2000년도 한국음향학회 학술발표대회 논문집 제 19 권 제 2(s) 호

발음 속도에 따른 지속시간 제한값의 보상

양태영†, 이충용, 윤대희, 차일환
연세대학교 전기전자공학과

A Compensation of the Duration Bounds According to Speaking Speed

Tae-Young Yang†, Chungyong Lee, Dae Hee Youn, and Il-Whan Cha
Dept. of Electrical and Electronic Eng., Yonsei Univ.
tyyang@radar.yonsei.ac.kr

요 약 문

본 논문에서는 제한 지속시간 모델링(bounded duration modeling)의 지속시간 제한값(duration bound)을 화자의 발음 속도에 따라 조절해주는 발음 속도 보상 알고리즘을 제안한다. 제안된 알고리즘은 두 번의 인식 과정을 수행하는데, 1차 인식 과정은 화자의 발음 속도를 추정하기 위한 과정이고, 2차 인식 과정이 인식 결과를 얻기 위한 과정이다. 1차 인식 과정에서 추정된 화자의 발음 속도에 따라, 지속시간 제한값을 증가, 또는 감소시킨 후, 2차 인식 과정에 사용한다. 제안된 알고리즘은 CHMM 기반의 한국어 연결 숫자음 인식 시스템에 적용되었으며, KAIST에서 제작된 4-7자리 연결 숫자음 데이터베이스인 DigitDB를 대상으로 성능을 평가하였다. 인식 실험 결과, 제안된 발음 속도 보상 알고리즘이 적용된 인식 시스템에서는 96.26% 단어 인식률을 얻어, 제안된 알고리즘이 적용되지 않은 인식 시스템의 94.72%보다 1.54% 향상된 인식 성능을 얻을 수 있었다.

I. 서 론

일반적인 상태 천이 확률(state transition probability)을 사용하는 HMM의 지속시간 확률(duration probability)는 지속시간이 증가함에 따라 지수적으로 감소하는 분포(geometric distribution)을 갖는다[1]. 이와 같은 지속시간 확률은 연결 숫자음을 비롯한 연속음 인식에서, 다량의 첨가(insertion) 오류를 일으키는 원인이 된다[2]. 따라서, 지속시간을 조절해주는 기법이 요구되며, 이러한 기법들을 지속시간 모델링(duration modeling)[3][4]이라고 한다. 지속시간 모델링 기법들 중 제한 지속시간 모델링(bounded duration modeling)[5]은 구현이 간단하면서도 좋은 성능을 보이는 장점이 있다.

제한 지속시간 모델링에서는 각 단어 및 상태(state)의 지속시간이 최소 제한값(low bound)과 최대 제한값(upper bound) 사이로 제한된다. 따라서, 지속시간의 최소 제한값과 최대 제한값에 의해 그 성능이 좌우된다. 만일, 인식 대상 화자의 발음 속도가 지나치게 빠르거나 느릴 경우, 올바른 인식을 위해서는 최소 제한값보다 짧은 지속시간이나 최대 제한값보다 긴 지속시간이 요구될 수 있다. 보다 넓은 지속시간 제한값을 주어서 이러한 문제를 해결하려 한다면, 전체적인 인식 성능이 저하될 우려가 있다. 따라서, 본 논문에서는 화자의 발음 속도를 추정하고, 이에 따라 지속시간의 제한값을 적절히 변경시켜주는 발음 속도 보상 알고리즘을 제안한다.

제안된 발음 속도 보상 알고리즘에서는 두 번의 제한 지속시간 모델링이 적용된 인식 과정을 수행한다. 1차 인식 과정의 인식 결과 단어로부터 얻은 각 단어의 지속시간을 학습 데이터로부터 구한 각 단어의 평균 지속시간과 비교하여, 인식 대상 화자의 발음 속도를 결정하는데, 발음 속도의 단위는 프레임 단위를 갖는다. 결정된 발음 속도를 제한 지속시간 모델링의 최소 및 최대 제한값에 더

함으로써, 발음 속도가 보상된 지속시간 제한값을 구한 후, 이 지속시간 제한값을 2차 인식 과정에 적용하여 최종 인식 결과를 얻는 알고리즘이다.

II. 발음 속도 보상 알고리즘

인식 화자의 발음 속도를 추정하고, 이에 따라 지속시간 제한값을 교정해주기 위해서는, 먼저 발음 속도의 늦고 빠름을 판단하는 기준이 되는, 각 인식 대상 단어의 평균 지속시간 $\overline{D_n}$가 요구된다. $\overline{D_n}$는 학습 데이터로부터 다음과 같이 구할 수 있다.

$$\overline{D_n} = \frac{1}{K}\sum_{k=1}^{K} D_n^k, \quad 1 \le n \le W. \tag{1}$$

여기서 W는 전체 단어 모델의 수, K는 학습 데이터에 존재하는 단어 n의 수이며, D_n^k는 k번째로 존재하는 단어 n의 지속시간이다.

1차 인식 과정은 인식 화자의 발음 속도를 추정하기 위하여 수행되는데, 인식 결과로 W_r개의 단어가 인식되었다고 가정하면, 각 결과 단어에 대한 W_r개의 지속시간을 얻을 수 있다. 인식 결과 문장 중 i번째 단어로 단어 n이 인식되었을 때, 그 지속시간 D_n^i로 나타낸다면, D_n^i와 각 단어의 평균 지속시간 $\overline{D_n}$ 의 차이를 구하여, 인식 결과 문장의 단어 수 W_r개의 발음 속도 SSR_n^i를 계산한다.

$$SSR_n^i = D_n^i - \overline{D_n}, \quad 1 \le i \le W_r. \tag{2}$$

W_r개의 SSR_n^i로부터 최종적인 화자의 발음 속도 SSR을 얻는 방법으로는 여러 가지 방법이 사용될 수 있으나, 본 논문에서는 미디언(Median) 평균을 취하는 방법을 선택하였다. 인식 결과 문장에 나타난 단어들은 올바른 인식 결과인지, 오인식인지를 판별할 수 없다. 첨가 오류가 일어난 단어의 경우 지속시간이 매우 짧은 경우가 많으며, 삭제 오류 주변의 단어는 매우 긴 지속시간을 갖는 경우가 많다. 따라서, SSR_n^i를 일반적인 평균을 취하여 SSR을 얻는다면, 인식 오류가 존재하는 문장의 경우, 올바르지 못한 발음 속도 SSR_n^i가 그대로 SSR에 영향을 주게 된다. 미디언 평균을 사용하면 위와 같은 비정상적인 발음 속도는 SSR의 결정에서 제외시킬 수 있다.

$$SSR = Median[SSR_n^i], \quad 1 \le i \le W_r. \tag{3}$$

추정된 화자의 발음 속도인 SSR은 지속시간 즉, 프레임의 단위를 갖으며, 학습 데이터들의 평균 발음 속도와 일치할 경우 0의 값을, 평균 발음 속도보다 빠르게 발음했을 경우 양수의 값을, 느리게 발음했을 경우 음수의 값을 갖게 된다.

화자의 발음 속도에 따른 지속시간 제한값들의 조절은 SSR를 더해줌으로써 가능하다. 발음 속도가 보상된 단어 n의 지속시간 최소 제한값과 최대 제한값을 $\widehat{l_n^w}$와 $\widehat{u_n^w}$로 나타내면,

$$\widehat{l_n^w} = l_n^w + SSR, \quad 1 \le n \le W \tag{4}$$

$$\widehat{u_n^w} = u_n^w + SSR, \quad 1 \le n \le W \tag{5}$$

와 같이 구할 수 있다. 여기서 W는 전체 단어의 수이다. 발음 속도가 보상된 상태 i의 지속시간 최소 제한값 $\widehat{l_n^s}$과 최대 제한값 $\widehat{u_n^s}$은 다음과 같이 구할 수 있다.

$$\widehat{l_i^s} = l_i^s + SSR/N, \quad 1 \le i \le N \tag{6}$$

$$\widehat{u_i^s} = u_i^s + SSR/N, \quad 1 \le i \le N. \tag{7}$$

여기서 N은 상태 i가 속한 단어의 총 상태 수이다.

발음 속도가 보상된 지속시간의 최소 제한값 $\widehat{l_n^s}$과 최대 제한값 $\widehat{u_n^s}$을 사용하여 두 번째 인식 과정을 수행하게 되며, 이로부터 얻은 인식 결과를 최종 인식 결과로 한다.

제안된 발음 속도 보상 알고리즘의 블록도는 그림 1과 같다.

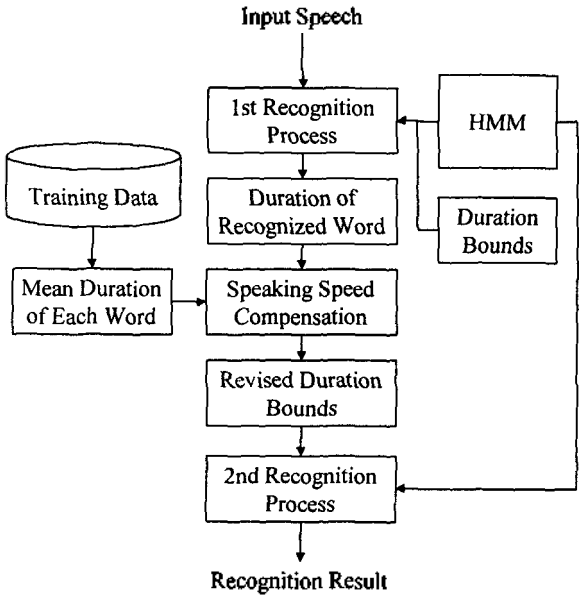

그림 1. 발음 속도 보상 알고리즘의 블록도

III. 연결 숫자음 인식 시스템

III.1. 연결 숫자음 데이터 베이스

인식 실험에 사용한 데이터 베이스는 한국과학기술원에서 제작한 DigitDB로써, 자연스럽게 발음한 한국어 연결 숫자음으로 0(영)부터 9(구)까지 10개의 숫자음과 0을 "공"으로 발음한 경우를 합한 11개의 단어로 구성되며, 한 문장은 3에서 7자리까지의 가변 자리 숫자음으로 구성되어 있는 데이터 베이스이다. DigitDB는 남성 화자 90명과 여성 화자 50명이 발음한 5,169 문장으로 이루어져 있는데, 이 중 남성 화자 60명과 여성 화자 33명이 발음한 3,440문장을 인식 시스템의 학습에 사용하였고, 나머지 남성 화자 30명과 여성 화자 17명이 발음한 1,729문장은 인식 시스템의 성능 평가에 사용하였다.

III.2. 음성 신호의 전처리 및 특징벡터 추출

16 kHz로 샘플링된 DigitDB 데이터를 $1-0.95z^{-1}$의 프리엠퍼시스(pre-emphasis) 과정을 거친 후, 20 ms의 길이를 갖는 해밍 윈도우(Hamming window)를 사용하여, 10 ms의 간격으로 처리하였다. 특징벡터는 3가지를 사용하였는데, 14차의 MFCC, 14차의 delta-MFCC와 delta-energy와 delta-delta-energy를 연결한 2차원의 벡터를 사용하였다.

III.3. HMM 구성

연결 숫자음 인식 시스템[6]은 CHMM을 기반으로 구축되었으며, 인식 단위는 triphone을 사용하였다. 각 triphone은 3개의 상태로 구성하였는데, 이 중 가운데 상태는 각 음소에 대해서 전후 음소에 관계없이 하나의 상태를 공유하도록 구성하였다. 음성 신호 사이의 묵음(silence) 구간을 인식하기 위해 묵음 모델도 구성하였는데, 묵음 모델에는 하나의 상태를 배정하였다. 11개의 숫자음과 묵음 모델을 포함하여 전체 상태 수는 181개로 이루어져 있으며, 각 단어는 이들 181개의 상태로부터 구성된다. 관찰 확률을 구하기 위한 Gaussian PDF는 각 상태당 MFCC와 delta-MFCC는 8개를 사용하였고, delta-energy와 delta-delta-energy를 연결한 2차원의 특징벡터의 경우는 4개를 사용하였다. 연속음 인식 알고리즘으로는 제한 지속시간 모델링이 결합된 One-Pass 알고리즘을 사용하였다.

IV. 실험 결과

제안된 발음 속도 보상 알고리즘의 성능을 평가하기 위하여 가변자리수 한국어 연결 숫자음에 대한 인식 실험을 수행하였으며, 그 결과는 표 1과 같다.

표 1. 인식 실험 결과 오인식률[%]

	ins	del	sub	word	sentence
no SSC	2.4	1.2	1.7	5.3	22.4
SSC	0.8	1.1	1.8	3.7	16.7

발음 속도 보상 기법이 적용되지 않은 기존의 인식 시스템(no SSC)[6]에서는 발음 속도가 느린 화자나 빠른 화자의 숫자음 문장들은 제대로 인식되지 못 하는 경우가 많았다. 특히, 발음 속도가 느린 화자의 "이"가 "이 이"로, "오"가 "오 오"로 분할되어 인식되는 예가 많았으며, 발음 속도가 빠른 화자의 경우에는 "이 이"가 "이"로, "오 오"가 "오"로, "이 일"이 "일" 등으로 모음이 합쳐져서 인식되는 예가 많이 발견되었다. 그러나, 제안된 발음 속도 보상 알고리즘이 적용된 인식 시스템(SSC)에서는 이러한

오인식들이 많이 감소되었다. 지속시간의 제한값을 변화시켰기 때문에, 치환 오류는 다소 증가하였으나, 발음 속도의 차이에 의해 많은 영향을 받았던 첨가와 삭제 오류는 감소하였고, 이 중, 특히 첨가 오류가 2.4%에서 0.8%로 크게 감소한 인식 결과를 얻을 수 있었다.

제안된 발음 속도 보상 알고리즘을 적용되어 인식 성능이 향상된 예를 그림 2와 그림 3에 보였다.

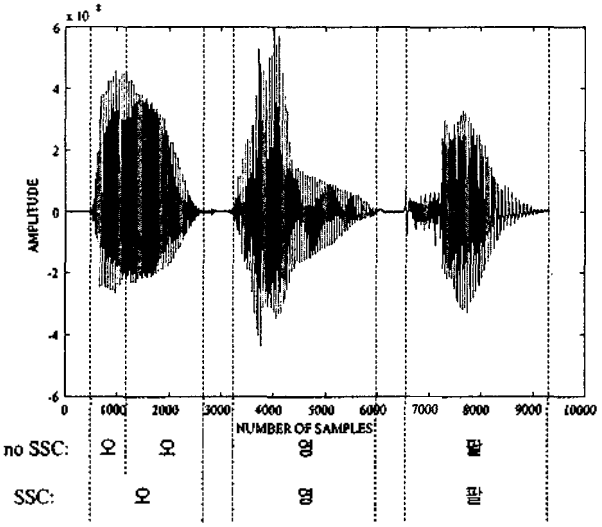

그림 2. 발음 속도가 느린 화자의 숫자음 문장에서 발생하던 첨가 오류가 교정되는 인식 결과의 예

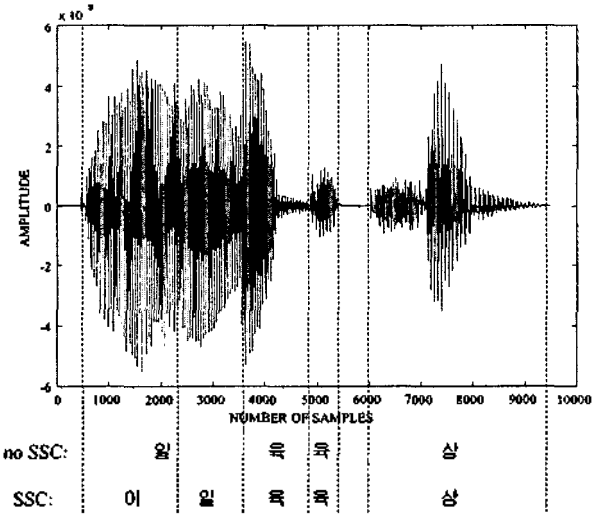

그림 3. 발음 속도가 빠른 화자의 숫자음 문장에서 발생하던 삭제 오류가 교정되는 인식 결과의 예

V. 결 론

본 논문에서는 인식 대상 화자의 발음 속도를 추정하고, 이에 따라 치한 지속시간 모델링의 지속시간 제한값을 조절해 주는 발음 속도 보상 알고리즘을 제안하였다. 제안된 알고리즘은 한국어 연결 숫자음 인식 시스템에 적용되었으며, 가변자리수 한국어 연결 숫자음 데이터 베이스인 DigitDB를 사용하여 성능을 평가하였다.

인식 실험 결과, 기존의 인식 시스템에서 많이 발견되었던 발음 속도가 지나치게 늦은 화자의 숫자음 문장에 대한 첨가 오류와, 발음 속도가 지나치게 빠른 화자의 숫자음 문장에 대한 삭제 오류들을 제안된 발음 속도 보상 알고리즘을 적용함으로써 크게 줄일 수 있었다. 단어 오인식률을 기준으로 살펴볼 때, 기존의 인식 시스템은 5.3%를 보였으나, 제안된 알고리즘을 적용한 인식 시스템의 경우 3.7%를 얻어, 1.6%의 성능 향상을 얻을 수 있었다.

참고 문헌

[1] David Burshtein, "Robust Parametric Modeling of Durations in Hidden Markov Models," in *Proc. of ICASSP*, vol. 1, pp. 548-551, Detroit, Michigan, USA, May 1995.

[2] Kevin Power, "Duration Modelling for Improved Connected Digit Recognition," in *Proc. Int. Conf. Spoken Language Processing*, vol. 2, pp. 885-888, Philadelphia, MA, USA, Oct. 1996.

[3] J. D. Ferguson, "Variable Duration Models for Speech," in *Proc. Symp. on the Application of Hidden Markov Models to Text and Speech*, pp. 143-179, Princeton, New-Jersey, Oct. 1980.

[4] M. J. Russel, R. K. Moore, "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition," in *Proc. ICASSP*, pp. 5-8, 1985.

[5] H. Y. Gu, C. Y. Tseng, L. S. Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations," *IEEE Trans. on SP*, vol. 39, no. 8, pp. 1743-1751, August 1991.

[6] 양태영, 이충용, 윤대희, 차일환, "제한 지속시간 모델링의 한계값 결정을 위한 반복적인 알고리즘," 한국음성과학회 제9회 학술발표대회 논문집, pp. 143-148, 제주대학교, 2000년 10월.