문자정보 기반 비디오 분할에서 성능 향상을 위한 음성신호처리

이용주*, 손종목*, 강경옥**, 배건성*
*경북대학교 전자 · 전기공학부,
**한국전자통신연구원 무선 · 방송기술연구소

Speech Signal Processing for Performance Improvement of Text-Based Video Segmentation

Yong Ju Lee*, Jong Mok Son*, Keun Sung Bae*, Kyeongok Kang**
*School of Electronic and Electrical Engineering, Kyungpook National University,
**Electronics and Telecommunications Research Institude

요 약

비디오 프로그램에서 영상 내에 포함되어 있는 문자정보는 동영상의 내용 검색 및 색인을 위한 비디오 분할에 사용될 수 있다. 일반적으로 장면 내에 포함되어 있는 문자들은 해상도가 낮고 글자 크기와 형태가 다양하기 때문에 추출과 인식이 어려울 뿐만 아니라 의도하지 않은 배경화면의 문자인 경우도 많기 때문에 내용기반 검색에는 사용되기가 어렵다. 그러나 비디오 내에 포함된 문자정보가 나타나는 시작 프레임과 끝나는 프레임을 검출하여 비디오 프로그램을 분할함으로써 내용기반 요약정보를 만들 수 있으며, 동영상의 내용 검색 및 색인에 사용할 수 있다.

일반적으로 문자정보의 추출에 의해서 비디오를 분할할 때 음성정보는 전혀 고려되지 않으므로 분할된 비디오 정보를 재생할 경우 음성신호가 단어 또는 어절/음절의 임의의 점에서 시작되고 끝나게 되어 듣기에 부자연스럽게 된다. 따라서 본 논문에서는 뉴스방송의 비디오 프로그램에서 문자정보가 포함되어 있는 비디오의 시작 프레임과 끝 프레임을 중심으로 그에 대응되는 구간의 음성신호를 검출한 후 이를 적절히 처리하여 분할 된 비디오를 재생할 때 음성신호가 보다 자연스럽게 들릴 수 있도록 하는 방법에 대해 연구하였다.

Ⅰ. 서 론

일반적으로 동영상의 내용검색 및 색인을 위한 방법으로 비디오를 분할 할 때 문자정보의 추출, 카메라의 움직임 등을 사용하게 된다 [1]. 음성정보를 고려하지 않은 이러한 방식으로 만들어진 요약정보는 비디오 재생시에 음성신호가 단어 또는 어절/음절의 중간부터 시작됨으로 인해 귀에 거슬리게 된다. 특히 길지 않은 영상정보가 계속적으로 연결되어 있다면 이러한 귀에 거슬리는 소리가 반복적으로 발생하게 된다. 뉴스방송의 비디오 프로그램에서 문자정보가 포함되어 있는 비디오의 시작 프레임과 끝 프레임의 경우 그 길이가 수초밖에 되지 않는 짧은 경우가 대부분이고, 이러한 요약정보를 연속적으로 저장할 경우 앞에서 언급한 것처럼 재생시에 귀에 거슬리는 음성신호가 반복적으로 발생하게 된다. 이러한 현상을 방지하면서 비디오를 분할하기 위해서는 분할된 비디오의 음성신호를 단어 또는 어절/음절 단위로 분할하고 이러한 음성분할을 기

준으로 하여 비디오를 재분할하는 방식을 사용해야 할 것이다. 본 논문에서는 이러한 관점에서 비디오 신호를 기준으로 분할된 음성신호를 분석하여 단어 또는 어절/음절이 끊어져서 나타나지 않도록 음성을 재분할하는 방법을 연구하고 실험결과를 제시하고자 한다.

연속적으로 발음되는 음성신호의 경우 단어와 단어사이 또는 어절과 어절 사이의 구간이 음성구간에 비해 매우 짧다. 본 연구에서는 어절/음절 사이의 짧은 묵음구간을 신호의 에너지와 피치정보를 이용하여 음성신호와 구분해 내는 방법을 연구하고, 실제로 뉴스방송의 비디오 프로그램에서 문자정보를 기준으로 분할된 프레임의 음성만을 추출하여 이를 대상으로 음성을 재분할하는 실험을 수행하였다. 그리고 이러한 실험을 통해 발생할 수 있는 문제점 및 보완사항을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 뉴스방송에서의 음성신호를 간략히 설명하고, 에너지와 피치정보를 이용하여 어절/음절 단위로 음성을 분할하는 방법을 설명한다. 3장에서는 뉴스방송의 음성을 대상으로 본 연구에서 제안한 재분할 방법을 이용한 실험의 결과를 제시하고 4장에서 결론을 맺는다.

II. 음성신호의 재분할 알고리듬

일반적인 뉴스방송의 음성은 다음과 같이 두가지로 나누어 볼 수가 있다. 앵커가 실내에서 발음하는 부분과 리포터가 실외에서 발음하는 부분이다. 그림 1은 실내에서 발음한 앵커의 음성과 리포터가 실외에서 발음한 음성을 나타낸 것이다. 실내에서 발음하는 경우에는 주변 잡음이 거의 없으므로 깨끗한 음성신호를 얻을 수 있다. 하지만, 실외에서 발음하는 경우에는 음성신호에 주변 잡음이 섞이고, 리포터가 아닌 주위 사람들의 음성이 섞이는 경우도 발생하게 된다. 따라서 음성신호의 에너지만으로는 어절이 끝나는 묵음 구간의 검출이 어렵게 된다. 본 연구에서는 음성신호의 피치 및 에너지 정보를 함께 이용하여 이러한 환경에서의 음성신호를 분할하는 실험을 수행하였다

뉴스방송에서 문자정보를 기준으로 분할한 비디오의 음성신호와 같은 조건의 음성신호를 대상으로 실험을 하기 위해 실제 뉴스방송의 비디오에서 문자정보가 나타나기 시작하는 프레임에서부터 문자정보가 사라지는 프레임까지의 음성신호를 수작업으로 검출하여 사용하였다. 음성신호는 16 kHz로 샘플링 된 신호이며, 주어진 음성신호의 시작점에서 부터는 뒤쪽으로, 끝점에서 부터는 앞쪽으로 적절한 묵음 구간을 검출하여 음성신호를 재분할 하고자 한다. 본 연구에서는 웨이브렛 변환을 이용하여 피치정보를 구했는데[2][3], 피치 및 에너지 정보를 이용하여 음성신호를 재분할 하는 과정은 다음과 같다.

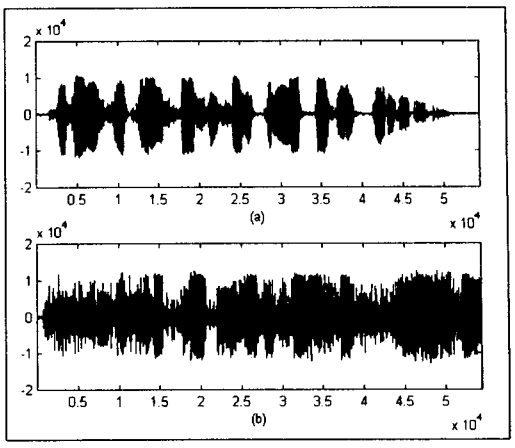


그림 1. 뉴스방송 음성의 예
(a) 실내에서의 음성
(b) 실외에서의 음성

Step 1: 피치검출을 위한 분석 프레임의 길이는 20 msec로 하여 10 msec씩 이동하면서 피치를 구한다. 피치가 4 프레임 이상 존재하지 않으면 잘못 구해진 피치로 간주하고 무시하며, 3 프레임을 기준으로 메디안 필터링을 수행하여 최종 피치궤적을 구한다.

Step 2: 프레임 길이를 10 msec로 하여 음성신호의 에너지를 구한다[4]. 현재 프레임을 기준으로 이전 50 프레임과 이후 50 프레임에서 최대에너지 값 E_{max}와 최소에너지 값 E_{min}을 구한다. 여기서 구한 최대/최소에너지 값의 차를 이용하여 기준값 E_{th}를 정하고, 이를 이용하여 E_{th} 이상의 값을 갖는 각 프레임을 음성 구간으로 판정한다. 기준값 E_{th}는 식 (1)과 같이 얻어진다.

$$E_{th} = (\log E_{max} - \log E_{min})/d_o + \log E_{min} \quad (1)$$

여기서 d_0는 음성신호의 잡음정도에 따라 다

른 값을 갖는 상수를 나타낸다. 즉, 최대에너지와 최소에너지 값의 차이가 적은 구간에서는 잡음이 포함된 구간으로 간주하여 d_0를 3.5로 정해주었고, 최대에너지와 최소에너지의 차이가 큰 구간에서는 잡음이 없는 구간으로 간주하여 d_0를 3으로 정해주었다

Step 3: 각 프레임 단위로 피치값이 zero이고 에너지 값이 E_{th} 보다 적으면 비음성 프레임으로 간주한다.

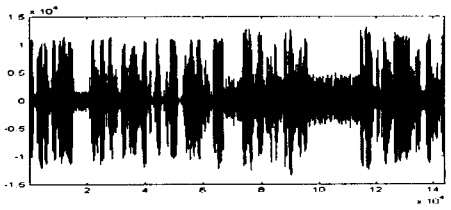
Step 4: (후처리과정) 연속되는 음성구간이 5 프레임 이하로 너무 짧으면 비음성 구간으로 판정하고, 또한 연속되는 비음성구간이 3 프레임 미만이면 음성구간으로 판정한다.

Step 5: 프레임 단위로 주어진 음성신호의 시작점에서 뒤쪽으로, 끝점에서 앞쪽으로 탐색하여 비음성 프레임이 연속 8개 이상인 구간을 찾아 재분할 된 음성신호의 시작/끝 프레임으로 결정한다.

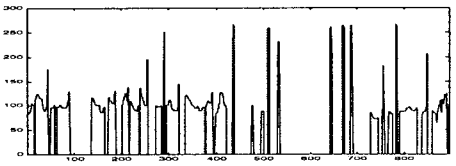
Ⅲ. 실험 및 검토

실제로 뉴스방송에서 문자정보가 나타나는 비디오 프레임의 시작부터 문자정보가 사라지는 프레임까지의 음성신호를 서로 다른 6장면의 내용에 대하여 음성신호를 수작업으로 검출하여 실험에 이용하였다. 그림 2는 뉴스방송에서 문자정보가 나타나기 시작해서 끝나는 구간에 대응되는 음성신호와 그의 피치 및 에너지 정보를 이용하여 묵음구간을 검출하고 음성신호를 재분할 하는 예를 보인 것이다. 그림 2(a)에서, 비디오 분할에 따라 나누어진 음성신호는 묵음 구간에서 나누어지지 않고 어절/음절의 임의의 점에서 끊어져 재생시에 듣기에 부자연스럽게 될 수 있음을 알 수 있다. 그림 2(b)~(e)는 음성신호의 분석에 의해 얻어지는 피치정보 및 에너지 값에 의한 음성/비음성 구간 판정결과와 최종적으로 재분할 되는 음성신호의 구간을 나타내고 있다.

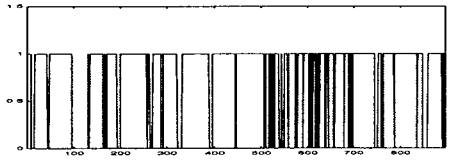
그림 3 및 4는 뉴스방송 프로그램에서 비디오 분할에서 얻어진 음성신호와 본 논문에서 제안한 음성신호처리 기법을 통해 재분할 된 음성신호를 나타낸 것이다. 그림 3은 /로 종료되면서 다시 그/의 음성신호인데 재분할 된 음성신호는 <종료되면서 다시 그>에 해당되며, 그림 4는 /투 대비 태세령과 정보감시 ..... 해군은 북한 서해안에 싣고온 미사일/의 음성신호인데 재분할 된 음성신호는 <정보감시 ..... 해군은 북한 서해안에 싣>에 해당된다. 각 그림에서 볼 수 있듯이 원래 음성신호는 처음과 마지막 어절/음절이 비디오 단위의 분할로 인해 손상되어 있다. 이러한 손상된 어절/음절은 음성의 재분할을 통해 제거되고 손상되지 않은 어절/음절이 장면의 처음과 마지막에 위치하게 됨을 볼 수 있다.



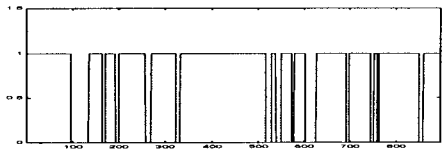
(a) 원래 음성



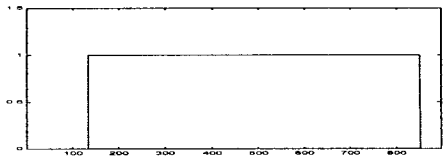
(b) pitch 정보



(c) 에너지에 의한 음성/비음성 구간 판정



(d) pitch와 에너지 정보를 이용한 음성/비음성 구간 판정



(e) 재분할 된 음성구간 영역

그림 2. 음성신호 분석 예

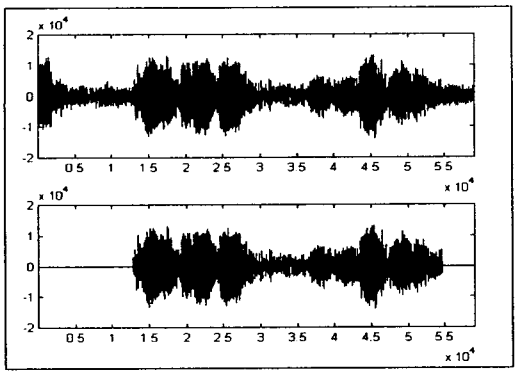


그림 3. 원래의 음성과 재분할한 후의 음성 (I)

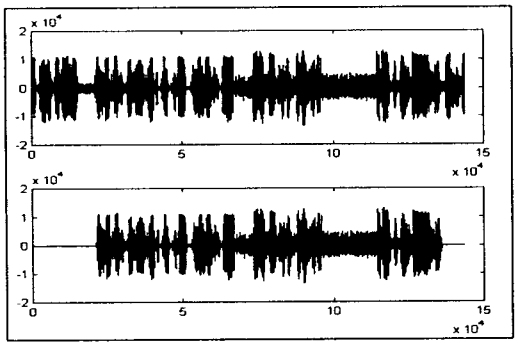


그림 4. 원래의 음성과 재분할한 후의 음성 (II)

표 1은 6개 장면의 비디오 분할 데이터에 대해 본 논문에서 제시한 방법으로 음성신호를 재분할한 결과를 재분할 하기전의 음성신호의 길이와 비교한 것이다. 원래 음성신호에 비해 재분할 한 후 음성신호의 길이는 평균 76% 정도로 줄어듬을 볼 수 있다. 따라서, 재분할 된 음성신호의 시작점과 끝점을 기준으로 비디오 정보를 요약할 경우 문자정보의 내용은 그대로 유지하면서 비디오 단위의 잘못된 분할에 의해 생긴 어절/음절의 손상을 방지하여 귀에 거슬리는 음성을 효과적으로 감소시킬 수 있다. 또한 저장에 필요한 비디오 프레임의 수도 음성신호의 길이가 줄어든 크기에 비례하여 감소되므로 보다 효율적으로 비디오 요약정보를 만들 수 있다.

표 1. 재분할 전/후의 음성신호의 길이 비교

구 분	원래 음성 샘플수	재분할 후의 샘플수	비 율 (%) (재분할/원래 음성)
장면 1	59261	41920	70.7
장면 2	144145	114880	79.7
장면 3	155890	107040	68.7
장면 4	132934	103200	77.6
장면 5	84886	63200	74.5
장면 6	264265	232480	88.0
평 균	140230.16	110453.33	76.53

Ⅳ. 결 론

일반적으로 문자정보의 추출에 의해서 비디오를 분할할 때 음성정보는 전혀 고려되지 않으므로 분할된 비디오 정보를 재생할 경우 음성신호가 단어 또는 어절/음절의 임의의 점에서 시작되고 끝나게 되어 듣기에 부자연스럽게 된다. 본 연구에서는 음성신호의 피치정보와 에너지를 이용하여 음성신호를 재분할함으로써 비디오 데이터의 문자정보 내용은 그대로 유지하면서 비디오 단위의 분할에 의해 생기는 음성신호의 어절/음절의 손상을 방지하여 재생시에 음성이 보다 자연스럽게 들릴 수 있도록 하였다. 또한, 본 연구에서 제안된 방법을 이용할 경우 저장에 필요한 비디오 프레임의 수도 음성신호의 길이가 줄어든 크기에 비례하여 감소되므로 보다 효율적으로 비디오 요약정보를 만들 수 있다.

본 논문에서 제시한 어절/음절 단위의 분할은 단어의 강세, 단어의 마지막 음소 등에 의한 영향으로 인해 여전히 조금씩의 부자연스러움을 가지게 됨을 알 수 있었다. 또한 방송뉴스의 경우 일반 낭독체/대화체 음성과는 달리 어절 사이의 묵음 구간이 폐쇄음 종성으로 끝나는 음절사이의 묵음구간 보다 짧은 경우가 많아 어절 단위의 구간 검출이 어려운 경우가 많았다. 따라서 향후에는 이런 문제점을 개선하여 어절단위로 검출할 수 있는 연구와, 뉴스방송의 캡션정보를 이용하여 문장단위의 음성구간을 검출하여 비디오를 분할하는 방법에 대한 연구가 필요하다.

본 연구는 한국전자통신연구원 방송기술연구부의 지원에 의해 수행되었습니다. 지원에 감사드립니다.

참 고 문 헌

[1] 나지훈, 조정원, 최병욱 "뉴스 영상에서의 자막영역 추출 및 문자인식", 신호처리합동학술대회 논문집, pp.393-396, 1999

[2] 손영호, 배건성 "웨이브렛 변환을 이용한 음성신호의 유/무성음/묵음 분류", 음성통신 및 신호처리 워크샵 논문집, pp.449-453, 1998

[3] 손영호, 석종원, 배건성, "EGG 신호와 웨이브렛 변환된 음성신호와의 Epoch 검출 및 비교", 신호처리합동학술대회 논문집, pp.743-746, 1997

[4] L. R. Rabiner, R. W. Schafer, Digital Processing of Speech signals, Prentice-Hall, 1978