분절 특징 HMM을 이용한 영어 음소 인식
(English Phoneme Recognition using Segmental-Feature HMM)

윤 영 선 †
(Young-Sun Yun)

요 약 본 논문에서는 여러 프레임 특징으로 표현되는 분절 특징(segmental feature) 표현 방법을 제안하고, HMM 개념 위에서 음향학적 모델과 그 알고리즘을 개발하여 HMM의 약점으로 지적되는 독립 관측 가정을 완화시키고자 한다. 제안된 특징 표현은 단일 프레임 특징이 음성 신호의 시간적 동적 특성(temporal dynamics)을 제대로 표현하지 못하기 때문에, 여러 프레임을 이용하여 음성 특징을 표현하도록 한다. 분절 특징은 다항식의 회귀 함수(polynomial regression function)에 의하여 관측 벡터의 궤적으로 표현되고, 이 특징을 패턴 분류에 사용하기 위하여 음성 신호의 궤적을 효과적으로 표현하는 분절 HMM (segmental HMM)을 이용한다. SHMM은 상태에서의 관측 확률을 외적 분절 변이와 내적 분절 변이로 세분하며, 외적 분절 변이는 장기적인 변화를, 내적 분절 변이는 단기적인 변화를 나타낸다. 음향학적 모델에서 분절 특성을 고려하기 위하여 외적 분절 변이는 분절의 확률 분포로 표현하고, 내적 분절 변이는 궤적의 추정 오차로 표현하도록 SHMM을 수정한 분절 특징 HMM(SFHMM; segmental-feature HMM)을 제안한다. SFHMM에서는 분절의 관측 확률을 분절 우도와 궤적의 추정 오차의 관계로써 표현하며, 추정 오차는 특정 상태에서의 분절의 우도에 대한 가중치로 고려될 수 있다. 제안된 방법의 유효성과 분절 특징의 특성을 살펴보기 위하여 TIMIT 자료를 이용하여 몇 가지 실험을 하였다. 이들 실험 결과에서, 제안된 방법이 기존의 HMM보다 매개 변수가 많더라도, 성능의 향상과 제안된 특징이 유연하고 정보를 많이 가진다는 점에서 의미가 있다고 하겠다.

키워드 : 음성 인식, 분절 특징, 분절 HMM, 분절 특징 HMM

Abstract In this paper, we propose a new acoustic model for characterizing segmental features and an algorithm based upon a general framework of hidden Markov models (HMMs) in order to compensate the weakness of HMM assumptions. The segmental features are represented as a trajectory of observed vector sequences by a polynomial regression function because the single frame feature cannot represent the temporal dynamics of speech signals effectively. To apply the segmental features to pattern classification, we adopted segmental HMM(SHMM) which is known as the effective method to represent the trend of speech signals. SHMM separates observation probability of the given state into extra- and intra-segmental variations that show the long-term and short-term variabilities, respectively. To consider the segmental characteristics in acoustic model, we present segmental-feature HMM(SFHMM) by modifying the SHMM. The SFHMM therefore represents the external- and internal-variation as the observation probability of the trajectory in a given state and trajectory estimation error for the given segment, respectively. We conducted several experiments on the TIMIT database to establish the effectiveness of the proposed method and the characteristics of the segmental features. From the experimental results, we conclude that the proposed method is valuable, if its number of parameters is greater than that of conventional HMM, in the flexible and informative feature representation and the performance improvement.

Key words : Speech Recognition, Segmental Feature, Segmental HMM, Segmental-feature HMM.

1. 서 론

1960년대 이후로 꾸준히 연구되고 있는 은닉 마코프 모델(HMM; hidden Markov model)은 시간적·공간적 변이를 잘 반영하는 이중 통계적 방법으로 음성 인식을 포함한 여러 분야에 널리 사용되고 있으며, 그 우수성은 잘 알려져 있다. 그러나 HMM이 음향학적인 음성 신호의 통계적 변이를 잘 모델링 한다고 할지라도, HMM을

† 종신회원 : 한남대학교 정보통신·멀티미디어공학부 교수
ysyun@mail.hannam.ac.kr
논문접수 : 2001년 6월 20일
심사완료 : 2001년 10월 16일

이용한 음성 인식 시스템의 성패여부는 인식 모델이 음성 신호를 얼마나 잘 모델링 하느냐에 달려 있다. 따라서 HMM이 효과적으로 음성 신호의 음향학-음소적 특징을 잘 반영하도록 하기위해서는 그 기본이 되는 이론적 배경에 대한 고찰이 필요하다. 일반적으로 HMM은 다음의 세 가지 가정에 기반하여 음성 신호를 모델링하고 있다[1].

- 부분적인 안정 상태(piecewise stationarity) : 음성 패턴은 부분적으로 안정적인 상태(stationary state)와 그 상태들 간의 순간적인 전이(instantaneous transition)로 구성되어 있다.
- 독립 가정(independence assumption) 또는 상태의 조건부 안정성 가정 : HMM의 특정 상태에 대한 특징 벡터의 관측 확률은 그 상태와 관측된 특징 벡터에 종속적이며, 이전이나 이후의 상태와 관측 벡터에는 독립적이라고 가정한다. 이것은 각 상태가 독립적이며 균등하게 분포(i.i.d.; independent and identically distributed)된 관측 벡터를 생성하는 안정적인 근원(stationary source)이라는 것을 의미한다. 이 가정은 상태가 공명음(sonorant)이거나 마찰음(fricative)과 같이 짧은 음을 표현하는 경우에는 어느 정도 합당하나 파열음(plosive)과 같이 긴 음을 표현하는 경우에는 인접한 관측 벡터들끼리 높은 상관관계를 가지고 있기 때문에 적합하지 않다고 알려져 있다[2].
- 상태에서의 지속 분포(state duration distribution) : 앞에서 언급한 독립 가정 때문에 여러 프레임이 한 상태에서 머무를 관측 확률은 오직 재귀 전이(self-loop transition) 확률에 의해 결정된다. 따라서 상태에서 머무를 시간이 오래되면 지속 확률은 기하급수적으로 감소하게 되며 올바르게 표현되지 못한다.

이러한 가정들은 HMM의 구현을 단순하게 하며, 처리 시간을 단축시키고 인식 성능을 어느 정도 향상시키는 효과가 있으나 처리 과정을 너무 단순화시켰기 때문에 실제 음성 신호의 표현에는 미흡하다고 보고되었다. 이중에서 독립 관측 가정을 손쉽게 보완하는 방법의 하나로써 여러 프레임에 걸친 시간적인 특성을 포함하도록 음향학적 특징을 개선하는 방법이 제안되었다[3]. 이 방법은 정적인 특징(stationary feature) 의 1차 또는 2차 미분(derivative)을 이용하여 동적인 특징(dynamic feature)을 표현하도록 하였다. 그러나 모델의 구조나 특성을 변화시키지 않고 데이타만을 수정하여 음성 신호를 표현하는 방법보다는 모델을 변화시켜 동적인 음성 신호에 적합하도록 하는 방법이 더욱 효과적이다. 그 이유는 HMM의 기본 가정이 프레임 단위의 특징에 기반하고 있으며 각 상태에서의 관측 벡터는 한 프레임만을 표현하고 있기 때문이다. 이 분석을 바탕으로 음성 특징의 표현에 단일 프레임 특징만을 이용하는 것보다는 프레임 특징들의 집합을 이용하고자 하는 연구가 진행되었다. 이러한 연구로는 HMM의 상태에서의 지속 시간을 다항식을 이용한 회귀 함수 (polynomial regression function)로 표현하거나, 프레임 특징(frame feature) 대신 분절 특징(segmental feature)을 이용하여 독립 관측 가정을 완화시키고자 하는 연구가 있다. 경향 HMM(trend HMM)이라고 불리는 HMM의 상태 모델링 연구는 1992년 Deng에 의하여 제안되었으며, 각 상태에서 관측 가능한 벡터를 지속 시간에 의한 회귀 함수로 예측하였다. 즉, 상태에서의 관측 벡터가 바로 이전의 관측과 상관관계가 있기 때문에, 음성이 관측되는 절대적인 시간[4] 또는 그 상태에서 머무는 시간[2]의 함수에 의해 그 변화량을 예측할 수 있다는 것이다. 이들 방법이 상태에서의 관측 벡터를 효과적으로 모델링하더라도 프레임 특징을 사용하고 있기 때문에 여전히 독립 관측 가정에서 벗어나지 못하고 있으며, 음성의 변이를 상태 수로 제약하고 있다. 분절 특징을 이용한 대표적인 연구로는 Gish가 발표한 다항식에 의한 분절 모델링(모수적 궤적 모델; parametric trajectory model)[5], Russell과 Gales가 독립적으로 연구한 분절 HMM(segmental HMM)[6,7], Ostendorf에 의한 분절 모델(segmental model)[8] 등이 있으며, 1996년에 Ostendorf 등에 의하여 포괄적인 정리가 이루어졌다[9]. 또한 국내에서도 음향학적 문맥 정보를 고려하여 성능을 향상시키고자 하는 연구가 진행되고 있다[14]. 그러나 기존 방법은 알고리즘이 복잡하여 변수의 추정이나 평가단계에서 계산 시간이 많이 걸리거나, 상태에서의 관측 확률 분포에 대한 특정 가정을 기반으로 하고 있다. 또한, 음소 단위의 전체 음향학적 벡터를 정규화하는 과정이 필요하여 연속 음성 인식에 부적합한 면도 있다. 이러한 문제점들을 해결하기 위하여 인식 모델은 HMM에 기반을 두고 분절 특징을 입력으로 하는 분절 특징 HMM (SFHMM; segmental-feature HMM)을 제안한다.

SFHMM은 분절의 특징을 다항식에 의한 궤적(trajectory)으로 표현하고, 추정된 궤적을 분절 HMM의 입력으로 사용한다. 각 궤적은 분절을 구성하는 프레임들 간의 인접 정보를 포함할 수 있도록 디자인 행렬을 개선하여 계산된다. 인접하는 프레임 정보를 구하면서 인식 단계에서의 정렬 문제를 해결하도록 개선된 디자인 행렬은 현재의 관측 벡터에 대해 대칭적이 되도록 하였으며, 분절 길이(segment length)와 무관하게 가변 분절을 모델링할 수 있도록 고려하였다. 기존의 연구들에서는 분절 길이가 변화될 수 있어 추정 과정이나 인식 과

정에서 많은 처리 시간을 필요로 하나, 본 연구에서는 실 응용에 적용할 수 있도록 분절의 길이와 궤적의 표현 차수(회귀차수; regression order)를 조정할 수 있게 하였다. 이들 값들은 학습 단계에서 알맞은 값으로 고정이 되며, 표현력을 결정하게 된다. SFHMM을 이용한 음성 인식 시스템은 프레임 특징을 이용하는 기존의 HMM의 입력 특징을 분절 특징으로 대치하고 상태에서의 관측 확률 값을 수정하여 사용하기때문에 기존 시스템의 인식 단위(recognition unit)와 동일한 형태의 인식 단위를 그대로 사용할 수 있다. 따라서 SFHMM의 변수 추정(parameter estimation)과 평가(evaluation) 문제는 일반 HMM의 추정·평가 문제와 유사하게 된다.

본 논문의 구성은 다음과 같다. 2장에서는 모수적 궤적 방법을 이용한 분절 모델링(segment modeling)과 분절 특징에 대해 소개하며, 3장에서는 분절 특징 HMM을 제안한다. 분절 특징 HMM은 기존의 분절 HMM에서 우도(likelihood)와 변수 추정 방법을 개선하였으며 분산의 다른 표현 방법을 고려한다. 4장에서는 제안된 방법의 유효성을 검증하기 위하여 TIMIT 자료를 이용하여 수행한 음소 인식 실험과 모음 분류 실험을 기술하며, 결과 및 특성에 대해 토의한다. 마지막으로 5장에서는 본 연구를 요약하고 결론을 맺는다.

2. 분절 모델링

음성 신호의 연속적인 음향 특징 벡터들의 관계는 특징 공간에서 궤적(trajectory)의 형태로 근사될 수 있다는 기본적인 생각에서 출발한 분절 모델링은 구현 방법에 따라 모수적(parametric) 또는 비모수적(non-parametric) 방식으로 분류된다. 모수적 방법에서는 특정 영역에 대하여 다항식을 이용하여 궤적을 추정하고, 그 영역에 대한 분포를 궤적 위의 점들로 표현한다. 반면, 비모수적 모델은 각각의 모델 영역에 대한 분포 변수들을 갖는다. 본 논문에서는 모수적 방법이 여러 음성 단위에서 궤적의 평활화 효과가 있어 잡음이나 환경 변화, 화자 변화에 강인한 성질을 보이기 때문에[9], 분절 모델링에 모수적 방법을 채택하였다.

2.1 모수적 모델링

모수적 궤적 방법에 대한 연구에서, Gish와 Ng는 1993년 지속 구간 프레임을 갖는 음성 분절 $\mathbf{C}$를 다음과 같이 표현하였다.

$$\mathbf{C}=\mathbf{Z}\mathbf{B}+\mathbf{E} \tag{1}$$

여기에서 각 프레임은 D차원의 특징 벡터로 구성되며, $\mathbf{Z}$는 분석에 사용된 음성 신호의 형태를 결정하는 $N\times R$ 크기의 행렬이다. 이 행렬은 관측하고자 하는 음성 분절의 시간적인 특성을 표현하며 디자인 행렬로 불린다. 또한, $\mathbf{B}$는 궤적 계수를 나타내는 $R\times D$행렬을 나타내며, $\mathbf{E}$는 궤적 추정 단계에서의 잔차 오차(residual error)를 표현한다. R은 궤적의 특징을 결정하는 회귀 차수(regression order)로써, 만약 $R=1$이면 평균(상수)을, $R=2$이면 일차 선형 시스템, $R=3$이면 2차 방정식의 궤적을 표현하며, R이 증가함에 따라 분절의 특징을 세밀하게 표현할 수 있다.

기존 연구에서는 음성 분절을 [0..1]로 정규화된 하나의 완전 궤적(complete trajectory)으로 표현하고 있기 때문에 분절의 경계를 반드시 알고 있어야 한다. 이러한 문제점을 해소하기 위하여, 본 논문에서는 입력 음성을 고정 길이의 분절로 분할하고 현재 관측 벡터가 분절의 중앙에 위치하도록 한다. 현재 관측 벡터를 중앙에 위치하기 때문에 구해진 분절 특징은 일반 프레임 특징과 마찬가지로 경계 문제를 고려하지 않아도 된다.

시간적인 특성을 표현하는 디자인 행렬에서 절대적인 시간 개념이 아니라 정규화된 시간 개념을 나타내도록 하면, 각 분절의 특징은 분절 길이에 무관하게 상대적인 시간을 표현할 수 있다. 또한 경계 문제를 해결하기 위하여 현재의 관측 벡터가 분절의 중앙에 위치하기 때문에 현재 시간을 0으로 표현하고, 이전 시간은 음(−)의 시간으로 표현하고 이후의 시간은 양(+)의 시간으로 표현하여 시간적인 특성을 반영하고자 한다.

위의 개념을 정리하여 모수적 궤적 모델에 적용하면 식 (1)은 다음식과 같이 변경된다. 초기의 모수적 궤적 모델은 음성 분절을 하나의 완전 궤적으로 표현하였기 때문에 시간 개념이 없었으나, 본 논문에서는 각각의 분절 특징들이 모여 하나의 음성 특징을 나타내기 때문에 시간 개념으로 표현된다.

$$\mathbf{C}_t=\mathbf{Z}\mathbf{B}_t+\mathbf{E}_t \tag{2}$$

여기에서 음성 분절 $\mathbf{C}_t$는 $N=2M+1$인 프레임을 분절 길이로 갖으며, 다음과 같이 관측 벡터들을 행렬로써 표현할 수 있다.

$$\mathbf{C}_t=\mathbf{Y}_{t-M}^{t+M}=\begin{bmatrix} y_{t-M,1} & \Lambda & y_{t-M,D} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} \\ y_{t,1} & \Lambda & y_{t,D} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} \\ y_{t+M,1} & \Lambda & y_{t+M,D} \end{bmatrix}=\begin{bmatrix} \mathbf{c}_{t-M} \\ \mathrm{M} \\ \mathbf{c}_t \\ \mathrm{M} \\ \mathbf{c}_{t+M} \end{bmatrix} \tag{3}$$

$$\mathbf{c}_\tau=\begin{bmatrix} y_{\tau,1} & \Lambda & y_{\tau,D} \end{bmatrix},\quad t-M\le\tau\le t+M$$

위 식에서 볼 수 있듯이 시간 t에서 현재 관측 벡터가 분절의 중앙에 오기 때문에 음성 분절의 앞부분과 뒷부

분은 시간 $t-1$이나 $t+1$의 음성 분절과 겹치게 된다. 따라서 위와 같은 분절을 분석하기 위한 새로운 디자인 행렬의 구상이 필요하다. 새로운 디자인 행렬은 인접한 분절 간의 전이 정보(transitional information)를 표현할 수 있으며, 또한 현재 관측 벡터가 중앙에 오도록 배치할 수 있어야 한다. 위의 조건을 만족하도록 디자인 행렬 Z는 다음과 같이 정의된다.

$$\mathbf{Z}=\begin{bmatrix} 1 & \left(-\frac{M}{2M}\right) & \left(-\frac{M}{2M}\right)^2 & \Lambda & \left(-\frac{M}{2M}\right)^{R-1} \\ 1 & \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ 1 & \left(-\frac{m}{2M}\right) & \left(-\frac{m}{2M}\right)^2 & \Lambda & \left(-\frac{m}{2M}\right)^{R-1} \\ 1 & \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ 1 & 0 & 0 & 0 & 0 \\ 1 & \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ 1 & \left(\frac{m}{2M}\right) & \left(\frac{m}{2M}\right)^2 & \Lambda & \left(\frac{m}{2M}\right)^{R-1} \\ 1 & \mathrm{M} & \mathrm{M} & & \mathrm{M} \\ 1 & \left(\frac{M}{2M}\right) & \left(\frac{M}{2M}\right)^2 & \Lambda & \left(\frac{M}{2M}\right)^{R-1} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_{t-M} \\ \mathrm{M} \\ \mathbf{z}_{t-m} \\ \mathrm{M} \\ \mathbf{z}_t \\ \mathrm{M} \\ \mathbf{z}_{t+m} \\ \mathrm{M} \\ \mathbf{z}_{t+M} \end{bmatrix} \quad (4)$$

여기에서 Z는 분절 길이로 정규화 된 상대적인 위치를 나타내기 때문에, 현재 관측 벡터의 앞부분 또는 다음에 오는 음향학적 특징을 궤적에 포함할 수 있다. 따라서 인접한 인식 단위(recognition unit, 일반적으로 음소 또는 단어 모델)들이 문맥 독립형(context indepen- dent type)으로 모델링되더라도, 생성되는 궤적은 부분적인 문맥 정보를 반영하게 된다. 각 열의 베이스 $\left(\frac{\tau-t}{2M}\right)$은 τ가 $t-M$부터 $t+M$까지 값을 가지므로, -0.5에서부터 0.5까지의 정규화 된 값을 갖는다. 비슷한 방법으로 궤적의 계수 행렬 $\mathbf{B}_t$는 다음과 같이 정의된다.

$$\mathbf{B}_t=\begin{bmatrix} b_{1,1} & \Lambda & b_{1,D} \\ \mathrm{M} & & \mathrm{M} \\ b_{R,1} & \Lambda & b_{R,D} \end{bmatrix}=\begin{bmatrix} \mathbf{b}_1 \\ \mathrm{M} \\ \mathbf{b}_R \end{bmatrix} \quad (5)$$

$$\mathbf{b}_i=\begin{bmatrix} b_{i,1} & \Lambda & b_{i,D} \end{bmatrix},\ 1\le i\le R$$

2.2 분절 특징

식 (2)~(5)와 같이 분절 모델이 주어지면, 다음 단계는 모델 변수들을 추정한다. 모든 오차가 i.i.d. 분포를 따른다면 궤적 계수 행렬 $\hat{\mathbf{B}}_t$는 선형 회귀 방식(linear regression approach)에 의하여 추정될 수 있다. 선형 회귀 방식을 적용하기 위해서 각 특징 차원은 다음과 같이 표현된다.

$$y_{\tau,i}=b_{1,i}z_{\tau,1}+b_{2,i}z_{\tau,2}+b_{3,i}z_{\tau,3}+\Lambda+b_{R,i}z_{\tau,R},\quad 1\le i\le D \quad (6)$$

여기에서 $z_{\tau,r}=\left(\frac{\tau-t}{2M}\right)^{r-1}$이다. 이 선형 회귀 방정식은 Singular Value Decomposition(*SVD*)에 의하여 쉽게 풀릴 수 있다. 만약 궤적 모델이 over-determined 시스템이라면($N>R$라면), *SVD*는 최소 자승 오류(least squared error)의 개념에서 최적의 근사 값을 구할 수 있다[10]. 그렇지 않고 행렬 연산에 의하여 궤적 계수 행렬 $\hat{\mathbf{B}}_t$를 구한다면 다음과 같이 구할 수 있다.

$$\hat{\mathbf{B}}_t=[\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{C}_t \quad (7)$$

여기에서 '는 행렬의 전치(matrix transpose)를 뜻한다.

$\hat{\mathbf{B}}_t$가 추정되면, 시간 t의 분절에 속하는 프레임별 잔차 오차를 더한 적합도(goodness-of-fit) x^2는 다음과 같이 계산된다.

$$x^2=\frac{\sum_{\tau=t-M}^{t+M}(\mathbf{c}_\tau-\mathbf{z}_\tau\hat{\mathbf{B}}_t)(\mathbf{c}_\tau-\mathbf{z}_\tau\hat{\mathbf{B}}_t)'}{N} \quad (8)$$

여기에서 N은 분절을 구성하는 프레임 수를 나타낸다. 위 식은 "만약 x^2가 작다면 음성 분절에 대해 궤적이 잘 추정되었다"는 것을 의미하므로, x^2가 0이라면, 음성 특징이 궤적에 의해 완벽하게 추정되었다는 것을 표현한다. 따라서, 입력 음성 신호는 궤적 계수와 추정 오차로써 표현될 수 있다. 즉, 음성의 특징을 나타내는 분절 특징은 분절에 대한 궤적 계수 행렬 $\hat{\mathbf{B}}_t$와 그 궤적의 적합도를 나타내는 x^2로 표현된다.

3. 분절 특징 HMM

전 장에서 제안된 분절 특징을 음향학적 모델에 적용하여 인식 과정에 사용하기 위해서는 HMM의 통계적 확률 표현을 수정하여야 한다. Gales와 Russell 등은 특정 상태와 대응되는 분절의 관측 확률을 그 분절의 평균과 평균이 주어졌을 때의 관측 확률로 세분하여 모델링하였다. 이 방법은 Ostendorf가 제안한 분절 모델과 달리, 일반 HMM의 구조 안에서 분절 모델을 적용하였기 때문에, 분절 HMM(Segmental HMM)이라 한다.

3.1 분절 HMM

분절 HMM은 음성 신호의 잠재된 궤적(또는 경향)에 대한 효과적인 표현 방법을 제안한다. Russell 등은 이 궤적을 고정 분산(constant variance, 1993) 또는 선형 시스템(linear system, 1995)으로 표현하는 가우스 통계적 과정(Gaussian stochastic process)으로 표현하고 있다. 고정 분산은 분절에서 관측되는 궤적은 시간에 관계 없이 일정하며, 정적인 궤적 특징으로 표현된다[1]. 이 접근 방법은 프레임 단위의 음성 분석 및 표현대신 분절 단위로 음성 신호를 표현한다는 점에서 의의를 지닌다. 그 후, 분절을 표현하는 궤적이 선형적인 특성을

보인다고 가정하고 통계적 분석을 통하여 궤적을 모델링하였다. 분절 HMM에서는 모델 λ의 상태 s_i에서 주어진 분절 $\mathbf{Y}=y_1,\ldots,y_T$에 대한 관측 확률은 표현 가능한 궤적 f_m에 대하여 다음과 같이 정의된다.

$$\Pr(\mathbf{Y}\mid s_i,\lambda)=\int \Pr(f_m\mid s_i,\lambda)\Pr(\mathbf{Y}\mid f_m,s_i,\lambda)\,df_m \qquad (9)$$

위식에서 $\Pr(f_m\mid s_i,\lambda)$은 상태 s_i에 표현되는 평균 궤적과 입력 음성에서 관측된 궤적 f_m과의 적합성을 나타내는 확률을 의미하며, 외적 분절 변이(extra-segmental variation)로 표현한다. 반면에 $\Pr(\mathbf{Y}\mid f_m,s_i,\lambda)$은 프레임 특징의 집합인 관측 벡터$\mathbf{Y}$가 주어지면, 이 벡터에서 궤적 f_m을 추정하고, 그 때의 추정 오차를 확률 값으로 표현한다. 이 확률 값은 추출된 궤적과 프레임 관측 벡터의 연관성을 표현하는 내적 분절 변이(intra-segmental variation)를 나타낸다. 외적 분절 변이는 화자의 특징이나 선택된 화자의 발음과 같이 장기적인 변이(long-term variability)를 의미한다. 그러나 내적 분절 변이는 연속적인 조음 현상이나 불규칙한 발성 등에서 보여지는 분절 내부의 단기적인 변이를 표현한다. 이 분절 변이 값들은 "모든 관측은 주어진 상태에 대해서는 독립적이나 그 분절에 대해서는 조건부 확률을 갖는다"는 가정에 기반하고 있다 [1,6,7].

모델 λ의 특정 상태에서 $N=2M+1$개의 프레임 집합으로 표현되는 분절의 관측 확률을 모수적 궤적 모델에 적용하면, 시간 t에서 관측 벡터 $\mathbf{C}_t=\mathbf{Y}_{t-M}^{t+M}$은 유일한 궤적 $\mathbf{Z}\hat{\mathbf{B}}_t$로 표현이 되므로 식 (9)는 다음과 같이 나타낼 수 있다.

$$\Pr(\mathbf{C}_t\mid s_i,\lambda)=\Pr(\mathbf{Z}\hat{\mathbf{B}}_t\mid s_i,\lambda)\Pr(\mathbf{C}_t\mid\mathbf{Z}\hat{\mathbf{B}}_t,s_i,\lambda) \qquad (10)$$

만약 혼합 모델(mixture model)이 사용된다면 주어진 상태의 분절 $\mathbf{C}_t$의 관측 확률은 상태의 모든 혼합 모델에 대한 확률을 더함으로써 구할 수 있다. 따라서, 식 (10)는 다음과 같이 일반 HMM의 혼합 모델과 같은 형태로 표현될 수 있다.

$$\begin{aligned}\Pr(\mathbf{C}_t\mid s_i,\lambda)&=\sum_{k=0}^{K-1}c_{ik}\Pr(\mathbf{C}_t\mid s_i,m_k,\lambda)\\&=\sum_{k=0}^{K-1}c_{ik}\Pr(\mathbf{Z}\hat{\mathbf{B}}_t\mid s_i,m_k,\lambda)\Pr(\mathbf{C}_t\mid\mathbf{Z}\hat{\mathbf{B}}_t,s_i,m_k,\lambda)\end{aligned} \qquad (11)$$

여기에서 c_{ik}는 상태 s_i의 k번째 혼합 밀도 m_k의 가중치를 나타낸다. 기존의 분절 HMM에서는 주어진 상태에 대한 분절의 확률은 가능한 궤적 정보를 모두 통합하여 표현하였으나[1], 본 논문에서는 일반적인 HMM의 혼합 모델과 유사하게 관측 가능한 궤적 정보의 가중치로 표현하였다.

분절에 대한 관측 확률을 계산할 때, 혼합 모델을 사용하지 않고 관측과 상태에 대한 결합 확률을 최대화하는 최적의 궤적으로 표현할 때에는 다음과 같이 구할 수 있다.

$$\begin{aligned}\Pr(\mathbf{C}_t\mid s_i,\lambda)&=\max_k\Pr(\mathbf{C}_t\mid s_i,m_k,\lambda)\\&=\max_k\Pr(\mathbf{Z}\hat{\mathbf{B}}_t\mid s_i,m_k,\lambda)P(\mathbf{C}_t\mid\mathbf{Z}\hat{\mathbf{B}}_t,s_i,m_k,\lambda)\end{aligned} \qquad (12)$$

3.2 우도(likelihood)의 개선

기존 분절 HMM에서는 각 변이에 적용된 확률적 분포 가정 때문에 외적 분절 변이를 표현하는 궤적은 고정 분산이나 선형 시스템으로 모델링 되었으며, 내적 분절 변이에 대한 분포는 대각 공분산(diagonal covariance)으로 표현되는 가우스 분포(Gaussian distribution)를 사용하였다[1,6]. 그러나 본 논문에서는 다음과 같이 외적 분절과 내적 분절에 대한 분포를 가정한다. 외적 분절의 분포는 평균 궤적과 그 분산으로 표현되며, 내적 분절은 분절에서의 궤적 추정 오차로 표현된다. 이것은 내적 분절의 변이가 상태에서의 궤적 관측 확률을 나타내는 외적 분절에 대한 가중치로 기여한다는 것을 의미한다. 따라서 내적 분절 변이는 추정된 궤적이 분절의 프레임 특징을 얼마나 잘 표현하는 가를 나타내는 척도를 의미한다.

모델 λ와 상태 s_i에서 괘적 $\mathbf{Z}\hat{\mathbf{B}}_t$의 외적 분절 확률은 궤적에서 복원된 프레임 특징의 정규 분포의 곱으로써 표현될 수 있다. 분절 분포는 모수적 방법을 통하여 표현되기 때문에 직접적인 매개 변수의 값에 의하여 구할 수가 없다. 따라서, 모수적 궤적은 프레임 특징으로 복원되어 궤적을 따라 형성된 점들의 정규 분포 $\Pr(\mathbf{z}_\tau\hat{\mathbf{B}}_t\mid\mathbf{z}_\tau\mathbf{B}_i)\sim\aleph(\mathbf{z}_\tau\mathbf{B}_i,\boldsymbol{\Sigma}_{\tau,i})$의 곱으로 그 분포를 표현한다.

$$\begin{aligned}\Pr(\mathbf{Z}\hat{\mathbf{B}}_t\mid s_i,\lambda)&=\Pr(\mathbf{Z}\hat{\mathbf{B}}_t\mid\mathbf{Z}\mathbf{B}_i,\boldsymbol{\Sigma}_i)\\&=\prod_{\tau=t-M}^{t+M}\aleph(\mathbf{z}_\tau\mathbf{B}_i,\boldsymbol{\Sigma}_{\tau,i})\end{aligned} \qquad (13)$$

여기에서 $\mathbf{B}_i$는 상태 s_i의 평균 궤적 계수 행렬을 나타내며, Σ_i는 평균 궤적의 분산을 표현한다. 즉, 분산 행렬로 이루어진 벡터이며 $\Sigma_i=\{\Sigma_{-M,i},\ldots,\Sigma_{0,i},\ldots,\Sigma_{M,i}\}$로 구성된다. 상태에서 관측되는 분절의 분산을 표현하는 방법에는 분절의 특징을 표현하는 궤적에 대해 공통 분산(common variance)을 이용하는 방법과, 시간적인 변화를 반영하는 시변 분산(time varying variance)을 이용하는 방법을 생각할 수 있다. 공통 분산은 분절 내의 시간적인 변화를 반영하지 못하기 때문에 고정 분산(fixed

variance)이라 한다. 이 경우, Σ_i의 각 원소는 분절 내의 모든 프레임 인덱스 $\tau-t$에 대해 동일한 분산 값을 갖는다. 반면, 시변 분산은 분절에서의 시간적인 역학(temporal dynamics)을 표현하기 때문에 상대적인 프레임 인덱스 $\tau-t$에 따라 분산 값은 변한다.

내적 분절 변이는 궤적의 추정 오차를 나타내며 상태 s_i와 독립이므로, 변이 확률은 다음과 같이 적합도 x^2를 이용하여 정의될 수 있다.

$$\Pr(\mathbf{C}_t \mid \mathbf{Z}\hat{\mathbf{B}}_t, s_i, \lambda) = \Pr(\mathbf{C}_t \mid \mathbf{Z}\hat{\mathbf{B}}_t) = \exp\left\{-\frac{1}{2}\chi^2\right\} = \exp\left\{-\frac{1}{2N}\sum_{\tau=t-M}^{t+M}(\mathbf{c}_\tau - \mathbf{z}_t\hat{\mathbf{B}}_t)(\mathbf{c}_\tau - \mathbf{z}_t\hat{\mathbf{B}}_t)'\right\} \tag{14}$$

따라서, 시간 t에서 상태 j의 분절에 대한 관측 확률은 다음과 같이 표현된다.

$$\begin{aligned} b_j(\mathbf{C}_t) &= P(\mathbf{C}_t \mid s_j, \lambda) \\ &= \sum_{k=0}^{K-1} c_{jk} b_{jk}(\mathbf{C}_t) = \sum_{k=0}^{K-1} c_{jk} P(\mathbf{C}_t \mid s_j, m_k, \lambda) \\ &= \sum_{k=0}^{K-1} c_{jk} \Pr(\mathbf{Z}\hat{\mathbf{B}}_t \mid \mathbf{Z}\mathbf{B}_{jk}, \Sigma_{jk}) \Pr(\mathbf{C}_t \mid \mathbf{Z}\hat{\mathbf{B}}_t) \\ &= \Pr(\mathbf{C}_t \mid \mathbf{Z}\hat{\mathbf{B}}_t) \sum_{k=0}^{K-1} c_{jk} \Pr(\mathbf{Z}\hat{\mathbf{B}}_t \mid \mathbf{Z}\mathbf{B}_{jk}, \Sigma_{jk}) \end{aligned} \tag{15}$$

여기에서 $\mathbf{B}_{jk}$와 Σ_{jk}는 상태 j의 혼합 밀도 m_k에 대응되는 궤적 모델을 나타낸다. $P(\mathbf{C}_t \mid \mathbf{Z}\hat{\mathbf{B}}_t)$는 상태 j에 독립적이며 분절에 대한 궤적의 추정 오차를 표현하고 있기 때문에, 내적 분절 변이는 주어진 상태의 외적 분절 변이 확률에 대해 적용되는 시변 가중치(time-varying weight)로 생각될 수 있다.

3.3 변수 추정

SFHMM의 변수 추정을 위하여 Baum-Welch 형태의 변수 추정 방법이 유도된다. $\gamma_t(j)$와 $\xi_t(j,k)$를 시간 t에서 상태 j에 존재할 사후 확률(posterior probability)과 상태 j의 혼합 밀도 m_k에 있을 사후 확률이라 하자. 그러면, 모델 λ와 관측 열 $\mathbf{C}_t$가 주어지면 $\xi_t(j,k)$는 다음과 같이 구할 수 있다.

$$\begin{aligned} \xi_t(j,k) &= P(s_t = j, k_t = k \mid \mathbf{C}_t, \lambda) \\ &= \frac{\sum_i \alpha_{t-1}(i) a_{ij} c_{jk} b_{jk}(\mathbf{C}_t) \beta_t(j)}{\sum_{i \in S_F} \alpha_T(i)} \end{aligned} \tag{16}$$

여기에서 a_{ij}는 상태 i에서 j로 가는 전이 확률(transition probability)을 나타내며, S_F는 최종 상태의 집합을 표현한다. 또한 $\gamma_t(j)$는 시간 t에서 상태 i까지의 전향 확률(forward probability)을 의미한다. $\gamma_t(j)$는 상태 j에서의 전향 확률과 후향 확률(backward probability)의 곱으로써 표현되기 때문에 일반 HMM과 동일하게 구할 수 있다. 따라서, $\gamma_t(i)$와 $\xi_t(j,k)$의 추정 후에, 상태 j의 k번째 혼합 밀도에 대한 가중치는 다음과 같이 추정된다.

$$\overline{c}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{17}$$

일반적인 HMM과 비슷하게, SFHMM에서도 특정 상태에서의 혼합 밀도에 대한 평균 궤적은 모든 분절에 대한 상태 j의 혼합 밀도 k에 머무를 기대치 $\xi_t(j,k)$와 궤적의 곱으로 나타내는 기대 평균치(expected average)로써 구할 수 있다.

$$\mathbf{Z}\overline{\mathbf{B}}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k) \mathbf{Z}\hat{\mathbf{B}}_t}{\sum_{t=1}^{T} \xi_t(j,k)} \tag{18}$$

SFHMM은 음성 분절을 고정된 길이만큼 분석을 하기 때문에 음성의 시작 부분과 끝부분을 제외하면 거의 대부분의 음성 분절에서 동일한 디자인 행렬 $\mathbf{Z}$를 사용한다. 따라서, 식 (18)의 양편에서 디자인 행렬을 생략할 수 있다. 디자인 행렬이 생략된 식 (18)은 궤적 계수 행렬 $\overline{\mathbf{B}}_{jk}$을 추정하는 식으로 변환된다.

$$\overline{\mathbf{B}}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k) \hat{\mathbf{B}}_t}{\sum_{t=1}^{T} \xi_t(j,k)} \tag{19}$$

$\overline{\mathbf{B}}_{jk}$가 추정되면 입력 분절 특징 $\mathbf{Z}\hat{\mathbf{B}}_t$과 평균 궤적 $\mathbf{Z}\overline{\mathbf{B}}_{jk}$과의 차이를 이용하여 분산을 구한다.

3.3.1 시변 분산

분산은 일반적으로 입력 벡터와 평균 벡터의 차이로써 구하므로, 평균 벡터가 분절 특징으로 표현되는 특징 표현 방법인 경우 분산 표현 방법은 매우 중요하다. 본 논문에서 특정 상태의 분산은 분절 특징의 분산을 의미하기 때문에, 입력 음성 분절에서 추정된 궤적과 주어진 상태의 평균 궤적의 차이로써 표현된다. 따라서 상태의 분절 특징에 대한 분산은 상태에서 관측 가능한 프레임 특징의 분산 열로 표현된다. 분절 길이가 N이라면 상태에서의 관측된 분절은 N개의 프레임 특징으로 구성되며, 각 프레임 특징의 분산은 분절 내에서의 상대적인 순서에 의해 정렬된다. 분절은 단일의 궤적으로 표현되기 때문에, 그 분절에 대한 분산은 추정된 궤적과 평균 궤적의 거리로써 표현되며, 두 궤적간의 거리는 각 궤적

에서 프레임별로 복원된 점들의 거리를 이용한다.
즉, 시변 분산은 다음과 같이 계산된다.

$$\overline{\Sigma}_{jk} = \{\overline{\Sigma}_{-M,jk},\ldots,\overline{\Sigma}_{0,jk},\ldots,\overline{\Sigma}_{M,jk}\}$$

$$\overline{\Sigma}_{n,jk} = \frac{\sum_{t=1}^{T}\xi_t(j,k)\{\mathbf{z}_n\hat{\mathbf{B}}_t - \mathbf{z}_n\overline{\mathbf{B}}_{jk}\}'\{\mathbf{z}_n\hat{\mathbf{B}}_t - \mathbf{z}_n\overline{\mathbf{B}}_{jk}\}}{\sum_{t=1}^{T}\xi_t(j,k)}, -M \le n \le M \quad (20)$$

여기에서 n은 분절 내에서의 상대적인 위치를 나타낸다. $\mathbf{z}_n\overline{\mathbf{B}}_{jk}$와 $\mathbf{z}_n\hat{\mathbf{B}}_t$는 각각 시간 t의 분절 특징으로 추정된 궤적과 상태 j와 혼합 밀도 k의 평균 궤적의 상대적인 프레임 특징을 의미한다. 시변 분산을 SFHMM에 적용하면 주어진 상태 j의 k번째 혼합밀도의 분절 특징에 대한 관측 확률은 다음과 같이 표현된다.

$$\Pr(\mathbf{Z}\hat{\mathbf{B}}_t \mid s_j, m_k, \lambda) = \Pr(\mathbf{Z}\hat{\mathbf{B}}_t \mid \mathbf{Z}\mathbf{B}_{jk}, \Sigma_{jk}) = \prod_{\tau=t-M}^{t+M}\aleph(\mathbf{z}_\tau\mathbf{B}_{jk}, \Sigma_{\tau-t,jk})$$
$$= \prod_{\tau=t-M}^{t+M}\frac{1}{(2\pi)^{D/2}|\Sigma_{\tau-t,jk}|^{1/2}}\cdot\exp\left\{-\frac{1}{2}(\mathbf{z}_\tau(\hat{\mathbf{B}}_t - \mathbf{B}_{jk}))\Sigma_{\tau-t,jk}^{-1}(\mathbf{z}_\tau(\hat{\mathbf{B}}_t - \mathbf{B}_{jk}))'\right\} \quad (21)$$

3.3.2 고정 분산

시변 분산 방법이 적용되면, 단일 혼합 밀도로 구성되는 한 상태를 표현하기 위해서는 궤적 계수 행렬 $\mathbf{B}$와 궤적의 분산을 나타내는 N개의 분산 행렬이 필요하다. 따라서 분절의 범위가 넓어지거나 혼합 밀도의 수가 증가하면 SFHMM을 표현하는 매개 변수의 수는 급격히 증가하게 되고, 더 많은 연산 시간을 필요로 한다. 이와 같은 문제점을 해결하기 위하여 매개 변수의 수를 줄이는 방법에 대한 연구가 필요하다. 특히 분절 특징의 분산을 표현하기 위해서는 각 상태의 혼합 밀도에 N개의 분산 행렬이 필요하므로, 대표적인 분산을 이용하여 표현한다면 많은 수의 매개 변수를 줄일 수 있을 것이다. 이 경우, 대표 분산을 선택하는 것은 그 방법에 따라 음성 인식 시스템의 성능에 많은 영향을 미칠 수 있기 때문에 본 연구에서는 분절 내에서의 각 프레임 특징의 분산을 평균하여 사용한다. 즉, 각 프레임 특징의 분산을 평균하여 사용하며 모든 프레임에 공통적으로 적용한다.

고정 분산은 분절 내의 모든 프레임 특징의 분산에 대한 평균으로써 얻어지며 다음과 같이 표현된다.

$$\overline{\Sigma}_{jk} = \frac{1}{N}\sum_{n=-M}^{M}\overline{\Sigma}_{n,jk} \quad (22)$$

위 식에서 $\overline{\Sigma}_{n,jk}$는 시변 분산 표현 방법에서와 같이 분절 내의 상대적인 프레임 특징의 분산을 의미한다. 따라서 식 (22)를 확장하여 표현하면 다음과 같이 나타낼 수 있다.

$$\overline{\Sigma}_{jk} = \frac{\sum_{t=1}^{T}\xi_t(j,k)\{\mathbf{Z}\hat{\mathbf{B}}_t - \mathbf{Z}\overline{\mathbf{B}}_{jk}\}'\{\mathbf{Z}\hat{\mathbf{B}}_t - \mathbf{Z}\overline{\mathbf{B}}_{jk}\}}{N\sum_{t=1}^{T}\xi_t(j,k)} \quad (23)$$

만약 고정 분산이 SFHMM에 적용되는 경우, 상태 j의 k번째 혼합밀도에서의 관측 확률은 다음과 같이 단순화될 수 있다.

$$\Pr(\mathbf{Z}\hat{\mathbf{B}}_t \mid s_j, m_k, \lambda) = \Pr(\mathbf{Z}\hat{\mathbf{B}}_t \mid \mathbf{Z}\mathbf{B}_{jk}, \Sigma_{jk})$$
$$= \frac{1}{(2\pi)^{ND/2}|\Sigma_{jk}|^{N/2}}\cdot\exp\left\{-\frac{1}{2}Tr\left\{(\mathbf{Z}\hat{\mathbf{B}}_t - \mathbf{Z}\mathbf{B}_{jk})\Sigma_{jk}^{-1}(\mathbf{Z}\hat{\mathbf{B}}_t - \mathbf{Z}\mathbf{B}_{jk})'\right\}\right\} \quad (24)$$

3.4 SFHMM의 조건에 따른 일반화

시변 분산이든 고정 분산이든지, 분절의 길이가 1인 경우, 즉 분절 특징 대신 프레임 특징이 사용된 경우 일반 HMM의 관측확률과 동일한 형태를 취한다. 그러나, 분절의 길이가 1보다 큰 경우에는 이들 방법들은 다른 분산 표현 방법을 취하게 되고, 그때의 관측확률은 서로 달라진다. 이와 같이 SFHMM은 조건에 따라 일반적인 HMM과 모수적 궤적 모델의 일반화 또는 확장으로 해석할 수 있다.

1. 만약 분절 길이 $N=1$이고 회귀 차수 $R=1$이라면, 시간 t의 분절 $\mathbf{Z}\hat{\mathbf{B}}_t$는 단일 프레임 특징 $\mathbf{c}_t$가 된다. 따라서 내적 분절 변이의 확률 $P(\mathbf{C}_t \mid \mathbf{Z}\hat{\mathbf{B}}_t, s_i, \lambda)$은 1이 되고 외적 분절 확률 $P(\mathbf{Z}\hat{\mathbf{B}}_t \mid s_i, \lambda)$은 추정된 궤적이 아닌 관측된 프레임 특징에 대한 가우스 분포를 따르게 된다. 이 경우, $\mathbf{C}_t$는 단일 프레임 특징을 표현하게 되고, 추정된 궤적 $\mathbf{Z}\hat{\mathbf{B}}_t$와 같게 된다. 따라서 추정 오차 또는 적합도 x^2는 0이 되며, $\mathbf{Z}\mathbf{B}_i$는 상태 s_i에 대한 평균 특징을 표시하게 된다. 그러므로, SFHMM은 연속 HMM과 완전히 같게 된다.
2. 만약 분절 길이 N이 주어진 모델에 대한 관측 열의 길이 T와 같다면, 각 음향학적 모델은 가변 분절 길이의 단일 상태 또는 단일 분절로 표현된다. 가변 길이를 갖는 입력 음성을 처리하기 위해서는 디자인 행렬 $\mathbf{Z}$의 크기를 관측 열의 길이 T가 되도록 확장하여야 한다. 이 경우 식 (7)에서 디자인 행렬은 k번째 관측 분절에 종속적인 디자인 행렬 $\mathbf{Z}_k$로써 표현되며 궤적 계수 행렬 $\hat{\mathbf{B}}_k$는 다음과 같이 계산된다.

$$\hat{\mathbf{B}}_k = [\mathbf{Z}_k'\mathbf{Z}_k]^{-1}\mathbf{Z}_k\mathbf{C}_k \quad (25)$$

각 모델은 단일 상태로 표현되기 때문에, 평균 궤적의 추적은 다항식에 의한 분절 모델과 동일하게 된다[11]. 즉,

$$\overline{\mathbf{B}} = \left[\sum_{k=1}^{K}\mathbf{Z}_k'\mathbf{Z}_k\right]^{-1}\left[\sum_{k=1}^{K}\mathbf{Z}_k'\mathbf{Z}_k\hat{\mathbf{B}}_k\right] \quad (26)$$

따라서 궤적 계수 행렬 $\hat{\mathbf{B}}_k$와 평균 궤적 계수 행렬이 위와 같이 조정된다면, SFHMM은 모수적 궤적

모델과 같이 가변 길이를 갖는 음성 분절을 모델링 할 수 있다.

4. 실험 및 결과

제안된 방법의 유효성을 입증하기 위하여 연속 음성 인식의 두 분야에서 분절 특징 HMM을 실험하였다. 하나는 문장을 39개의 음소 군으로 인식하여 음소 열을 생성하는 음소 인식 실험이며, 다른 하나는 문장에서 추출한 16개의 모음을 인식하는 모음 분류 실험이다. 음소 인식 실험에서는 분절 특징의 특성을 살펴보기 위하여 연속 HMM과 성능 비교를 한다. 연속 HMM은 정적 특징과 동적 특징 모두를 사용하며, SFHMM에서는 분절 특징만을 사용한다.

두 분야의 음성 인식 실험에는 동일한 전처리 과정이 사용되었다. 신호 처리를 위하여 20ms 프레임 크기를 갖는 음성 파형을 선-강조(pre-emphasis)하여 10ms씩 프레임을 이동시켜가면서 분석하였다. 특징 추출 단계에서는 각 프레임에 해밍 창(Hamming window)을 적용하여 푸리에(Fourier) 변환을 시켰으며, 24 채널의 필터 뱅크(Filter bank)에 멜 변환(Mel transform)을 적용하였다. 변환된 멜 필터 뱅크 에너지를 이산 코사인 변환(DCT; Discrete Cosine Transform)을 통하여 12차의 계수와 로그 에너지를 구하여 음성 인식에 사용하는 특징으로 사용하였다. 이들 특징은 연속 HMM인 경우에는 동적인 특징을 갖도록 델타(Δ) 또는 델타-델타($\Delta\Delta$ 또는 Δ^2) 특징을 계산하는데 사용되며, SFHMM인 경우에는 분절 특징을 구하는데 사용된다.

4.1 데이타베이스

모든 실험에는 TIMIT 데이타베이스가 사용되었다. TIMIT 데이타베이스는 화자 독립 음성 인식기를 구현할 수 있도록 학습 자료와 평가 자료를 구분하여 구축이 되었으며, 총 6,300문장을 포함하고 있다. 미국의 지리학적으로 분리된 8 지방에서 선택된 630명의 화자가 각 10문장씩 발성을 하였으며, 남성 화자가 약 70%정도를 차지하고 대부분 백인 성인이 주류를 이룬다.

TIMIT 자료에 사용된 문장들은 SRI에서 고안된 2개의 방언, MIT에서 제안한 음성학적으로 조밀하게 구성된 450 문장, TI에서 선택한 음성학적으로 다양하게 구성된 1,890문장으로 구성되었으며 다음과 같은 3개의 군으로 분류된다.

- 모든 화자에게 공통적으로 발성된 2개의 SA 문장
- MIT의 450문장에서 추출된 5개의 SX 문장
- TI가 선택한 문장에서 무작위로 추출된 3개의 SI 문장

이들 문장들은 다시 학습과 실험을 위하여 분리되었으며 학습에는 462명, 평가에는 TIMIT 데이타베이스가 권장하는 168명 또는 24명의 화자로 구성되었다. 24명의 화자가 평가에 사용될 경우에는 8개의 방언에서 남자 2명 여자 1명의 3명의 화자가 선택되어 5개의 SX 문장과 3개의 SI 문장을 발성하도록 하였으며, 이 집합을 "핵심 평가 집합(core test set)"이라 하였다. 반면에 평가 집합을 확대하여 학습에 참여하지 않은 모든 화자가 각각 8문장(5개의 SX문장과 3개의 SI문장)씩을 발성한 평가 집합을 완전 평가 집합(complete test set)이라 한다. 이 경우에는 총 화자수가 168명이 되며 1,344 문장이 평가에 사용된다. 그러나, 본 연구에서는 TIMIT 데이타베이스가 권장한 완전 평가 집합에서 제외된 2개의 SA 문장을 포함하여 실험하여 총 1,680문장을 평가에 사용하였다. 그 이유는 기존의 연구와 음성 인식 시스템의 성능 비교를 위하여 비슷한 실험 조건을 구축하였기 때문이다.

4.2 음소 인식

음소 인식 분야에서는 48개의 음소 인식 모델이 사용되었다. TIMIT 데이타베이스에는 약 60여개 이상의 음소가 있으나 일반적으로 연속 음성 인식에는 대표적인 61음소 또는 48음소를 사용한다. 본 연구에서는 1989년에 K.F.Lee가 제안한 48표준 음소를 채택하여 학습에 사용하였으며, 인식 과정에서는 다시 39개 음소 군으로 분류하여 인식 성능을 평가하였다[12]. 각 음소는 문맥 독립(context independent)형으로 표현되었으며, 탐색과정에서 음소 이진 언어모델(phone bigram)을 이용하였다.

4.2.1 분절 특징의 특성

이 실험의 목적은 분절 특징을 이용한 분절 특징 HMM과 프레임 특징을 이용하는 일반 HMM의 성능을 비교하는 데 있다. 분절 특징은 여러 프레임 특징들로 구성되어 있으며 분절 길이와 회귀 차수등에 따라 성능의 변화가 발생하기 때문에 다양한 조건의 특징 조합을 통하여 성능을 분석하였다. 앞장에서 설명한 바와 같이 단일 분절 길이와 단일 회귀 차수를 이용하는 경우의 분절 특징 HMM이 일반 HMM과 같기 때문에, 이 조건의 분절 특징 HMM을 기본 시스템으로 삼았다. 따라서, 본 절에서의 실험은 단일 분절 길이와 회귀 차수를 갖는 SFHMM과 비단일 분절 길이와 회귀 차수를 갖는 SFHMM의 성능 비교가 된다.

분절 특징의 특성을 살펴보기 위해서 두 가지 조건의 특징 집합을 이용하여 성능 비교를 하였다. 먼저 기본 시스템은 12차의 MFCC와 로그 에너지로 13차 기본 특징, 그리고 1차 미분 계수(Δ)를 포함한 26차의 특징을

이용하였으며, SFHMM은 13차의 특징을 이용한 분절 특징(궤적)을 이용하여 성능을 비교하였다. 표 1은 기본 시스템의 성능 (정확도)과 다양한 특징 조건에 따른 SFHMM의 성능을 보여주고 있다.

표 1 SFHMM과 HMM의 음소 인식 정확도 (기본 시스템은 13차의 기본 벡터와 일차 미분 계수를 사용하고, SFHMM은 13차 특징에 기반한 다른 조건의 분절 길이(N)과 회귀 차수(R) 사용, M은 혼합 밀도의 수를 의미함)

시스템	*M=1*	*M=2*
기본 시스템	52.8	56.1
N=3, R=2	51.0	54.0
N=3, R=3	51.5	54.3
N=5, R=2	51.6	54.6
N=5, R=3	52.9	55.8
N=5, R=4	53.1	56.3
N=5, R=5	53.2	56.4

이들 실험에서 분절 길이와 회귀 차수가 같은 경우에는 궤적의 추정 오차, 즉 적합도는 0이 된다. 실험 결과 분절 길이가 5이고 회귀 차수가 3이상인 경우에는 SFHMM의 성능이 일반 HMM보다 우수함을 알 수 있다.

다음으로는 분절 특징의 기본이 되는 특징 집합을 26차로 확장하여 실험하였다. 즉, 기본 시스템은 13차의 기본 특징과 1차 미분 계수(Δ), 2차 미분 계수(Δ^2)를 이용한 총 39차의 특징을 이용하였으며, SFHMM은 13차의 기본 특징과 1차 미분 계수인 26차의 특징에 기반한 분절 특징을 사용하여 실험하였다. 실험 결과는 표 2에 있다.

표 2 SFHMM과 HMM의 음소 인식 정확도 (기본 시스템은 13차의 기본 벡터와 일·이차 미분 계수를 사용하고, SFHMM은 26차 특징에 기반한 다른 조건의 분절 길이(N)과 회귀 차수(R) 사용, M은 혼합 밀도의 수를 의미함)

시스템	*M=1*	*M=2*
기본 시스템	52.6	57.0
N=3, R=2	54.4	58.1
N=3, R=3	54.7	58.5
N=5, R=2	54.6	58.7
N=5, R=3	55.6	59.9
N=5, R=4	55.6	60.1
N=5, R=5	55.6	60.1

두 번째 실험에서는 제안된 SFHMM의 성능이 항상 일반 HMM보다 우수함을 알 수 있다. 이들 실험 결과로부터, 정적 특징(stationary feature)에 기반한 궤적 특징을 이용하도록 제안된 방식이 동적 특징(dynamic feature)을 이용한 방법과 마찬가지로 음성 신호의 역학을 표현할 수 있다고 생각할 수 있다. 특히 두 번째 실험에서 알 수 있듯이 SFHMM에서 1차 미분 계수를 이용하여 분절 특징을 추출한 경우 성능이 우수하였기 때문에, 다음 실험에서는 정적 특징뿐만 아니라 동적 특징인 1차 미분계수에 기반한 분절 특징을 사용하고자 한다.

4.2.2 분절 길이와 회귀 차수

이전 실험에서 정적 특징과 동적 특징에 기반한 분절 특징이 효율적으로 음성 패턴을 표현할 수 있다는 것을 알 수 있었다. 본 절에서는 궤적을 표현하는 데 있어서 분절 길이와 회귀 차수의 상관관계를 살펴보고자 한다.

기본 시스템은 전 절과 동일하게 26차의 기본 특징을 이용한 연속 HMM이며, SFHMM은 이들 특징으로부터 추출한 분절 특징을 이용하였다. 서로 다른 분절 길이와 회귀 차수의 효과를 검사하기 위하여, 분절 길이와 회귀 차수를 변경하면서 다양하게 실험하였다. 실험 환경은 전절에서 사용된 환경과 동일하며, 인식 결과는 표 3과 4에 정리되어 있다.

표 3 단일 혼합 밀도를 이용한 SFHMM의 성능 변화

SFHMM	%Corr.	%Acc.	%Subs.	%Del.	%Ins.	%Err
기본시스템	58.0	52.8	29.5	12.5	5.2	47.2
N=3,R=2	59.4	54.4	28.5	12.1	4.9	45.6
N=3,R=3	59.6	54.7	28.3	12.1	4.9	45.3
N=5,R=2	58.9	54.6	28.0	13.1	4.3	45.4
N=5,R=3	60.0	55.6	27.5	12.5	4.4	44.4
N=5,R=4	60.1	55.6	27.4	12.5	4.5	44.4
N=5,R=5	60.1	55.6	27.4	12.6	4.5	44.4

표 4 이중 혼합 밀도를 이용한 SFHMM의 성능 변화

SFHMM	%Corr.	%Acc.	%Subs.	%Del.	%Ins.	%Err
기본시스템	62.5	56.1	27.3	10.2	6.4	43.9
N=3,R=2	63.7	58.1	26.3	10.0	5.6	41.9
N=3,R=3	64.1	58.5	26.1	9.9	5.6	41.5
N=5,R=2	63.8	58.7	25.7	10.5	5.1	41.3
N=5,R=3	65.0	59.9	24.9	10.1	5.1	40.1
N=5,R=4	65.3	60.1	24.8	9.9	5.2	39.9
N=5,R=5	65.2	60.1	24.8	10.0	5.1	39.9

위 표에서 보인 것처럼 SFHMM에서는 분절 길이가 증가하고 회귀 차수가 높을수록 성능이 향상되었다. 이 결과로부터 서로 다른 분절 길이와 회귀 차수는 음성 인식 시스템의 성능에 영향을 준다는 것을 알았다. 특히 동일한 분절 길이에 대하여 회귀 차수가 증가하면 인식률(percent correct)이 증가하고 치환 오류(substitution error)가 감소한다. 이것은 동일한 분절 길이에서 높은 회귀 차수에 의해 변별력(discrimination)이 높아졌다고 할 수 있다. 반면 동일한 회귀 차수에 대해 분절 길이를 증가시키면 삽입 오류(insertion error)와 치환 오류가 감소하지만, 삭제 오류(deletion error)는 어느 정도 증가한다. 그렇지만, 이 경우에도 감소된 오류가 증가된 오류보다 크기 때문에 음성 인식의 성능 척도인 정확도(accuracy)는 증가한다. 따라서 치환 오류와 삽입 오류의 감소에서 동일한 회귀 차수인 경우 분절 길이의 증가는 전이 정보(transitional information)가 증가한다고 생각할 수 있다.

기본 시스템에 비해서 선형 궤적 시스템(linear trajectory system)인 경우 3.3%에서 5.9%까지 오류가 감소하였으며, 이차 궤적(quadratic trajectory)인 경우 4.0%에서 8.7%까지 오류가 감소하였다. 또한 실험 결과에서 상태를 표현하는 혼합 밀도의 수가 작은 경우에는 분절 길이가 증가하고 회귀 차수가 높아지더라도 더 이상 성능이 향상되지 않음을 알 수 있다. 즉, 표 3에서 분절 길이가 5이고 회귀 차수가 2 이상인 경우에 모두 성능이 같다. 이것은 긴 분절이 높은 회귀 차수에 의해 모델링되는 경우, 분절의 변이가 증가하기 때문에 더 많은 혼합밀도가 필요함을 의미한다. 그러나 혼합 밀도의 수를 증가시키더라도 표 4에서 처럼 회귀 차수가 3이상인 경우에는 마찬가지로 성능이 동일하다. 따라서 혼합 밀도의 수와 SFHMM의 회귀 차수나 분절 길이가 상관관계가 있다는 것을 알 수 있다. 만약 더 많은 수의 혼합 밀도를 이용한다면 SFHMM의 성능 차이도 발생할 것이다.

4.3 모음 분류

이 절에서는 다른 종류의 분산 추정 방법의 효과를 살펴보는 실험에 대해 기술한다. 서로 다른 분산 표현 방법을 갖는 SFHMM의 성능을 비교하기 위하여, 기존 연구와 비슷한 환경이 되도록 모음 분류 실험을 하였다 [11, 13]. 모음 분류 실험을 위해서 16개의 모음을 추출하였으며, 16개의 모음은 13개의 단모음(*/iy, ih, ey, eh, ae, aa, ah, ao, ow, uw, uh, ux, er/*)과 3개의 복모음(*/ay, oy, aw/*)으로 구성되었다. 이들 모음들은 문맥상의 어떤 제약도 주지 않은 상태에서 TIMIT 데이타베이스의 발음 기호로부터 추출되었다. 실험에 사용된 특징은 26차 기본 특징 벡터로부터 추출한 분절 특징을 사용하였다. 문장에서 모음을 추출하여 학습에 41,429개의 모음을 사용하였으며, 평가에는 15,119개의 모음을 이용하였다.

분산 표현 방법에 따른 성능 변화를 살펴보기 위하여 SFHMM의 혼합 밀도의 수와 분절 길이, 회귀 차수를 조정하며 실험하였다. 시변 분산이 적용된 경우, 상태에서의 분산은 분절의 각 프레임에 대해 계산되었으며, 고정 분산을 이용한 경우, 분절의 분산은 분절 내부의 프레임들에 대한 평균치로 표현되었다. 시변 분산에 대한 성능의 변화는 표 5에, 고정 분산에 대한 성능의 변화는 표 6에 보인다.

표 5 시변 분산을 이용한 SFHMM의 모음 분류율

시스템	*M=1*	*M=2*
기본 시스템	55.23	58.44
N=3, R=2	56.86	59.56
N=3, R=3	57.21	59.64
N=5, R=2	58.77	60.43
N=5, R=3	58.79	60.84
N=5, R=4	58.76	60.88
N=5, R=5	58.74	60.86

표 6 고정 분산을 이용한 SFHMM의 모음 분류율

시스템	*M=1*	*M=2*
기본 시스템	55.23	58.44
N=3, R=2	56.12	59.90
N=3, R=3	56.12	59.87
N=5, R=2	56.99	59.90
N=5, R=3	56.86	60.14
N=5, R=4	56.70	61.10
N=5, R=5	56.68	60.94

이들 실험 결과에서 두개의 혼합 밀도가 사용된 경우, 61.1%의 인식률을 보였다. 이 결과는 기존 연구[11,13]에서 보고한 최고의 성능(66.0%[11], 66.2%[13])보다 저하된 값을 보이고 있다. 비록 본 연구의 결과가 기존의 연구보다 낮은 성능을 보이더라고 데이타 집합이나 혼합 밀도의 수, 평균 궤적에 대한 회귀 차수 등이 다르기 때문에 상대적인 비교는 가능하다. Gish와 Ng등이 1996년에 발표한 실험 결과는 완전 공분산(full covariance) 집합과 8개의 혼합 밀도를 사용하였으며[11],

Fukada 등이 발표한 실험결과는 최고 9개의 혼합 밀도를 사용하였기 때문이다[13].

시변 분산이 단일 혼합밀도의 SFHMM에 적용된 경우, 분절 길이와 회귀 차수가 증가됨에 따라 성능이 향상되었다. 그러나 고정 분산을 이용한 경우, 시변 분산을 이용한 경우와 같은 뚜렷한 성능 향상을 보이지 않고 있다. 그러나, SFHMM의 혼합 밀도를 증가하면 두 분산 표현 방식의 성능 차이는 현저히 줄어든다. 심지어 고정 분산을 이용한 경우의 SFHMM의 성능이 시변 분산을 적용한 경우보다 높은 경우도 있다.

이런 결과는 기존 연구에서 발표된 결과와 다른 양상을 보이며, 분절 길이에 영향때문으로 생각한다. 기존 연구에서는 완전 궤적을 이용하였기 때문에 각 분절은 음소와 같은 음향학적 모델에 대응된다. 즉, 분절 길이가 SFHMM보다 크다는 것을 의미한다. 따라서 긴 분절을 이용하여 음성을 분석한다면 고정 분산은 분절의 변이를 충분히 표현하지 못한다. 그러나, 만약 분절의 길이를 SFHMM과 같이 작게 사용한다면 적절한 혼합 밀도가 주어진 경우 분절의 변이를 충분히 표현할 수 있을 것이다. 따라서 SFHMM을 이용하여 음성을 표현하고자 할 때 작은 분절을 사용하고 다중 혼합 밀도를 이용한다면, 성능의 저하 없이 고정 분산 방식을 SFHMM에 적용할 수 있을 것이다.

4.4 결과 및 토의

본 장에서 설명한 실험들로부터 SFHMM이 음소 인식이나 모음 분류에서 기존의 HMM보다 성능이 뛰어남을 알 수 있었다. 기존의 프레임 특징에 기반한 분절 특징의 효과를 살펴보기 위하여 정적 특징과 동적 특징을 모두 이용한 HMM과 정적 특징에 기반한 분절 특징을 이용한 SFHMM의 성능을 비교하였다. 성능 비교 결과, 긴 분절 길이와 높은 회귀 차수를 이용한 경우, 분절 특징을 이용한 SFHMM의 성능이 기존의 HMM보다 우수한 성능을 보임을 알 수 있었다. 또한 SFHMM이 정적 특징과 동적 특징에 기반한 분절 특징을 사용한 경우 성능은 더 향상되었다. 따라서, 이와 같은 실험에서 분절 특징은 음성 처리에서 효과가 있음이 입증되었다.

SFHMM의 특성을 조사하기 위하여 분절 길이와 회귀 차수를 변경하면서 음소 인식 실험을 하였다. 이 실험에서 회귀 차수가 증가하면 변별력이 증가하고 분절 길이가 길어지면 전이 정보량이 증가한다는 결과를 얻었다.

마지막으로 성능의 저하 없이 SFHMM을 표현하는데 필요한 매개 변수의 수를 줄이고자 시도하였다. 이 실험에서 다른 종류의 분산 표현 방법을 이용하여 성능을 비교하였다. 시변 분산은 상태에서 관측 가능한 분절의 각 프레임의 분산을 표현한 방법이고, 고정 분산 방식은 상태의 분산을 분절 내의 모든 프레임 분산에 대해 평균하여 얻는 방식이다. 이들 두 분산 표현 방식을 이용하여 SFHMM의 성능을 비교한 결과, 단일 혼합 밀도가 사용된 경우, 시변 분산 방식을 이용한 SFHMM이 고정 분산 방식을 이용한 경우보다 성능이 우수하였으나 다중 혼합 밀도가 사용된 경우에는 두 방식의 성능 차이는 미미함을 알 수 있다. 따라서 고정 분산 방식은 분절 길이가 작은 경우에 어느 정도 효과적이다고 생각할 수 있다.

제안한 방식으로 인해 성능 향상이 이루어졌다고 하더라도 매개 변수가 많기 때문이라고 여겨질 수 있다. 표 7에서는 HMM과 SFHMM의 매개 변수를 비교한다.

표 7 HMM과 SFHMM의 상태 표현에 필요한 매개 변수의 수 비교 (D: 특징 차수, R: 회귀 차수, N: 분절 길이, V: 분산 차수 ($D\times D$ 또는 $1\times D$))

시스템	평균 벡터	분산
HMM	$1\times D$	V
SFHMM (시변분산)	$R\times D$	$N\times V$
SFHMM (고정분산)	$R\times D$	V

디자인 행렬은 분절 길이와 회귀 차수가 정의되면 디자인 행렬의 크기가 고정되고 각 항은 자동적으로 결정되기 때문에 매개 변수의 비교에서 생략하였다. 이 표에서 알 수 있듯이 고정 분산이나 시변 분산이 적용된 경우에 SFHMM의 매개 변수의 수는 일반 HMM보다 많음을 알 수 있다. 그러나, 동적 특징은 프레임 특징의 집합에 대해 평균 변이를 포함하지만 분절 특징은 분절에서의 특징 변화의 경향을 표현하기 때문에 분절 특징은 정적 특징과 동적 특징보다 많은 정보를 포함하고 있다고 할 수 있다.

일반적으로 특징 벡터의 차수를 조절하여 일반 HMM의 매개 변수 수와 SFHMM의 매개 변수 수를 동일하게 할 수 있다. 만약 HMM이 정적 특징과 이들 특징의 일차 미분 계수를 이용하고, SFHMM을 고정 분산 방식을 이용하여 일차 선형 궤적으로 표현한 경우에는 두 시스템의 매개 변수가 동일하게 된다. 이 경우에 별도의 실험을 추가하지 않고 기존의 실험으로부터 상대적인 성능 비교를 할 수 있다. 기존 실험에서 혼합 밀도가 두 개인 경우 고정 분산 방식과 시변 분산 방식의 성능이 비슷함을 알 수 있다. 이 사실과 표 1에서 고정 분산 방

식의 성능을 유추할 수 있다. 그 경우 동일한 매개 변수의 수인 경우에 SFHMM의 성능은 HMM의 성능보다 저하됨을 알 수 있다. 이러한 성능 저하는 평균 궤적의 변이로 설명할 수 있다. 동적인 특성을 나타내는 미분 계수는 분절의 평균 변이를 나타내는 단일 항목으로 표현되며, 분절 특징은 다항식의 궤적으로 표현된다. 분절 특징으로 표현되는 궤적의 변이는 여러 프레임 특징의 변이로 표현되기 때문에, SFHMM에서의 평균 궤적의 변이는 HMM의 평균 벡터의 변이보다 크게 된다. 따라서 SFHMM에서 각 혼합 밀도의 평균 변이를 축소하기 위하여 혼합 밀도의 수를 증가하거나 인식 모델 수를 증가시킴으로써 두 시스템의 성능 차이를 줄일 수 있다. 위의 결과로부터 동일한 매개 변수의 수에서는 성능이 저하되더라도 기존 방법보다 많은 정보를 표현하는 새로운 모델에 대한 연구는 지속되어야 한다고 생각한다. 이와 같은 관점에서 제안한 방식이 기존 HMM보다 매개 변수의 수가 많더라도 특징의 표현 방법이나 인식 방법에서 의미를 지닌다고 할 수 있다.

5. 결 론

본 논문에서는 모수적 궤적 방식에 의한 분절 특징을 이용하는 음성 인식의 새로운 모델링 방식을 제안하였다. 현재 프레임에 대칭적이 되도록 디자인 행렬을 설정함으로써 인접한 음향학적 단위들간의 전이 정보를 표현하였다. 또한 SFHMM의 인식 모델을 HMM과 비슷한 방식으로 제안하고 HMM의 개념을 이용하여 매개 변수 재 추정 알고리즘을 제안하였다. 제안한 방법은 분절 길이와 회귀 차수 등 SFHMM의 구성 요소를 조절하여 일반 HMM 또는 모수적 궤적 방식으로 동작한다.

SFHMM의 유효성과 다항식의 회귀 함수에 기반한 분절 특징의 특성을 살펴보기 위하여 몇 가지 실험을 하였다. 첫째 실험은 정적 특징과 동적 특징을 이용하는 일반 HMM과 정적 특징에 기반한 분절 특징을 사용하는 SFHMM의 성능을 비교하였다. 실험 결과 제안한 분절 특징은 정·동적 특징을 조합한 경우와 같이 효율적임을 알 수 있었다. 또한, 분절의 변이에 대한 특성을 살펴보기 위하여 분절 길이와 회귀 차수를 변경하며 음소 인식 실험을 하였다. 실험으로부터 동일한 분절 길이에 대하여 회귀 차수를 증가시키면 변별력이 증가하고, 동일한 회귀 차수에 대하여 분절 길이를 증가시키면 전이 정보량이 증가함을 알 수 있었다. 긴 분절이 사용되고 높은 회귀 차수가 사용된 경우, 일반적인 HMM보다 성능이 지속적으로 향상되어 제안한 SFHMM이 인접한 음성 프레임간의 시간적인 상관관계를 잘 표현하고 명확한 문맥 정보가 도입되지 않더라도 부분적인 문맥 정보를 표현할 수 있다고 할 수 있다. 마지막 실험에서는 성능의 저하 없이 매개 변수의 수를 줄이고자 시도하였다. 매개 변수 수를 줄이기 위하여 특정 상태의 분절 변이를 표현하는 다양한 분산 표현 방법을 고려하였다. 분절의 분산 표현 방법으로 분절을 구성하는 각 프레임에 대해 개별적인 분산을 표현하는 시변 분산과 하나의 공통 분산으로 표현하는 고정 분산을 제안하고, 각 분산 표현방법을 이용하여 성능 비교를 하였다. 비교 결과 단일 혼합 밀도를 이용한 경우에는 시변 분산을 이용한 시스템이 고정 분산을 이용한 경우보다 성능이 우수하였으나, 혼합 밀도의 수를 증가시킨 경우 두 시스템의 성능 차이는 작았다. 이런 결과로부터 작은 분절 단위에서는 고정 분산이 어느 정도 효과가 있음을 알 수 있었다.

제안한 방식이 기존의 HMM 방식보다 성능이 우수하다고는 하나, 현재로서는 많은 매개 변수를 필요로 한다. 따라서 추후 연구로 성능의 저하없이 매개 변수를 줄이는 연구가 지속되어야 한다. 또한 일반적인 어휘 사전 모델에서는 음소 모델을 단순한 병합에 의하여 단어와 같은 상위 언어 모델을 구축하기 때문에, 분절 특징에 알맞은 모델에 대한 연구가 필요하다고 본다. 본 연구는 연속 음성 인식 과정에서 중요한 역할을 담당하는 음향학적 모델의 성능 향상에 주안점을 두고 있기 때문에 객관적인 성능 평가를 위하여 널리 사용되고 있는 영문 데이타베이스를 이용하였다. 그러나 한국어는 발성 방법이나 언어학적 관점에서 영어와는 많이 다르기 때문에 한국어의 특성에 맞는 음성 인식 시스템의 개발에 관한 연구도 필요하리라 본다.

참 고 문 헌

[1] Holmes, W.J. and Russell, M.J., "Probabilistic-trajectory segmental HMMs," Computer Speech and Language, vol 13, pp. 3-37, 1999.

[2] Deng, L. and Aksmanovic, M. and Sun, Du. and Wu, J., "Speech recognition using hidden Markov models with polynomial regression functions as non-stationary states," IEEE Trans. on Speech and Audio Proc., vol. 2, no. 4, pp. 507-520, 1994.

[3] Furui, S. "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 34, no. 1, pp. 52-59, 1986.

[4] Deng, L. "A generalized hidden Markov model with state-conditioned trend functions of time for speech signal," Signal Processing, vol. 27, pp. 65-78, 1992.

[5] Gish, H. and Ng, K. "A segmental speech model with application to word spotting," In International Conference on Acoustics, Speech and Signal Processing 1993, vol. 2, pp. 447-450, Minneapolis, Minnesota, 1993.
[6] Russell, M. "A segmental HMM for speech pattern modeling," In International Conference on Acoustics, Speech and Signal Processing 1993, vol. 2, pp. 499-502, Minneapolis, Minnesota, 1993.
[7] Gales, M.J.F. and Young, S.J. "The Theory of Segmental Hidden Markov Models," CUED/F-INFENG/TR 133, Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England, 1993.
[8] Ostendorf, M. and Roukos, S. "A stochastic segment model for phoneme-based continuous speech recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 37, no. 2, pp. 1857-1869, 1989.
[9] Ostendorf, M. and Digalakis, V. and Kimball, O.A. "From HMM's to Segmental Models: A Unified View of Stochastic Modeling for Speech Recognition," IEEE Trans. on Speech and Audio Processing, vol. 4, no. 5, pp. 360-378, 1996.
[10] Press, W.H. and Teukolsky, A.A. and Vetterling, W.T. and Flannery, B.P. Numerical Recipes in C, 2nd Ed. Cambridge University Press, pp. 671-680, 1992.
[11] Gish, H. and Ng, K. Parametric trajectory models for speech recognition. In International Conference on Spoken Language Processing 1996, pp. 466-469, Philadelphia, Oct. 1996.
[12] Lee, K. and Hon, H. Speaker-independent phone recognition using hidden Markov models, IEEE Trans. On Acoustics, Speech and Signal Processing, vol. 37, no 11, pp.1661-1648, Nov. 1989.
[13] Fukada, T. and Sagisaka, Y. and Paliwal, K. Model Parameter Estimation For Mixture Density Polynomial Segment Models, In International Conference on Acoustics, Speech and Signal Processing 1997, Munich, Germany, pp. 1403-1406, April 1997.
[14] 최인정, HMM에 기반한 음성 인식에서 음향학적 문맥 정보의 결합, 박사학위 논문, KAIST, 1999.

윤 영 선
1990년 2월 한국과학기술원 전산학과(학사). 1992년 2월 한국과학기술원 전산학과(석사). 1992년 3월 ~ 1995년 7월 (주)핸디소프트 주임연구원. 1995년 9월 ~ 2001년 2월 한국과학기술원 전산학과(박사). 2000년 7월 ~ 2001년 9월 (주)보이스피아 감사. 2001년 3월 ~ 현재 한남대학교 정보통신·멀티미디어 공학부 조교수. 관심분야는 음성 인식, 패턴 인식, 음성 정보 검색 등