ETRI 방송뉴스음성인식시스템 소개

박 준
ETRI 음성/언어정보연구센터 음성인터페이스연구팀

Introduction of ETRI Broadcast News Speech Recognition System

Jun Park
Voice Interface Research Team, Speech/Language Information Research Center
Electronics and Telecommunications Research Institute
E-mail : junpark@etri.re.kr

Abstract

This paper presents ETRI broadcast news speech recognition system. There are two major issues on the broadcast news speech recognition: 1) real-time processing and 2) out-of-vocabulary handling. For real-time processing, we devised the dual decoder architecture. The input speech signal is segmented based on the long-pause between utterances, and each decoder processes the speech segment alternatively. One decoder can start to recognize the current speech segment without waiting for the other decoder to recognize the previous speech segment completely. Thus, the processing delay is not accumulated. For out-of-vocabulary handling, we updated both the vocabulary and the language model, based on the recent news articles on the internet. By updating the language model as well as the vocabulary, we can improve the performance up to 17.2% ERR.

I. 서론

방송뉴스는 배경 음향 및 잡음이 많이 포함되어 있으며, 다양한 주제에 대하여, 여러 사람이 각기 다른 발화 스타일로 발성하므로, 음성인식 기술을 적용함에 있어 가장 어려운 분야 중 하나라고 할 수 있다. 1990년대 초부터 시작한 미국 DARPA의 음성인식 기술개발과제에서도 기술 축적이 상당히 이루어진 후 마지막 단계에서 실제로 방영된 뉴스를 대상으로 성능평가를 실시하였다.

음성인식의 성능평가뿐 아니라 방송콘텐츠에 대한 음성인식 기술은 자막방송의 자동화에 직접적으로 활용할 수 있다. 우리나라에서는 1999년부터 자막방송을 시작하여, 2004년 8월 기준으로 KBS 1/2, MBC, SBS, EBS의 5대 공중파 방송사가 15%에서 40%까지 자막방송을 하고 있고, 케이블방송인 KTV는 10%정도를 실시하고 있다.[1] 현재 2~4초의 지연시간, 99%이상의 정확률로 세계 최고수준을 나타내고 있지만, 속기타자 전문인력의 양성에 장시간이 소요되어 자막방송 확대에 어려움을 겪고 있는 상황이다.

본 논문에서는 ETRI가 개발하고 있는 한국어 방송뉴스 음성인식시스템을 소개한다.[2],[3] 제 II장에서는 방송뉴스 음성인식에 사용된 시스템의 구조 및 기술적 요소를 간략히 기술하고, III장에서는 방송뉴스 콘텐츠의 특성을 분석한다. 이어서 IV장에서는 방송뉴스에 음성인식 기술을 적용함에 있어 주요 고려사항과 이에 대한 해결방안을 기술하고, 마지막으로 제V장에서 결론으로 끝맺는다.

II. ETRI 대어휘 연속음성인식시스템

방송뉴스 음성인식에 사용된 ETRI의 대어휘 처리 연속음성인식시스템의 구조는 그림 1과 같다.[4],[5]

음성 신호는 16kHz 샘플링, 16 비트로 양자화하고 프레임 크기는 16 msec, 프레임간 이동은 10 msec이다. 특징 벡터로는 멜 켑스트럼을 사용하고, 1차 및 2차 델타 특징벡터를 포함하는 총 39차 특징벡터로부터 LDA (linear discriminant analysis)를 적용하여 24차 특징벡터를 생성한다. 묵음을 포함하는 40개 기본 음소를 사용하며, 음소 문맥으로 단어 내에서는 좌, 우 양방향에 대해서 최대 2개 음소까지 고려하고, 단어간 경계에서는 1개의 주변음소를 고려하며, 각각의 음소는 세 개의 스테이트를 갖는 left-to-right 구조의 HMM으로 모델링한다. 문맥 부류화 (context clustering)를 하기 위하여 모음, 자음, 마찰음 등의 47개의 상세 분할된 음소 카테고리를 결정 트리의 분류기준으로 사용하고, 결과적으로 총 3,000개의 세논을 사용하며, 각각의 세논은 16차 가우시안 혼합함수 (Gaussian mixtures)를 갖는 각기 고유의 코드북을 가진다.

인식 단위로는 의사형태소를 사용한다.[5],[6] 의사형태소는 형태소 단위를 수정하여 발음이 유지되도록 하였으며, 짧고 자주 발생하는 형태소를 병합하여 인식 오류를 감소시킨다. 형태소 분석기를 이용하여 텍스트 코퍼스 문장내의 어절을 의사 형태소 단위로 분할하며 발음 사전은 형태소 기반의 발음열 생성기 (grapheme-to-phoneme converter)를 통해 자동적으로 생성하고[7], 단어사전의 어휘 수는 65,000개까지 수용한다.

인식된 내용 중 영어 알파벳, 기호, 숫자에 대한 표기는 후처리에서 HTG (hypothesis to grapheme)를 통해 자동으로 복원한다.[8]

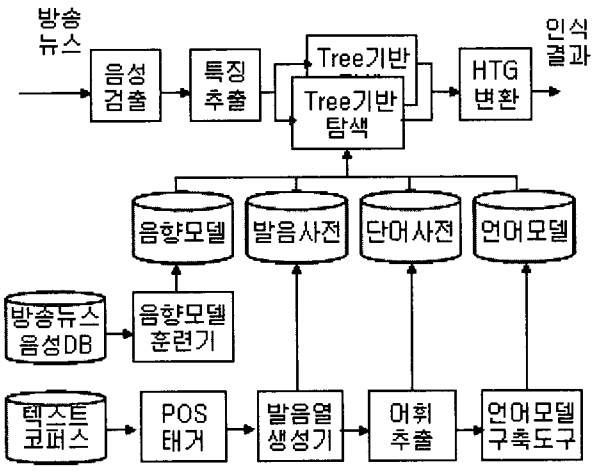


그림 1. ETRI 대어휘 연속음성인식시스템 구조

III. 방송뉴스 콘텐츠 분석

3.1 발성자별 콘텐츠 분석

방송뉴스의 내용은 발성자를 기준으로 앵커, 리포터, 인터뷰의 세 가지 부분으로 구분할 수 있다. 앵커는 대부분 음향통제가 잘되어 있는 방송 스튜디오에서 발성하므로 그 음성신호에 잡음이 거의 없다. 또한 발화와 발화간 0.5초 이상의 비교적 일정한 휴지구간이 존재한다. 리포터는 대부분의 경우 현장에서 발성한다. 조용한 환경도 포함하나, 스포츠 경기장이나 헬기 내부 등 다양한 음향환경에서 여러 가지 잡음이 포함된다. 그러나, 앵커와 더불어 리포터는 발성훈련이 잘되어 있어 발성 자체는 매우 명료하다.

그런데, 방송뉴스 인식에서 문제가 되는 것 중의 하나가 인터뷰 부분이다. 인터뷰는 리포터의 경우와 같이 다양한 음향환경을 포함하며, 상대에 따라서 사투리, 구어체, 잡음, 발성의 불명확성 등 여러 모로 인식에 있어서 난해한 요인들을 다수 포함하고 있다. 실제 방송뉴스에서도 인터뷰의 경우 발성이 명료하지 않을 때 화면 하단에 자막을 보여주고 있다. 방송뉴스의 내용을 문장 단위로 나타낼 때 앵커, 리포터 및 인터뷰가 차지하는 비중은 표 1과 같으며, 인터뷰의 비중도 16%정도로 상당한 부분을 차지하고 있다. 한편, 성별 분포를 살펴보면, 앵커의 경우 남성과 여성의 비중이 유사하나, 리포터와 인터뷰의 경우에는 여성보다 남성이 월등히 많음을 알 수 있다.

표 1. 발화자 별 방송뉴스 문장 분포 (단위: 문장수,%)

KBS (2001.12)	앵커	리포터	인터뷰	합계
남성	2363	8801	2125	13289 (83.7%)
여성	1072	1079	440	2591 (16.3%)
합계	3,435 (21.6%)	9,880 (62.2%)	2,565 (16.2%)	15,880 (100%)
SBS* (2001.11)	앵커	리포터	인터뷰	합계
남성	1,125	5,588	1,279	7,992 (79,9%)
여성	859	778	377	2,014 (20.1%)
합계	1,984 (19.8%)	6,366 (63.6%)	1,656 (16.5%)	10,006 (100%)

(*) KBS 데이터는 2001년 12월, SBS 데이터는 2001년 11월 방영분을 대상으로 하였으며, SBS의 경우 지역방송시간대 및 스포츠뉴스, 일기예보가 포함되지 않았음

3.2 음성인식 관점에서의 방송뉴스 음성 특성

방송뉴스 특성상 음성인식 기술을 적용하는데 있어서 어려움을 초래하는 요인은 다음과 같다.

- 실시간 처리 요구: 뉴스는 글자 그대로 새로운 정보를 제공하는 것이 목적이므로 그 내용을 사전에 파악하기 어려우며, 따라서 뉴스가 방송되는 시점에서야 그 내용을 접근할 수 있다. 또한 그 내용이 일방적으로 전달되므로 음성인식 수행상의 지연시간을 고려하여 입력신호의 속도를 조정하는 것은 불가능하다. 즉 모든 작업을 실시간으로 처리하여야 한다.
- 신규 어휘의 빈번한 출현: 새로운 정보를 전달함에 따라서 나타나는 어휘에도 새로운 것이 많이 출현한다. 특히, 인명이나 기관명, 지명 등 고유명사는 매일 새로운 것들이 발생한다. 현재의 음성인식 방식으로 기본적으로 인식어휘 대상을 미리 등록하고 이 안에서 인식을 수행하므로 새로운 어휘에 대한 대처방안이 요구된다.
- 고속 발성 속도: 방송뉴스의 주요 발성자인 앵커와 리포터들은 보다 짧은 시간에 신속하게 정보를 전달하고자 기본적으로 상당히 빠른 속도로 발성한다. 이는 음성인식의 입장에서 단위시간 당 입력정보의 양이 상대적으로 적음을 의미한다.
- 인터뷰 발성의 다양성: 앵커와 리포터의 발성 내용과 달리 인터뷰에서는 다양한 발화자가 발성을 하며, 통상적인 방송뉴스의 발성패턴에서 크게 벗어나 대화체의 발성이 많으며, 방언이나 매우 비정형적인 문형도 자주 나타난다.
- 다양한 음향환경 및 배경 잡음: 방송뉴스의 상당부분을 차지하는 리포터와 인터뷰의 경우 현장취재가 대부분으로 현장의 다양한 음향환경과 잡음을 포함한다. 주변 잡음이 심할 경우 롬바드효과로 인하여 리포터 음성의 피치가 높아지며, 정상적인 발성 패턴에서 벗어나게 된다.

반면, 음성인식의 관점에서 작업을 수월하게 하는 방송 뉴스 음성신호 특성도 다음과 같이 존재한다.

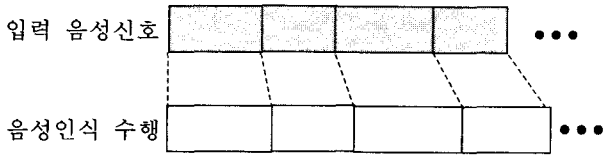
- 통제된 음성신호: 방송뉴스의 음성신호는 전체적으로 매우 정형화 되어 있다. 신호의 최대 크기는 전 구간에 걸쳐서 일정하게 유지되며, 발성되는 문장간 휴지 구간도 일정한 간격을 나타내고 있다. 또한 현장감을 위한 배경음향도 상황별로 보도 초기에 강하게 나타나지만, 발성이 시작되면 상대적으로 배경음향의 크기가 줄어드는 경우가 많다.
- 훈련된 발성 및 정형 문장: 뉴스의 대부분을 차지하는 앵커와 리포터의 발성이 표준발음에 준하여 명료하게 발성되고 있다. 또한 문장도 문법에 준하여 발성되고, 또한 방송뉴스에 특이한 정형화된 문형도 반복적으로 많이 나타난다.

IV. 방송뉴스 음성인식 처리

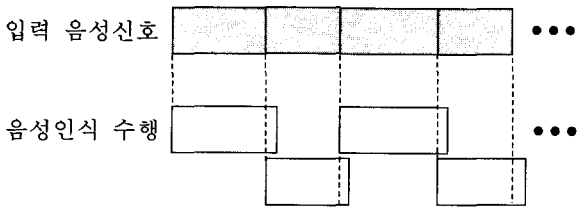
3.1 실시간 음성인식처리를 위한 디코더 이중화

앞에서 살펴본 바와 같이 방송뉴스가 실시간으로 송출되어 이에 따라 음성인식 작업도 실시간으로 수행되어야 한다. 하지만, 처리 대상 음성신호와 동시에 음성인식을 수행한다 하더라도, 그림 1.a)에서 보는 음성신호가 끝나기 이전에 음성인식 수행을 완료하는 것은 불가능하며, 음성신호가 끝난 후 음성인식 작업 수행에 어느 정도 지연시간이 존재하게 된다. 이러한 지연시간은 음성신호의 길이와 비교하여 작다고 하더라도 방송뉴스가 진행됨에 따라 계속 누적이 되고, 결과적으로 실시간 처리가 불가능하게 된다.

이러한 지연시간 누적문제를 해결하기 위하여 ETRI에서는 디코더를 이중화 하여 인식을 수행한다. 즉, 그림 1.b)와 같이 들어오는 음성신호를 일정 길이 범위 안에서 시간적으로 분할하고, 분할된 음성구간을 이중화된 디코더가 번갈아 처리함으로써 지연시간 누적을 방지하였다. 이러한 구조의 잇점을 실질적으로 살리기 위하여 2개의 CPU가 장착된 시스템을 사용하였다.



a) 단일 디코더의 경우 지연시간 발생



b) 이중 디코더의 경우 지연시간 누적 방지

그림 2. 디코더 이중화를 통한 인식 지연시간 방지

3.2 음성분할

위에서 언급한 디코더 이중화 처리를 위해서는 먼저 입력음성신호를 적당한 크기로 분할하는 음성분할 기능 구현이 필요하다. 아울러 음성신호가 아닌 구간은 제거하고 음성 구간만 디코더로 전달하는 것이 요구된다.

방송뉴스에는 음향, 음성, 음악 등 다양한 소리가 혼재하여 있는 반면, 일정한 형식을 따르는 규칙성도 또한 지니고 있다. 앵커는 대부분 잡음이 거의 없는 상태

에서 발성하며, 발화간의 사이도 의미전달을 명확하게 하기 위하여 대부분 500msec이상의 휴지 구간이 존재한다. 또한, 현장에서 리포터가 발성할 경우 발성시간에 앞서 주변 잡음구간이 500msec에서 1-2초간 존재하는 경우가 발성할 경우가 많다. 또한 전체 음성신호가 잘 제어되어 있어, 음성신호나 잡음 신호의 크기가 정규화 되어 있어 이들을 분류하는 임계치 설정이 용이하다. 이러한 관측을 토대로 음성 신호의 상태를 다음 5가지로 구분하여 상태 천이도를 기반으로 음성 구간을 추출한다.

- 대기상태(WAIT모드): 음성 판정 대상 구간을 기다리는 상태
- 탐지상태(DETECTED모드): 음성으로 분류할 수 있는 신호가 나타난 상태
- 음성상태(SPEECH모드): 음성으로 판정이 되어 구간이 끝나는 시간을 기다리는 상태
- 비음성상태(NONSPEECH모드): 음성이 아닌 구간으로 판정된 상태
- 임시 휴지상태(BREAK모드): 잠정적으로 음성이 끝났다고 판정한 상태로서 다시 음성상태로 돌아가거나, 대기상태로 이동

각 상태를 이동하는 판단은 음성구간 신호 및 잡음구간신호의 크기에 대한 임계치를 실험적으로 설정한 후 음성신호로부터 추출한 여러 가지 이벤트에 따라 수행한다. 예를 들어, 대기모드에서 30msec이상 신호가 나타날 경우 탐지모드로 이동하고, 탐지모드에서 비정상적으로 큰 신호가 지속되거나 단시간에 음성신호에서 나타날 수 없는 크기변화가 빈번하게 나타나는 경우에는 비음성상태로 이동하며, 이를 통과하면 음성상태로 천이한다.

이와 같이 음성구간을 추출함에 있어, 중요한 점은 음성이 음성 분할 단계에서 비음성으로 잘못 분류되는 경우가 가급적 생기기 않도록 하는 것이다. 그리고, 잡음의 존재 등으로 음성구간이 예외적으로 길어지는 경우 분할 기준을 강화하여 강제로 끊어주거나, 음성구간이 지나치게 짧게 검출되는 경우는 일정 길이가 되도록 분할 기준을 완화하는 등 분할 기준을 가변적으로 운용한다.

3.3 인터넷 검색을 통한 신규 어휘 수용

뉴스의 내용은 여러 시간대에 걸쳐 반복 방영되며, 방영된 방송뉴스의 대본이나, 인터넷 매체 및 기존 신문의 기사의 내용은 인터넷 상에서 용이하게 접근할 수 있다. 이러한 점에 착안하여, 인터넷에 올라와 있는 뉴스의 내용을 주기적으로 검색하여 새로운 어휘를 검출하고, 이를 음성인식시스템의 단어사전 을 보완할 수 있다. 또한, 새로 나온 뉴스의 텍스트를 기반으로 언어모델을 구성하고 이를 기존의 언어모델에 반영함으로써 신규 어휘 문제를 상당부분 해결할 수 있다.[9]

신규 어휘는 인식시스템의 단어사전에 포함여부로 용이하게 검출할 수 있다. 그런데, 인식시스템의 단어사전을 무제한으로 크게 운용할 수는 없으므로 65,000 개로 제한하고 있다. 이에 따라, 신규 어휘를 추가함에 있어, 최신 기사 어휘에 더 큰 가중치를 두고 통합하고, 이중에서 빈도수를 고려하여 고빈도 어휘를 추출, 사전을 구성하였다.

신규로 수집한 코퍼스를 이용하여 언어모델을 갱신하는 방법으로는 선형보간 방법을 사용하였다. 그런데, 새로운 코퍼스의 양에 따라 갱신되는 언어모델의 성능이 영향받게 된다. 기반 시스템 대비 여러 가지 코퍼스 구성의 경우에 대한 인식성능 및 OOV(사전 외 단어)수는 표 2와 같다.

표 2. 언어모델 갱신에 사용하는 코퍼스 종류에 따른 인식성능 및 OOV갯수 변화

순번	사용 코퍼스	단어오류율* (%)	OOV 갯수
1	기반시스템	17.4	70
2	1일분 방송뉴스	15.1	40
3	1개월분 방송뉴스	15.3	23
4	1일치 방송뉴스 +방송뉴스 추가	14.8	35
5	1일치 방송뉴스 +방송뉴스/신문기사 추가	14.4	27

(*) 의사형태소 단위

표 2의 결과는 2003년 10월 14일의 KBS 9시뉴스에서 스포츠뉴스와 일기예보를 제외한 앵커, 리포터 및 인터뷰를 모두 포함하는 239문장에 대하여 실험한 결과인데, 순번1의 경우는 언어모델 갱신을 하지 않은 경우로서 기반시스템의 성능을 나타낸다. 순번2, 3의 경우는 각각 과거 1일분과 1개월분 방송뉴스기사로서 언어모델을 갱신한 경우이다. 순번 4의 경우 이전 1일분의 방송뉴스와 아울러, 이전 1개월분의 방송뉴스 기사를 주제단위로 분류하고 유사도를 측정하여 1일분 방송뉴스와 가장 가까운 기사 100개를 추가하여 실험한 결과이다. 순번 5의 경우는 4의 경우에 추가하여 신문기사 1개월분 중 유사 기사 100개도 함께 포함한 경우이다.

기반시스템에 비하여 모든 경우에 오류율이 감소하였으며, OOV갯수도 감소함을 알 수 있다. 특히, 뉴스의 내용이 수시로 바뀌기 때문에, 3의 경우에서 보는 바와 같이 1개월분의 기사를 그대로 사용할 경우 1일분의 경우와 비교할 때 오히려 인식오류율은 증가하였다. 그러나, 기존 1개월분의 기사 중에서 이전 1일분의 기사와 유사한 기사를 선별하여 추가하면 4의 경우에서 보는 바와 같이 인식오류율이 감소하며, 5의 경우와 같이 유사한 신문기사를 더 사용하는 경우는 더욱 향상됨을 보이고 있다.

V. 결론

음성인식 기술을 방송뉴스에 적용함에 있어, 실시간 처리와 신규어휘 수용방안이라는 두가지 주요 현안이 있다. ETRI 방송뉴스음성인식시스템에서는 실시간 처리문제에 대하여, 이중화된 디코더를 도입하여, 인식 지연시간의 누적을 방지함으로써 실시간 처리가 가능하게 하였다. 또한, 새로이 출현하는 어휘에 대해서는 인터넷에 게재된 뉴스 텍스트를 수집, 분석하여 추가 대상 어휘를 추출하여 단어사전에 추가하였으며, 아울러 언어모델에도 최신기사의 언어모델을 반영함으로써 신규 어휘 출현에 따르는 성능 저하를 최소화하였다.

방송뉴스는 다양한 음향환경에서 다수의 발화자가 여러 불특정 주제에 대하여 다양한 말투로 발성하는 음성으로 구성되므로, 음성인식 기술을 적용함에 있어 가장 어려운 분야 중 하나라고 할 수 있다. 그러나, 위에서 살펴본 바와 같이 발성 훈련이 잘 된 앵커와 리포터의 음성이 뉴스의 많은 부분을 차지하고 있으며, 이에 대한 인식성능도 상당 수준에 이르고 있어, 방송 내용별로 선별적인 활용 가능성을 보여주고 있다.

방송뉴스를 포함하여 방송 콘텐츠는 실로 다양한 내용으로 구성되어 있어, 음성인식 기술에 대한 성능평가 척도로 계속 남아있을 것이다. 그러나, 잘 통제된 음향환경과 명료한 발성이 차지하는 비중도 커서 음성인식 기술의 활용가능성도 높은 분야이다. 이미 실용화된 방송 프롬프터를 시작으로, 보다 정확한 음성/비음성구별, 남/여 성별식별 등 여러 요소기술의 성능을 확보함으로써 자막방송 자동화, 방송콘텐츠 자동 색인/검색 등 여러 가지 응용이 가능할 것이다.

참고문헌

[1] http://www.koreasteno.com, (주)한국스테노

[2] 박준, 김승희, 이영직, 양재우., "방송 뉴스 자막 처리 시스템 개발", *제17회 음성통신 및 신호처리 학술대회*, 2000

[3] Seunghi Kim, Jong Jin Kim and Jun Park, "Introduction to Korean Broadcast News Transcription System," *Proc. ICSP*, pp. 669-672, 2001

[4] O. Kwon, K. Hwang, and J. Park, "Korean Large Vocabulary Continuous Speech Recognition of Newspaper Articles," *Proc. ICSP*, pp. 333-336, 1999

[5] Oh-Wook Kwon and Jun Park, "Korean Large Vocabulary Continuous Speech Recognition with Morpheme-based Recognition Units," *Speech Communication*, Vol.39, No. 3-4, pp.287-300, 2003

[6] Oh-Wook Kwon, Kyuwoong Hwang and Jun Park, "Korean Large Vocabulary Continuous Speech Recognition Using Pseudomorpheme Units," *Proc. EUROSPEECH 99*, Budapest, Hungary, pp. 483-486, 1999

[7] Jehun Jeon, Sunhwa Cha, Minhwa Chung, Jun Park, and Kyuwoong Hwang, "Automatic Generation of Korean Pronunciation Variants by Multistage Applications of Phonological Rules," *Proc. ICSLP'98*, Sydney, Australia, 1998

[8] 박경현, 김승희, 박준, "한국어 연속음성 인식을 위한 HTG 변환," *제19회 음성통신 및 신호처리 학회*, pp. 71-74, 2002

[9] 김현숙, 전형배, 김상훈, 최준기, 윤승, "방송 뉴스 인식을 위한 언어 모델 적응," *말소리*, Vol.51, pp.99-116, 2004