결정트리기반 음성인식 시스템에서의음소지속시간 사용방법

구 명완, 김 호경

KT 서비스개발연구소

A phoneme duration modeling in a speech recognition system based on decision tree state tying

Myoun-Wan Koo, Ho-Kyoung Kim

KT Service Development Research Laboratory, mwkoo@kt.co.kr, hokyoung@kt.co.kr

Abstract

In this paper, we propose a phoneme duration modeling in a speech recognition system based on disicion tree state tying.

We assume that phone duration has a Gamma distribution. In a training mode, we model mean and variance of each state duration in context-independent phone model based on decision tree state tying. In a recognition mode, we get mean and variance of each context-dependent phone duration form state duration information obtaind during training mode.

We make a comparative study of the proposed meth with conventinal methods. Our method results in good performance compared with conventional methods.

I. 서론

HMM(hidden Markov model) 기술이 음성인식 시스템에 사용하기 시작한 후 20년 이상이 되었으며 음소를 HMM의 기본 모델로 보편적으로 사용하기 시작된지도 10년 이상이 되었다. 그러나 음소는 주위 음소의 종류에 따라 음가가 달라지므로 이를 정확하게 표현하기 위한 방법으로 문맥독립 음소 대신 문맥종속 음소모델을 사용하게 되었다. 대표적인 문맥종속 음소 모델은 유사한 문맥종속 모델을 합치는 방식인 unit reduction모델이다. 이 방식은 문맥종속 모델로 트라이폰(triphone)을 사용할 경우 유사모델을 줄이기 위한 방식으로 트라이 폰의 빈도수를 고려하는 것이다. 이 방식은 인식을 위한 훈련데이타에 존재하는 트라이 폰의 빈도수를 고려하여 빈도수가 적은 트라이 폰을 합치는 방식이므로 음운적인 특징을 고려하지 않았으며 트라이 폰 내부의 상태를 고려하지 않고 전체 음소를 즐여서 공유하는 방식이다.

최근에는 트라이 폰 내부의 상태를 음성, 음운학적 정보를 이용하여 모델링하는 방안이 제안되었다. 이 방식은 문맥독립 음소 단위 내의 매 상태(B, M, E)마다 좌,우측의 음소의 음성, 음운 현상을 고려한 규칙 노드로 구성되는 트리를 만든 후 이 트리의 리프 노드(leaf node)에 HMM 파라미터를 저장하는 방안이다. 이 방안은 유니트 감소 알고리즘에 비해서 메모리를 적게 사용하면서 성능이 우수해서 최근에 많이 사용되고 있다.

본 논문은 음소지속시간 정보를 인식기에 사용하기 위한 방안에 대한 것이다. 특히 문맥종속 음소 지속시간 정보를 결정트리 기반 음성인식 시스템에 이용하기 위한 방식으로 상태단위 지속시간 모델을 하여 저장한 후 인식과정에서는 음소지속시간 정보를 변환시켜 사용하는 방안을 제안하고자 한다. 먼저 KT 서비스개발연구소의 VAD 시스템에 대한 소개를 하고 KT 음성인식기인 HUVOIS에 대한 설명을 하고자 한다. 결정트리기반 음성인식 시스템에서 문맥종속 음소

지속 시간 모델링을 하는 방안에 대해서 설명하고 시뮬레이션을 통한 인식률을 비교하고자 한다. 그리고 마지막에 결론을 맺는다.

II. KT VAD 시스템

VAD 시스템이란 전화망을 통하여 음성으로 사람 이름을 말을 하면 음성을 인식하여 사람이 있는 사무실 혹은 핸드폰으로 전화가 자동으로 걸리게 하는 시스템이다. KT 서비스개발연구소에서는 VAD시스템을 개발하여 1999년부터 연구소(02-526-5114) 내에서 서비스를 제공하였으며 2001년부터는 본사(02-750-5114) 및 유관기관에 서비스를 확대하고 있다. 특히 사내 인사시스템과 연동되어 인사이동이 있을 경우에도 자동으로 인식명을 갱신하고 있다. 또한 동일한 사원이름이 있을 경우에는 조직이름을 확인함으로써 사용자의 편리성을 도모하였다.

현재 KT에서 운용중인 VAD 시스템은 그림 1. 에 나타나 있다. 사내 PABX에 붙어져 있으며 내선(5114) 혹은 외선번호(02-526-5224)를 다이알을 하면 다이알 로직 전화보드가 자동으로 인지하여 음성인식을 위한 대화 과정이 수행된다. 사용할 수 있는 단어은 사원이름 혹은 사원이름 + 핸드폰 이며 음성이 인식되면 사무실 혹은 핸드폰으로 자동으로 전화가 걸리게 된다

III. HUVOIS 인식기

KT에서 자체 개발한 HUVOIS 음성인식기에 대해서 소개하고자 한다. HUVOIS는 HMM 파라미터를 생성하는 훈련프로그램과 훈련된 HMM 파라미터를 이용하여 인식하는 인식기로 나누어 진다. 음성인식기는 인식모듈과 검증(Verification)모듈로 나누어지면 인식모듈은 비터비 빔(Viterbi beam)검색 알고리즘을 사용한다. 검증모듈은 인식기의 출력을 검증하기 위하여 반음소(anti-phoneme)모델을 사용한다.

3.1 특징추출

음성은 매 10msec 단위로 분석이 되며 특징은 12차 LPC(linear predictive coding)기반 멜 켑스트론, 델타 및 델타델타 켑스트론, 그리고 델타 및 델타델타 에너지로 구성되는 38차의 특징을 사용한다.

3.2 음소모델

음소모델은 그림 2. 와 같이 사용한다. 전체 7 개의 상태로 이루어져 있으며 상태변화에 대한 출력확률은 B, M, E로 나누어진다.

3.3 결정트리 기반 상태 모델

결정트리 기반 상태모델이란 그림 3.에 나타난 바와 같이 문맥종속 음소모델을 위하여 음소를 여러 개 만드는 것이 아니고 음소를 구성하고 있는 매 상태 주위의 음소분류를 통해서 상태를 여러 개 만들어 주는 것을 말한다. 그림 3은 음소모델의 가운데 상태인 M을 모델링 하는 것을 보여 준다. 이 상태는 5종류의 문맥종속 상태(leaf node)로 나누어 지는데 이 분류는 R-central consonant(우측음소가 central-consonant인가?), R-nasal(우측 음소가 비음인가?),와 같이 질문 노드의 가,부에 따라 최종 상태를 결정하는 리프 노드를 찾게된다. KT- HUVOIS는 158개의 질의 셋을 사용하고 있다.

IV. 문맥종속 음소 지속시간 모델링

결정트리기반 상태 모델링을 사용하였을 경우에는 상태 단위로 HMM 파라미터를 공유하기 때문에 상태단위로 지속시간을 모델링하여 저장해야 한다. 그러나 상태지속 시간 정보는 상태의 지속 시간이 너무 짧기 때문에 안정성이 부족하여 성능향상에 크게 기여하지 못한다.

그래서 본 논문에서는 훈련과정에서는 상태단위 지속시간을 구해서 저장하고 인식과정에서는 음소단위 지속시간을 사용하는 방안을 제안하고자 한다. 이를 위해서 상태지속 시간의 확률 분포는 감마(Gamma) 분포를 갖는다고 가정하면 각 상태(B, M, E)의 평균과 분산값과 문맥종속 음소 지속시간의 평균, 분산 값을 다음과 같은 수식이 성립한다.

$$\begin{aligned} &E[\text{문맥종속 음소 지속시간}] \\ &=E[\text{B상태 지속시간}] + E[\text{M상태 지속시간}] \\ &\quad +E[\text{E상태 지속시간}] \end{aligned} \tag{1}$$

$$\begin{aligned} &Var[\text{문맥종속 음소 지속시간}] \\ &=Var[\text{B상태 지속시간}] + Var[\text{M상태 지속시간}] \\ &\quad +Var[\text{E상태 지속시간}] \end{aligned} \tag{2}$$

여기서 E[ ]는 평균을 의미하며, Var[ ]은 분산을 의미한다.

즉 상태지속 시간 정보로부터 문맥종속 음소지속

시간 정보를 쉽게 구하기 위해 상태 B, M, E 각각의 지속시간을 독립된 랜덤프로세서라고 가정하고 문맥종속 음소 지속시간을 상태 랜덤프로세서의 합이라고 가정하여 상기 (1), (2) 식이 성립하도록 랜덤프로세서의 확률 분포로 감마 함수로 정의한다.

V. 인식실험

VAD시스템의 인식실험은 소프트웨어의 버전과 사용하는 정보에 따라 다양하게 수행되었다. 먼저 문맥종속음소 모델방식으로 음소기반 상태방식을 수행한 방식과 결정트리기반 상태결합 방식을 사용하였을 경우의 인식실험을 수행하였다[1]. 그 결과 표 1. 나타난 바와 같이 결정트리기반 상태결합 방식을 사용할 경우가 음성인식률이 약간 높았다. 또한 그때 사용한 상태 개수 및 상태를 나타내는 데 필요한 믹스춰(mixture) 개수도 작았다[1][2].

다음에는 음소길이 정보를 사용하여 음성 인식률을 사용하였다[3]. 사용된 길이 정보는 음소단위로 모델링하였으며 길이정보를 모델링하기 위하여 감마함수를 사용하였다. 표2 에는 길이 정보를 사용하였을 경우와 사용하지 않을 경우의 인식률을 비교하였다. 그 결과 길이 정보를 사용하면 인식률이 약간 상승되고 있음을 알수 있었다. 또한 문맥종속 음소 길이 정보를 사용하였을 경우에 인식 결과를 표 3에 나타내었다. 문맥종속 음소 지속시간 정보를 이용하면 89.47%로 성능이 향상되었다.

VI. 결론

본 논문에서는 KT에서 개발한 VAD 시스템을 소개하고 이 시스템 개발에 필요한 HUVOIS 인식기를 기술하였다. 특히 인식기의 버전에 따른 다양한 인식실험을 수행하였으며 결정트리 기반 상태 모델링 방식을 사용한 인식 시스템에서 음소지속 정보를 이용한 인식 알고리즘을 제한하였다. 제한된 알고리즘을 사용했을 경우와 기존 방식을 이용했을 경우 성능을 비교하였다. 제안된 방식을 이용했을 경우가 89.47%로 가장 높은 성능을 나타내었다.

참고문헌

[1] 박성준 구명완,전주식, "결정트리 모델링 기반의 음성 인식기", 제 17회 음성통신 및 신호처리 학술대회 17권 1호, pp. 175-178 , 2000.

[2] S.J. Young, J.J. Odell, P.C. Woodland, "*Tree-based state tying high accuracy acoustic modeling*", Proc. Of the DARPA Speech and Natural Language Processing Workshop, Plainsboro, pp. 307-312, 1994

[3] 김호경,구명완, "음소길이정보를 이용한 음성인식 무인자동교환 서비스", 제 15회 신호처리 합동학술대회,pp.274, 2002.

[4] 구명완, 김재인, 정영준, 김호경, "트리기반 발음사전을 이용한 VAD 시스템", 음향학회 추계학술대회, 2002. 11월

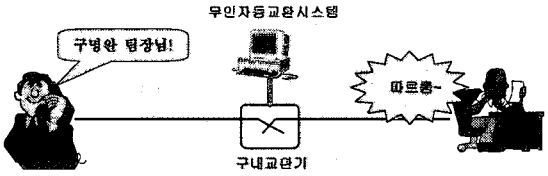

그림 1. KT-HUVOIS를 이용한 VAD시스템

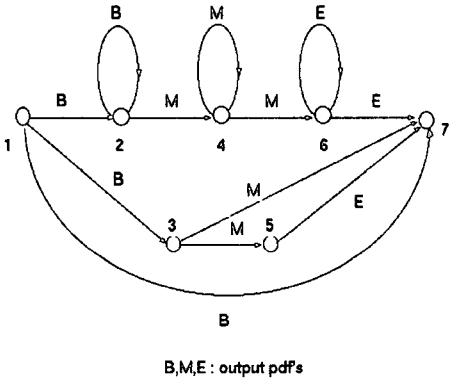

그림 2. 음소 모델

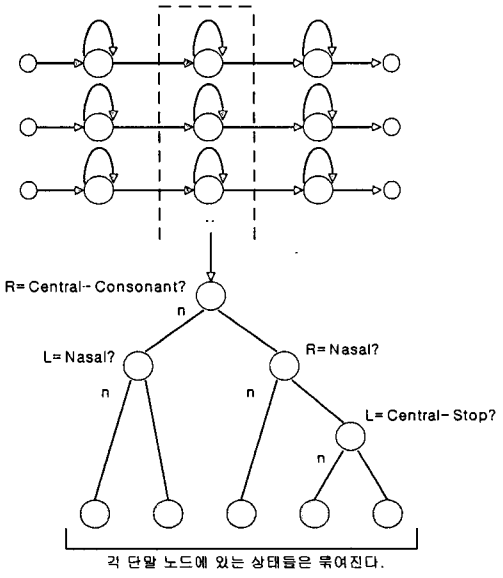

그림 3. 결정트리 기반 상태 모델

표 1. 결정트리기반 상태결합과 음소기반 상태결합.
(인식대상 단어 : 2413)

	음소기반	결정트리
인식률(3,923실험 단어)	88.66%	89.22%

표 2. 음소길이 정보 사용유무에 따른 성능비교.

	문맥종속 음소 지속시간	문맥독립 음소 지속시간
인식률(3,923실험 단어)	89.47%	89.40%
틀린단어 갯수	413	416

표 3. 문맥독립 음소 지속시간 및 문맥종속 음소 지속시간 정보를 이용한 성능비교

	길이정보 사용	길이정보 사용 않음
인식률(3,923실험 단어)	89.40%	89.22%