모수적 궤적 기반의 분절 HMM을 이용한 연속 음성 인식

Continuous Speech Recognition based on Parmetric Trajectory Segmental HMM

윤 영 선*, 오 영 환*

(Young-Sun Yun*, Yung-Hwan Oh*)

요 약

본 논문에서는 음성 패턴을 효율적으로 모델링하고자 분절 특징(segmental feature)을 이용하여 은닉 마코프 모델(hidden markov model)의 일반적인 형식에 기반한 새로운 모수적 궤적 모델 (parametric trajectory model)을 제안한다. 일반적으로 벡터의 열로써 표현되는 분절은 관측 열의 궤적(trajectory)으로 표현된다. 이 궤적은 연속적인 프레임들의 전이 정보(transitional information)를 표현하는 디자인 행렬을 이용하여 얻어지며, 다항식의 회귀 함수(polynomial regression function)로써 나타낼수 있다. 이러한 궤적을 HMM에 적용하기 위해서 프레임 특징 대신 분절의 특성을 표현하는 궤적으로 대치하고 우도(likelihood) 계산에 궤적들의 비교에 의한 확률 값을 반영시켜야 한다. 본 논문에서는 궤적간의 유사도를 측정하는 분절 우도(segment likelihood)와 모델을 구성하는 궤적 변수의 추정 알고리즘을 제안한다. 임의의 분절에 대한 관측 확률은 제안된 분절 우도와 궤적의 추정 오차(estimation error of trajectories)의 곱으로써 표현된다. 궤적의 추정 오차는 상태에서 주어진 분절 우도의 가중치로 표현될 수 있으며, 이 가중치는 궤적과 대응되는 분절의 적합도를 표현하는 확률을 나타낸다. 본 논문에서 제안된 모델은 일반적인 HMM과 모수적 궤적 모델의 일반화(generalization) 또는 확장(extension) 모델로 생각될 수 있다. 본 모델의 성능을 평가하기 위하여 TIMIT 데이터에 기반한 실험을 한 결과, 분절 길이(segment length)와 회귀 차수(regression order)가 변할수록 일반적인 HMM에 비하여 뚜렷한 성능향상이 있음을 알 수 있었다.

핵심 용어: 모수적 궤적 모델, 분절 HMM, 모수적 궤적 분절 HMM, 음성 인식

ABSTRACT

In this paper, we propose a new trajectory model for characterizing segmental features and their interaction based upon a general framework of hidden Markov models. Each segment, a sequence of vectors, is represented by a trajectory of observed sequences. This trajectory is obtained by applying a new design matrix which includes transitional information on contiguous frames, and is characterized as a polynomial regression function. To apply the trajectory to the segmental HMM, the frame features are replaced with the trajectory of a given segment. We also propose the likelihood of a given segment and the estimation of trajectory parameters. The obervation probability of a given segment is represented as the relation between the segment likelihood and the estimation error of the trajectories. The estimation error of a trajectory is considered as the weight of the likelihood of a given segment in a state. This weight represents the probability of how well the corresponding trajectory characterize the segment. The proposed model can be regarded as a generalization of a conventional HMM and a parrametric trajectory model. The experimental results are reported on the TIMIT corpus and performance is show to improve significantly over that of the conventional HMM.

Key words: Parametric trajectory model, Segmental HMM, Parametric trajectory segmental HMM, Speech recognition
투고 분야: 음성처리(2.5)

Ⅰ. 서 론

1960년대 이후로 널리 연구되고 많이 사용되는 은닉 마코프 모델(HMM; hidden Markov model)은 시간적·공간적인 특징을 잘 반영하는 이중 통계적 방법으로 음성 인식을 포함한 다양한 분야에서 성공적으로 적용되고 있다. 그러나, HMM이 음향적인 음성 신호의 통계적인 변이를 잘 모델링하고, 다른 방법에 비하여 성능이 우수하지만, HMM의 이론을 구성하는 기본 가정에 대한 검토가 필요하다. HMM의 기본 가정 [1, 2]은

*한국과학기술원 전산학과
접수일자 : 1999년 10월 6일

크게 두 가지로 표현되는데, 실제 음성 신호의 특성을 제대로 표현하지 못하고 있다. 첫째 가정은 각각의 관측 시간에서 새로운 상태(state)는 바로 이전의 상태에 대해서만 조건 확률을 갖는다는 1차 마코프 가정이다. 이 가정으로 인하여 상태에서의 출현 확률은 그 상태에서의 지속시간에 따라 지수적으로(exponentially) 감소하게 되므로, 음성 신호의 시간적 구조를 제대로 표현하지 못하게 된다. 이러한 문제점을 해결하기 위하여 상태에서의 지속 시간을 모델링하기 위한 다양한 방법들이 제시되었다. 다음으로 관측 독립(observation independent) 가정을 들 수 있다. 이것은 임의의 상태에서 관측 벡터는 그 상태에서 생성된(관측된) 벡터에 독립적으로 출현한다는 것을 뜻한다. 그러나, 인접한 관측 벡터들은 서로 밀접하게 관련되어 있어, 실제적으로 상태에 대한 관측 벡터들의 출현을 제대로 모델링하지 못하게 된다. 또한, 음성 신호의 특징 중에서 인접한 특징들에 의해 생성되는 정보가 중요하다는 연구[3,4,5,6]가 발표되고 있어, 시계열상에서의 포괄적이고 유연한 음성 특징의 모델링에 대한 연구가 필요하다는 것을 알 수 있다.

이러한 관점에서 HMM의 약점을 보완하기 위한 연구들이 진행되어 왔는데, 대표적인 연구 방식으로는 비정상 상태(non-stationary state) 해석에 의한 지속 시간 모델링(duration modeling)[7], 분절 모델(segmental model)[8], 다항식에 의한 궤적 모델(polynomial or parametric trajectory model) [3, 4], 분절 HMM(segmental HMM) [5] 등이 있다. 이들 연구 방법들은 전통적인 HMM의 약점을 보완하기 위하여 프레임 수에 대한 회귀 함수나 분절 특징을 이용하여 지속 시간 또는 관측 독립 가정을 보완한다. 먼저 분절 HMM은 "주어진 상태에서 발생된 모든 관측들은 그 상태와는 독립적이며, 관측들이 속해 있는 분절에 대해 조건 지어진다"고 가정한다[6]. 기존의 연구들은 분절을 대표하는 궤적을 시간과는 독립적인 상수 분산으로 정의하거나 [3], 다양한 지속시간에 대한 일차 선형 시스템으로 해석하였다[9]. 따라서, 분절을 표현하는 궤적에 대한 가정이 변경되면, 주어진 상태에서의 분포 변수들은 다시 정의되어야 한다. 이와 달리 HMM에 기반하여 상태에서의 평균 관측 벡터를 지속시간에 의하여 추정하는 비정상 상태 HMM(non-stationary state HMM) 또는 경향 HMM(trended HMM)은 주어진 상태를 프레임 수 [10]나, 상대적인 프레임 위치에 의한 회귀 함수 (regression function)로 정의하고 있다[7]. 이들 모델링 방법이 효과적으로 지속시간을 표현한다 할지라도 프레임 특징을 사용하고 있으며, 여전히 관측 독립 가정이 사용되고 있어, HMM의 약점을 보완했다고 하기 어렵다. 궤적을 이용한 모델링 방법에서는 분절에서의 관측인 실제 관측을 이용하는 것이 아니고, 추정된 전역 궤적(global or complete trajectory)의 선형 샘플링으로 표현한다. 이 모델의 경우 분절 특징에 대한 평활화 효과를 보이고 있어 잡음이나 화자의 개인성 등에 의한 음성의 변형을 효과적으로 모델링할 수 있다[3,4]. 그러나, 분절을 전역 궤적으로 표현하기 위해서는 분절의 길이를 알고 있어야 하는 문제점이 있다. 또한, 여러 관측 벡터들을 이용하여 회귀 함수나 평균 궤적을 추정하기 위해서는 다중 관측 벡터들을 하나의 벡터로 연장하여 추정해야 하기 때문에, 계산 시간이 길어지고 복잡도가 증가하게 된다.

본 논문에서는 여러 연구에서 사용되고 있는 분절 기반의 접근 방식을 이용하면서, 문제점으로 지적되고 있는 경계문제(boundary problem)나 제한된 궤적의 표현 방법을 완화시키고자 한다. 제안된 방법은 일반적인 HMM에 기반하여 유연성 있고 확장이 쉽도록 모수적 궤적 모델과 분절 HMM을 결합한다(PTSHMM;parametric trajectory segmental HMM). PTSHMM은 분절의 특징을 다항식에 의한 궤적으로 표현하고, 추정된 궤적을 분절 HMM의 입력으로 사용한다. 각 궤적은 분절을 구성하는 프레임들간의 인접 정보를 포함할 수 있도록 디자인 행렬을 개선하여 구한다. 인접하는 프레임 정보를 반영하기 위하여, 제안된 디자인 행렬은 현재의 관측 벡터에 대하여 대칭적인 형태를 띄게 되며, 분절 길이에 종속적이지 않도록 시간에 대하여 정규화 된 형태를 갖게 된다. 기존의 연구와 달리, 본 연구에서는 분절의 길이를 고정시켜, 평균 궤적은 기대 평균(expected average)값으로 쉽게 구할 수 있도록 하여 궤적을 추정하는 시간이 감소되도록 하였다. 또한, 궤적의 추정 오차를 분절간의 외적 분절 변이(extra-segmental varition)에 대한 가중치로 사용한다. PTSHMM의 추정(estimation)과 평가(evaluation) 문제는 일반적인 HMM의 프레임 특징을 분절 특징으로 대치하고, 그에 대한 가중치로 추정 오차를 이용하기 때문에 일반 HMM의 추정, 평가 문제와 유사하게 된다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 모수적 궤적 방법을 이용한 분절 모델링(segment modeling)에 대하여 소개하고, 인접하는 음향학적 정보를 반영하도록 새로운 디자인 행렬을 제안한다. 또한, 분절 HMM의 기본 개념과 표현 방법을 소개한다. 3장에서는 PTSHMM의 상태에서의 우도 (likelihood) 계산법을 제안하고, 그 변수들을 추정하는 알고리즘을 설명한다. 다음으로는 PTSHMM이 주어지는 조건에 따라 일반적인 HMM 또는 모수적 궤적 모델로 표현될 수 있음을 설명한다. 4장에서는 제안된 방법의 유효성을 검증하기 위하여 TIMIT 자료에서의 음소 인식 성능을 연속

HMM과 비교하며, 마지막으로 결론을 맺도록 한다.

Ⅱ. 음성 분절 모델링

음성 신호의 연속적인 음향 특징 벡터들간의 관계는 특징 공간에서 궤적의 형태로 근사 될 수 있다. 이러한 생각은 여러 분절 모델의 기본이 되었으며, 모수적(parametric) 또는 비모수적(non-parametric) 방식으로 구현되었다. 모수적 방법에서는 특정 영역에 대하여 다항식을 이용하여 궤적을 추정하고, 그 영역에 대한 분포는 궤적 위의 점들로서 표현이 된다. 반면, 비모수적 궤적 모델은 각각의 모델 영역에 대한 분포 변수들을 갖는다. 본 논문에서는 모수적 방법이 여러 음성 단위에서 궤적의 평활화 효과를 가져오기 때문에 잡음이나 환경 변화, 화자 변화에 강인하다는 장점이 있어[8], 분절 모델링에 모수적 방법을 채택하였다.

2.1. 분절 모델링

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

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

여기에서 각각의 프레임은 D차원의 특징 벡터로 구성되어 있으며, $\mathbf{Z}$는 사용된 윈도우의 형태를 결정하는 $N \times R$ 크기의 디자인 행렬이다. 또한, $\mathbf{B}_t$는 궤적 계수를 나타내는 $R \times D$ 행렬을 나타내며, $\mathbf{E}$는 궤적 추정에서의 잔차 오차(residual error)를 표현한다. R은 궤적의 특징을 결정하는 회귀 차수 (regression order)을 나타낸다. 만약 R=1이면 평균 (상수)을, R=2이면 일차 선형 시스템, R=3이면 2차 방정식의 궤적을 표현하며, R 이 증가함에 따라 분절의 특징을 세밀하게 표현할 수 있다.

기존의 연구에서와 같이 음성 분절을 전역 궤적으로 모델링는 경우, 분절의 범위는 [0~1]로 정규화 되기 때문에, 분절의 경계를 반드시 알아야 한다. 이러한 문제점을 해소하기 위하여, 우리는 입력된 음성을 고정된 분절로 분할하고 그 고정 분절을 모수적 궤적으로 표현하여 분절 HMM에 적용시키고자 한다. 분절 HMM을 적용하게 되면, 평가단계(evaluation phase)에서 Viterbi 알고리즘을 이용하여 쉽게 분절 경계를 조정할 수 있게 된다. 분절의 길이를 고정시켰기 때문에 시작과 끝 부분을 제외한 거의 모든 음성 분절은 동일한 디자인 행렬 형태를 이용할 수 있어, 추정 단계에서는 각 궤적의 기대 평균 값으로 평균 궤적을 예측할 수 있다.

만약 정규화 된 디자인 행렬을 적용한다면, 각 분절의 특징은 분절 길이에 독립적이도록 상대적인 시간을 표현할 수 있을 것이다. 그러나 분절의 길이를 모르기 때문에 Gish와 Ng(1993)처럼 정규화 된 디자인 행렬을 이용한 전역 궤적을 분절 HMM의 입력으로 적용하기 어렵다. 따라서, 본 논문에서는 디자인 행렬의 열이 현재의 관측벡터에 대해 대칭적이고, 시간 축에 대하여 정규화 되도록 하여 경계의 조정 문제를 해결하도록 한다. 이때, 가변적인 분절 길이를 이용하지 않고 고정적인 분절 길이를 이용함으로써 정규화 시 미리 분절 경계를 알고 있어야 하는 제약을 완화시킨다.

분절길이 $N = 2M+1$ 프레임인 음성 분절이 주어지면, 관측 벡터들은 다음의 행렬로서 표현된다.

$$\mathbf{C}_t = \mathbf{Y}_{t-M}^{t+M} = \begin{bmatrix} y_{t-M,1} & \cdots & y_{t-M,D} \\ \vdots & \vdots & \vdots \\ y_{t,1} & \cdots & y_{t,D} \\ \vdots & \vdots & \vdots \\ y_{t+M,1} & \cdots & y_{t+M,D} \end{bmatrix} = \begin{bmatrix} \mathbf{c}_{t-M} \\ \vdots \\ \mathbf{c}_t \\ \vdots \\ \mathbf{c}_{t+M} \end{bmatrix} \tag{2}$$

$$c_\tau = \begin{bmatrix} y_{\tau,1} & \cdots & y_{\tau,D} \end{bmatrix}, \quad t-M \le \tau \le t+M$$

위 식에서 볼 수 있듯이 시간 t 에서 현재 프레임 벡터가 분절의 중앙에 오기 때문에 음성 분절의 앞부분과 뒷부분은 시간 t -1이나 t +1의 음성 분절과 겹치게 된다. 따라서, 위와 같은 분절을 분석하기 위한 새로운 디자인 행렬의 구상이 필요하다. 새로운 디자인 행렬은 인접한 분절간의 전이 정보(transitional information)를 표현할 수 있으며, 또한 현재의 프레임벡터가 중앙에 오도록 배치할 수 있어야 한다. 위의 조건을 만족할 수 있도록 디자인 행렬 $\mathbf{Z}$는 다음과 같이 정의될 수 있다.

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

$$\mathbf{z}_\tau = \begin{bmatrix} 1 & \left(\frac{\tau-t}{2M}\right) & \left(\frac{\tau-t}{2M}\right)^2 & \cdots & \left(\frac{\tau-t}{2M}\right)^{R-1} \end{bmatrix} \tag{3}$$

여기에서 $\mathbf{Z}$는 분절 길이로 정규화 된 상대적인 위치를 나타내기 때문에, 현재 관측 벡터의 앞부분 또는 다음에 오는 음향학적 특징은 궤적에 포함될 수 있다. 따

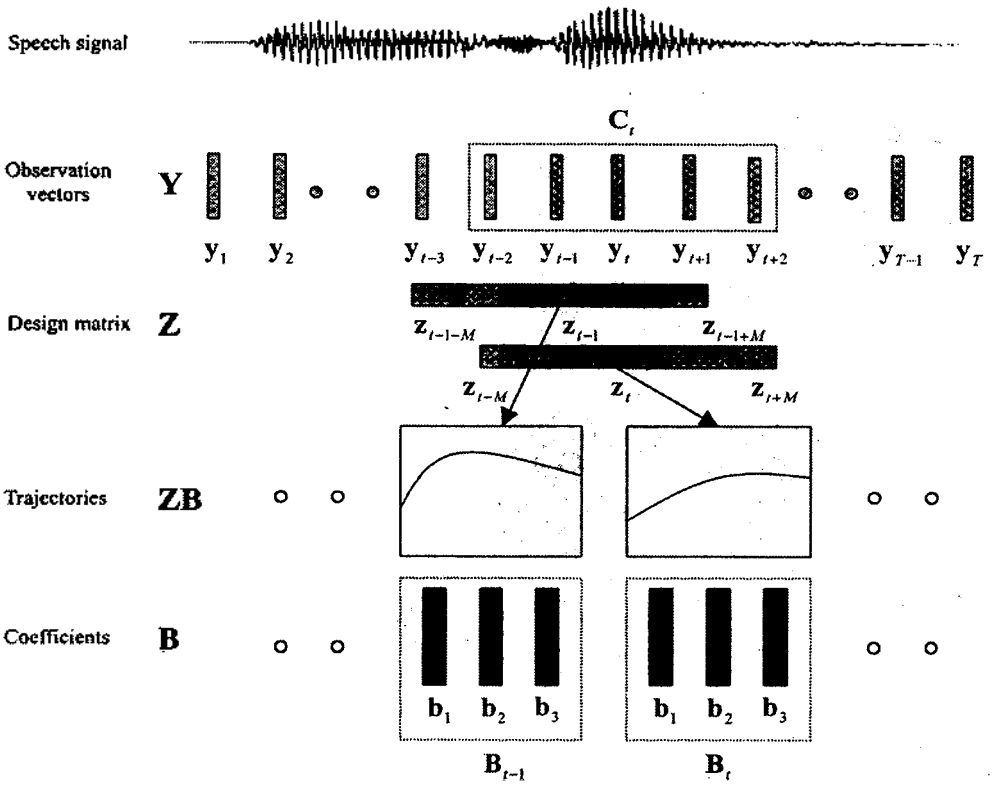


그림 1. 관측 벡터와 디자인 행렬, 그리고 대응되는 궤적과 궤적 계수 행렬간의 관계
Fig. 1. The relation among the observation vectors, the design matrix, the trajectories and their coefficient.

라서, 인접한 단위들이 문맥 독립 모델(context independent model)로 모델링 되더라도, 생성되는 궤적은 부분적인 문맥 정보를 반영하게 된다. 디자인 행렬의 각 열에서의 기본 $\left(\frac{\tau-t}{2M}\right)$은 τ가 $t-M$ 에서부터 $t+M$ 까지의 값을 갖게 되므로, -0.5에서부터 0.5까지의 정규화 된 값을 갖는다. 비슷한 방법으로 궤적의 계수 행렬 $\mathbf{B}_t$는 다음과 같이 정의 된다.

$$\mathbf{B}_t = \begin{bmatrix} b_{1,1} & \cdots & b_{1,D} \\ \vdots & & \vdots \\ b_{R,1} & \cdots & b_{R,D} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_R \end{bmatrix}$$

$$\mathbf{b}_i = \begin{bmatrix} b_{i,1} & \cdots & b_{i,D} \end{bmatrix}, \quad 1 \le i \le R \tag{4}$$

그림 1은 음성 신호가 입력되었을 때의 시스그램 흐름을 보여주며, 그때의 관측 벡터와 디자인 행렬, 대응되는 궤적과 계수 행렬간의 관계를 나타낸다.

식 (1)~(4)와 같이 분절 모델이 주어지면, 다음 단계는 모델 변수 들을 추정하는 것이다. 모든 오차가 독립적이고 균등하게 분포되어 있다면 (i.i.d.; independent and identically distributed) 궤적 계수 행렬 $\hat{\mathbf{B}}_t$는 선형 회귀 방식 (linear regression approach)에 의하여 추정될 수 있다. 선형 회귀 방식을 적용시키기 위하여는 각 특징 차원에 대하여, 다음의 다항식을 고려할 수 있다.

$$y_{\tau,i} = b_{1,i}z_{\tau,1} + b_{2,i}z_{\tau,2} + b_{3,i}z_{\tau,3} + \Lambda + b_{R,i}z_{\tau,R}, \quad 1 \le i \le D \tag{5}$$

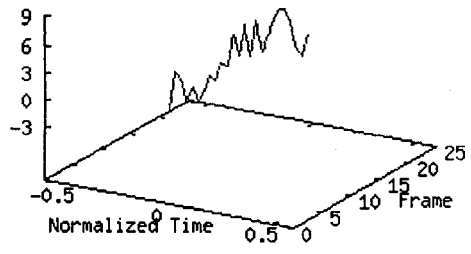
여기에서 $z_{\tau,r} = \left(\frac{\tau-t}{2M}\right)^{r-1}$ 이다. 이러한 선형 회귀 방정식은 Singular Value Decomposition(SVD)에 의하여 쉽게 풀릴 수 있다. 만약 궤적 모델이 over-determined 시스템이라면, 즉, $N > R$ 라면, SVD는 최소 자승 오류(least squared error)의 개념에서 최적의 근사 값을 구할 수 있다[11]. 그렇지 않고 행렬 연산에 의하여 궤적 계수 행렬 $\hat{\mathbf{B}}$를 구한다면 다음과 같이 구할 수 있다.

$$\hat{\mathbf{B}} = [\mathbf{Z}'\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{C} \tag{6}$$

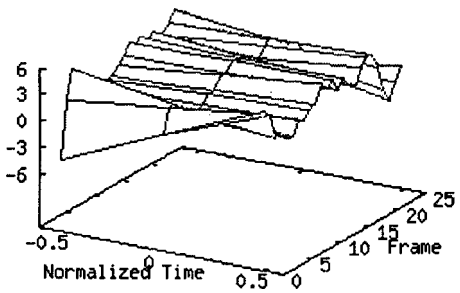
여기에서 $'$는 행렬의 전치(matrix transpose)를 뜻한다. 위와 같은 방법으로 궤적 계수행렬을 추정하기 위해서는, 시작 부분과 끝부분이 다른 부분의 분절 길이보다 작기 때문에 음성 신호의 시작부분과 끝부분에 대하여 디자인 행렬을 조정해야 한다.

그림 2는 0차 MFCC 계수의 시간적 변이와 N=3, 5 일 때, 대응되는 궤적 계수로부터 복원된 분절 특징을 보여주고 있다.

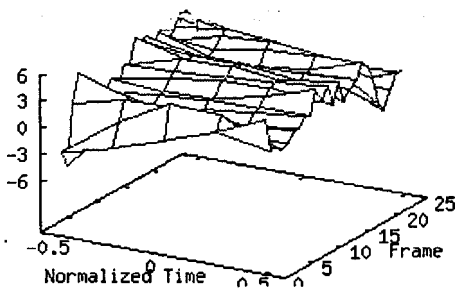
$\hat{\mathbf{B}}_t$가 추정된 후, 시간 t에서의 분절에 속하는 프레임별 잔차 오차를 더하여 적합도(goodness of fit) χ^2를 다음과 같이 계산할 수 있다.



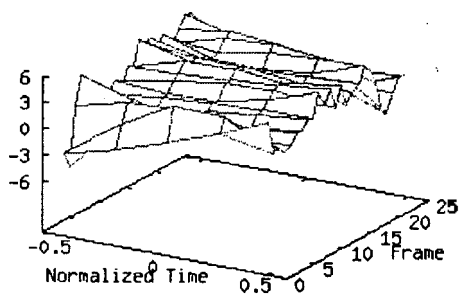
(a) 원래의 0차 특징



(b) N=3, R=2



(c) N=3, R=3



(d) N=5, R=3

그림 2. /oy/ 모음 분절에 대한 0차 MFCC계수의 시간적 변이와 분절 길이 N=3과 N=5이고 회귀 차수 R=2또는 3인 경우, 그에 대응되는 궤적으로부터 복원된 프레임 특징

Fig. 2. Temporal variation of the 0th MFCC coefficient for /oy/ vowel segment, and the corresponding segmental trajectory features with N=3 and 5, respectively. N means the segment length and R the regression order for trajectories.

$$\chi^2 = \frac{\sum_{\tau=t-M}^{t+M}\left(\mathbf{c}_\tau - \mathbf{z}_\tau\hat{\mathbf{B}}_t\right)\left(\mathbf{c}_\tau - \mathbf{z}_\tau\hat{\mathbf{B}}_t\right)'}{N} \tag{7}$$

여기에서 N은 분절을 구성하는 프레임 수를 나타낸다. 위의 식의 의미는 "만약 χ^2가 작다면 분절에 대해 궤적이 잘 추정되었다"는 것을 나타낸다. 따라서, χ^2가 0이라면, 완벽한 추정을 의미하게 된다. 이들 변수에 대한 예측이 끝나면, 각각의 분절 특징은 분절에 대한 계수 행렬 $\hat{\mathbf{B}}_t$와 적합도 χ^2로 표현된다.

2.2. 분절 HMM

분절 HMM은 음성 신호의 잠재된 궤적을 효과적으로 표현하는 방법이라고 생각되고 있다. Russell 등은 궤적을 고정 분산(constant variance, 1993) 또는 선형 시스템(linear system, 1995)으로 표현하는 가우시안 통계적 과정(Gaussian stochastic process)으로 표현하고 있다. 이러한 경우에 모델 λ의 상태 s_i에서 주어진 분절 $\mathbf{Y} = y_1, \Lambda, y_T$에 대한 관측 확률은 표현가능한 궤적 f_m에 대하여 다음과 같이 정의된다.

$$P(\mathbf{Y} \mid s_i, \lambda) = \int P(f_m \mid s_i, \lambda) P(\mathbf{Y} \mid f_m, s_i, \lambda)\, \mathrm{dfm} \tag{8}$$

여기에서 $P(f_m \mid s_i, \lambda)$은 f_m이 상태에 대응되는 평균 분절 특징과 궤적 f_m과의 적합성을 나타내는 확률을 나타내며, 외적 분절 변이(extra-segmental variation)라 한다.

반면에 $P(\mathbf{Y} \mid f_m, s_i, \lambda)$은 궤적 f_m이 주어졌을 때, 관측 분절 $\mathbf{Y}$이 제대로 표현되는지를 나타내는 확률 값을 표현하며, 특정 궤적과 분절 특징의 연관성을 표현하는 내적 분절 변이(intra-segmental variation)를 나타낸다. 외적 분절 변이는 화자의 특징이나 선택된 화자의 발음과 같이 장기적인 가변성(long-term variability)을 가리킨다. 그러나 내적 분절 변이는 연속적인 조음 과정이나 불규칙한 변이 등에서 발생되는 분절 내부의 단기적인 가변성을 나타내고 있다. 이 두 가지 분절 변이는 "모든 관측은 주어진 상태에 대해서는 독립적이나 그 분절에 대해서는 조건 지어진다"는 가정에 기반하고 있다[5,6,9].

$$\begin{aligned} P(\mathbf{C}_t \mid s_i, \lambda) &= \int P(f_m \mid s_i, \lambda) P(\mathbf{C}_t \mid f_m, s_i, \lambda)\, \mathrm{dfm} \\ &= P(\mathbf{ZB}_t \mid s_i, \lambda) P(\mathbf{C}_t \mid \mathbf{ZB}_t, s_i, \lambda) \end{aligned} \tag{9}$$

이런 특징을 갖는 상태에서의 관측 확률을 궤적 모델에 기반하여 고려하면, 시간 t에서의 관측 벡터 $\mathbf{C}_t = \mathbf{Y}_{t-m}^{t+m}$은 단 하나의 궤적 계수 행렬 $\mathbf{B}_t$로 표현이 되기 때문에 식 (8)은 다음과 같이 작성할 수 있다.

만약 혼합 모델(mixture model)이 사용된다면 주어진 상태에서의 분절 $\mathbf{C}_t$의 관측 확률은 상태에서의 모든 혼합 모델에 대한 확률을 더함으로써 구할 수 있다. 따라서, 식 (9)는 다음과 같이 혼합 모델로 표현할 수 있다.

$$P(\mathbf{C}_t \mid s_i,\lambda) = \sum_{k=0}^{K-1} c_{ik} P(\mathbf{C}_t \mid s_i,m,\lambda) = \sum_{k=0}^{K-1} c_{ik} P(\mathbf{ZB}_t \mid s_i,m,\lambda) P(\mathbf{C}_t \mid \mathbf{ZB}_t,s_i,m_k,\lambda) \tag{10}$$

여기에서 c_{ik}는 상태 s_i에서의 k번째 혼합 m_k의 가중치를 나타낸다.

그러나, 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.

$$P(\mathbf{C}_t \mid s_i,\lambda) = \max_k P(\mathbf{C}_t \mid s_i,m_k,\lambda) = \max_k P(\mathbf{ZB}_t \mid s_i,m_k,\lambda) P(\mathbf{C}_t \mid \mathbf{ZB}_t,s_i,m_k,\lambda) \tag{11}$$

Ⅲ. 모수적 분절 HMM

PTSHMM은 분절을 고정된 지속 시간을 이용하여 표현한 다항식의 궤적으로 모델링한다. 따라서, 분절을 표현하는 궤적 계수 행렬 $\mathbf{B}_t$와 적합도 χ^2가 추정되면, 이들 변수들은 음성 분절의 우도 (likelihood) 값을 계산하는데 사용된다. 본 장에서는 모델에 대한 새로운 분절 우도 값과 일반적인 HMM 개념 위에서 변수들을 추정하는 알고리즘을 제안한다.

3. 1. 분절 우도(segment likelihood)

기존의 분절 HMM에서는 적용되는 분포의 가정으로 인하여 외적 분절 변이에 대한 궤적은 고정 분산이나 선형 시스템으로 제한되었으며, 내적 분절 변이에 대한 분포는 대각선의 공분산으로 표현되는 가우시안 분포(Gaussian distribution)을 사용하였다[5,9]. 그러나 본 논문에서는 다음과 같이 외적 분절에 대한 분포를 정의하고자 한다. 외적 분절의 분포는 평균 궤적과 그 분산으로 표현되며, 내적 분절은 분절에서의 궤적 추정 오차로 정의한다. 이것은 내적 분절의 변이가 상태에 주어진 궤적의 관측 확률을 나타내는 외적 분절에 대한 가중치로 기여한다는 것을 의미한다. 여기에서 가중치는 주어진 궤적인 분절의 특징을 얼마나 잘 표현하는 가에 대한 척도를 나타낸다.

모델 λ와 상태 s_i가 주어지면 궤적 $\mathbf{ZB}_t$의 외적 분절 확률은 다음과 같이 계산된다.

$$P(\mathbf{ZB}_t \mid s_i,\lambda) = P(\mathbf{ZB}_t \mid \mathbf{ZB}_i,\Sigma_i) = \prod_{\tau=t-M}^{t+M} \frac{1}{(2\pi)^{D/2} \mid \Sigma_{\tau-t,i} \mid^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z}_\tau(\mathbf{B}_t-\mathbf{B}_i))\sum_{\tau-t,i}^{-1}(\mathbf{z}_\tau(\mathbf{B}_t-\mathbf{B}_i))'\right\} \tag{12}$$

여기에서 $\mathbf{B}_i$는 상태 s_i에서의 평균 궤적 계수 행렬을 나타내며, Σ_i는 궤적의 분산을 표현한다(분산 행렬로 이루어진 벡터이다). 또는 $\Sigma_{n,i}$, $-M \le n \le M$는 상태 s_i에서의 평균 궤적 위의 점들에 대한 분산을 표시한다. 이때 n은 궤적에 대응되는 분절 내에서의 프레임 인덱스를 나타낸다. 상태에서 관측되는 분절의 분산을 표현하는 방법에는 분절의 특징을 표현하는 궤적에 대해 공통적으로 분산(common variance)을 이용하는 방법과, 시간적인 종속성을 반영하는 시변 분산(time varying variance)을 이용하는 방법으로 구분할 수 있다. 공통분산은 분절에서 시간적인 변이를 반영하지 못하기 때문에 고정 분산(fixed variance)이라 표현하기도 한다. 이 경우, Σ_i의 각 원소는 분절내의 모든 프레임 인덱스 $\tau-t$에 대해 동일한 분산값을 갖는다. 즉, $\Sigma_i = \{\Sigma_{-M,i},\ldots,\Sigma_{0,i},\ldots,\Sigma_{M,i}\}$, $\Sigma_{n,i} = \Sigma_{m,i}, \forall n,m$으로 표현할 수 있다. 그러나, 시변 분산은 분절에서의 시간적인 역학(temporal dynamics)을 표현하기 때문에 식 (12)에서의 $\Sigma_{\tau-t,i}$는 분절에서의 상대적인 프레임 인덱스 $\tau-t$에 따른 프레임 특징의 분산을 표현한다.

내적 분절 변이는 궤적의 추정 오차를 나타내기 때문에 상태 s_i와 독립적이므로, 변이는 다음과 같이 적합도 χ^2를 이용하여 정의될 수 있다.

$$P(\mathbf{C}_t \mid \mathbf{ZB}_t,s_i,\lambda) = P(\mathbf{C}_t \mid \mathbf{ZB}_t) \approx \exp\left\{-\frac{1}{2}\chi^2\right\} = \exp\left\{-\frac{1}{2N}\sum_{\tau=t-M}^{t+M}(\mathbf{c}_\tau-\mathbf{z}_\tau\mathbf{B}_t)(\mathbf{c}_\tau-\mathbf{z}_\tau\mathbf{B}_t)'\right\} \tag{13}$$

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

$$b_j(\mathbf{C}_t) = P(\mathbf{C}_t \mid s_j,\lambda) = \sum_{k=0}^{K-1} c_{jk} b_{jk}(\mathbf{C}_t) = \sum_{k=0}^{K-1} c_{jk} P(\mathbf{C}_t \mid s_j,m_k,\lambda) = \sum_{k=0}^{K-1} c_{jk} P(\mathbf{ZB}_t \mid \mathbf{ZB}_{jk},\Sigma_{jk}) P(\mathbf{C}_t \mid \mathbf{ZB}_t) = P(\mathbf{C}_t \mid \mathbf{ZB}_t)\sum_{k=0}^{K-1} c_{jk} P(\mathbf{ZB}_t \mid \mathbf{ZB}_{jk},\Sigma_{jk}) \tag{14}$$

여기에서 $\mathbf{B}_{jk}$와 Σ_{jk}는 상태 j에서의 혼합 밀도 m_k에 대응되는 궤적 모델을 나타낸다.

위식에서 볼 수 있듯이, $P(\mathbf{C}_t \mid \mathbf{ZB}_t)$는 상태 j에 독립적이며 분절에 대한 궤적의 추정 오차를 표현하고 있기 때문에, 내적 분절 변이는 주어진 상태에서의 외적

분절 변이 확률에 대해 기여하는 시변 가중치(time-varying weight)로 생각될 수 있다.

3. 2. 변수 추정(Parameter estimation)

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

$$\xi_t(j,k) = P(s_t = j, k_t = k \mid \mathbf{C}_t, \lambda) = \frac{\sum_i \alpha_{t-1}(i) a_{ij} c_{jk} b_{jk}(\mathbf{C}_t) \beta_t(j)}{\sum_{i \in S_F} \alpha_T(i)} \tag{15}$$

여기에서 a_{ij}는 상태 i에서 j로 가는 전이 확률(transition probability)을 나타내며, S_F는 최종 상태의 집합을 표현한다. 또한 $\alpha_t(i)$는 시간 t에서 상태 i까지의 전향 확률(forward probability)을 의미한다. 만약 PTSHMM에서의 $\gamma_t(j)$가 일반 HMM과 같다면, $\gamma_t(i)$와 $\xi_t(j,k)$의 추정 후에, 상태 j에서의 k 번째 혼합 밀도에 대한 가중치는 다음과 같이 추정될 수 있다.

$$\bar{c}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{16}$$

일반적인 HMM과 비슷하게, PTSHMM에서도 특정 상태에서의 혼합 밀도에 대한 평균 궤적은 모든 분절에 대해 상태 j의 혼합 밀도 k에 머무를 기대치와 그 기대치와 궤적의 곱으로 나타내는 기대 평균치(expected average)로써 구할 수 있다.

$$\mathbf{Z}\bar{\mathbf{B}}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k) \mathbf{Z}\mathbf{B}_t}{\sum_{t=1}^{T} \xi_t(j,k)} \tag{17}$$

PTSHMM은 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 (17)의 양편에서 디자인 행렬을 생략할 수 있다. 디자인 행렬이 생략된 식 (17)은 궤적 계수 행렬 $\bar{\mathbf{B}}_{jk}$을 추정하는 식으로 변환된다.

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

$\bar{\mathbf{B}}_{jk}$가 추정되면 입력 분절 특징과 평균 궤적 $\mathbf{Z}\bar{\mathbf{B}}_{jk}$과의 차이를 이용하여 분산을 구한다. 분산 추정 방법에서는 분절의 모든 프레임에 공통적으로 적용되는 고정 분산과 분절 내부의 시간적 가변성을 반영하는 시변 분산으로 구분할 수 있다.

고정 분산 방식에서, 상태 j의 분산은 평균 궤적 $\mathbf{Z}\bar{\mathbf{B}}_{jk}$이나 궤적 계수 $\bar{\mathbf{B}}_{jk}$에 의하여 계산된다. 만약 평균 궤적이 사용된다면 켑스트럼(cepstrum)이나 MFCC등과 같은 특징 영역에서 분산을 추정하게 된다. 이것은 추정된 분산이 관측된 특징 벡터의 변이를 반영한다는 것을 의미하며, 다음과 같이 얻어질 수 있다.

$$\bar{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k) \{(\mathbf{C}_t - \mathbf{Z}\bar{\mathbf{B}}_{jk})'(\mathbf{C}_t - \mathbf{Z}\bar{\mathbf{B}}_{jk})\}}{\sum_{t=1}^{T} \xi_t(j,k)} \tag{19}$$

그러나, 위식의 분산은 외적 분절 변이를 표현하기 위해 추정된 값이지만, 분절 $\mathbf{C}_t$와 평균 궤적 $\mathbf{Z}\bar{\mathbf{B}}_{jk}$와의 차이를 나타내는 궤적 추정 오차인 내적 분절 변이도 포함되어 있다. 따라서, 식 (19)에서 분절 $\mathbf{C}_t$를 추정된 궤적 $\mathbf{Z}\bar{\mathbf{B}}_{jk}$으로 치환하여 계산한다.

$$\bar{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k) \{(\mathbf{Z}\mathbf{B}_t - \mathbf{Z}\bar{\mathbf{B}}_{jk})'(\mathbf{Z}\mathbf{B}_t - \mathbf{Z}\bar{\mathbf{B}}_{jk})\}}{\sum_{t=1}^{T} \xi_t(j,k)} \tag{20}$$

여기에서 $\mathbf{Z}\mathbf{B}_t$와 $\mathbf{Z}\bar{\mathbf{B}}_{jk}$는 시간 t에서 주어진 분절의 궤적과 상태 j의 k 번째 혼합 밀도의 평균 궤적을 나타낸다.

이와 별도로 평균 궤적을 이용하여 분산을 구하지 않고, 평균 궤적 계수 행렬을 직접 이용하여 분산을 구할 수 있다. 이는 식 (12)에서 알 수 있듯이 외적 분절 변이를 표현하는 확률을 계산할 때, 디자인 행렬이 공통적으로 이용되며, 계수와 분산 간의 계산 후에 디자인 행렬이 적용되고 있기 때문이다. 따라서 분산을 구할 때 궤적 위의 점들에 대한 분포 대신 궤적 계수에 대한 분포를 이용할 수 있다. 즉, 궤적 계수는 디자인 행렬 $\mathbf{Z}$에 의하여 변환되기 때문에 궤적 계수로도 분절의 시간적 가변성이 표현된다고할 수 있다. 이 경우 추정된 분산은 프레임 특징에 대한 분산 (variance of the features)이 아니고 궤적에 대한 분산(variance of the trajectories)으로 명명된다. 궤적에 대한 분산은 다음과 같이 구할 수 있다.

$$\bar{\Sigma}_{jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k) \{(\mathbf{B}_t - \bar{\mathbf{B}}_{jk})'(\mathbf{B}_t - \bar{\mathbf{B}}_{jk})\}}{\sum_{t=1}^{T} \xi_t(j,k)} \tag{21}$$

분절의 길이가 길어지면, 고정 분산 방식은 외적 분절 변이를 충분히 반영하지 못한다. 그러므로 분절에서의 시간적 가변성을 반영하기 위하여, 시간에 따라 변화되는 분산을 고려해볼 수 있다. 분절에서의 각 프레임은 시간적인 순서를 가지고 있기 때문에 시변 분산은 주어진 분절로부터 추정된 궤적의 특징과 평균 궤적을 따라 생성된 특징의 차이를 이용하여 구해져야 한다. 따라서, 시변 분산은 분절의 상대적인 위치에 따라 값이 달라진다.

$$\overline{\Sigma}_{n,jk} = \frac{\sum_{t=1}^{T} \xi_t(j,k)\{\mathbf{z}_n(\mathbf{B}_t - \overline{\mathbf{B}}_{jk})\}'\{\mathbf{z}_n(\mathbf{B}_t - \overline{\mathbf{B}}_{jk})\}}{\sum_{t=1}^{T} \xi_t(j,k)} \quad (22)$$

여기에서 $\mathbf{z}_n\mathbf{B}_t$와 $\mathbf{z}_n\mathbf{B}_t$는 각각 시간 t에서 주어진 분절의 궤적과 상태 j의 k 번째 혼합 밀도로부터 복원된 프레임 특징을 나타낸다.

앞에서 언급한 세가지 분산의 추정 방법은 분절 길이가 1인 경우, 즉, 프레임 특징을 사용하는 경우에는 일반 HMM의 분산 추정 방법과 동일하다. 그러나, 분절 길이가 1보다 큰 경우에는 이들 방식은 서로 다른 분산을 갖게 되고 외적 분절 변이에 대한 확률 $P(\mathbf{ZB}_t | s_t, \lambda)$도 달라진다.

3. 3. 일반화(generalization)

PTSHMM의 조건을 다음과 같이 정의하면, 일반적인 HMM과 모수적 궤적 모델의 일반화 또는 확장으로 해석할 수 있다.

1. 만약 분절 길이 N =1이고 회귀 차수 R =1이라면, 내적 분절 변이의 확률 $P(\mathbf{C}_t | \mathbf{ZB}_t, s_t, \lambda)$은 1이 되고 외적 분절 확률 $P(\mathbf{ZB}_t | s_t, \lambda)$은 추정된 궤적이 아닌 관측된 프레임 특징에 대한 가우시안 분포를 따르게 된다. 이 경우, $\mathbf{C}_t$는 단일 프레임 특징을 표현하게 되고, 추정된 궤적 계수 $\mathbf{B}_t$와 같게 된다. 따라서 추정 오차 또는 적합도 χ^2는 0이 되며, $\mathbf{B}_t$는 상태 s_t에 대한 평균 특징을 표시하게 된다. 그러므로, PTSHMM은 연속 HMM과 완전히 같게 된다.
2. 만약 분절 길이 N이 주어진 모델에 대한 관측 열의 길이 T와 같게 된다면, 각 음향학적 모델은 가변 분절 길이의 단일 상태 또는 단일 분절로 표현된다. 이 경우, PTSHMM은 모수적 궤적 모델(parametric trajectory model)처럼 가변 길이를 갖는 음성 분절의 역학을 모델링할 수 있다.

Ⅵ. 실험 및 결과

제안한 방법이 유효한지를 검사하기 위하여 TIMIT 자료에 기반한 화자 독립 음소 인식(speaker-independent phoneme recognition) 실험을 하였다. 학습에는 8개 방언 사용자로 분류된 462명의 화자가 10번씩 발성한 자료를 사용하였으며, TIMIT 자료에서 제시한 완전 테스트 집합(complete test set)을 이용하여 테스트하였다. 완전 테스트 집합은 학습에서와 마찬가지로 8개 방언의 사용자 168명의 화자가 8번씩 발성한 자료로서 총 1344 문장을 이룬다. 학습과 평가에 사용된 음성 신호는 20msec의 분석 구간, 10 msec의 구간 이동 조건에서 12차의 MFCC와 로그 에너지를 구한 후 다시 각 특징의 1차 미분 계수를 포함한 총 26차의 특징으로 표현되어 사용된다. 이렇게 구해진 음성 신호의 특징은 일반 HMM의 입력으로 사용되거나 PTSHMM의 분절 특징인 궤적을 구하는데 사용된다.

성능 평가를 위하여 표준 48개의 문맥 독립 음소 모델(context-independent phoneme model)을 이용하고, 음소 2진 언어 모델 (phoneme bigram)이 사용되었다. 각 모델은 모델의 평균 프레임 길이에 따라 상태 수가 2개 또는 3개인 좌-우향 HMM을 사용하여 학습이 되었으며, 각 상태는 5개의 가우시안 혼합 모델로 표현이 되었다. 음소 인식 후에, 48개의 음소를 Lee가 정의한 방법대로 39개의 음소로 인식 결과를 통합하여 분석을 하였다[13].

먼저 분산의 추정 방법에 따른 성능의 차이를 비교하기 위하여 기본 시스템으로 연속 HMM을 이용하였고, PTSHMM에 대해서는 분산의 추정 방법을 달리하여 비교하였다. 기본 시스템인 연속 HMM으로는 분절 길이 N =1과 회귀 차수 R =1인 경우의 PTSHMM을 이용하였다. 분산 추정 방법에 따른 성능 평가를 위하여 나머지 PTSHMM은 동일하게 N =3과 R =2를 적용하였다. 실험 결과는 표1에 정리되었다.

표1에서 제시된 실험결과는 기존의 연구와 비슷한 양상을 보인다[4, 12]. 기존의 연구에서와 같이 시변 분

표 1. 서로 다른 분산 추정 방법에 따른 음소 인식 성능, 기본 시스템은 PTSHMM 에서 N=1, R=1인 경우이며, 분산 추정 방법에 따른 성능 평가를 위한 조건은 N=3, R =2로서 같음(FVF:특징에 대한 고정 분산, FVT:궤적에 대한 고정 분산, TVV:시변 분산)

Table 1. Phoneme recognition results for the different variance estimation. N=1 and R=1 for the baseline system. N=3 and R=2 for PTSHMM(FVF:Fixed Variance of the Features, FVT:Fixed Variance of the Trajectories, TVV:Time-Varying Variance).

(단위:%)

PTSHMM	Correct	Accuracy	Substitution	Deletion	Insertion	Error
기본시스템	62.0	56.0	28.0	10.0	6.0	44.0
FVF	32.4	31.1	40.5	27.1	1.3	68.9
FVT	46.1	43.4	37.2	16.7	2.7	56.6
TVV	64.9	59.2	26.1	9.2	5.4	40.8

산 추정 방법이 고정 분산 추정 방법에 비하여 월등한 성능 향상을 보였다. 기존의 연구에서는 실험이 모음 분류(vowel classification)에만 적용되어 성능의 차이가 그리 크지 않았으나, 본 실험에서는 모음 뿐만이 아니고 자음도 포함하며, 연속된 음소들에 대한 위치 조정(phone alignment)을 포함하기 때문에 모델들간의 혼동성이 증가하여 추정 방법에 따른 성능의 차이가 컸다. 심지어 고정 분산을 이용한 두 가지 추정 방법 모두 분절특징을 이용하였음에도 불구하고, 기본 시스템보다 낮은 성능을 보였다. 이것은 고정 분산 방법이 시간적, 공간적 변이(temporal and spatial variation)를 제대로 표현하지 못하기 때문에 분절의 특징 변화를 충분히 반영하지 못하였다는 것을 의미한다. 특징에 대한 고정 분산 추정 방법이 사용된 경우, 분절에서의 프레임에 대한 시간적 종속성(temporal dependence)은 분절의 길이에 대한 고려가 없어서 약해지게 된다. 이것은 분절 내부의 모든 프레임에 대하여 공통적인 분산을 적용하므로써 프레임간 유지되는 시간적 종속성 정보가 약해진다는 뜻으로 해석할 수 있다. 궤적에 대한 분산 추정 방법이 사용된 경우 공간적 변이에 대한 모델링이 약해지게 된다. 궤적은 시간적인 종속성 위에서 표현되는 특징이므로 궤적 계수에 의한 분산 추정은 시간적 정보보다는 공간적 정보를 더 많이 표현하게 된다. 공간적 정보를 표현하는 궤적 계수는 디자인 행렬 Z에 의한 변환에 의하여 시간적 정보를 반영하게 된다. 즉, 궤적 공간에서의 공간적 정보의 표현력 약화는 특징 공간에서의 시간적 정보의 약화를 가져와 분절의 특징 변화를 제대로 표현하지 못한다고 해석할 수 있다.

다음으로는 PTSHMM의 분절 길이와 회귀 차수의 변화에 따른 성능 비교를 하였다. 첫번째 실험과 마찬가지로 기본 시스템은 PTSHMM에서 분절 길이 N을 1로, 회귀 차수 R을 1로 정하였다. 분절 길이 N과 회귀 차수 R의 변화에 따라 N=3,5일 때와 R=2,3일 때의 PTSHMM에 대하여 각각 실험하였으며, 그 결과는 표 2에 보인다.

표 2. 서로 다른 분절의 길이와 회귀 차수에 따른 음소 인식 성능, 기본 시스템은 PTSHMM 에서 N=1, R=1로 설정되었다.

Table 2. Phoneme recognition results for the different segment length and regression order. N=1 and R=1 for the baseline system.

(단위:%)

PTSHMM	Correct	Accuracy	Substitution	Deletion	Insertion	Error
기본시스템	62.0	56.0	28.0	10.0	6.0	44.0
N=3,R=2	64.9	59.2	26.1	9.2	5.4	40.8
N=3,R=3	65.2	59.7	25.8	9.1	5.5	40.3
N=5,R=2	64.9	59.9	25.6	9.5	5.0	40.1
N=5,R=3	65.6	60.6	25.2	9.2	5.0	39.4

동일한 분절 길이일 때, 즉, N=3이거나 5일 때, 회귀 차수 R이 2에서 3으로 증가하면 인식률은 N=3 일 때 64.6%에서 65.2%로, N=5일 때 64.9%에서 65.6%로 증가하고 있다. 반대로 동일한 회귀 차수에서, 즉, R=2일 때와 R=3일 때 N=3에서 5로 분절 길이가 증가되면, 삽입 오류가 각각 5.4%에서 5.0%로, 5.5%에서 5.0%로 감소함을 알 수 있다. 따라서, 삭제 오류가 약간 증가하더라도 전체적인 정확도 측면에서는 성능이 향상되었다. 이것은 PTSHMM의 경우, 동일한 분절 길이 조건에서 회귀 차수가 증가하면 변별력(discrimination power)이 증가하고, 동일한 회귀 차수에서 분절 길이가 길어지면 전이 정보의 양이 증가되었기 때문이라고 생각한다.

정확도(percent accuracy) 측정에 의한 전체적인 성능 평가에서, 기본 시스템인 경우 56.0%, 일차 선형 시스템(linear trajectory system)의 PTSHMM인 경우 59.2%(N=3, R=2), 이차 궤적 시스템(quadratic trajectory system)인 경우 60.6%의 성능을 보여 기본 시스템에 비해 궤적 모델에 의한 분절 HMM을 이용한 제안된 방법이 각각 7.3%와 10.4%의 오류율 감소를 보였다. 이러한 결과는 제안된 방법이 음성 단위에 대한 전이 특성을 충분히 표현한다는 것을 보여주고 있어 궤적 기반의 PTSHMM이 일반 HMM보다 시간적 종속성을 잘 표현한다고 할 수 있다.

V. 결 론

본 논문에서는 다항식으로 표현되는 궤적을 분절 특징으로 이용하는 연속 음성 인식을 위한 새로운 모델링 방법을 제안하였다. 제안된 방법은 현재 프레임에 대하여 대칭되는 디자인 행렬을 이용하여, 이웃하는 음성 단위간의 전이 정보를 표현할 수 있다. 또한, PTSHMM에 대한 기본적인 수학적인 분석을 하였으며, HMM의 개념 안에서 Baum-Welch 형태의 변수 재추정 알고리즘을 제안하였다. PTSHMM은 분절 특징으로 모수적 궤적 방법을 사용하였으며, HMM의 개념 안에서 분절 HMM으로 모델링하였기 때문에, 일반적인 HMM과 모수적 궤적 모델의 일반화 또는 확장으로 볼 수 있다. 또한 각 상태에서의 변이를 모델링하기 위하여 평균 궤적의 분포에 의한 외적 분절 변이와 특정 분절에 대한 궤적의 적합도로 내적 분절 변이를 표현하였다. 외적 분절 변이는 추정된 궤적과 평균궤적과의 차이에 의한 다중 변량 가우시안 분포로 모델링되었으며, 고정 분산이나 시변 분산을 이용하여 확률을 계산하였다. 분산 추정 방법에 따른 실험 결과, 프레임간의 시간적 변이를 반영한 방법이 고정된 분산 방법에 비하여 좋은 성능을 보여 분절에서 시간적 종속성(temporal dependence)의 중요성을 파악하였다. 또한,

분절 표현의 변이 특성을 파악하기 위하여, PTSHMM에서의 분절 길이와 회귀차수의 변화에 따른 실험을 하였다. 실험 결과, 동일한 분절 길이에서 회귀 차수가 높아지면 변별력이 높아짐을 알 수 있었고, 동일한 회귀 차수에서는 전이 정보의 양이 많아짐을 알 수 있었다. 실험 결과에서 알 수 있었듯이 분절 길이가 길어지고 회귀 차수가 높아질수록 PTSHMM의 성능은 높아짐을 알 수 있었다. 따라서, 우리는 PTSHMM이 명시적으로 문맥 종속적인 모델링(Context Dependent modeling)을 하지 않더라도, 음성 프레임간의 시간적 관계를 충분히 표현할 수 있으며 부분적인 문맥 종속 정보, 즉, 전이 정보를 모델링할 수 있다고 판단한다.

참 고 문 헌

1. L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. IEEE Transactions on ASSP, 37(8):1214-1225, 1989.
2. X. D. Huang, Y. Ariki, M. A. Jack, Hidden Markov models for speech recognition, Edinburgh, UK:Edinburgh University, 1990.
3. H. Gish, K. Ng, A segmental speech model with application to word spotting. In International Conference on Acoustics, Speech and Signal Processing 1993, volume II, pages 447-450, Minneapolis, Minnesota, April 1993.
4. H. Gish, K. Ng. Parametric trajectory models for speech recognition. In International Conference on Spoken Language Processing 1996, volume I, pages 466-469, Philadelphia, PA, October 1996.
5. M. Russell. A segmental HMM for speech pattern modeling, In International Conference on Acoustics, Speech and Signal Processing 1993, volume II, pages 499-502, Minneapolis, Minnesota, April 1993.
6. M. J. F. Gales, S. J. Young. Segmental hidden Markov models. In European Conference on Speech Communication and Technology 1993, pages 1579-1582, Berlin, Germany, September 1993.
7. L. Deng et al. Speech recognition using hidden Markov models with polynomial regression functions as non-stationary states. IEEE Trans. on Speech and Audio Processing, 2(4):507-520, 1994.
8. M. Ostendorf et al. From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition. IEEE Trans. on Speech and Audio Processing, 4(5):360-378, 1996.
9. W. J. Holmes, M. J. Russell, Speech recognition using a linear dynamic segmental HMM. In International Conference on Acoustic, Speech and Signal Processing, Detroit, Michigan, pages 1611-1614, 1995.
10. L. Deng. A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal. Signal Processing, 27, 65-78, 1992.
11. W. H. Press, A. A. Teukolsky, W.T. Vetterling, B. P. Flannery. Numerical Recipes in C, 2nd Ed. Cambridge University Press, pp. 671-680, 1992.
12. T. Fukada, Y. Sagisaka, K. K. Paliwal. Model Parameter Estimation For Mixture Density Polynomial Segment Models. In Proceedings of IEEE International Conference on Acoustic, Speech and Sisnal Processing, April 1997, Munich, GerMany, pp. 1403-1406, 1997.
13. K. F. Lee, H. W. Hon, Speaker-independent phone recognition using hidden Markov models. IEEE Trans. On Acoustics, Speech and Signal Processing, 37(11), 1661-1648, 1989.

▲윤 영 선(Yun Young-Sun)

1990년 2월 : 한국과학기술원 전산학과 졸업(공학사)
1990년 2월 : 한국과학기술원 전산학과 졸업(공학석사)
1992년 3월~1995년 7월 : (주) 핸디소프트 기술연구소 주임연구원
1997년 9월~현재 : 한국과학기술원 전산학과 박사과정
※주 관심 분야 : 음성인식

▲오 영 환(Oh Yung-Hwan)
한국음향학회지 제17권 6호 참조