한국어 연속음성 인식을 위한 단어 결합 모델링에 관한 연구

A Study on Word Juncture Modeling for Continuous Speech Recognition of Korean Language

최 인 정*, 은 종 관*

(In Jeong Choi*, Chong Kwan Un*)

요 약

본 논문에서는 단어 조음결합의 음성학적 모델을 이용한 한국어 연속음성 인식에 관해 연구한다. 조음결합 현상에 의한 성능 감소를 줄이기 위해 단어내에서의 전이뿐만 아니라 단어간의 전이를 모델링하는 context-dependent (CD) 단위를 사용한다. 모든 경우에서 각 단어의 첫 음소는 앞에 올 수 있는 모든 단어의 마지막 음소에 의해 지정되며, 각 단어의 마지막 음소도 유사한 방법으로 지정된다. Hidden Markov model(HMM) 파라미터들의 강인성을 개선하기 위해 공분산 행렬을 평활화한다. 또한 음성 단위들 사이의 분별력을 높이기 위해 position-dependent 단위를 사용한다. 실험 결과들은 개선된 조음결합 모델을 사용함으로서 intra-word 단위만을 사용하는 기본 인식 시스템에 비해 성능을 상당히 개선할 수 있음을 보여주었다.

ABSTRACT

In this paper, we study continuous speech recognition of Korean language using acoustic models of word juncture coarticulation. To alleviate the performance degradation due to coarticulation problems, we use context-dependent units that model inter-word transitions in addition to intra-word transitions. In all cases the initial phone of each word has to be specified for each possible final phone of the previous word similarly for the final phone of each word. To improve the robustness of the HMM parameters, the covariance matrix is smoothed. We also use position-dependent units to improve the discriminative power between units. Simulation results show that when the improved models of word juncture coarticulation are used, the recognition performance is considerably improved compared to the baseline system using only intra-word units

I. 서 론

음성의 발음 형태에 따라 음성 인식을 분류하면 크게 고립단어와 연속어 인식으로 나눌 수 있다. 초기의 음성인식 시스템은 대개 고립단어 인식을 그 목표로 하였으나 최근에는 화자독립 연속어 인식 시스템의 개발에 연구의 초점이 맞추어져 있다. 연속음성 인식이 중요한 이유는 인간과 기계사이의 통신 수단으로 음성이 사용될 때, 다양한 응용범위, 원하는 통신 속도 및 자연스러움을 얻으려면 오직 연속어를 통해서만 가능하기 때문이다.

*한국과학기술원 전기및 전자공학과
통신연구실
접수일자 : 1994년 3월 3일

연속음성에서는 단어내의 조음결합뿐만 아니라 낱말 경계에서의 조음결합이 발생하며, 단어의 경계가 명확하지 않으므로 더욱 정확한 모델링이 필요하다. 93년도에 발표된 기본 인식 시스템 [1][2][3]에서는 좌우에 인접한 음소의 정보를 포함하는 triphone을 인식단위로 사용하여 단어내에서 일어나는 조음결합 현상을 모델링하였으며, 단어 경계에서는 인접한 하나의 음소 정보만을 포함하는 diphone을 사용하였다. 본 논문에서는 단어 경계부분의 개선된 조음결합 모델링을 제공하는 단어 결합 모델링에 대하여 다룬다. 단어 결합 모델링은 단어 경계 영역에서의 조음결합 현상을 모델링하기 위하여 단어간에 묵음이 존재할 경우와 그렇지 않은 경우에 대해 inter-word 단위를 사용하여 모두 고려해 주는 방법이다. 또한 본 논문에서는 다음과 같은 두가지 측면에서 단어 결합 모델링을 개선하여 인식 성능을 개선하였다. 첫번째는 단어 결합 모델링의 단점인 인식단위의 학습성 문제를 개선한 방법이다. Inter-word 단위를 사용할 경우 한정된 양의 학습데이타에 비해 학습해야 할 인식단위의 수가 크게 증가되어 HMM 파라미터의 통계적 신뢰성이 문제가 된다. 이 문제를 해결하기 위해 연속 분포 HMM에서 중요한 파라미터인 공분산을 평활화하여 파라미터의 강인성을 향상시킨다. 두번째는 단어 결합 모델링에서의 인식단위간의 분별력을 증가시켜 인식 성능을 개선하는 방법이다. 음소는 단어내에서 차지하는 위치에 따라 스펙트럼 특성의 차이가 크므로 이러한 영향을 반영하기 위하여 intra-word 단위와 inter-word 단위를 독립적으로 학습시킨다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 개선된 방법들과의 비교를 위한 기본 인식 시스템의 개요와 각 부분의 특성을 살펴보고, 3장에서는 단어 결합 모델링과 개선된 방법들에 대하여 다룬다. 4장에서는 컴퓨터 모의실험 결과 및 이에 대한 토의를 하고, 마지막으로 5장에서 결론을 맺는다.

Ⅱ. 기본 인식 시스템

기본 인식 시스템은 93년도에 발표된 100단어 연속음성 인식 시스템으로서 intra-word 단위만을 사용하여 단어 모델을 구성하였다. 발표 당시에는 filter bank 계수를 특징벡터로 사용하여 인식실험을 하였다. 이 논문에서는 켑스트럼 계수를 사용하여 재실험하였으며, 기본 인식 시스템의 구성에 대해 간략히 요약하면 다음과 같다.

음성 신호는 표본화, 잡음 검출, pre-emphasis 처리 등을 포함하는 전처리 과정을 거친다. 다음 과정으로 음성 신호는 학습과 인식에 사용될 수 있도록 특징 벡터의 나열로 변환된다. 본 논문에서 사용된 특징 벡터는 12차의 켑스트럼(cepstrum) 계수와 1, 2차 미분 계수로서, 한 프레임이 36차원의 벡터로 표현되어 진다.

인식 시스템에서 음성 신호를 모델링하기 위해 사용한 방식은 left-to-right 형태의 연속 분포 HMM이다. 기본 인식 시스템에서는 subword 단위로 HMM을 구성하였으며, 각 subword 모델은 3개의 상태로 구성된다. 출력 확률 분포를 추정하기 위해서는 6개의 mixture를 갖는 Gaussian mixture density를 이용하였다.

기본 인식 시스템에서는 CD phone-like unit(PLU)를 인식단위로 사용하였다. CD PLU는 단어내에서 심한 변이성을 보이는 음성을 정확히 묘사하기 위한 수단으로 제안되었다[4][5]. 각각의 context-independent (CI) PLU에 대해 그 단위가 일어날 수 있는 모든 가능한 context를 표현함으로써 CD PLU를 얻을 수 있다. 일반적으로 PLU p에 대한 CD PLU는 p_L-p-p_R로 나타낸다. 여기서 p_L은 바로 이전의 PLU이며, p_R은 바로 다음의 PLU이다. Subword PLU 모델을 학습하기 전에 모든 학습 문장에 대한 발음사전을 구성해야 한다. 기본 인식 시스템에서는 단어의 시작점에서 right-context PLU(예, *$-s-a*)를, 단어의 끝에서는 left-context PLU(예, *i-l-$*)를 사용하였다. 여기서 달러 표시($)는 "don't care"를 나타낸다. Right-context와 left-context PLU를 single-context PLU로 총칭한다. 단어 내부에서는 double-context PLU(예, *s-a-i*)를 사용하여 단어의 발음사전을 구성하였다. 그리고 어떤 두 단어 사이 및 문장의 처음과 끝에서 묵음을 고려하지 않았다. 이러한 방법으로 191개의 CD PLU가 구성되었다.

모든 문장들을 subword 단위들의 나열로 나타낸 후 segmental K-means 알고리즘[6]을 사용하여 학습하였다. 이 알고리즘은 단어나 문장 등의 음성을 미리 지정해준 발음 사전과 함께 입력으로 받아 자동적으로 음성을 분할하여 학습하며, 이러한 과정을 반

학하여 HMM 파라미터를 추정한다.

인식을 위해 사용된 탐색 알고리즘은 one-pass 탐색 알고리즘이다[7]. 전체 탐색공간에 대한 고려로 생길 수 있는 계산량의 비효율성을 개선하기 위해 가능성이 있는 beam size의 경로만을 고려하는 beam 탐색 기법을 채택하였다. 또한 이 시스템에서는 단어의 첨가와 삭제의 균형을 맞추기 위해 word insertion penalty를 사용하였다[4]. 단어간의 천이마다 누적된 likelihood 값에 word insertion penalty를 추가하는 방법을 택하였다.

Ⅲ. 단어 결합 모델링

3.1 단어 결합 모델링

단어사이의 경계에서 일어나는 조음결합 현상은 연속적으로 발음된 단어들의 처음과 끝부분에 대한 음향학적 변이성이 주요한 원인이 되고 있다. 만약 이러한 문맥적 변이성이 인식 시스템에서 충분히 반영되지 못한다면, 이로 인한 오류가 생길 수 있다. 이러한 문제의 한 해결책으로 제시된 것이 단어 경계 영역에 대해 더 정확한 음운학적 표현을 제공하는 단어 결합 모델링 방법이다[8].

많은 연속 음성인식 시스템들이 inter-word CD PLU를 모델링하여 인식에 적용하였다[9][10]. 각 단어의 첫 음소는 이전 단어의 마지막 음소에 대해, 유사하게 각 단어의 마지막 음소는 이어지는 단어의 첫 음소에 대해 지정되어야 한다. 그러므로 일반적으로 주어진 어휘에 대해, inter-word CD PLU로 모델링하지 않았을 경우보다 훨씬 더 많은 모델이 존재하게 된다.

학습 과정을 통하여 신뢰성 있는 모델을 얻기 위하여 각 단위에 대한 충분한 학습데이타가 필요하다. Inter-word CD PLU가 고려될 때 학습 데이타의 부족 현상은 더욱 심각해진다. 이러한 문제를 해결하기 위하여 가장 간단한 방법인 unit reduction rule을 적용할 수 있다. Unit reduction rule은 PLU의 발생 빈도수가 부족할 경우 상세성이 떨어지지만 충분히 학습이 가능한 모델로 대체해서 학습시키는 방법이다[9]. 이 인식 시스템에서는 발생 빈도수의 임계값으로 30을 사용하여 규칙을 적용하였다.

Inter-word CD PLU의 학습에서 가장 큰 변화는 문장의 모델을 구성하는 방법이다. 기본 인식 시스템

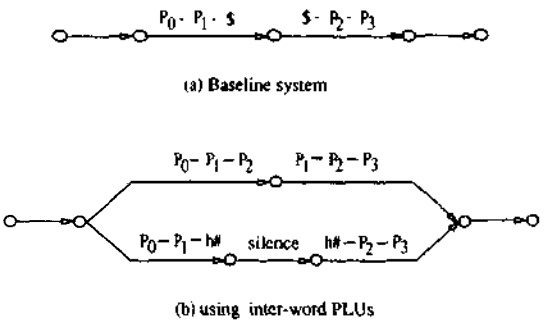

그림 1. 두 단어 사이의 연결 형태.
Fig 1. Connection between two words.

에서는 단어의 양 끝에 single-context PLU가 사용되므로 단어간의 불연속성이 없이 단어 모델을 연결하여 문장의 모델을 구성할 수 있다. 단어결합 모델링에서 단어 경계 부분을 모델링하기 위해서는, 즉 두 단어를 연결하기 위하여 추가되는 경로가 필요하다. 즉, 두 단어사이에 묵음의 존재가 허용되는 경로가 추가된다. 이 경로에서는 묵음의 context가 사용된다. 그림 1은 두 단어 w_1과 w_2 사이의 연결을 보여주고 있다. 그림 1의 (a)는 기본 인식 시스템에서의 단어 연결을, 그리고 (b)는 inter-word CD PLU를 사용할 때의 두 단어 사이의 연결을 나타내고 있다. 여기서 P_0, P_1은 w_1의 마지막 두 음소이고, P_2, P_3는 w_2의 첫번째 두 음소이다. 그리고 h#은 묵음의 PLU에 대한 기호이다.

학습을 하기전에 unit reduction rule을 적용하여 CD PLU를 구성하므로, 단어 경계부분에서는 double-context PLU뿐만 아니라 no-context(예, $-p-$)와 single-context PLU가 존재한다. 그러므로 학습 과정이 복잡하게 될 수 있다. 이 시스템에서는 가장 상세한 PLU만을 사용하여 망을 구성한 후, segmental K-means 알고리즘을 사용하여 학습하였다. 단어 사이의 묵음은 CI PLU로서 하나의 state를 가지는 모델을 사용하였고, 문장의 맨 처음과 끝에서의 묵음은 CD PLU로서 3개의 state를 가지는 모델을 사용하였다.

인식시에는 어떤 단어들이 주어진 단어 앞에 올지, 아니면 뒤따라 올지를 미리 알수가 없다. 그러므로 각 단어들은 모든 가능한 시작과 끝의 segment로 표현되어져야 한다. 이러한 segment들은 double-con-

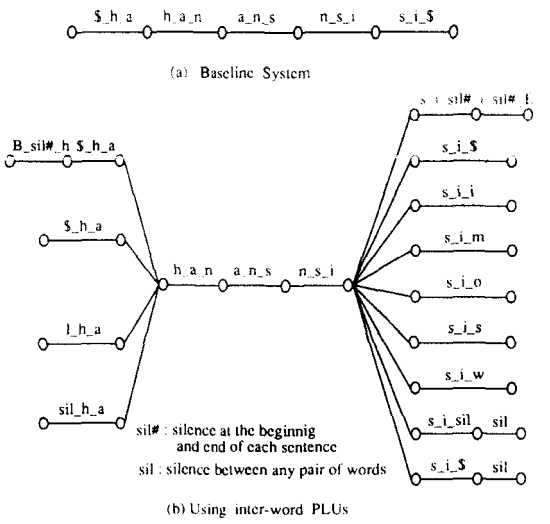

그림 2. CD PLU를 사용한 단어의 음성 표기(/한시/의 경우)
Fig 2. Phonetic transcription of word "hansi", using CD PLUs

text PLU뿐만 아니라 no-context와 single-context PLU를 포함한다. Inter-word CD PLU를 사용하면 단어 사이의 연결이 더 복잡하게 된다. 이것은 각 단어들이 여러개의 시작과 끝 부분을 가지며, 하나의 음소로 구성된 단어들은 예외로 취급해야 하기 때문이다. 두 단어 사이의 묵음은 그 자체로서 한 단어로 표현될 수 있다. 그러나 이것은 두 단어 사이의 문법적 적합성을 검토할 때 단어 사이의 묵음을 무시해야 하므로 문법의 구현을 어렵게 만든다. 이러한 어려움을 피하기 위하여 묵음이 단어들의 마지막 segment로 추가되어진다. 다음과 같은 두가지 형태의 PLU 뒤에 묵음이 따라올 수 있다. 즉, right context로 묵음을 가지는 PLU와 "dont't care"($)를 가지는 PLU들이다. 한 예로서 /한시/에 대한 단어 모델은 그림 2에서와 같이 표현된다. 그림 2(a)는 기본 인식 시스템에서 구성된 단어 모델이며, (b)는 inter-word CD PLU를 사용한 경우에 구성된 단어 모델이다. 그림에서 보듯이 inter-word PLU를 사용할 경우 각 단어에 대한 모델이 매우 복잡해지는 반면, 그만큼 단어 사이의 조음화 현상을 잘 모델링하므로 인식 성능을 개선할 수 있다.

3.2 HMM 모델의 학습성 개선

한정된 양의 학습 데이타를 사용하여 HMM 파라미터를 추정할 때 통계적인 변동성의 차이 문제가 발생할 수 있다. 특히 inter-word CD PLU와 같은 상세 [illegible] 사용할 때에 학습 모델에서는 파라미터들이 정확 추정되어질 가능성이 더욱 높아진다. 본 논문에서는, HMM 모델의 강인성을 개선하기 위하여 연속 분포 HMM에서 중요한 파라미터인 공분산을 평활화하였다.

출력 확률 분포를 추정하기 위하여 큰 수의 mixture를 사용할 경우 어떤 cluster에는 작은 수의 표본 데이타가 할당되어 진다. 이러한 경우 신뢰성있게 모델 파라미터를 추정하기 어려우며, 특히 공분산 값들은 0에 가깝게 추정되어진다. Gaussian 밀도의 피크 값은 공분산 행렬의 determinant의 제곱근에 역비례하므로 매우 가파른 피크를 갖는 Gaussian 분포가 된다. 한 mixture 성분의 가파름 정도를 정량화하기 위하여, mixture Gaussian 밀도의 평균 피크 값에 한 그 성분 밀도의 피크 값의 비가 사용된다. M개의 mixture 성분을 사용하여 출력 확률 분포를 추정할 때 i번째 cluster에서의 공분산 행렬을 Σ_i라 하면, i번째 성분의 상대적인 첨예도 R_i는 식 (1)과 같이 주어진다.

$$R_i = \frac{1/|\Sigma_i|^{\frac{1}{2}}}{1/\left(\prod_{k=1}^{M}|\Sigma_k|^{\frac{1}{2}}\right)^{\frac{1}{M}}} \quad (1)$$

만일 R_i가 주어진 임계값 이상이면, 그 성분은 첨예 정도가 심한 것으로 본다. 이러한 첨예한 피크를 갖는 분포를 평활화하기 위하여 하나의 mixture 성분을 사용하여 추정한 파라미터가 사용된다. 하나의 mixture 성분을 사용하여 추정된 공분산 행렬 Σ는 식 (2)와 같이 얻어질 수 있다.

$$\Sigma = \sum_{i=1}^{M} c_i \Sigma_i + \sum_{i=1}^{M} \sum_{j=i+1}^{M} c_i c_j (\mu_i - \mu_j)(\mu_i - \mu_j)' \quad (2)$$

여기서 c_i와 μ_i는 각각 i번째 cluster의 표본 집합의 평균 벡터와 가중치이며, $'$는 행렬의 transpose를 나타낸다. 평활화는 Σ_i와 Σ의 선형 보간으로 정의된다[11]. 평활화된 공분산 행렬을 $\hat{\Sigma}_i$라 하면 아래의 식 (3)과 같이 얻어진다.

$$\hat{\Sigma}_i = \lambda \Sigma_i + (1-\lambda)\Sigma \quad (3)$$

여기서 λ는 보간 파라미터이며 0과 1 사이의 값을 갖는다. 일반적으로 학습 데이타의 양과 인식 단위의 다양성에 따라 적절한 λ 값을 선택할 수 있다. 또한 멕스트럼 계수와 1, 2차 미분 멕스트럼 계수의 공분산 행렬을 분리하여 정확화 할 수 있다. 즉 멕스트럼 계수의 공분산 행렬은 균일하게 정확화하고, 미분 멕스트럼 계수의 공분산 행렬은 첨에도 R_i와 비교하여 선택적으로 정확화한다.

3.3 Position-dependent PLU

인식 성능의 향상을 위해 모델의 상세성을 개선하는 방법이 있다. 단어 결합 모델링에서는 context에 관계한 기본적인 음성 단위를 사용하였다. 그러나 단어내에 존재하는 PLU의 스펙트럼 특성과 단어 경계에 나타나는 PLU의 특성이 다를 수 있다. 심지어 같은 context를 가지는 PLU 사이에도 인식 단위가 나타나는 위치에 따라 특성의 차이가 생긴다. 따라서 본 논문에서는 intra-word PLU와 inter-word PLU를 독립적으로 모델링하는 방법을 사용하였다. 이렇게 얻어진 것을 position-dependent PLU [9]라 한다. 이 시스템에서는 전체 학습 데이타에 대해 342개의 PLU가 얻어지며, 이 중에서 163개는 inter-word 단위고 나머지는 intra-word 단위이다. 분할과 인식을 위해 망을 구성하는 방법은 앞에서 서술한 방법과 같으나, 단어 경계에서는 inter-word 단위만 나타나고 단어 내부에서는 intra-word 단위만 나타날 수 있다는 점이 다르다.

Ⅳ. 인식 실험 및 결과

4.1 데이타베이스

인식 대상이 되는 문장은 날짜, 요일, 시간과 관련된 단어로 구성되어 있는 시간구이다. 각 문장은 그림 3의 finite state network(FSN)에 근거하여 작성되었는데, 예를들면 "월요일 오후 세시 삼십분" 또는 "오월 십일 열시 오분"등과 같은 것이 있다.

모두 102개의 어휘로 구성되어 있으며, 20대 남, 녀 화자 90명이 자연스럽게 발음한 다양한 문장을 녹음하여 학습 및 인식에 사용하였다. 녹음은 무한반향의 방송국에서 하였으며, 사용된 A/D 변환기의 resolution은 16 bit이고, sampling rate은 16 kHz였다. 학습에는 남자 33명, 여자 37명의 데이타를 임의로 선정하여 사용하였으며, 인식 실험에는 학습에 참가하지 않은 남, 녀 각각 10명이 발음한 139개의 연속어 문장을 이용하였다.

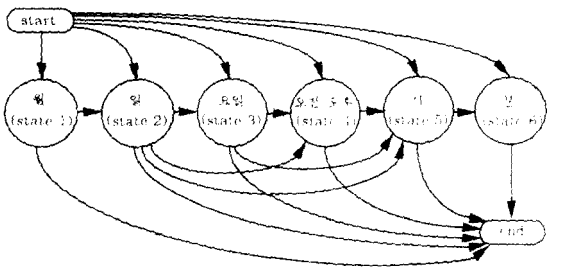

그림 3. Finite State Network을 이용한 문법.
Fig 3. Grammar using Finite State Network

이 연구에서는 문법을 적용하지 않은 경우와 FSN의 문법을 적용한 경우 각각에 대하여 인식 실험을 하였다. 문법의 결합은 연속어 인식에서는 필수적이라 할 수 있는데, 인식 대상의 문법적 어려움의 척도로는 perplexity라는 정의를 사용한다 [12]. Perplexity는 정성적으로 임의의 단어뒤에 올 수 있는 평균적인 후보 단어의 수라고 할 수 있으며, 문법이 적용되지 않을 경우는 인식 어휘의 수가 perplexity가 된다. 본 시스템의 경우 문법이 적용되지 않은 경우는 perplexity가 102이고, FSN의 문법을 적용한 경우의 perplexity는 약 26이다.

4.2 실험 결과 및 고찰

연속어의 인식률은 단어 단위의 인식률과 문장 단위의 인식률로 나타내며, 단어 단위의 인식률은 아래와 같이 주어진다.

단어인식률 =

$$\frac{\text{총단어수} - (\text{치환단어} + \text{삭제단어} + \text{첨가단어})}{\text{총단어수}} * 100(\%)$$

인식 실험에서 계산량을 줄이기 위하여 beam search 기법을 적용하였다. Beam size를 150으로 하였으며, 각 시간마다 활성 노드의 수에 따라 beam size를 변화시키는 방법을 사용하였다. 또한 첨가와 삭제에러의 균형을 맞추기 위하여 word insertion penalty를

사용하였다. 인식 실험을 통해 50에서 100사이의 word penalty 값을 설정하여 사용하였다.

표 1은 CI PLU로 모델링한 경우와 CD PLU로 모델링한 기본 인식 시스템에 대한 인식 실험 결과를 보여주고 있다. 이 실험에서는 word penalty 값을 사용하지 않았다. 문법을 적용한 경우 CI PLU로 단어 모델을 구성했을때 83.63%의 단어 인식률을 얻었으며, CD PLU로 단어 모델을 구성했을때의 단어 인식률은 90.75%이었다. 결과에서 보듯이 좌우의 context 정보를 포함하는 상세한 인식단위를 사용함으로서 단어내에서 발생하는 조음결합 현상을 잘 모델링할 수 있다.

다음은 단어 결합 모델링을 이용한 인식 실험을 하였다. 표 2는 inter-word PLU를 이용하여 얻은 결과와 기본 시스템에서의 결과를 비교하고 있다. 이 실험에서 사용된 PLU 목록을 얻기 위하여 30의 임계값을 사용하여 unit reduction rule을 적용하였으며, word penalty 값으로 100을 사용하였다. Intra-word CD PLU를 사용하여 90.61%의 단어 인식률을 얻었으며, intra-word 단위와 inter-word 단위를 모두 사용하여 모델링하였을 때는 91.78%의 단어 인식률을 얻었다. 인식 결과를 통해 확인할 수 있는 사실은 inter-word 단위를 사용하여 단어 경계부분에 대해 더 상세히 모델링 함으로서 인식 시간이 증가한다는 단점이 있으나 오인식률이 약 10%정도 개선되었다.

한정된 양의 학습 데이타, 상세한 인식 단위, 큰 수의 mixture 성분 사용 등으로 인하여 모델의 파라미터들이 부족 추정될 가능성이 있다. 이러한 문제를 보완하기 위하여 공분산 행렬의 값을 평활화하는 방법을 사용하였다. 이러한 방법의 영향을 확인하기 위하여 다음과 같은 세가지 경우에 대하여 인식 실험을 하였다.

Case 1: 모든 공분산 행렬은 $\lambda=0.5$의 보간 파라미터를 사용하여 평활화한다.

Case 2: 먼저 모든 공분산 행렬에 대해 임계치 100을 사용하여 첨예도 K와 비교한 후, 임계치 이상의 공분산 행렬에 대해서만 $\lambda=0.5$를 사용하여 평활화한다.

Case 3: *Case 2*에서 켑스트럼 계수의 공분산 행렬과 1, 2차 미분 켑스트럼 계수의 공분산 행렬을 분리하여 평활화한다. 즉 켑스트럼 계수의 공분산 행렬은 균일하게 평활화되고, 미분 계수의 공분산 행렬은 첨예도와 비교하여 선택적으로 평활화된다.

각 실험에 대한 실험 결과는 표 3에 주어진다. 실험 결과로부터 첨예도에 의한 선택적 공분산 파라미터에 대한 평활화 방법을 적용함으로서 인식 성능을 상당히 개선할 수 있었다. 특히 환경에 둔감한 동적 특성 벡터에 비해 화자와 문맥에 따라 큰 변화를 보여주는 순시적인 특성 벡터에 대하여 더 심하게 평활화함으로서 93.78%의 단어 인식률을 얻었으며, 이는 기본 인식 시스템에 비해 단어 오인식률이 34%나 개선된 것이다.

표 4는 intra-word 단위와 inter-word 단위를 독립적으로 학습시켜 얻어진 position-dependent PLU를 이용한 인식 결과를 보여주고 있다. FSN의 문법을

표 1. 기본 인식 시스템에서의 인식 실험 결과
Table 1. Test Results in Baseline System

항 목	CI PLU		intra-word CD PLU	
	No Grammar	FSN	No Grammar	FSN
대치 에러(%)	18.58	15.12	11.46	8.97
삭제 에러(%)	0.35	0.35	0.07	0.00
첨가 에러(%)	4.21	0.90	5.32	0.28
단어 인식률(%)	76.86	83.63	83.08	90.75
문장 인식률(%)	51.94	60.36	62.19	72.20

표 2. inter-word CD PLU를 이용한 인식 결과
Table 2. Test Results Using Inter-word CD PLUs

항 목	CI PLU	intra-word CD PLU	inter-word CD PLU
단어 인식률(%)	84.18	90.16	91.78
문장 인식률(%)	60.82	71.98	75.40

표 3. 공분산 평활화에 의한 인식 결과
Table 3. Test Results Using Covariance Smoothing

항 목	*Case 1*	*Case 2*	*Case 3*
단어 인식률(%)	92.96	93.23	93.78
문장 인식률(%)	77.68	78.59	80.41

표 4. position-dependent PLU를 이용한 인식 결과
Table 4. Test Results Using Position-dependent PLUs

항 목	Position-dependent PLUs
단어 인식률(%)	95.03
문장 인식률(%)	84.74

식할 때의 95.03%의 단어 인식률을 얻었다. PLU가 단어에서 차지하는 위치에 따라 다르게 나타나는 스펙트럼 특성의 차이를 반영하여 인식단위간의 분별력을 크게 함으로서 기본 인식 시스템에 비해 47%, 단어 결합 모델링에 비해 40%의 단어 오인식률을 개선하였다. 단어 내부와 단어 경계에 같은 CD PLU가 많이 존재할 경우에 그들 사이의 분별을 위해 독립적으로 모델링하는 것이 효과적임을 볼 수 있다.

V. 결 론

본 논문에서는 연속음성 인식에서 중요한 문제인 단어 내부뿐만 아니라 단어 사이에서 일어나는 조음결합 현상을 효과적으로 모델링하기 위한 단어 결합 모델링에 관해 연구하였다. 단어 내에서 일어나는 조음결합 현상은 CD PLU를 인식 단위로 사용함으로서 효과적으로 모델링될 수 있다. 단어 경계 영역에서 일어나는 조음결합 현상은 inter-word CD PLU를 사용하여 모델링하였다. Intra-word 단위와 inter-word 단위 모두를 사용하여 단어 모델을 구성한 인식 시스템은 intra-word 단위만을 사용하여 모델링한 기본 인식 시스템보다 10% 정도의 단어 오인식률을 개선하였다. 또한 다음과 같은 두가지 방법을 사용하여 단어 결합 모델링을 개선함으로서 인식 성능을 향상시켰다. 먼저 한정된 학습 데이타, 상세한 인식 단위의 사용 등으로 인한 모델 파라미터의 통계적 신뢰성 결여 문제를 보완하기 위하여 공분산 행렬을 평활화하였다. 이 방법에 의해 모델 파라미터가 부족 추정되는 것을 방지함으로서 기본 인식 시스템에 비해 34%의 오인식률을 개선하였다. 마지막으로 inter-word CD PLU와 intra-word CD PLU를 독립적으로 모델링하여 구성한 position-dependent CD PLU를 사용하였다. 이 방법은 subword 모델이 단어의 어느 부분에 위치하느냐에 따라 나타나는 스펙트럼 특성의 차이를 반영해 주는 것으로서 95.03%의 단어 인식률을 보였다. 현재는 1000 단어 이상의 인식 시스템에 대해 인식 시간의 큰 증가없이 단어 결합 모델링의 효과를 살릴수 있는 모델링 방법에 대해 연구되고 있다.

참 고 문 헌

1. 은 종관 외, 연속 음성 인식 시스템 개발연구, 한국과학기술원, 1992.
2. 김 도영 외, "한국어 연속 음성 인식 시스템 개발," 제10회 음성통신 및 신호처리 워크샵 논문집(제 SCAS 10권 1호), 1993.
3. 김 도영 외, "연속분포 HMM을 이용한 한국어 연속 음성 인식 시스템 개발," 한국음향학회지, Vol. 13, No. 1, pp. 24-31, 1994.
4. C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, "Acoustic Modeling for Large Vocabulary Speech Recognition," *Computer Speech and Language*, Vol. 4, No. 2, pp. 127-165, Apr. 1990.
5. R. Schwartz, Y. L. Chow, O. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context Dependent Modeling for Acoustic-phonetic Recognition of Continuous Speech," *Proc. of ICASSP*, pp. 1205-1208, Tampa, FL, U.S.A., Mar. 1985.
6. L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A Segmental K-Means Training Procedure for Speech Recognition," *IEEE Trans. on ASSP*, Vol. 38, No. 12, pp. 2033-2045, Dec. 1990.
7. H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," *IEEE Trans. on ASSP*, Vol. 32, No. 2, pp. 263-271, Apr. 1989.
8. E. P. Giachin, C. H. Lee, L. R. Rabiner, A. E. Rosenberg, and R. Pieraccini, "On the Use of Inter-word Context-dependent Units for Word Juncture Modeling," *Computer Speech and Language*, Vol. 6, No. 3, pp. 197-213, Apr. 1992.
9. C. H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini, and A. E. Rosenberg, "Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition," *Computer Speech and Language*, Vol. 6, No. 2, pp. 103-107, Apr. 1992.
10. D. B. Paul, "The Lincoln Robust Continuous Speech Recognition," *Proc. of ICASSP*, pp. 449-452, Glasgow, Scotland, May 1989.
11. Y. Zhao, "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units," *IEEE Trans. on ASSP*, Vol. 1, No. 3, pp. 345-361, July 1993.
12. F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of IEEE*, Vol. 64, No. 4, pp. 532-556, Apr. 1976.

▲최 인 정(In Jeong Choi) 1969년 1월 14일생

1991년 2월 : 한양대학교 전자공학과 졸업(공학사)

1994년 2월 : 한국과학기술원 전기 및 전자공학과 졸업(공학석사)

1994년 3월 ~ 현재 : 한국과학기술원 전기 및 전자공학과 박사과정

▲온 종 관(Chong Kwan Un) : 10권 3호 참조