소음문장 제거를 위한 음소지속시간 사용

구 명완, 김 호경, 박 성준, 김 재인
KT 서비스개발연구소

The Usage of Phoneme Duration Information for Rejecting Garbage Sentences

Myoung-Wan Koo, Ho-Kyoung Kim, Sung-Joon Park, Jae-In Kim
Service Development Laboratory, KT
E-mail : {mwkoo, hokyoung}@kt.co.kr

Abstract

In this paper, we study the usage of phoneme duration information for rejection garbage sentence. First, we build a phoneme duration modeling in a speech recognition system based on dicicion tree state tying. We assume that phone duration has a Gamma distribution. Next, we build a verification module in which word-level confidence measure is used. Finally, we make a comparative study on phoneme duration with speech DB obtained from the live system. This DB consistes of OOT(out-of-task) and ING(in-grammar) utterences. the usage of phone duration information yields that OOT recognition rate is improved by 46% and that another 8.4% error rate is reduced when combined with utterence verification module.

Ⅰ. 서론

최근에 운율(prosody)의 하나인 지속시간을 HMM(hidden Markov model) 시스템에 적용하는 방안에 대한 연구가 최근에 수행되어 지고 있다[1]. 지속시간을 인식기에 적용하기 위한 연구는 두가지 방향으로 이루어져 왔다. 첫 번째 방안은 인식 알고리즘에 인식시간의 증가 없이 적용하는 방안에 대한 연구를 하는 것이고[2][3][4] 두 번째 방안은 지속시간을 모델링하는 기본단위와 모델방안에 대한 연구이다[5][6]. 첫 번째 방안으로 인식과정에 지속시간을 사용하는 방안과 N-best 문장을 찾은 후에 후처리과정에 음소지속시간 정보를 이용하는 방안이 제시되었다[2]. 두 번째 방안으로는 지속시간을 사용하는 기본단위로 상태, 음소, 음절 및 단어 등을 사용하는 방안에 대한 연구를 수행하였으며 모델링하는 방안으로는 감마함수, 가우시안 함수등을 사용하는 방안이 연구되어 왔다[5][6].

본 논문에서는 지속시간의 기본단위로 음소를 사용하였으며 이를 위한 모델링 함수로 감마함수를 사용하였다. 또한 수정 비터비 함수를 사용하여 지속시간을 인식과정에 직접 적용하였다. 후처리 과정에서는 검증단계를 수행하여 인식과정에서 나온 결과를 검증하였다. 인식실험에 사용한 음성 DB는 서비스되고 있는 인식시스템에서 녹음한 음성을 사용하였으며 특히 소음데이터를 중심으로 길이정보를 사용하였을 경우의 인식률과 인증 단계를 거쳤을 경우의 인식률을 비교고찰하였다. 먼저 II 장에서는 지속시간을 모델링하는 장안에 대한 설명을 하고 III 장에서는 KT의 음성인식기인 HUVOIS를 소개한다 IV 장에서는 인식실험에 대해서 토의하고 V 장에서 결론을 맺는다.

Ⅲ. 지속시간 모델

지속시간의 기본단위로 음소를 선택하였고 지속시간을 감마함수를 모델링하였다. 그런데 KT가 사용하는 음성인식 알고리즘은 결정트리기반 상태 모델링이므로

음소단위 지속시간정보를 직접 사용할 수 없다. 즉 결정트리기반 상태 모델링을 사용하였을 경우에는 상태단위로 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 훈련과정에서는 상태단위 지속시간을 구해서 저장하고 인식과정에서는 음소단위 지속시간을 사용하는 방안이 제안되었다[1]. 먼저 상태지속 시간의 확률분포도 감마(Gamma) 분포를 갖는다고 가정하면 각 상태(B, M, E)의 평균과 분산값과 문맥종속 음소 지속시간의 평균, 분산 값을 다음과 같은 수식이 성립한다.

E[문맥종속 음소 지속시간]
=E[B상태 지속시간] + E[M상태 지속시간]
+E[E상태 지속시간] (1)

Var[문맥종속 음소 지속시간]
=Var[B상태 지속시간] + Var[M상태 지속시간]
+Var[E상태 지속시간] (2)

여기서 E[ ]는 평균을 의미하며, Var[ ]은 분산을 의미한다. 즉, 상태지속 시간 정보로부터 문맥종속 음소 지속 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.

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 인식 모듈은 비터비 빔(Viterbi beam)검색 알고리즘을 사용한다. 검증모듈은 인식기의 출력을 검증하기 위하여 반음소(anti-phoneme)모델을 사용한다[7][8][9][10].

3.1 특징추출

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

3.2 음소모델

음소모델은 전체 7 개의 상태로 이루어져 있으며 상태변화에 대한 출력확률은 B, M, E로 나누어진다.

3.3 결정트리 기반 상태 모델

결정 트리 기반 상태모델이란 문맥종속 음소모델을 위하여 음소를 여러 개 만드는 것이 아니고 음소를 구성하고 있는 매 상태 주위의 음소분류를 통해서 상태를 여러 개 만들어 주는 것을 말한다. KT- HUVOIS는 158개의 질의 셋을 사용하고 있다.

3.4 수정된 비터비 알고리즘

음소기반 지속시간을 인식 모듈에 적용하기위해선 수정된 비터비 알고리즘이 필요하다. 음소가 끝나는 시간 t 에서의 음소지속시간에 의한 로그 라이크리후드 정보를 D(t) 라고 하면, 수정된 비터비 디코더는

$$V(t+1) = V(t) + O(t) + w* D(t) \quad (3)$$

가 된다. 여기선 V(t)는 시간 t 에서는 로그 라이크리후드이며 O(t)는 HMM파라미터의 천이 및 출력함수의 로그값이고 w는 음소지속시간정보를 위한 무게값(weighting)이다.

3.5 검증 모듈(verification module)

검증모듈에서는 인식결과를 검증하는 단계이며 인식결과를 단어단위로 검증한다. 검증하는 방법은 단어단위의 확신 값(confidence value)을 사용한다[11]. 이 값이 미리 정해진 임계값(confidence threshold)보다 작으면 이 단어는 검증이 실패한 것으로 간주하고 다른 단어를 선정한다. 그리고 임계값보다 큰 단어만 선정하여 문장을 구성한다. 단어단위 확신 값은 이 단어를 구성하고 있는 음소단위 확신 값의 함수로 구성되며 음소단위 확신 값은 음소를 구성하고 있는 LLR(log likelihood ratio)의 시그모이드 함수(sigmoid function) 값 이며 매 음소의 끝에서 구해진다. LLR은 프레임 단위의 확신 값이며 매 프레임 단위로 HMM 모델의 라이크리후드 값과 반 음소 모델에 의한 라이크리후드 값과의 비이다[11].

IV. 인식실험

4.1 음성 DB

음성 DB는 현재 KT 서비스개발본부에서 사용 중인 음성 이름 인식기를 통해서 얻은 실제 음성으로 구성되어 있다. 현재 KT 이름 인식기는 "구명완" 이라고 말하면 구명완 사무실로 전화가 되면 "구명완 핸드폰"으로 말하면 핸드폰으로 전화를 걸어주는 시스템이다. 이 시스템은 이름 혹은 이름+ 핸드폰(휴대폰 등)이 포함되는 문장도 인식할 수 있다. 핵심단어의 종류는 2416단어이며 5000개 이상의 음성이 얻어 졌다. 이 음성 중 5584개의 음성이 인식실험용 토큰으로 사용되었다. 인식실험용 토큰 중 575개가 문법에 맞지 않는 OOT(out of task)로 구분이 되었으며 5009개의 토큰이 ING(in grammar)로 구분이 되었다. 여기서 OOT는 ""음" "에 저"등과 같은 단어와 인식대상에 포함되어 있지 않는 사람이름 등과 같이 인식 문법에 맞지 않는 단어로 구성되어 있다는 것을 말한다. 이러한 토큰이 입력이 되면 소음으로 인식이 되거나 거절되어야 한다. 표 1 은 음성 DB의 구성도를 나타 낸 것이다.

4.2 실험

인식실험은 먼저 3.4 에서 설명한 음소 지속시간 정보에 사용하는 무게값의 변화에 따른 인식 실험을 수행 하였다. 그림 1에는 무게값에 따른 인식률 변화를 나타 내었다. 사용된 음소지속시간은 문맥종속음소를 감마함수로 모델링한 값을 사용하였다[1]. 무게값이 높아 질수록 에러율은 감소 하였고 최적의 성능은 무게값이 5일 경우 13.1%의 에러율이 얻어 졌다. 그림 2 에서는 무게 값이 5일 경우에 확신 임계값 변화에 따른 인식률의 변화를 음소지속시간을 사용하지 않을 경우와 비교하였다. 인식결과를 보면 음소지속시간을 사용하지 않고 검증 모듈을 사용하였을 경우 에러율은 14.2% 이었으나 음소길이 정보를 사용하면 13.1%로 감소되었음을 알 수 있었다. 이것은 음소길이 정보가 8.4%의 인식률을 향상 시켰음을 나타 낸 것이다. 또한 음소길이 정보를 사용하지 않을 경우의 OOT인식률은 -4.% 였으나 음소지속시간정보를 사용하였을 경우에는 46.61%로 향상되어 음소지속시간정보가 소음문장들을 제거하는데 매우 효율적이다는 것을 알 수 있다. 여기서 음소지속시간정보를 사용하지 않을 경우 마이너스 인식률이 나온 것은 소음문장에 대한 인식결과가 단어의 추가가 일어날 경우에 발생하는 것이다. 표 2 에서는 ING,OOT에 대해서 음소길이 정보를 사용할 경우와 사용하지 않을 경우의 성능을 나타 낸 것이다. 길이 정보를 사용할 경우에는 OOT의 성능을 향상시킴을 알 수 있다. 이 과정은 보통인식기의 검증모듈과정에서 수행되는 것이다. 즉 검증모듈과정이 없더라도 음소길이 정보를 사용함으로서 소음문장을 많이 제거함을 알 수 있다. 그림 2를 관찰하면 길이 정보를 사용하게 됨에 따라 검증모듈의 영향이 작게 나타남을 알 수 있다.

V. 결론

본 논문에서는 음소지속시간 정보를 사용함에 따른 인식률의 변화를 연구하였다. 음소지속시간이 소음문장을 얼마나 인식 할수 있는지를 검토하였고 후처리 과정인 검증모듈과의 관계도 관찰 하였다. 음소지속시간정보만 사용하여도 소음문장을 사용하지 않을 경우와 비교하여 약 46%이상의 성능향상을 가져 왔으며 검증모듈을 사용하였을 경우 음소지속시간정보를 이용하면 약 8.4%의 오인식률이 감소됨을 알 수있었다.

참고문헌

[1] 구명완,김호경, "결정트리 기반 음성인식 시스템에서의 음소지속시간 사용방법", 대한음성학회 창립 25주년 기념 학술대회, pp. 197-200, Nov. 15 2002

[2] R. Schwarts et al., "New method for the N-best sentence hypothes within the BYBLOS speech recognition system", Proc. ICSSSP 1992, pg 1-4.

[3] Anastasios Anastasakos et al., "Duration modeling in large vocabulary speech recognition," Proc. ICASSp 1995, pg. 628-631

[4] Xue Wang et al., "Integration of context-dependent duration knowledge into HMM-based speech recognition," Proc. ICSLP 1996, pg. 1073-1076

[5] Nestor Becerra Yoma et al., "On including temporal constraints in Viterbi Alignment for speech recognition in noise", IEEE Trans. Speech and Audio Proc., 9(2): 179-182, 2001

[6] David Burshtein, "Robust parametric modeling of durations in hidden Markov models", IEEE Trans. Speech and Audio Proc, 4(3):240-242, 1996

[7] 박성준 구명완,전주식, "결정트리 모델링 기반의 음성 인식기", 제 17회 음성통신 및 신호처리 학술대

회 17권 1호, pp. 175-178 , 2000.

[8] 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

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

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

[11] Myoung-Wan Koo et al., "Speech recognition and utterance verification based on a generalized confidence score", IEEE Trans. Speech and Audio Proc., 9(6): 821-832, 2001.

표 1. 음성 DB 구조

	ING	OOT	전체
토큰수	5009	575	5,584
비(%)	89.7	10.3	100

표 2. OOT에 따른 인식성능

	토큰수	No Dur.(%)	Dur. (%)
ING	5009	87.58	88.56
OOT	575	-4.0	46.61
전체	5,584	78.15	84.24

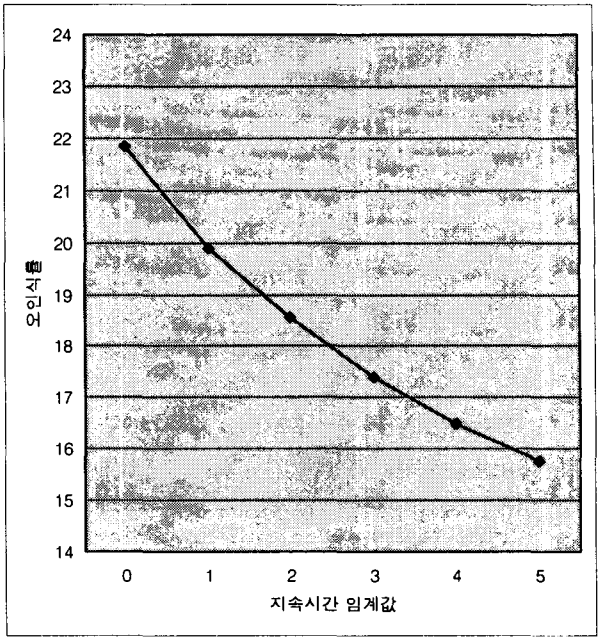

그림 1. 지속시간 임계값에 따른 오인식률

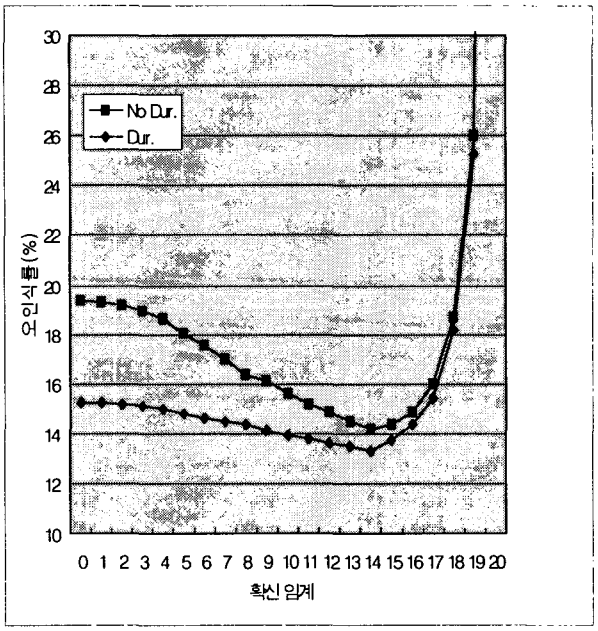

그림 2. 확신값에 따른 오인식률 변화