16bits/sample로 양자화 하였다. .

"슛" 단어 검출 실험은 경기 2에 포함된 "슛"단어를 훈련 데이터로 사용하였고, 경기 3에서 "슛" 단어가 발생하는 구간 8부분을 각각 10초의 길이로 잘라내어 검출 실험에 사용하였다. 표 1은 8부분에 대한 "슛" 검출 결과를 나타낸다.

표 1. "슛" 검출 결과

문장수 (길이)	단어수	검출수	오검출수
8 (각 10초)	8	4	5

"슛"이 짧게 발음된 부분에서는 관측확률이 낮게 나타나서 검출되지 않은 부분이 발생하였고, 해설자에 의해 동시에 다른 단어가 발음되는 경우에도 "슛"이 검출되지 않는 경우가 발생하였다. 정확히 발음된 경우에도 관측확률이 정해진 기준값보다 작게 나타나는 부분이 발생한 곳도 있었으며, 검출된 핵심어의 빈도수에 비해 다수의 오검출 횟수가 나타남을 볼 수 있다. 이는 HMM 모델 훈련에 사용된 데이터가 적어 경기 상황에 따라 다양한 형태로 나타나는 "슛" 음성을 제대로 모델링 하기에는 부족했고, 주변잡음의 크기가 경기에 따라 변하는 스포츠 경기의 특징을 잘 모델링 하지 못했기 때문이라고 생각된다.

V. 결 론

본 연구에서는 축구경기 비디오에서 효율적인 요약

정보 생성에 이용될 수 있는 오디오 신호의 핵심사운드로 주심의 호르라기 소리 및 아나운서의 "슛" 음성을 정의하고 이를 검출하는 방법에 대해 실험하였다. 실제 축구경기의 경우 관중함성, 휘파람소리, 북소리 등의 다양한 잡음의 영향으로 인해 호루라기 소리 검출시 많은 미검출과 오검출이 발생하였다. 또한, "슛" 단어 검출 실험에서도 HMM을 작성하기 위한 훈련 데이타가 충분하지 못하였고 특히, 배경잡음으로 인해 적절한 HMM을 작성하지 못하여 많은 오검출이 발생한 것으로 생각된다. 따라서, 현재 제시한 방법으로는 스포츠 비디오의 오디오 신호를 비디오 요약정보 생성에 효율적으로 이용하기가 어렵다고 본다. 그러므로, 앞으로 heuristic한 신호처리 기법을 통한 특징 파라미터 추출과 배경잡음에 강인한 음성인식 및 핵심어 검출 기법에 대한 연구를 통해 스포츠 비디오에서의 핵심사운드 검출에 대한 연구를 계속할 계획이다.

본 연구는 한국전자통신연구원 방송기술연구부의 지원에 의해 수행되었으며, 지원에 감사드립니다.

참고 문헌

[1] 이용주, 배건성, "스포츠 비디오에서 호루라기 소리 검출을 위한 특징 파라미터 추출", 제17회 음성통신 및 신호처리 대회, 논문집, pp 151-154, vol. 17, No. 1, 2000, 8

[2] Dharanipragada, S, Roukos, S. "New Word Detection in Audio-Indexing", IEEE ASRU, pp 551-557, 1997

[3] Wilpon, J.G, Miller L.G, Modi, P. "Improvements and Applications for Key Word Recognition Using Hidden Markov Modelling Techniques", IEEE ASSP, pp 309-312 vol.1, 1991

[4] S. Sunil, Supriyo Palit and T.V. Sreenivas, "HMM based fast keyword spotting algorithm with no garbage models",IEEE ICSP, Sep, 1997

[5] Yariv Ephraim, David Malah, "Speech Enhancement Using a Minimun Mean-square Error Short-Time Spectral Amplitude Estimator", IEEE ASSP, vol.32, No.6, Dec, 1984