모음 기반 화자 식별 모델을 이용한 화자 인덱싱

금지수*, 박찬호**, 이현수*

경희대학교 컴퓨터공학과*, 부천대학 인터넷응용과**

Speaker Indexing using Vowel Based Speaker Identification Model

Ji Soo Kum*, Chan Ho Park**, Hyon Soo Lee*

Dept. of Computer Engineering, Kyung Hee University*
Dept. of Internet Information Science, Bucheon College**
tbno@cann.khu.ac.kr, chpark@hangil.bucheon.ac.kr, leehs@khu.ac.kr

요약

본 논문에서는 음성 데이터에서 동일한 화자의 음성 구간을 찾아내는 화자 인덱싱(Speaker Indexing) 기술 중 사전 화자 모델링 과정을 통한 인덱싱 방법을 제안하고 실험하였다.

제안한 인덱싱 방법은 문장 독립(Text Independent) 화자 식별(Speaker Identification)에 사용할 수 있는 모음(Vowel)에 대해 특징 파라미터를 추출하고, 이를 바탕으로 화자별 모델을 구성하였다. 인덱싱은 음성 구간에서 모음의 위치를 검출하고, 구성한 화자 모델과의 거리 계산을 통하여 가장 가까운 모델을 식별된 결과로 한다. 그리고 식별된 결과는 화자 구간 변화와 음성 데이터의 특성을 바탕으로 필터링 과정을 거쳐 최종적인 인덱싱 결과를 얻는다. 화자 인덱싱 실험 대상으로 방송 뉴스를 녹음하여 10명의 화자 모델을 구성하였고, 인덱싱 실험을 수행한 결과 91.8%의 화자 인덱싱 성능을 얻었다.

1. 서론

화자 인덱싱 기술은 음성 데이터에서 동일한 화자의 음성 구간을 찾아내는 기술로 다수의 화자 음성에서 특정 화자의 음성만을 찾아내고 내용을 파악하여 방송 뉴스나 토론 등의 문서 요약과 자료 검색 등에 응용할 수 있다[1][2].

현재의 화자 인덱싱 연구 방법들은 크게 화자 모델을 사전에 구성하고 음성 구간을 모델과 비교하는 방법과, 화자의 음성 특성과 인원 수 등의 사전 정보를 알지 못하는 상태에서 온라인으로 화자 모델을 구성하고 인덱싱을 수행하는 방법으로 분류할 수 있다[3][4].

그러나 온라인 상태에서 화자 모델과 인덱싱을 수행하거나, 화자의 변화구간만을 찾아내는 방법은 찾아낸 동일한 화자의 음성 구간에 대해 어느 화자의 음성 구간인지 구별하기 위해 별도의 화자 정보 추출 과정과 식별 단계가 필요하다. 따라서, 본 연구에서는 음성 데이터에 참여하고 있는 화자의 정보를 사전에 추출하여 화자 모델을 구성하고, 문장 독립형 화자 식별에 사용할 수 있는 모음 구간을 추출하여 구성한 화자 모델과 거리를 계산하여 인덱싱을 수행하고 결과를 평가하였다.

본 논문의 구성은 2절에서 화자 식별 모델을 이용하는 화자 인덱싱 시스템 구성을 설명하고, 3절에서는 화자 모델 구성 방법과 식별 방법에 대해 기술하였다. 그리고 4절에서는 인덱싱 실험과 결과 분석을 하였으며, 마지막으로 5절에서는 보다 높은 인덱싱을 위한 향후 연구 방향에 대해 기술하였다.

2. 제안한 화자 인덱싱 시스템 구성

화자 식별 모델을 이용한 화자 인덱싱 시스템은 그림 1과 같이 끝점 검출된 음성 구간에서 모음의 위치를 찾아내는 단계와 찾아낸 모음을 인식하는 단

계, 그리고 인식한 결과를 바탕으로 화자 모델과 비교하여 식별하는 단계로 구성된다. 식별된 결과는 화자 구간 변화와 음성 데이터의 특성에 따른 필터링 과정을 거쳐 최종적인 인덱싱 결과를 얻는다.

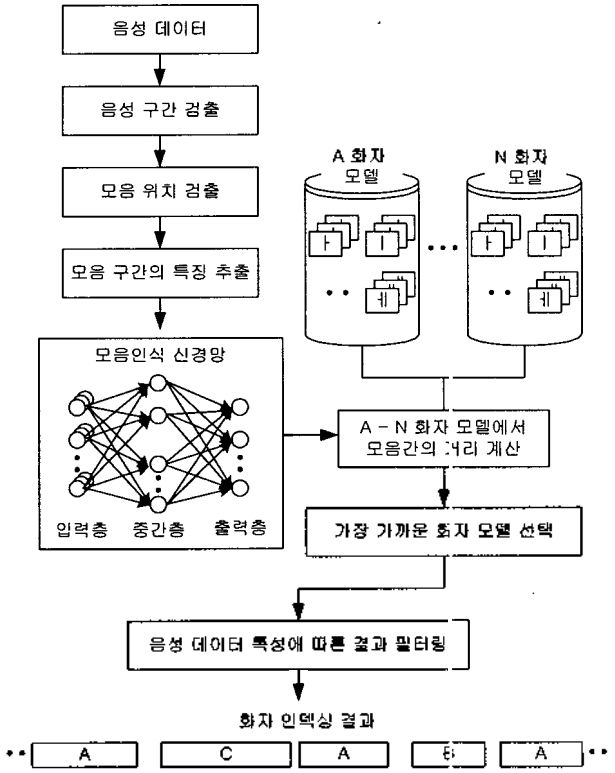


그림 1. 제안한 화자 인덱싱 시스템 구성

2.1 음성 구간 검출

음성 구간 검출은 음성 데이터에서 묵음 구간을 제거하므로 인덱싱에 불필요한 구간을 줄일 수 있다. 본 연구에서는 음성언어 처리에서 널리 사용되는 단구간 에너지(Short-Term Energy)와 단구간 영교차율(Short-Term Zerocrossing)을 측정하고 임계값을 적용하여 음성 구간과 묵음 구간을 구분하였다.

2.2 모음 위치 검출

모음 위치 검출 방법은 모음 검출을 기반으로 하는 문장 독립형 화자 인식 시스템이나 음성인식 시스템에서 다양하게 연구되고 있다[5]. 본 연구에서는 그림 2에서와 같이 음성 구간 검출에 적용했던 단구간 길이보다 짧은 길이로 단구간 에너지와 영교차율을 측정하고, 단구간 에너지 측정에 의해 나타난 피크(Peak)들 사이의 시간 차이와 평균 에너지, 영교차율에 임계 조건을 적용하여 모음 위치를 검출하였다.

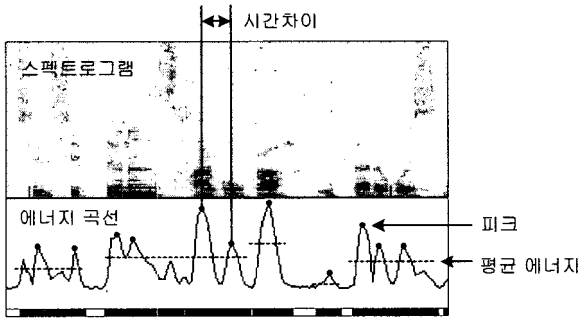


그림 2. 임계 조건을 적용한 모음 검출

2.3 모음인식 신경망

모음 위치가 검출되면 검출된 위치를 기준으로 전후에 있는 프레임에서 특징 파라미터를 추출한다. 그리고 추출된 파라미터는 모음 인식을 위한 에러 역전파(Error Back-Propagation) 신경망의 입력 값으로 사용하고, 신경망은 학습 속도 향상을 위해 적응적으로 학습률을 변화시키면서 학습한다.

3. 화자 식별 모델 구성 및 인덱싱

화자 식별을 위한 모델은 HMM이나 신경망 등 여러 가지 방법으로 구성할 수 있으나, 본 연구에서는 문장 독립의 화자 식별을 수행할 수 있고 화자의 음성 특성에 따른 모델 재구성 등을 고려하여 모음을 화자 식별 단위로 하여 특징 파라미터들 사이의 거리를 측정 방법으로 사용하였다.

3.1 화자 식별 모델 구성

화자 모델 구성을 위해 음성 데이터에서 특정 화자의 음성 구간을 찾고, 모음 검출과 인식 과정을 통해 모델 구성을 위한 화자별 모음 데이터를 저장하였다. 저장한 모음 데이터는 여러 단어에서 k번 발음된 모음으로 m개의 프레임 길이이며, n차의 특징 파라미터로 구성되어 있다. 저장한 모음에 대해 프레임별 n차의 특징 파라미터를 차수별로 비교한 결과 큰 변화없이 일정한 범위 내에서 값들이 분포하고 있으며, 동일한 모음에 대해서 화자별 비교를 했

을 때 차수별 차이를 확인할 수 있었다.

화자 모델은 모음 검출 구간의 전후 프레임에 대해 n차의 LPC 캡스트럼(Cepstrum) 계수를 구하고, 프레임들에 대한 차수별 평균을 구하여 화자별로 k번 발음한 모음들로 구성하였다. 모델 구성은 한국어 음성에서 주로 사용되는 모음으로 단순 모음 7개 /ㅣ, ㅔ, ㅏ, ㅜ, ㅡ, ㅗ, ㅓ/와 j-계 이중 모음 4개 /ㅑ, ㅕ, ㅛ, ㅠ/, 그리고 ㅟ-계 이중 모음 /ㅞ/로 모두 12개를 사용하였다.

3.2 화자 모델에 의한 식별

음성 데이터에서 모음 검출과 인식이 수행되면 구성한 화자 모델에서 화자별 모음 사이의 거리를 계산 하여 최소의 값을 갖는 모델을 식별 결과로 한다. 거리 계산에는 각 화자별로 k번 발음한 모음에 대해 유클리디안(Euclidean) 거리를 계산하였다.

for 전체 화자 모델 중에서 인식된 모음에 대해
  for k번 발음하여 구성한 모음에 대해

$$Dist(x, y_k) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

  end
end

3.3 식별 결과의 필터링

제안한 방법에서 화자 인덱싱은 끝점 검출된 음성 구간에 대한 식별 결과를 바탕으로 이루어진다. 식별 결과에 있어서 끝점 검출된 음성 구간에 포함되어 있는 모음이 모두 동일한 화자의 음성으로 식별되는 경우도 있지만, 일부의 모음이 다른 화자의 음성으로 식별되는 경우도 발생한다. 이러한 경우에는 그림 3과 같은 세 가지 방법의 필터링을 적용하였다. 첫 번째 방법은 끝점 검출된 음성 구간에서 가장 많이 나타나는 화자를 식별 결과로 하는 것이고, 두 번째 방법은 끝점 검출된 음성 구간에 동일한 수의 화자가 나타나거나 동일한 화자와 그렇지 않은 화자의 수가 같은 경우에 대해 식별된 모델 다음으로 거리가 가까운 모델을 찾아 필터링 한다. 마지막 방법은 음성 구간의 길이가 짧을 경우 인접한 음성 구간과의 시간 차이를 계산하여 화자 변화가 일어나지 않을 정도의 길이라면 인접한 음성 구간의 인덱싱 결과를 바탕으로 필터링 한다.

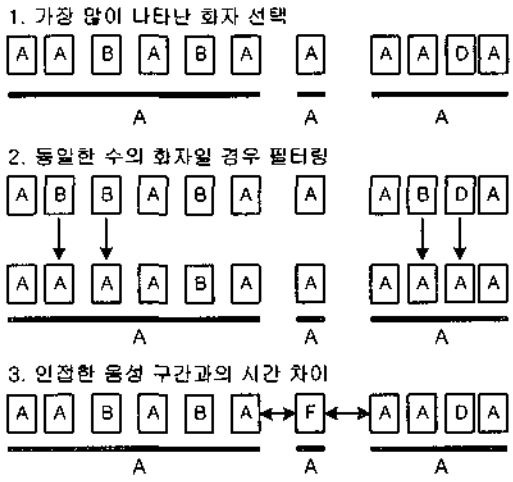


그림 3. 식별 결과의 필터링 방법

4. 인덱싱 실험 및 결과 분석

4.1 음성 DB 및 분석 조건

제안한 화자 인덱싱 시스템의 성능을 평가하기 위하여 방송 뉴스 40분 분량의 데이터 4회분을 수집하였다. 수집한 데이터 2회분에서 남녀 앵커를 포함하여 10명의 화자가 12개의 모음을 3회씩 발음한 데이터를 추출하여 화자 모델을 구성하였다. 인덱싱을 위한 음성데이터 분석 조건은 표 1과 같고, 나머지 2회분의 데이터에서 화자 모델 구성에 포함된 화자의 음성을 추출하여 인덱싱 실험에 사용하였다.

표 1. 음성데이터 분석 조건

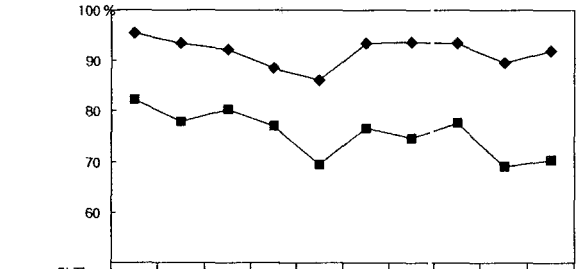
구분	분석 조건
샘플링 주파수	16KHz
양자화	16Bit
음성 검출 프레임 길이	400샘플(25ms)
모음 검출 프레임 길이	256샘플(16ms)
창함수	해밍(Hamming)
특징 파라미터	LPC 캡스트럼 16차

4.2 모음인식 신경망 구성

모음인식 신경망은 모델 구성에 참여한 화자와 참여하지 않은 화자의 음성 데이터에서 12개의 모음에 대해 10회씩 발음한 데이터를 가지고 학습하였다. 신경망의 학습률을 적응적으로 변경하였고, 중간 뉴런 수 80개에서 가장 좋은 성능을 얻었다.

4.3 인덱싱 결과 및 분석

인덱싱 결과는 화자 10명의 음성 데이터를 대상으로 끝점 검출된 전체 단어에 대해 필터링 하지 않은 결과와 필터링 과정을 수행한 결과를 화자별 비율로 나타냈다.



화자	A	B	C	D	E	F	G	H	I	J
필터링	95.5	93.4	92.2	88.5	86.2	93.4	93.7	93.5	89.7	91.9
1차결과	82.3	78	80.4	77	69.5	76.7	74.7	77.7	69	70.3

그림 4. 화자 10명에 대한 인덱싱 결과

화자 10명의 음성데이터에 대해 필터링 과정을 수행하지 않았을 경우에는 평균 75.56%, 필터링 과정을 수행하였을 경우에는 평균 91.8%의 인덱싱 결과를 얻었다. 인덱싱 결과를 분석한 결과 화자 모델을 구성하는데 사용했던 화자의 음성 데이터와 인덱싱 실험에 사용했던 음성 데이터의 환경 차이가 보완해야할 사항으로 분석되었다. 특히, 앵커와 같이 동일한 장소에서 발음한 음성과 현장을 돌아다니며 발음한 기자의 음성데이터 사이에서 인덱싱 성능 차이가 나타났다. 또한 한 명의 화자 음성만 존재하지 않고 여러 사람의 음성이 섞여 있는 경우와 배경음이 같이 녹음되어 있는 경우도 인덱싱 성능에 영향을 미쳤다.

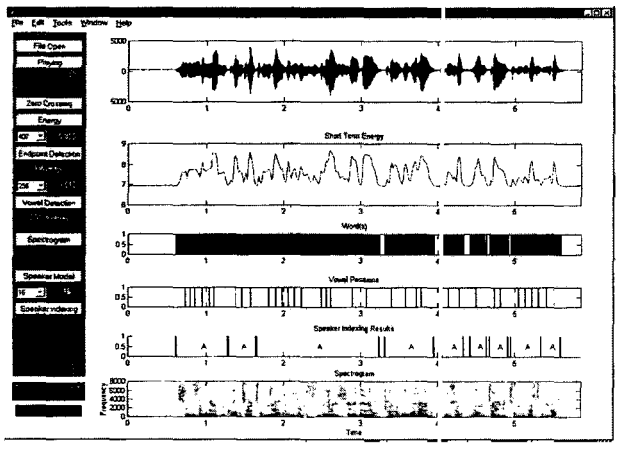


그림 5. 구현한 화자 인덱싱 프로그램

5. 결론 및 향후 연구 방향

본 연구에서는 음성 데이터에서 화자의 음성 특성을 나타낼 수 있는 음성 구간을 찾고, 화자 모델을 구성하여 인덱싱을 수행하였다. 제안한 인덱싱 방법에서 화자의 음성 특성 구간을 모음 구간으로 선택하였는데, 이것은 문장 독립의 화자 인식에서 많은 모델을 필요로 하지 않으며, 현재 음성인식에서 많이 사용되는 음소 단위의 인식기와 결합하여 음성인식과 인덱싱 과정을 동시에 수행하기 위해서이다.

향후 연구 방향으로는 인덱싱 결과에서 나타났던 화자 모델 구성과 인덱싱 실험 환경의 차이 극복이며, 또한 여러 사람의 음성이 섞여있는 경우와 배경음이 섞인 경우에 대해 연구할 계획이다.

참고문헌

[1] P. Delacourt, C.J. Wellekens, "DISTBIC: Aspeaker-based segmentation for audio data indexing", Speech Communication 32, 2000

[2] Deb K. Roy, "Speaker Indexing Using Neural Network Clustering of Vowel Spectra", International Journal of Speech Technology, Vol 1. No 2. 1997

[3] M. Nishida, Y. Ariki, "SPEAKER INDEXING FOR NEWS ARTICLES, DEBATES AND DRAMA IN BROADCASTED TV PROGRAMS", ICMCS 99, Vol 2, 1999

[4] L. Wilcox, F. Chen, D. Kimber, V. Balasubramaman, "SEGMENTATION OF SPEECH USING SPEAKER IDENTIFICATION", ICASSP 94, Vol 1, 1994

[5] Nikos Fakotakis, Anastasios Tsopanoglou and George Kokkinakis, "A text-independent speaker recognition system based on vowel spotting", Speech Communication 12, 1993