모음길이 비율에 따른 발화속도 보상을 이용한 한국어 음성인식 성능향상

박 준 배*, 김 태 준*, 최 성 용*, 이 정 현**
*인하대학교 전자계산공학과, **인하대학교 컴퓨터 공학부

An Improvment of Korean Speech Recognition Using a Compensation of the Speaking Rate by the Ratio of a Vowel length

*Jun Bae Park, *Tae Jun Kim, *Seong Yong Choi, **Jung Hyun Lee
[*Dept, **School] of Computer Science and Engineering, Inha University
*[tohope, hope673]@nlsun.inha.ac.kr, **jhlee@inha.ac.kr

Abstract

The accuracy of automatic speech recognition system depends on the presence of background noise and speaker variability such as sex, intonation of speech, and speaking rate. Specially, the speaking rate of both inter-speaker and intra-speaker is a serious cause of mis-recognition. In this paper, we propose the compensation method of the speaking rate by the ratio of each vowel's length in a phrase. First, the number of feature vectors in a phrase is estimated by the information of speaking rate. Second, the estimated number of feature vectors is assigned to each syllable of the phrase according to the ratio of its vowel length. Finally, the process of feature vector extraction is operated by the number that assigned to each syllable in the phrase. As a result, the accuracy of automatic speech recognition was improved using the proposed compensation method of the speaking rate.

I. 서론

음성인식 시스템은 일반적으로 학습과정과 정합과정의 두 단계 처리과정으로 이루어진다. 따라서 학습과정과 정합과정 사이의 상태가 불일치 할 경우 음성인식 시스템의 성능에 크게 영향을 줄 수 있다. 이와 같은 학습과정과 정합과정 상태의 불일치 요인으로는 입력 장치, 배경잡음, 그리고 화자의 성별, 억양, 발화속도와 같은 화자 변이성 등이 있다. 특히 발화속도의 변화는 음성인식 시스템의 성능에 크게 영향을 미치는 것으로 알려져 있다. 이는 화자간의 발화속도 변화나 동일 화자내의 발화속도 변화가 음성신호 변이성의 원인이기 때문이다[1]. 최근 이와 같은 발화속도 변화에 따른 오인식률을 줄이기 위한 시도가 다음과 같이 이루어지고 있다. (1)HMM의 상태 천이 확률들을 적절하게 수정하거나, (2)보통의 발화속도, 느린 발화속도, 그리고 빠른 발화속도의 음성 말 뭉치를 수집하여 발화 속도 의존적인 음향적 모델을 설계하는 방법, (3)혹은 다른 발화속도를 가지는 음성의 스펙트럼 분석을 이용한 시도 등이 있다[2].

본 논문은 발화속도의 변화에 따른 오인식률을 줄이기 위해 어절 내 평균 음절 발화속도와 모음의 길이 비율을 이용하여 특징 벡터 열을 추출하는 발화속도 보상 방법을 제안한다. 각 음성 분석구간에서 평균 음절 발화속도와 모음 길이 정보를 이용하여 발화속도에 보다 강인한 음성 특징 벡터 열을 추출하여 발화속도 변화에 따른 음성인식 시스템의 오인식률을 줄일 수 있다.

II. 관련연구

II.I 발화속도 개요

발화속도의 중요성이 알려지면서 많은 연구가 이루어지고 있지만, 아직 실질적인 발화속도의 정의에 대해서는 단일한 의견이 나와 있지 않다. Pfitzinger는 발화속도를 전역(Global) 발화속도, 국부(Local) 발화속도, 그리고 상대적(Relative) 발화속도로 구별하여 설명하고 있다[3,4,5].

상대적 발화속도 (Relative Speaking Rate)
상대적 발화속도 측정은 주어진 음성과 언어학적으로 동일한 기준 음성과의 대응되는 부분의 비율을 이용하는 방법이다.

전역 발화속도 (Global Speaking Rate)
전역 발화속도는 단위시간에 발화된 음절들이나 모라(Mora)와 같은 특정 음성학적 단위들의 개수에 의해서 정의된다.

국부 발화속도 (Local Speech Rate)
국부 발화속도는 전역 발화속도와는 다르게 명확히 정의되어 있지 않으며, 일반적으로 음성파형이나 주파수 스펙

트럼을 통해서 얻어지는 분절들의 지속시간으로 나타낸다.

II.II 주요 발화속도 측정 단위

평균 음소 속도 (Average Phone Rate)

음소 속도는 음소 지속시간의 역으로 나타낸다. 발화된 음성 내에서 음소들 간의 경계를 자동으로 결정하는 것은 매우 어렵기 때문에 음소들의 개수에 대한 평균 음소 속도를 이용한다[6]. 정밀하게 발화 속도를 측정할 수 있지만, 음소 경계를 결정하는 것은 인식과정과 비슷한 계산 량을 필요로 한다[7].

단어 속도 (Word Rate)

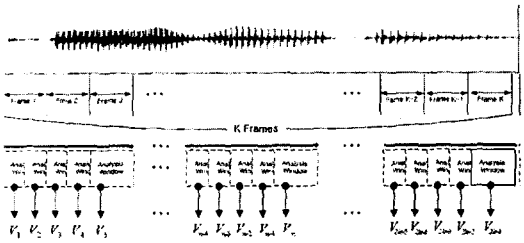
가장 단순한 발화 속도 측정 단위로써 1분 혹은 1초간 발화된 단어의 개수로 정의된다[6]. 단어 속도는 각 단어에 대한 길이 및 구조의 예측이 불가능하고 단어들 사이의 단락(pause)의 불확정성 때문에 정확한 값을 얻기 힘들다.

평균 음절 속도 (Average Syllable Rate)

음절 속도는 음소 속도와 마찬가지로 음절 지속시간의 역으로 나타낸다. 음절 속도의 경우 음절의 개수가 많고 각 음절마다의 길이가 틀리기 때문에 음절 속도 보다 평균 음절 속도를 이용하는 것이 더욱 효과적이다[8]. 음절은 발화시 인간이 인지할 수 있는 가장 기초 적인 단위로써 하나의 음절에는 반드시 모음 하나를 포함하고 있다. 본 논문에서는 음절 속도를 이용하여 발화 속도를 측정한다.

II.III 일반적인 특징벡터열 추출과정

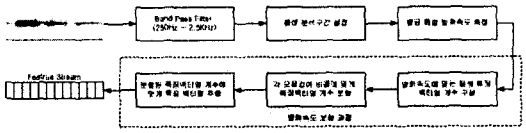
일반적인 특징 벡터열 추출과정은 [그림 1]과 같이 분석창을 인접한 프레임으로 동일한 크기 만큼씩 이동시키며 특징 벡터열을 추출한다. 이와 같은 경우 분석창이 이동하는 크기가 일정하기 때문에 빠른 발화속도를 가지는 음성신호에 대해 인식과정에서 사용할 충분한 특징 벡터열을 얻을 수 없다. 반대로 느린 발화속도의 경우 상대적으로 많은 특징 벡터열을 갖게 되는 문제점이 있다.



[그림 1] 일반적 특징 벡터열 추출 과정

III. 발화속도 보상

본 논문은 발화속도 보상을 위해, 평균 음절 발화속도 정보와 어절 내 음절의 비율 정보를 알아내고, 이 정보들을 이용하여 특징 벡터 열을 추출하는 보상 방법을 제안한다. [그림 2]는 제안한 발화속도 보상 방법을 나타낸다. 먼저 입력된 음성 신호에 대해 잡음성분을 제거하고 음절의 핵심부를 검출이 용이하도록 대역 통과 필터(Band Pass Filter)에 통과시킨다[9]. 필터를 통과한 음성신호에 대해 음성 분석 구간들을 설정하고 해당 분석구간의 평균 음절 발화속도를 측정한다. 본 논문에서 제안하는 실질적인 발화속도 보상 과정은 [그림 2]에서 점선으로 이루어진 블록으로 크게 3가지 단계로 이루어진다.



[그림 2] 발화속도 보상 과정

III.I 발화속도에 맞는 음성 특징벡터열 개수 구성

[그림 2]의 발화속도 보상 과정은 다음 식들을 이용하여 구성된다.

$$SR = \frac{AvgSRate}{Standard\ \ AvgSRate} \tag{1}$$

$$Fnum' \approx Fnum \times SR, \quad if\,(SR > 1) \tag{2}$$

$$Fnum' = Fnum, \quad if\,(SR \le 1) \tag{3}$$

AvgSRate : 분석구간의 평균 음절 발화 속도
Standard AvgSRate : 기준 평균 음절 발화속도
SR : 기준 발화속도와 분석구간의 발화속도의 비율
Fnum : 기존 특징벡터열 개수
Fnum' : 발화속도에 의한 특징 벡터열 개수

*Fnum*은 발화속도를 고려하지 않았을 때, 생성할 특징 벡터열 개수를 의미하며, 음성 분석구간의 전체 길이를 분석창 크기로 나누고 프레임간 중첩 비율을 적용하여 구할 수 있다. *Fnum'*는 발화속도를 고려하여 구성한 새로운 특징 벡터열 개수를 나타내며, 통계적으로 측정된 평균 음절 발화속도와 음성분석 구간의 평균 음절발화속도와의 비율을 나타내는 식(1)에 따라 식(2)과 식(3)으로 나타낼 수 있다. 즉, 기준 발화속도에 비해 빠른 발화속도를 가지는 경우 더 많은 특징 벡터 열을 생성할 수 있도록 *SR*을 반영해 주고 느린 발화속도의 경우에는 기준 발화속도에서 생성할 특징 벡터열 개수를 그대로 적용시킨다. 분석창의 크기는 고정이기 때문에 느린 발화속도를 가지는 음성 분석구간에 *SR*을 반영해줄 경우 특징 벡터 열을 추출하는 과정에서 인접한 프레임간의 연관성이 없어지고 노이즈 요소가 추가되기 때문이다.

III.II 모음길이 비율 측정

발화속도를 고려하여 구성된 *Fnum'*은 음성 분석구간에서 추출할 벡터열의 총 개수이다. 각 음성 분석구간은 한 개 이상의 모음으로 구성되어 있으며 분석구간 내 모음의

비중이 다르기 때문에 *Fnum'*을 모음의 비중에 맞게 분할해야 한다. 본 논문에서는 모음구간의 정보를 포만트 주파수 정보와 바크 척도를 이용하여 알아낸다. 먼저 하나의 음성 분석구간에 대해 프레임별 포만트 주파수인 *F1, F2, F3* 정보를 구하고, 식(4)의 청각 모델인 바크 척도(Bark Scale)를 이용하여 *B1,B2,B3*로 정규화 시킨다[10]. 정규 화된 바크척도의 |*B2-B1*|,|*B3-B2*| 변화량을 이용하여 모음구간의 정보를 알아낸다[11].

$$B_i = 13\tan^{-1}(0.76F_i/1000) + 3.5\tan^{-1}(F_i/7500)^2 \qquad (4)$$

F_i : i번째 포만트 주파수

식(5)은 모음의 비율정보를 나타내며 상대적으로 짧은 모음이 보다 높은 비율 정보를 갖도록 음성 분석구간의 전체 길이에 대해 각 모음구간 길이의 역수로 나타낸다. 구해진 모음의 비율정보와 생성할 특징 벡터열 개수를 분석 구간 내에 짧은 모음구간에서 보다 많은 특징 벡터열 개수를 할당함으로써 성능을 향상 시킨다.

$$R_i = \frac{L_T}{L_i} \qquad (5)$$

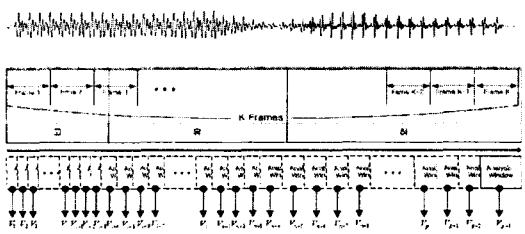
L_i : i번째 모음 구간 길이

III.III각 모음 길이 비율에 맞게 특징 벡터열 개수 분할

식(6)은 음성 분석구간의 발화속도를 고려한 특징 벡터열 개수 *Fnum'*와 각 모음 길이 비율 R_i를 가중치로 이용하여 모음 구간별로 생성할 특징 벡터열 개수를 나타낸다. 따라서 $Fnum_i$은 음성 분석구간의 발화속도 정보와 모음 길이 비율 정보가 반영된 음성 특징 벡터열의 개수이다.

$$Fnum_i = Fnum' \times \frac{R_i}{\sum_{i=1}^{N} R_i} \qquad (6)$$

$Fnum_i$: i번째 모음에 분할될 특징벡터열 개수



[그림 3] 발화속도 보상에 의한 특징 벡터열 추출과정

[그림 3]은 '고유하다' 음성 중 음성 분석구간 설정에 의해 설정된 분석구간 '고유하' 음성에 대해 발화속도 보상과정을 적용한 결과를 보여주고 있다. 발화속도 측정에 의해 결정된 추출할 특징 벡터열 개수가 각 모음 길이 가중치에 따라 분할되었음을 볼 수 있다. '고' 발음이 분석구간에서 상대적으로 차지하는 비율이 낮기 때문에 보다 많은 특징 벡터 열이 추출 되는 결과를 볼 수 있다. 반대로 상대적으로 높은 비율을 차지하고 있는 '하' 발음의 경우 보다 적은 특징 벡터 열이 추출 되는 결과를 볼 수 있다.

VI. 실험 및 결과

VI.I 실험 환경

성능 평가를 위해, 본 논문에서 제안하는 발화속도 보상 방법을 기존의 고립단어 음성인식 엔진에 적용하였을 때와 적용하지 않았을 때의 인식률값을 비교 실험을 하였다. 실험용 데이터는 50개의 컴퓨터 명령어를 남, 여 각 3명이 '느린 속도', '보통 속도', 그리고 '빠른 속도'로 5번씩 낭독한 음성을 이용하였으며, 기준이 되는 평균 음절 속도는의 값을 이용하였다[8].

VI.II 실험 결과

[표 1]은 발화속도 보상 방법을 기존의 음성인식 엔진에 적용하지 않았을 때의 인식률을 보여주고 있다.

[표 1] 발화속도 보상을 하지 않은 경우의 인식률 (단위 %)

성별	화자	느린속도	보통속도	빠른속도	평균
남	화자1	81.1	87.6	68.7	79.1
	화자2	80.5	84.7	70.9	78.7
	화자3	82.6	86.9	63.3	77.6
여	화자4	80.2	86.1	73.5	80.0
	화자5	79.9	85.9	68.5	78.1
	화자6	83.2	87.2	72.2	80.9
평균		81.3	86.4	69.5	79.0

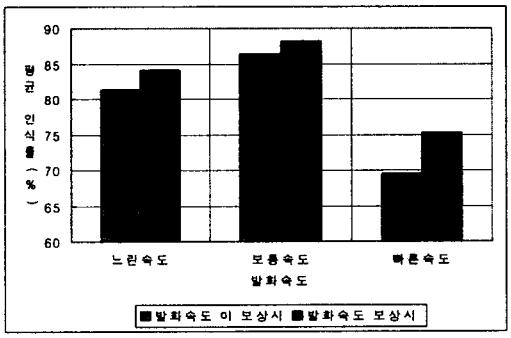
일반적으로 느린 발화속도 때와 빠른 발화속도를 가질 때의 인식률이 보통 발화속도때의 인식률보다 낮아지는 것을 알 수 있다. 특히, 빠른 발화속도의 경우 인식률이 급격하게 떨어지는 것을 알 수 있다.

[표 2]는 본 논문에서 제안하는 모음 길이 비율에 따른 발화속도 보상 방법을 적용하였을 때의 인식률을 보여주고 있다.

[표 2] 발화속도 보상을 적용한 경우의 인식률(단위: %)

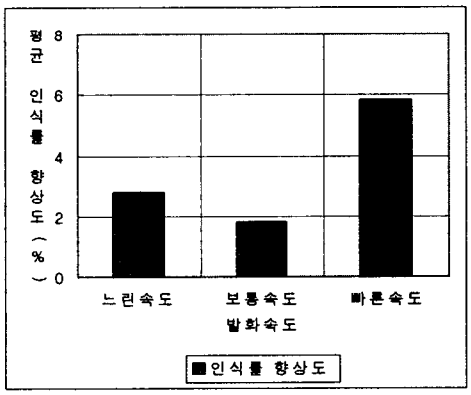
성별	화자	느린속도	보통속도	빠른속도	평균
남	화자1	85.5	88.9	75.7	83.4
	화자2	82.7	87.5	77.5	82.6
	화자3	86.6	89.1	70.0	81.9
여	화자4	83.5	87.7	75.6	82.3
	화자5	82.3	86.5	73.1	80.7
	화자6	84.1	89.2	79.8	82.5
평균		84.1	88.2	75.3	82.5

[표 2]의 결과는 [표 1]의 결과와 비교해 느린 발화속도와 빠른 발화속도의 경우에 향상된 결과를 보여주는 것을 알 수 있다. [그림 4]는 각 발화속도에 대해서 제안한 발화속도 보상 방법을 적용하였을 때와 적용하지 않았을 때의 평균 인식률을 그래프로 보여주고 있다.



[그림 4] 발화속도 보상법에 의한 평균 인식률 향상도

[그림 5]는 각 발화속도 구간에서 본 논문에서 제안한 발화속도 보상법에 의한 인식률 향상도를 그래프로 보여준다. 제안한 발화속도 보상 방법이 빠른 발화속도에서 보다 효과적임을 알 수 있다.



[그림 5] 각 발화속도에 따른 평균 인식률

V.결론 및 향후과제

일반적으로 발화속도는 화자의 고유한 특징으로 화자간에 차이가 있으며 동일 화자의 경우에도 감정이나 음성 구간별로 차이가 발생한다. 본 논문에서는 이와 같은 발화속도의 변화에 의한 음성인식 시스템의 성능저하를 막기 위해 평균 음절 발화속도와 모음 길이 비율을 이용한 보상 방법을 제안하였다. 제안한 보상 방법은 발화속도 비율에 따른 음성 특징 벡터열 개수를 구성하는 단계와 각 모음 길이 비율에 맞게 음성 특징 벡터열 개수를 분할하는 과정으로 이루어져 있다. 제안된 발화속도 보상방법을 이용하여 기존 고립단어 음성인식 시스템에 적용하여 실험한 결과 평균 3.5%의 향상된 결과를 얻었다. 향후 연구 과제로는 보다 세밀한 음절구간 검출을 위한 연구가 필요하며, 아직까지 단일한 정의가 없는 발화속도 정의 및 다양한 보상방법에 대해 지속적인 연구가 필요하다.

참고문헌

[1] T. Pfau and G. Ruske, "Estimating the Speaking Rate by Vowel Detection,"Proc., ICASSP '98, pp. 945-948, Seattle, Washington, May 1998.

[2] M. Richardson, M, Hwang, A. Acero and X. Huang,"Improvements on Speech Recognition for Fast Talkers,"Proc., the Eurospeech Conference, Budapest, Sep., 1999.

[3] H. R. Pfitzinger, "Local Speech Rate as A Combination of Syllable and Phone Rate," Proc., ICSLP '98, vol. 3, pp. 1087-1090, Sydney. 1998.

[4] H. R. Pfitzinger, "Two Approaches to Speech Rate Estimation," Proc., SST '96, pp.421-426, Adelaide. 1996.

[5] S. Ohno, M. Fukumiya and H. Fujisaki., "Quantitative Analysis of the Local Speech Rate and Its Application to Speech Synthesis," Proc., ICSLP '96, vol. 4, pp. 2254-2257, Oct 1996.

[6] M. A. Siegler, *Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition*, Thesis, Carnegie Mellon University, 1995.

[7] M. A. Siegler, and R. M. Stern, "On The Effects Of Speech Rate In Large Vocabulary Speech Recognition Systems," Proc., CASSP '95, pp. 612-615, Detroit, Michigan, May 1995.

[8] 김재범, 이홍규, 이정현, "한국어 연속음 인식을 위한 발화속도 측정기의 설계 및 구현," 한국정보처리학회 96 추계학술대회 vol. 3, pp. 755-759, 1996.

[9] H. R. Pfitzinger, S. Burger and S. Heid, "Syllable Detection in Read and Spontaneous Speech," Proc., ICSLP '96, vol. 2, pp. 1261-1264. Philadelphia. Oct. 1996.

[10] X. Huang, A. Acero and H. W. Hon, *Spoken Language Processing*, Prentice Hll PTR, 2001.

[11] A. K. Sydal, H. S. Gopal, "A Perceptual Model of Vowel Recognition Based on the Auditory Representation of American English Vowels," Journal of the Acoustic Society of America, vol. 79, pp. 1086-1100, 1986.