발화속도 적응적인 한국어 연속음 인식기

김 재 범[†] · 박 찬 규[††] · 한 미 성[††] · 이 정 현[†††]

요 약

본 논문에서는 발화속도 측정과 이를 통한 보상방법을 통하여 성능 향상된 한국어 연속음 인식 시스템을 제안한다. 연속음 인식은 다양한 조음화 현상과 발화속도의 변화로 인하여 고립단어 인식에 비하여 어렵다. 따라서, 연속음 인식을 위해서는 조음화 현상과 발화속도의 변화를 모델링할 수 있는 방법이 필요하다.

본 논문에서는 발화속도를 포만트의 변화율로서 측정하였고, 이 정보를 이용하여 빠른 발화에서는 상대적으로 많은 특징벡터를 발생시켜 보상을 시도하였다. 또한 조음화 현상을 모델링하기 위하여 한국어의 다이폰 집합을 514개로 정의하였고, 훈련을 위한 음성 DB로는 ETRI의 445 단어 DB를 사용하였다. 이러한 방법을 결합한 한국어 연속음 인식기를 DHMM (Discrete Hidden Markov Model)으로 구현하여 인식률이 향상됨을 보였다.

Adaptive Korean Continuous Speech Recognizer to Speech Rate

Jae-Beom Kim[†] · Chan-Kyu Park[††] · Mi-Sung Han[††] · Jung-Hyun Lee[†††]

ABSTRACT

In this paper, we presents automatic Korean continuous speech recognizer which is improved by the speech rate estimation and the compensation methods. Automatic continuous speech recognition is significantly more difficult than isolated word recognition because of coaticulatory effects and variations in speech rate. In order to recognize continuous speech, modeling methods of coaticulatory effects and variations in speech rate are needed.

In this paper, the speech rate is measured by change of formant, and the compensation is performed by extracting relatively many feature vectors in fast speech. Coarticulatory effects are modeled by defining 514 Korean diphone set, and ETRI's 445 word DB is used for training speech material. With combining above methods, we implement automatic Korean continuous speech recognizer, which shows improved recognition rate, based on DHMM(Disctrete Hidden Markov Model).

1. 서 론

인간이 기계와 통신을 하는 방법으로서 가장 바람직한 것은 인간의 음성을 통신 수단으로 사용하는 것이다. 이러한 목적을 달성하기 위해서 기계가 음성을 인식하게 하는 음성인식 기술이 빠르게 발전해 왔다. 이러한 음성인식 기술의 최종 목표는 임의의 화자가 발성한 연속적인 음성을 실시간에 높은 인식률로 인식하는 시스템의 개발이다[21]. 여기에 있어서 다른 입력수단인 키보드, 마우스 등에 비해 경쟁력을 갖기 위해서는 모든 사용자와 모든 환경에서 높은 인식률을 유지할 수 있어야 한다. 그러나 이러한 화자에 독립적인 연속음을 인식하는 것은 대단히 어려운 일이

※본 연구는 인하대학교 96년도 연구비 지원에 의하여 수행되었음.
† 정 회 원:LG정보통신 중연구소
†† 준 회 원:인하대학교 전자계산공학과
††† 종신회원:인하대학교 전자계산공학과
논문접수:1996년 12월 19일, 심사완료:1997년 4월 30일

므로 기존의 인식기는 좀 더 제한적인 기능을 갖는 것이 연구되어 왔다. 이러한 인식기는 대상화자의 수에 따라 화자종속, 화자독립으로 나뉘고, 음성의 형태에 따라 고립단어 인식, 연결음 인식, 연속음 인식으로 나뉠 수 있다. 현재까지는 한국어에 대하여 화자독립 고립단어인식, 화자독립 연결음인식에 대한 많은 연구가 수행되었고, 고립단어 인식 시스템은 이미 하드웨어로 구현할 수 있는 단계에 이르렀으나, 연속음 인식은 아직 시작 단계에 있다[6].

연속음 인식이 어려운 이유는 매우 많지만, 그중에서 대표적인 것으로는 발화속도의 다양한 변화와 조음화 현상(coarticulation effect)을 들 수 있다[10]. 이 때문에 인식단위 선택시 다양한 조음화 현상을 포함하고 있는 문맥 의존적인 단위(즉, diphone, triphone)에 대한 연구가 활발히 진행되고 있다[9]. 하지만 이에 비해 발화속도에 대한 연구는 적은 실정이다. 기존의 연속음 인식 시스템에서는 확률모델을 이용하여 전체속도에 대한 평균에 의해 발화속도를 고려하고 있지만[4], 발화속도의 부분변화나 강세같은 음성의 길이변화는 대부분의 시스템에서 무시되어 왔고, 이로 인한 인식률의 저하는 피할 수 없었다[5]. 즉, 빠른 발화에서는 학습에 사용된 정규패턴이 몇 샘플 정도 밖에 발생되지 않고, 느린 발화에서는 중복적으로 많이 발생되기 때문에 발화속도를 고려하지 않고 학습된 모델에서의 인식율이 떨어지게 된다. 따라서 자연발화에서 발화속도의 변화를 인식시에 반영하는 것이 필요하다.

기존의 발화속도에 대한 대부분의 연구에서, 인식도중[4, 15] 또는 인식 후[1] 음소 길이 모델링을 통하여 빠른 발화를 보상하려고 시도하여 왔다. 하지만 이들은 대부분 음소길이의 통계치를 측정하고 이것의 평균치로써 음소의 길이 모델링을 시도하였다. 그러나, 실제 인식에서 문제가 되는 부분은 발화의 전체속도가 아니라 부분속도의 변화이므로 인식률 향상에 크게 기여하지 못했다[5, 14].

따라서 본 논문에서는 인식 이전에 발화속도를 측정하고, 이 측정된 정보를 이용, 인식 시에 보상하는 방법을 제안한다. 발화속도는 음절속도와 언절단위의 평균 음절속도 두가지를 측정하는데, 우선 LP 다항식의 근을 구하여 여기에서 포만트를 계산해내고, 이 포만트의 변화량을 구하여 결정한다. 따라서 포만트의 변화가 큰 부분을 음절의 경계로 판단하고 음절의 속도를 계산해 낸다. 그 후에 이러한 정보를 인식시에 분석창 이동율의 결정에 사용하여 보상을 하였다. 즉, 빠른 발화에서는 특징벡터를 상대적으로 많이 추출하고, 느린 발화에서는 특징벡터를 상대적으로 적게 추출하여, 발화속도에 적응적으로 특징벡터를 추출하였다. 보상효과의 실험을 위하여 DHMM(Discrete Hidden Markov Model)기반의 연속음 인식기를 구현하여 발화속도에 적응된 인식과 발화속도를 고려하지 않은 시스템을 비교하여 수행하였다.

2. 발화속도 측정 및 보상방법

2.1 발화속도 측정단위

가장 단순한 발화속도 측정단위로는 단어속도를 생각할 수 있다. 단어속도는 분당 발화된 단어의 개수로 정의된다.

$$\text{단어속도} = \frac{\text{단어의 개수}}{\text{1분동안 발화된 총 크기}} \tag{2-1}$$

하지만 이는 단어안에 음절수가 많으면 음절의 속도는 올라가는데 단어속도는 낮아지는 문제점이 있고, 단어사이의 휴지의 길이가 지배적이므로 다소 부적절하다. 또한 발화속도의 부분변화를 측정하려면 단어속도보다 더 상세한 정보가 필요하다[14]. 다른 단위로는 음소속도가 있다. 음소속도는 음소길이의 역수로 정의된다. 하지만 이는 발화단위 안에서 정확한 음소 경계를 자동으로 결정하는 것이 매우 어렵다. 따라서 몇 개의 음소에 대한 평균 음소속도는 에러를 줄일 수 있으며 다음과 같이 구해진다.

$$\text{평균음소속도} = \frac{\text{음소의 개수}}{\text{각 음소 길이의 총합}} \tag{2-2}$$

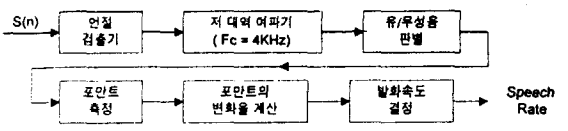
이것은 가장 정밀하게 발화속도를 측정할 수 있지만, 음소경계를 결정하는 것이 인식과 비슷한 정도의 많은 계산량을 필요로 하고 음소경계 결정시 에러율도 크므로 신뢰할 만한 결과를 얻기 힘들다[14]. 또 다른 측정단위로는 음절속도가 있다. 이는 초당 음절의 수로 정의하는데, 이 또한 연속발화안에서 음절경계의 결정이 어려우므로 식 (2-3)가 같이 평균 음절 속도를 정의한다.

$$\text{평균음절속도} = \frac{\text{음절의 개수}}{\text{발화단위의 총 크기}} \cdot f_s \qquad (2\text{-}3)$$

여기서 f_s는 표본화 주파수이며, 발화단위의 총 sample의 숫자이다. 이는 한국어에서는 하나의 음절에 반드시 하나의 모음이 포함된다는 점과 언절은 휴지로 구분이 용이하고 계산량도 많지 않으므로 가장 타당하다 할 수 있다. 따라서 본 논문에서는 평균음절속도 측정기를 목표로 한다.

2.2 발화 속도의 측정

본 논문에서 제안하는 발화속도 측정기는 다음 (그림 1)과 같다.



(그림 1) 발화속도 측정과정
(Fig. 1) Process of Speech Rate Estimation

입력되는 문장 데이터에 대하여 영교차율과 대수 에너지를 이용하여 끝점 검출을 함으로써 연속발화 음성에서 언절 단위를 분리한다. 이 분리된 언절은 포만트의 변화만을 목적으로 함으로 차단 주파수가 4KHz인 저대역 여파기를 거쳐 피치 또는 잡음 성분을 제거한다. 이렇게 처리된 언절 데이터에 대하여 영교차율과 대수 에너지를 이용하여 유/무성음을 판별한다. 이러한 유/무성음 판별의 결과는 무성음 구간의 포만트 변화율 상관계수를 음절 숫자의 검출에서 제외하는 과정에서 사용한다.

2.2.1 포만트 계산

포만트 궤적은 우선 선형예측분석에 의해 행해진다. 선형예측 분석법은 앞의 2.1.1 절에서 이미 언급한 것과 같이 음성코딩과 계수분석을 하기 위해 사용되는 가장 일반적인 방법중의 하나이다. LPC 계수는 잔여신호의 평균 에너지를 최소로 하는 과정에서 구할 수있다. LPC 계수를 구하는 방법은 크게 3가지로 자기상관법(Auto-correlation Method), 공 분산법(Covariance Method), 사다리법(Lattice Method)등이 있다[11]. 본 논문에서는 LPC 계수를 구하기 위해 자기상관법의 해법인 Levinson-Durbin 알고리즘을 적용하였다.

이렇게 구한 LPC계수 a_i는 성도의 진동특성을 표현하고 있고, 위 식(3)의 $A(z)$의 근을 구함으로서, 포만트를 계산할 수 있다. 이때 역필터는 다음과 같이 표현할 수 있다.

$$A(z) = A_s(z) \cdot A_u(z)$$
$$= \prod_{i=1}^{q/2} (1 - b_i z^{-1})(1 - b_i^* z^{-1}) \prod_{i=q+1}^{p} (1 - c_i z^{-1}) \qquad (2\text{-}4)$$

여기서 $A_u(z)$은 공액쌍을 갖지 못한 불안정근과 실수근으로 이루어진다. LPC 분석에 의한 $S(z)$의 다항식은 $B(z)$이 되고, 최대 q/2개의 포만트를 가진다. LPC 분석에서는 불안정근의 소거에 의한 차수의 감소로 충분히 높은 차수로 분석할 때 충실도를 갖는 분석이 가능하다[8]. 포만트는 성도의 공진 특성을 모델링한 것으로, 공진주파수, 공진주파수의 크기와 대역폭의 추정이 필요하다. 포만트 주파수 F는 (2)식의 안정화된 근 $B(z)$에서 실수부와 허수부의 분리에 의해 다음과 같이 구해진다.

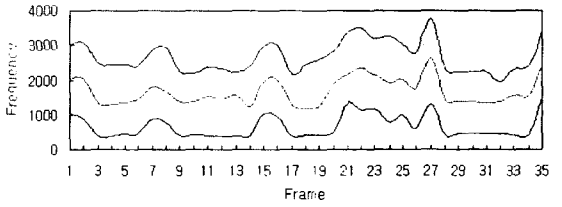
$$F_i = \frac{1}{2\pi T} \tan^{-1}\left[\frac{Im(z_i)}{Re(z_i)}\right]$$

$$B_i = -\frac{1}{2\pi T} \log\left[Re(z_i)^2 + Im(z_i)^2\right] \qquad (2\text{-}5)$$

여기서 T는 표본화 시간이다. 구해진 포만트 주파수와 대역폭에서 크기순으로 5개를 선택하여 포만트 주파수로 설정한다.

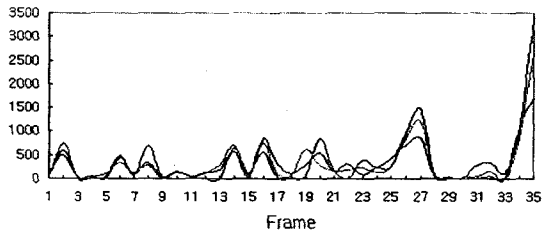
2.2.2 발화속도의 결정

위에서 구해진 포만트중에서 지배적인 영향을 미치는 제3 포만트까지 만을 가지고 속도를 측정한다. 445 단어 DB와 10명의 임의의 화자를 통계적으로 분석해 본 결과, 느린 화자의 경우 평균 음절 속도가 3.9 음절/sec, 빠른 화자의 경우 5.6 음절/sec이므로 평균적으로 4.7 음절/sec이고, 한 음절은 $200 \pm 25ms$, 표준편차는 0.5~1.5 음절/sec이다. 따라서 이점에 착안하면 모음구간에서는 최소한 60ms 동안 기울기의 변화가 거의 없을 것이다.



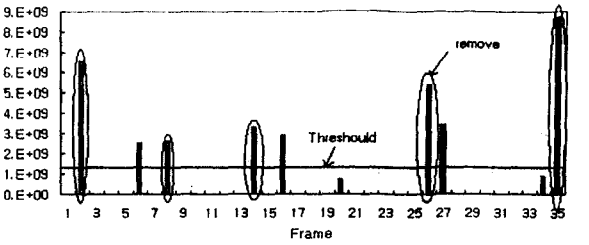
(그림 2) 포만트의 궤적
(Fig. 2) Formant Trajectory

(그림 2)는 '연속적인'이라는 연속발화의 포만트의 궤적을 보여주고 있다. 이때 안정구간이 나오는데 이는 모음구간이고, 음절의 중심으로 판단할 수 있다. 또한 세가지 포만트가 모두 변할 때 구강의 모양이 변했다고 판단할 수 있으므로 세가지 포만트의 변화율을 구하여 보면 (그림 3)과 같다.



(그림 3) 포만트의 변화율
(Fig. 3) Slope of Formants

그후 이들의 상관계수를 구하면 (그림 4)와 같다. 따라서 이 상관계수값으로 음절의 개수를 판단한다. 이때 맨앞의 세 프레임과 맨뒤의 세 프레임을 제외시켰는데, 그 이유는 언절의 시작과 끝에서는 구강의 모양의 변화가 아주 크므로 이와 비교할 때 다른 변화는 상대적으로 작아지기 때문에 문턱값 결정이 어려워지기 때문이다. 그 후에 언절의 크기로 음절의 개수를 나누어 평균 음절속도를 계산해 낸다. 이때 문턱값의 결정이 문제가 되는데 본 연구의 실험환경에서는 반복된 실험을 통해 상관계수의 최대값의 1/8의 값에서 가장 정확한 결과를 얻을 수 있었다. 또한 검출된 경계가 너무 가까운 경우는 유/무성음 판별 오차에 의한 오분석이므로 자음 구간으로 판단하여 큰값을 제외시켰다.



(그림 4) 포만트 변화율의 상관계수
(Fig. 4) Correlation Coefficients of Slope of Formants

2.3 보상기법

일반적으로 음성신호의 특징벡터 열은 일정구간의 음성데이터에 분석창을 씌운 다음, 해당 프레임의 특징벡터를 추출하고 분석창을 일정 길이만큼 이동해 나감으로 얻어진다. 이때, 분석창이 이동하는 길이가 매 프레임마다 일정하게 되는데 이는 빠른 발화에서는 부적절한 요소를 내포하고 있다. 따라서, 앞절에서 얻어진 발화속도를 고려하여 분석창을 가변적으로 이동하여 특징벡터의 발생률을 달리한다. 이로 인하여 빠른 발화에서는 보다 많은 특징벡터를 발생시키고, 느린 발화에서는 적은 수의 특징벡터를 발생시킴으로써 빠른 발화를 보상한다. 인간도 청취도중에 빠른 발화가 나타나면 조금더 주위를 기울이게 되는데 이는 빠른 발화에서는 비교적 적은 수의 특징벡터가 발생하므로 특징벡터를 보다 많이 추출하려는 것이라 할 수 있다.

각각의 시간 T_f, $2T_f$,…,MT_f는 특징벡터 추출의 기본구간을 결정하고 각구간의 길이는 T_f로 일정하다. 각각의 기본구간 내에서는 그 구간에서 추출되어야 하는 특징벡터의 수가 정해지면, 분석창이 자리잡는 위치가 결정된다.

그 구간에서 추출되어야 할 특징벡터의 수는 다음에 의해 결정된다.

$$n = n' \times \frac{\text{측정된음절속도}}{\text{평균음절속도}} \qquad (2\text{-}6)$$

$$\text{단, } n' = \frac{\text{언절의크기}}{\text{분석창의크기}}$$

즉, n'는 발화속도를 고려하지 않을 때의 특징벡터의 발생률을 의미하고, 통계적으로 측정된 평균 음절

속도와 측정된 음절속도의 비율로서 이를 통해 빠르다, 혹은 느리다를 판단한다. 이 비율을 n′에 곱하여 추출할 특징벡터의 개수를 결정한다.

구간 $[m\ T_f,\ (m+1)T_f]$에서 추출되어질 특징벡터의 수를 n이 결정되면, 첫 번째 분석창의 중앙은 $(m+1/2n)\ T_f$에 위치하고 T_f/n의 길이만큼 이동하게 된다.

3. 제안된 연속음 인식 시스템

고립단어 인식에 비해 연속음 인식은 매우 어렵다. 그 이유는 연속음의 선천적 성질에 기인한다. 첫째로 인식 단위의 경계 구분이 어렵고, 둘째로 조음화 현상(coarticulate effect)이 더욱 심해져서 같은 소리가 다양한 문맥속에서 각기 다르게 발음되는 결과가 나타난다. 셋째로는 발화 속도의 변화가 심해져서 훈련 데이터에 사용된 정규패턴이 감소하거나 늘어난다. 연구에 의하면 고립어에서 연속음으로 갈 때 에러율은 280% 정도 증가한다고 보고되고 있다[10]. 그러나 이러한 문제와 성능 저하에도 불구하고 연속음 인식은 매우 중요하다. 왜냐하면 궁극적으로 인간과 컴퓨터의 상호작용이 원활해지려면 연속음을 통해야만 하기 때문이다.

본 논문에서 제안한 발화속도 측정과 이를 이용한 보상기법에 의한 연속음 인식 시스템을 다음과 같이 구성하였다.

3.1 인식단위

고립단어 인식 시스템에서 연속음 인식 시스템으로 전환하려면 우선 연속음의 다양한 조음화 현상을 모델링할 수 있는 방법이 필요하다. 따라서 연속음 인식에서 가장 중요한 것은 인식단위의 선택이다. 이러한 인식 단위는 연속음의 다양한 조음화 현상을 모델링할 수 있는 구조를 가져야 한다. 보다 정확하게 말하면, 좋은 인식단위는 훈련성(trainability)과 일관성(consistency)을 가져야 한다[9]. 훈련성은 학습에 충분할 정도로 데이터가 빈번히 발생해야 함을 의미하며, 일관성은 같은 인식단위에 대해서 일관된 독특한 특성을 가져야 함을 의미한다.

초기의 시스템에서 사용되던 인식단위인 단어는 일관성을 갖는 반면, 단어수가 증가함에 따라 훈련데이터 양도 선형적으로 증가하여 훈련성에 문제가 나타난다. 또한 다른 인식단위로서 음소는 훈련에 필요한 충분한 양의 데이터를 얻을 수 있으나, 인접한 음소에 민감하게 반응하므로 일관성이 없어진다[9]. 다른 인식단위로 음절을 생각할 수 있으나 음절사이의 연속된 변화를 모델링하기 어려우며, 분류할 클래스의 수도 많다. 따라서 CMU(Carnegie Mellon University)의 화자 독립, 연속음 인식 시스템인 SPHINX에서는 같은 성질의 클래스들을 합하여 일반화된 트라이폰이라는 인식단위를 새롭게 정의하여 높은 인식결과를 얻었다[10]. 트라이폰은 한 음소와 그 음소와 연결된 왼쪽과 오른쪽의 문맥을 말하는데 이는 음소인식에 있어 확률적인 학습방법을 사용하여 하나의 음소를 인식할 때, 하나의 음소를 주위의 문맥에 따라 분리하여 그 음소를 인식하는 방법으로 비록 인식대상수는 많지만 높은 인식률을 얻을 수 있어서 많은 음성인식 시스템에서 사용해 왔다. 그러나 한국어의 경우, 본 연구에서 분류해 본 결과 훈련할 트라이폰이 12가지 그룹으로 나뉘고, 그 개수가 36,584개에 이르며, 이에 따른 인식시의 계산량도 너무 커진다. 뿐만 아니라, 발생빈도가 적은 트라이폰이 대부분이라 훈련 데이터의 확보가 상당히 어렵다.〈표 1 참조〉

〈표 1〉 한국어 트라이폰의 분류
〈Table 1〉 Classification of Korean Triphones

그 룹	구성 음소	예	개 수
C_iVC_f	자음(초성)+모음+자음(초성)	가즈+ㅏ	18*17*18 = 5508
C_iVV	자음(초성)+모음+모음	가야	18*17*17 = 5202
C_iVC_f	자음(초성)+모음+자음(종성)	각	18*17*7 = 2142
VC_iV	모음+자음(초성)+모음	아자	17*18*17 = 5202
VVC_f	모음+모음+자음(초성)	아이즈+ㅏ	17*17*18 = 5202
VVC_f	모음+모음+자음(종성)	아아	17*17*7 = 2023
VC_iC_f	모음+자음(종성)+자음(초성)	아ㄱㄴㅣ	17*7*18 = 2142
VC_fV	모음+자음(종성)+모음	앙아	17*7*17 = 2023
C_fC_iV	자음(종성)+자음(초성)+모음	가+ㄱ자	7*18*17 = 2142
C_iVC_f	자음(종성)+모음+자음(초성)	가+ㅇ아지	7*17*18 = 2142
C_iVV	자음(종성)+모음+모음	가+ㅇ아이	7*17*17 = 2023
C_iVC_f	자음(종성)+모음+자음(종성)	가+ㅇ앞	7*17*7 = 833

〈표 2〉다이폰 그룹(긴자음 : ㄴ, ㄹ, ㅁ, ㅇ)
〈Table 2〉 Class of diphone

그룹	구성 음소	다이폰 개수
V	모음	17
CV	자음 + 모음	306(=18*17)
VC	모음 + 자음	119(=17*7)
CC	긴자음 + 자음	72(=4*18)

따라서, 본 논문에서는 앞에서 살펴본 문제점을 해결하고 인식대상의 수를 어느 정도 줄이면서, 음절이 발달한 한국어 음성인식에서 음절사이의 연결도 모델링이 가능하고, 음절보다 그 개수가 적은 다이폰(diphone)을 선택하였다. 다이폰 레벨에서의 인식은 하나의 음소와 그 음소와 연결된 오른쪽 음소에 해당하는 영역을 하나의 다이폰으로 인식해낸다. 이러한 다이폰은 음성인식 단위가 갖추어야 할 조건을 가장 잘 만족한다(표 2 참조). 따라서 본 연구에서는 한국어의 특성을 고려하여 다이폰의 집합을 〈표 3〉과 같이 정의하였다.

〈표 3〉 인식단위의 특성 비교
〈Table 3〉 Characteristic comparison of Recognition Unit

인식 단위	Consistency	Trainability
음소	나쁨	좋음
단어	좋음	나쁨
다이폰	좋음	양호
트라이폰	좋음	어려움
음절	양호	좋음

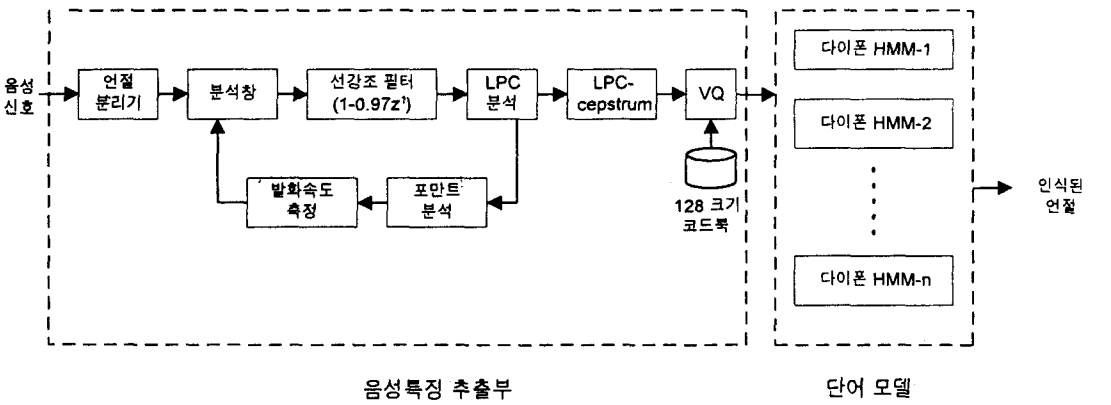
다이폰을 2개 이하의 음소로 이루어 진 단위로 볼 때, 자음만으로 이루어진 그룹(CC)과 모음만으로 이루어진 그룹(VV)이 제외되었는데 이것은 분석구간의 크기를 고려할 때, 이들은 그 구간내에서 특징을 추출해낼 수 없기 때문이다. 다이폰은 4개의 그룹으로 분류된다. CV형과 VC형이 주로 존재하지만 V와 CC도 빈번하게 발생하고 이들은 연속음절을 이어주는데 중요한 역할을 한다. 한국어 모음을 그 신호의 유사성에 따라 17개의 그룹(아, 어, 오, 에:애, 이, 야, 여, 요, 유, 외:왜:웨, 우, 와, 워, 예:애, 의, 으, 위)으로 분류하였다. 또한 종성의 경우 끝소리 규칙에 의해 7개의 대표음(ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅇ)으로만 발음되므로 V형 17개, C(초성자음)형 18개, C(종성자음)형 7개로 분류하였다.

CC의 경우 두자음중 하나는 반드시 유음이나 비음이어야 한다. 이는 다이폰을 강하게 결속된 두 개의 음소로 볼 때 모음과 유사한 성질의 긴자음일 경우만 자음과의 결합 가능성이 높기 때문이다. 이에 따라 한국어의 경우 가능한 다이폰의 조합의 수는 514개로 정의된다.

3.2 제안된 연속음 인식기

발화속도 측정에 의한 전체 연속음성 인식 시스템의 구조는 (그림 5)와 같으며, 음성신호의 특징 추출 부분, 단어 모델 부분의 두 부분으로 나뉜다.

음성신호 특징 추출 부분에서는, 먼저 선강조 필터(preemphasis filter)를 사용하여 성문에서의 −12dB의 손실과 입술의 방사 특성에 의한 +6dB의 이득으로 인한 총 −6dB의 에너지 손실을 보상해 준다[8]. 그 이후에 LPC 분석과 포만트 계산에 의해 발화속도를 결정하여, 분석창의 이동률을 구한다. 이후 다시 LPC분석과 LPC-cepstrum 분석을 수행하고, 이 특징벡터를



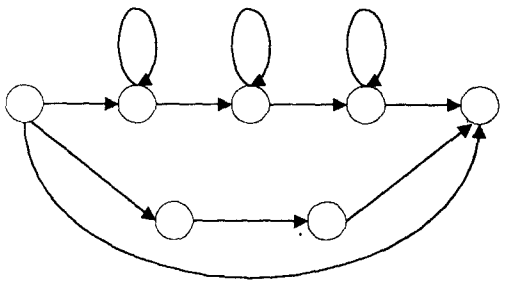
(그림 5) 전체 연속음 인식 시스템의 구조
(Fig. 5) Structure of Overall Continous Speech Recognizer

HMM의 입력벡터로 사용하기 위해서 128크기를 갖는 코드북 탐색을 통해 벡터 양자화한다.

단어 모델에서는 연속음의 조음화 현상을 모델링하기 위하여 다이폰 단위의 모델로 HMM의 한 종류인 DHMM (Discrete Hidden Markov Model)을 사용하였다. 이 DHMM의 각각의 내부구조는 (그림 6)과 같다.

현재 연속음 인식실험에서 가장 좋은 인식율을 보이는 것으로 보고되고 있는 구조로서[10], 이러한 구조의 HMM을 학습시키기 위해 한국전자 통신 연구소에서 학술용으로 공개한 445 단어 DB와 611 단어 DB를 다이폰 단위로 hand-segmentation하여, HMM의 학습 알고리즘인 forward-backward iteration을 한 번 수행하여 514개의 다이폰을 초기화하여 생성한 후, 이 다이폰 모델을 결합하여 445개의 단어모델을 생성하였다.

그 다음에 이 단어 모델을 다시 단어단위로 forward-backward iteration을 수행하여 훈련시켰다. 이렇게 훈련된 단어 모델은 인식 시에 직접 사용된다.



(그림 6) 다이폰 레벨 HMM의 내부구조
(Fig. 6) Internal Structure of Diphone Level HMM

인식 시에는 특징벡터 추출부에서 추출된 입력특징 벡터열과 시스템의 단어모델 사이의 유사도를 측정하여 어떤 단어가 가장 유사한가를 결정하는 과정[17]을 거쳐 마지막으로 인식된 언절이 출력된다.

4. 실험 및 평가

4.1 실험 환경

실험은 133MHz의 펜티엄 PC에서 16비트 사운드 카드로 수행하였다. 음성신호는 표본화 주파수 16KHz, 양자화 해상도 16Bit로 샘플링하였고, $1-0.97z^{-1}$의 선달함수를 갖는 선강조 필터(preemphasis filter)를 사용하였다. 프레임 크기는 20ms, 기본 오버랩은 10ms, 분석창은 해밍창(hamming window)을 사용하여 추출된 음성 샘플에서 LPC계수를 구하였다. LPC 차수는 16차이고, 여기에서 LPC-cepstrum 계수를 16차로 구하였다. 이 16개의 LPC-cepstrum계수를 128크기의 코드북으로 벡터 양자화(Vector Quantization) 하였다. 벡터 양자화에는 LBG 알고리즘을 사용하였다[16].

4.2 음성 DB

모델 학습을 위한 음성 DB로는 한국 전자 통신연구소에서 학술용으로 공개한 445 단어 DB와 611 단어 DB를 사용하였다. 이 DB는 국민학교 교과서에 사용된 고빈도 단어 4084 개로부터 음성학적으로 균형적인 단어 445 개를 선정한 한국어 고립단어 음성자료로서, 22명의 남성화자가 2회 발성한 것이다. 611 단어 DB는 445 단어 DB에 CV음절 144개와 숫자음 22개를 추가한 것이다. 이 DB로부터 다이폰 단위로 hand-segmentation하여 HMM 학습에 사용하였다. 연속음 인식을 위해서는 단어 DB가 아닌 연속음 DB를 사용해야 하지만, 학습에 필요한 방대한 양의 연속 음성 DB 구축은 많은 비용과 시간, 노력이 요구되고 양질의 DB 구축이 어렵기 때문에, 단어 DB에서 연속음의 성질이 약한 시작부분과 끝부분의 다이폰을 제외하고 연속음의 성질을 가지고 있는 단어 중간에서 다이폰을 추출하였다.

4.3 실험내용 및 평가

실험은 발화속도 측정의 성능 평가와 연속음 인식 시 보상에 의한 성능향상 여부를 실험하였다.

발화속도 측정기의 성능 평가를 위하여 실험은 남성화자 3명과 여성화자 3명에게 신문사설에서 무작위로 발췌한 20개의 문장을 연속 발화하게 하여 낭독음성에 대해 수행하였다. 이때 화자에게는 문장과 함께 3가지 지문이 주워졌다. '빠르게(2배정도 빠르게) 읽어 주세요', '보통 빠르기로 읽어 주세요', '느리게(2배정도 느리게) 읽어 주세요'. 이렇게 3가지 속도에 대한 발화속도 측정결과는 〈표 4〉과 같다.

〈표 4〉에서 보면 빠른 발화에서 68.9%의 성능을 보이고, 보통과 느린 발화속도에 대해서는 75.0%, 81.7%

〈표 4〉 발화속도 측정 결과

〈Table 4〉 Results of Speech Rate Estimation

화 자	빠른 속도	보통 속도	느린 속도	평 균
화자1(남)	79.1%	82.3%	87.5%	82.9%
화자2(남)	69.2%	79.7%	86.3%	78.4%
화자3(남)	62.1%	75.4%	81.8%	73.1%
화자4(여)	68.4%	72.4%	80.1%	73.7%
화자5(여)	69.2%	70.7%	79.2%	73.0%
화자6(여)	65.6%	69.5%	75.8%	70.3%
평균	68.9%	75.0%	81.7%	75.2%

의 성능을 보였다. 또한 남성의 측정오류가 여성의 것보다 작은데 이는 여성의 발화속도가 일반적으로 남자보다 조금 빠르기 때문이다. 또한 포만트가 비슷한 음절이 연이어 발화된 경우에는 포만트의 변화율이 작으므로 음절 경계를 검출해 내지 못했다.

이 정보를 이용하여 보상하는 인식 실험은 두가지로 수행하였다. 우선 445 단어 DB와 611 단어 DB로 인식실험을 수행하였고, 다른 하나는 실험실에서 일반 화자의 발화를 녹음하여 인식 실험을 수행하였다. 먼저 학습된 다이폰 HMM을 가지고 학습에 참가시키지 않은 3명분의 445 단어 DB와 2명분의 611 단어 DB와 학습에 참가한 화자 5명분의 445 단어 DB로 인식 실험을 수행한 결과는 다음 〈표 5〉와 같다.

연속음 실험은 남성화자 10명에게 445 단어로 구성된 20개의 실험 문장을〈표 6. 참조〉 연속 발화하게 하

〈표 5〉 445 단어 DB 인식 실험 결과

〈Table 5〉 Results of Recognition Experiment in 445 Word DB

학습 참가 여부	화자	발화속도 미 보상시	발화속도 보상시	성능 향상
학습 참여	화자1	86.2%	92.1%	4.9%
	화자2	82.3%	83.8%	1.5%
	화자3	85.6%	89.0%	4.4%
	화자4	81.4%	84.1%	2.7%
	화자5	79.7%	82.5%	2.8%
학습 불참	화자6	65.3%	73.3%	8.0%
	화자7	58.6%	69.2%	10.6%
	화자8	61.2%	67.7%	6.5%
	화자9	57.8%	65.2%	7.4%
	화자10	52.3%	59.9%	7.6%
평 균		71.0%	76.7%	5.6%

〈표 6〉 연속음 인식 실험 결과

〈Table 6〉 Results of Continous Speech Rrcognition Exp-reiment

화자	발화속도 미 보상시			평균	발화속도 보상시			평균	성능 향상			평균
	빠른 발화	보통 발화	느린 발화		빠른 발화	보통 발화	느린 발화		빠른 발화	보통 발화	느린 발화	
화자1	43.8	52.1	53.0	49.6	51.3	54.3	55.1	53.6	7.5	2.2	2.1	3.9
화자2	44.1	51.9	53.5	49.8	53.7	53.8	54.6	54.0	9.6	1.9	1.1	4.2
화자3	38.5	48.5	50.2	45.7	49.8	55.3	53.5	52.9	11.3	6.8	3.3	7.1
화자4	39.2	50.6	52.1	47.3	52.8	53.0	53.6	53.1	13.6	2.4	1.5	5.8
화자5	40.6	53.2	57.4	50.4	53.3	56.2	58.9	53.8	12.7	3.0	1.5	5.7
화자6	42.3	52.3	54.5	49.7	56.2	56.3	57.0	56.5	13.9	4.0	2.5	6.8
화자7	37.4	49.5	55.1	49.1	53.1	54.3	60.6	54.1	15.7	4.8	5.5	8.6
화자8	41.8	50.7	55.3	49.2	54.7	55.0	55.7	55.1	12.9	4.3	0.4	5.8
화자9	43.2	49.9	51.8	48.3	55.0	55.6	56.1	55.5	11.8	5.7	4.3	7.2
화자10	37.6	48.6	49.8	45.3	53.8	54.8	54.9	54.5	16.2	1.2	5.1	7.5
평균	40.8	50.7	53.2	48.2	53.3	54.8	56.0	54.2	12.5	3.63	2.73	6.2

〈표 7〉 445 단어로 구성된 실험문장

〈Table 7〉 Sample Sentences composed of 445 Word

문장1	교육은 교과서를 통해서 이루어진다고 한다.
문장2	그 아주머니는 이 아파트의 아홉번째 주인이다.
문장3	대체로 스포이드 사용법은 샤알레 보다 쉽다.
문장4	뜻밖에도 다람쥐는 높이뛰기를 하는 습관이 있다.
문장5	마음씨가 따스한 아주머니가 교회에 간다.
문장6	민주주의 나라에서는 많은 사람들이 자유로운 습관을 가지고 있다.
문장7	사회가 발전할수록 의식주에 대한 수요는 증가한다.
문장8	시장에서는 센티미터가 더 쓸모있다.
문장9	스탠드가 꺼지면 방안은 어두워진다.
문장10	시바이찌는 민주주의는 자유롭다고 말했다.
문장11	여섯 번째인 이 대회는 옛날 보다 훨씬 효과적으로 개최된다.
문장12	옛날에는 미처 중요하지 않은 것들이 오늘날에는 예술품으로 된다.
문장13	옛날 임진왜란은 왜적이 쳐들어온 것이다.
문장14	오히려 옛날보다 오늘날 불편한 것이 많다고 생각하는 사람이 있다.
문장15	우산이끼는 따스한 곳에서 자라난다.
문장16	우체국에서 근무하시는 집배원 아저씨는 피로이 있다.
문장17	인도네시아는 인쇄술이 끊임없이 발전한다.
문장18	최무선과 의병들은 쳐들어오는 적을 숲속에서 붙잡았다.
문장19	쿠웨이트는 석유산업으로 많은 외화를 벌어들인다.
문장20	풀숲에 사는 꾀꼬리는 평야에 사는 까마귀보다 더 오래 산다.

여 실험실 환경에서 녹음한 낭독 음성에 대하여 수행하였다. 이때 지문은 앞과 동일하다. 이렇게 3가지 속도에 대한 발화속도 측정결과는 〈표 7〉과 같다.

실험결과 445 단어 DB와 611 단어 DB에 대해서는 발화속도를 고려하여 보상시 약 5.6%의 인식율 향상에 참여하지 않았고, 녹음 환경이 잡음이 많은 실험실 환경이기 때문에 인식율 저하의 원인이 되었다. 이는 연속음성 DB의 구축과 화자적응화에 대한 연구, 그리고 잡음환경하에서의 인식 및 보상방법 등에 대한 연구가 필요하다. 세번째로는, 벡터 양자화 모듈의 양자화 에러를 들 수 있다. 이것은 코드북의 크기를 256으로 늘리고, 또한 다중 코드북을 사용함으로써 줄일 수 있을 것이다.

5. 결 론

본 논문에서는 발화속도를 측정하여 이것을 인식시에 보상하는 방법과 이것을 적용한 연속음 인식 시스템을 제안하였다. 실험결과 본 논문에서 구현한 발화속도 측정기는 간단한 포만트의 계산만으로 75.2%의 정확도로 평균 음절속도를 측정하였고, 이 정보를 이용하여 보상을 실시한 연속음 인식 시스템의 경우 발화속도를 고려하지 않았을 때와 비교하여 약 6.2%의 인식율 향상이 있었다.

발화속도 측정시 사용되는 LPC계수는 인식과정에서 특징벡터 추출시 필요하므로 추가적인 연산량의 증가없이 계산할 수 있다. 따라서 연속음 인식시스템의 실시간 응용에 이용될 수 있을 것이다.

일반적으로 빠른 발화는 화자의 고유한 성질로 이해되어 왔다. 따라서 발화속도를 정확히 측정할 수 있으면 연속음 인식에서 화자 적응적인 성능 향상을 기대할 수 있다. 뿐만 아니라, 발화속도의 부분적인 변화는 의미의 강조나 감정의 표현같은 중요한 정보를 가지고 있다. 그러므로 이러한 정보는 음성인식 시스템뿐만 아니라 더 나아가서는 언어이해를 위한 시스템에 중요한 정보로 이용될 수 있을 것이다.

향후 과제로는, 보다 정확한 발화속도 측정을 위해서 음소속도 측정기와 이것의 연산량을 감소시키기 위한 연구가 계속되어야 할 것이고, 측정된 발화속도의 다양한 보상방법에 대한 연구가 필요하다. 연속음 인식 시스템의 성능 향상을 위해서는 연속음의 동적 변화를 반영할 수 있는 특징벡터 추출에 대한 연구와, 잡음환경 하에서의 인식방법과 이를 위한 보상방법, 문장모델을 생성하기 위하여 문법의 정확한 정의를 위한 연구가 필요하다.

참 고 문 헌

[1] A. Anastasakos, R. Schwartz, H. Shu, "Duration Modeling in Large Vocabulary Speech Recognition," IEEE Intenational Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 628-631, 1995.

[2] A. V. Oppenheim, R. W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 1989.

[3] B. S. Atal, S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," Jounals of Acoustic Society of America, vol. 50, pp. 637-655, 1971.

[4] D. Burstein, "Robust Parametric Modeling of Durations in Hidden Markov Models," IEEE Intenational Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 548-551, 1995.

[5] D. O'Shaughnessy, "Timing Patterns in Fluent and Disfluent Spontaneous Speech," IEEE Intenational Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 600-603, 1995.

[6] H. R. Kim, K. W. Hwang, Y. M. Ahn, J. H. Ryu, "A Continuous Speech Recognition System UsingFinite State Network and Viterbi Beam Search for Automatic Interpretation," IEEE Intenational Conference on Acoustics, Speech, and Signal Processing, vol 1, pp. 117-120, 1995.

[7] H. Wakita, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," IEEE Transactions on Audio and Electroacoustis, vol. AD-21, no. 5, pp. 417-427, Oct. 1973.

[8] J. D. Markel, A. H. Gray, *Linear Prediction of Speech*, Springer Verlag, New York, 1976.

[9] K. F. Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continous

Speech Recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, no. 4, January 1990.

[10] K. F. Lee, H. W. Hon, R. Reddy, "An Overview of the SPHINX Speech Recognition System," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, no. 1, January 1990.

[11] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood cliffs, N. J., 1978.

[12] L. R. Rabiner, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood cliffs, N. J., 1993.

[13] L. R. Rabiner, "A Tutorial on Hidden Markov Moedls and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, February 1989.

[14] M. A. Siegler, R. M. Stern, "Onthe Effects of Speech Rate in large Vocabulary Speech Recognition Systems," IEEE Intenational Conference on Acoustics, Speech, and Signal Processing, vol 1, pp. 612-615, 1995.

[15] N. Suaudeau, R. Andre-Obrecht, "An Efficient Combination of Acoustic and Supra-Segmental Informations in a Speech Recognition System," IEEE Intenational Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 65-68, 1995.

[16] R. M. Gray, "Vector Quantization," IEEE ASSP Magagine, April 1984.

[17] S. Furui, M. M. Sondi, *Advances in Speech Signal Processing*, Marcel Dekker, Inc., 1992.

[18] 구명완, "음성인식 기술의 현황과 전망," 전자공학회지, 제20권, 5호, pp. 548-556, 1993.

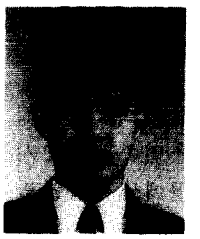
[19] 김경회, 이근배, 이종혁, "한국어 음성 언어 처리를 위한 음소 단위 인식과 형태소 분석의 결합," 정보과학회 논문지, 제 22 권, 10 호(B), pp. 1488-1498, 1995.

[20] 김성겸, *한국어 연속음 인식을 위한 자연어 후처리 시스템의 설계 및 구현*, 인하대학교 공학석사 학위논문, 1995.

[21] 안수길, "한국에서의 음성 신호 처리 기술의 현황과 전망," 제12회 음성 통신 및 신호처리 워크샵논문집, SCAS-12권 1호, pp. 17-23, 1995.

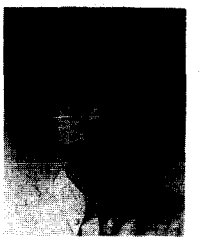
김 재 범

1995년 인하대학교 전자계산공학과 졸업(학사)
1997년 인하대학교 대학원 전자계산공학과(공학석사)
1997년~현재 LG정보통신 중앙연구소 연구원
관심분야:음성신호처리, 자연어처리



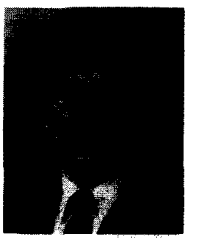
박 찬 규

1996년 인하대학교 전자계산공학과 졸업(학사)
1996년~현재 인하대학교 대학원 전자계산공학과 석사과정
관심분야:음성신호처리, 자연어처리



한 미 성

1996년 인하대학교 전자계산공학과 졸업(학사)
1996년~현재 인하대학교 대학원 전자계산공학과 석사과정
관심분야:정보검색, 자연어처리, 음성신호처리



이 정 현

1977년 인하대학교 전자공학과 졸업(학사)
1980년 인하대학교 대학원 전자공학과(공학석사)
1988년 인하대학교 대학원 전자공학과(공학박사)
1979년~1981년 한국전자기술연구소 시스템 연구원
1984년~1988년 경기대학교 조교수
1989년~현재 인하대학교 전자계산공학과 교수
관심분야:자연어처리, 음성신호처리, HCI