분절특징 HMM의 특성에 관한 연구*

윤영선(한남대), 정호영(ETRI)

<차 례>

<Abstract>

A Study on the Characteristics of Segmental-Feature HMM

Young-Sun Yun, Ho-Young Jung

In this paper, we discuss the characteristics of Segmental-Feature HMM and summarize previous studies of SFHMM. There are several approaches to reduce the number of parameters in the previous studies. However, if the number of parameters decreased, the performance of systems also fell. Therefore, we consider the fast computation approach with preserving the same number of parameters. In this paper, we present the new segment comparison method to speed up the computation of SFHMM without loss of performance. The proposed method uses the three-frame calculation rather than the full (five) frames in the given segment. The experimental results show that the performance of the proposed system is better than that of the previous studies

* 주제어: 음성 인식(speech recognition), 분절 특징 HMM(segmental-feature HMM), 분절 특징 (segmental feature)

* 본 연구는 한국전자통신연구원의 2002년도 위탁과제의 결과물입니다.

1. 서 론

음성 인식 분야뿐 아니라 다양한 분야에서도 은닉 마코프 모델(HMM; hidden Markov model)은 구현하기 쉽고, 유연한 모델링 능력과 높은 성능 때문에 널리 사용되고 있으며, 그 사용 분야는 점차 넓혀지고 있다. 그러나 HMM은 뛰어난 성능에도 불구하고 채택하고 있는 가정으로 인해 음성 신호의 시간적 종속성을 제대로 표현하지 못한다고 알려지고 있다. 이런 단점을 보완하기 위하여 여러 연구 방식이 소개되었으며, 대표적인 방법으로는 분절 모델(Gales 1993, Ostendorf 1996)을 이용한 방법과 궤적 방식(Gish 1996, Holmes 1999)을 적용한 방법, 그리고 선형 회귀 방식을 이용한 동적 특성(Furui 1985)을 표현한 방법이 있다. 이들 방식은 프레임 특징(frame feature) 대신에 여러 프레임으로 구성된 분절 특징(segmental feature)을 이용하거나 여러 프레임의 회귀 함수에 의한 매개 변수(coefficient) 또는 평균값으로 표현하고 있다. 기존 연구는 음성 신호의 동적 특성을 반영하기 위하여 분절의 길이에 제한을 두지 않고 궤적의 확률 분포와 추정 오차를 통계적 방법으로 정량화시켰기 때문에 계산량이 많다는 약점과, 분절 길이가 변하는 특성 때문에 경계 문제(boundary problem)가 발생하여 연속 음성 인식에 적용하기 어렵다는 문제점이 있었다. 이러한 문제점을 완화하고 성능을 향상시키기 위해, 여러 프레임 특징을 모수적 궤적(parametric trajectory)방식을 이용한 분절 특징으로 표현하고 인식 모델에 적용한 분절 특징 HMM(SFHMM; segmental-feature HMM)이 제안되었다(윤영선 2000a, 2000b, 2002). 그러나 제안된 분절 특징 HMM의 성능이 우수하다 할지라도 기존의 HMM을 이용한 음성 인식 시스템에 비하여 여전히 많은 계산량과 매개 변수의 수가 문제점으로 지적되고 있다. 따라서 본 연구에서는 기존의 SFHMM 연구를 정리하고 그 특성을 파악한 후에 계산량을 감소시키기 위한 방안으로서 분절 내의 일부 프레임만 계산에 반영하고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존 연구에서 제안된 분절 특징 HMM을 간략히 소개하며, 3장에서는 분절 특징 HMM에 대한 특성들을 정리하고, 마지막 절에서 인식 시간을 빠르게 하기 위한 부분 프레임 계산을 제안한다. 4장에서는 제안된 방식의 유효성을 검증하기 위한 실험 및 결과를 보이며, 마지막으로 본 연구의 요약 및 결론을 맺도록 하겠다.

2. 분절 특징 HMM

본 장에서는 분절 특징 HMM에 대해 간략히 요약하도록 한다. 분절 특징 HMM은 분절 분포를 프레임 특징의 확률 분포로 표현하지 않고, 각 분절을 다항식으로 표현하고 그 다항식의 확률 분포로써 모델링하는 방법이다. 이 방법은 모수적 방

법이 평활화(smoothing) 효과를 내포하고 있기 때문에 잡음 환경에서도 좋은 성능을 보일 것으로 기대되며, 기존의 HMM에서 쉽게 확장될 수 있다. 또한, 기존의 분절 모델 방식이 가변적인 분절 길이를 채택하여 인식 과정에서 많은 계산 시간이 필요하다는 점을 개선하고자 고정된 분절 길이를 이용하여 특징 표현 방법과 인식 모델을 분리하였다 (윤영선 2002).

2.1. 분절 특징

Deng(1992)은 HMM의 상태에서의 출력 확률을 표현하기 위하여 절대적인 시간에 대한 다항식으로 상태의 평균 변화를 모델링 하였으며, 모델을 개선하여 상태에서의 지속 시간으로 관측 확률의 변이를 표현하였다(Deng 1994). 이 방법은 음성 특징을 모수적인 방법으로 표현한 것이 아니라, 특정 상태에서의 관측 확률을 모수적 방법으로 예측하였다. 다음으로 Gish와 Ng가 핵심어 검출(word spotting)에서 단어의 경계가 결정이 된 경우, 모수적 방식에 의하여 검출된 단어를 검증하는데 사용하였다 (Gish 1993, Gish 1996). 전자의 경우는 다항식을 이용하여 HMM의 상태 모델링을 향상시켰으나 여전히 프레임 특징에 기반을 두어 HMM의 약점으로 지적되고 있는 독립 관측(independence observation) 가정을 완화시키지 못하였다. 후자의 경우에는 주어진 패턴 전체를 하나의 특징 표현으로 모델링하여 패턴 분류에 사용하였다. 이와 같이 여러 프레임 특징으로부터 얻어지는 특징 표현은 분절 특징(segmental feature)이라 불린다.

분절 특징 HMM은 HMM의 독립 관측 가정을 완화시키기 위하여 음성 신호를 분절 특징으로 모델링하였으며, 연속 음성 인식에 쉽게 사용할 수 있도록 음성 패턴을 고정된 길이의 분절들의 열로써 표현하였다. 각 분절은 인접한 분절들과 중첩될 수 있으므로 기준점을 분절의 중앙에 두었다. 이것을 고려하여 고정길이를 갖는 분절을 표현하면 다음과 같이 나타낼 수 있다.

(1)

$$C_t = ZB_t + E$$

위 식에서 C_t와 B_t는 각각 시간 t에서의 음성 분절과 궤적 계수를 나타낸다. 궤적으로 표현되는 분절 특징은 주어진 분절 안에서 적용할 프레임의 범위와 표현 형태를 나타내는 디자인 행렬(design matrix) Z와 궤적 계수 B_t의 곱으로 표현된다. 원래의 음성 분절 C_t와 표현된 궤적 ZB_t의 차이로 인하여 발생되는 잔차 오차(residual error) E_t는 독립적이며 균일하게 분포(independent and

identically distributed)되어 있다고 가정한다. 이 식에서 각 프레임은 D차원의 특징 벡터로 표현되며, 음성 분절 C_t는 $N\times D$ 행렬, Z와 B_t는 각각 $N\times R$, $R\times D$ 차원의 행렬을 나타낸다.

주어진 음성 분절이 $N=2M+1$의 프레임으로 구성된다고 하면, 입력 벡터 $Y=y_1,\ldots,y_N$는 다음과 같이 분절 단위로 표현될 수 있다.

(2)

$$C_t = Y_{t-M^{t+M}} = \begin{bmatrix} y_{t-M} \\ \vdots \\ y_t \\ \vdots \\ y_{t+M} \end{bmatrix}$$
$$y_\tau = [y_{\tau,1} \cdots y_{\tau,D}], \quad t-M\le\tau\le t+M.$$

분절 표현에서 알 수 있듯이 현재 시간 t의 분절 C_t의 기준점은 분절 중앙에 있는 프레임 특징 y_t가 되기 때문에, $t-1$ 또는 $t+1$시간의 분절과 중첩될 수 있다. 이와 같은 음성 분절을 표현하기 위하여 음성 신호의 프레임 특징 열의 범위를 조절하는 디자인 행렬 Z는 다음과 같이 정의될 수 있다.

(3)

$$Z = \begin{bmatrix} 1 & \left(-\frac{M}{2M}\right) & \cdots & \left(-\frac{M}{2M}\right)^{R-1} \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & \left(\frac{M}{2M}\right) & \cdots & \left(\frac{M}{2M}\right)^{R-1} \end{bmatrix}$$
$$z_\tau = \left[1 \left(\frac{\tau-t}{2M}\right) \cdots \left(\frac{\tau-t}{2M}\right)^{R-1}\right],$$

여기에서 z_τ는 디자인 행렬 Z의 τ번째 행 벡터(row vector)를 나타낸다. 디자인 행렬 Z가 현재 시간 t에 대해서 분절 길이로 정규화된 상대적인 위치 정보를 나타내기 때문에, 디자인 행렬을 이용하여 계산된 궤적은 바로 이전에 관측된 특징 벡터와 다음에 오는 관측 벡터를 포함하게 된다. 이와 비슷한 방법으로 궤적 계수 행렬 B_t는 다음과 같이 정의된다.

(4)

$$\boldsymbol{B}_t = \begin{bmatrix} \boldsymbol{b}_{1'} \\ \vdots \\ \boldsymbol{b}_{R'} \end{bmatrix}$$
$$\boldsymbol{b}_{i'} = [\, b^t_{i,1} \cdots b^t_{i,D}], \quad 1 \le i \le R.$$

음성 분절 $\boldsymbol{C}_t$와 디자인 행렬 $\boldsymbol{Z}$가 주어지면 추정되는 궤적 계수 행렬 $\widehat{\boldsymbol{B}}_t$는 선형 회귀(linear regression) 방정식이나 다음과 같은 행렬 연산에 의하여 계산될 수 있다.

(5)

$$\widehat{\boldsymbol{B}}_t = [\, \boldsymbol{Z}'\boldsymbol{Z}]^{-1}\boldsymbol{Z}'\,\boldsymbol{C}_t,$$

여기에서 '는 행렬의 전치(transpose)를 의미한다. 궤적 계수 행렬 $\widehat{\boldsymbol{B}}_t$가 추정되면, 최적 적합도 (goodness-of-fit) χ^2는 시간 t의 분절을 구성하는 모든 프레임 특징에 대한 잔차 오차를 더하여 계산된다.

(6)

$$\chi^2_t = \frac{1}{N}\sum_{\tau=t-M}^{t+M}(\, \boldsymbol{y}_\tau - \boldsymbol{z}_\tau\,\widehat{\boldsymbol{B}}_t)(\, \boldsymbol{y}_\tau - \boldsymbol{z}_\tau\,\widehat{\boldsymbol{B}}_t)',$$

여기에서 $\boldsymbol{y}_\tau$와 $\boldsymbol{z}_\tau$는 음성 분절과 디자인 행렬의 행 벡터(row vector)를 의미하며, 분절의 길이는 $N=2M+1$이다. 위 식에서 최적 적합도 χ^2가 작은 값을 나타내면 추정된 궤적 계수가 원래의 음성 분절을 잘 표현하고 있다는 것을 나타낸다. 이와 같은 과정을 통행 입력 음성 분절에 대응되는 궤적 계수 $\widehat{\boldsymbol{B}}_t$와 최적 적합도 χ^2가 인식 시스템의 특징으로 사용된다.

2.2. 분절 우도

분절 HMM(segmental HMM)에서는 특정 상태(state)에서의 분절 관측 확률을 외적 분절 확률(extra-segmental probability)과 내적 분절 확률(intra-segmental probability)의 곱으로 표현하였다(Holmes 1999). 외적 분절 확률은 화자의 특성이나 특정 음에 대한 발음의 변이와 같은 장기적인 변이를 나타내고, 내적 분절 확률은 연속된

조음 현상이나 불안정한 요소에 의해 발생되는 단기적인 변이 현상을 표현한다. 기존의 분절 HMM에서는 분절의 길이가 가변적이기 때문에 외적 분절 변이와 내적 분절 변이를 정확히 추정하기가 어렵다. 먼저 외적 분절 변이를 고정된 표현 방식(기울기와 중간 값)을 이용하여 음성 신호로부터 추정하고 다시 구해진 외적 분절 변이를 이용하여 내적 분절 변이를 계산하는 과정으로 변수들을 추정한다. 따라서 학습 시간이 오래 걸린다는 단점이 존재한다. 이러한 문제점을 해결하기 위하여 SFHMM에서는 분절 특징 표현과 인식 단계를 분리하여 독립시켰다.

입력 음성으로부터 추출된 분절 특징을 인식 단계에 적용하기 위하여 SFHMM에서는 외적 분절 변이를 상태에서의 평균 궤적으로 표현하고, 내적 분절 변이는 입력 음성을 분절 특징으로 변환한 경우의 궤적 추정 오차로 나타낸다. 시간 t에서 관측 벡터열 $\boldsymbol{C}_t$가 단일 궤적 $\boldsymbol{ZB}_t$로 표현된다면, 모델 λ의 상태 s_i에서 발생하는 $\boldsymbol{C}_t$의 관측 확률은 외적 분절 확률과 내적 분절 확률로써 다음과 같이 표현된다.

(7)

$$P(\boldsymbol{C}_t|s_i,\lambda)=P(\boldsymbol{Z}\hat{\boldsymbol{B}}_t|s_i,\lambda)P(\boldsymbol{C}_t|\boldsymbol{Z}\hat{\boldsymbol{B}}_t,s_i,\lambda).$$

따라서 시간 t에서 상태 j의 분절 관측 확률은 상태 j의 평균 궤적 $\boldsymbol{Z\beta}_j$와 분산 $\boldsymbol{\Sigma}_j$를 이용하여 다음과 같이 표현할 수 있다.

(8)

$$b_j(\boldsymbol{C}_t)=P(\boldsymbol{C}_t|s_j,\lambda)=P(\boldsymbol{Z}\hat{\boldsymbol{B}}_t|\boldsymbol{Z\beta}_j,\boldsymbol{\Sigma}_j)\cdot P(\boldsymbol{C}_t|\boldsymbol{Z}\hat{\boldsymbol{B}}_t),$$

위 식에서 $P(\boldsymbol{Z}\hat{\boldsymbol{B}}_t|s_i,\lambda)\approx P(\boldsymbol{Z}\hat{\boldsymbol{B}}_t|\boldsymbol{Z\beta}_j,\boldsymbol{\Sigma}_j)$는 장기적인 변이의 외적 분절 확률을 나타내고, $P(\boldsymbol{C}_t|\boldsymbol{Z}\hat{\boldsymbol{B}}_t,s_i,\lambda)\approx P(\boldsymbol{C}_t|\boldsymbol{Z}\hat{\boldsymbol{B}}_t)$는 단기적인 변이를 나타내는 내적 분절 확률을 의미한다. 내적 분절 변이는 음성 분절 $\boldsymbol{C}_t$에서 추정된 궤적 $\boldsymbol{Z}\hat{\boldsymbol{B}}_t$에 관련되고 모델 λ의 상태 j와 무관하기 때문에 모델 관련 변수를 생략할 수 있다.

(9)

$$P(\boldsymbol{Z}\hat{\boldsymbol{B}}_t|\ \boldsymbol{Z}\boldsymbol{\beta}_i,\ \boldsymbol{\Sigma}_i) = \prod_{\tau=t-M}^{t+M} \frac{1}{(2\pi)^{D/2}|\Sigma_{\tau-t,i}|^{1/2}} \cdot \exp\left\{-\frac{1}{2}\{\boldsymbol{z}_\tau(\hat{\boldsymbol{B}}_t-\boldsymbol{\beta}_i)\}\Sigma^{-1}_{\tau-t,i}\{\boldsymbol{z}_\tau(\hat{\boldsymbol{B}}_t-\boldsymbol{\beta}_i)\}'\right\},$$

(10) $$P(\boldsymbol{C}_t|\ \boldsymbol{Z}\hat{\boldsymbol{B}}_t) = \exp\{-\frac{1}{2}\chi_t^2\},$$

외적 분절 변이에서 사용되는 분산 $\boldsymbol{\Sigma}_j$는 채택된 가정에 따라 각 프레임별로 계산되는 분산 수열을 나타내거나(시변 분산) 분절 내의 모든 프레임에 공통적으로 적용된 단일의 공통 분산(고정 분산)을 의미한다. 고정 분산을 채택한 경우, 단일 혼합 밀도(single mixture) 환경에서는 시변 분산 시스템보다 성능이 저하되나 혼합 밀도의 수가 증가할수록 성능이 향상되며 시변 분산 시스템보다 성능이 뛰어난 것으로 보고되고 있다(윤영선 2002).

3. 분절 특징 HMM의 특성

본 장에서는 분절 특징 HMM의 특성을 정리하고, 그 개선 방향을 제안한다. 제안하는 방법은 분절 특징의 특성을 고려하여 우도 계산 시에 분절에 해당하는 모든 프레임에 대하여 확률 값을 계산하지 않고 일부 프레임만 계산에 포함되도록 한다.

3.1. 분절 조건에 따른 일반화

분절 특징 HMM의 경우, 분절의 길이가 1인 경우 분절 특징 대신 프레임 특징이 사용된다. 이 경우, 분절 특징 HMM은 일반 HMM의 관측확률과 동일한 형태를 띤다. 또한 분절의 길이가 1보다 큰 경우에 단일 상태로 모델링하게 되면, 기존의 모수적 궤적 모델 (parametric trajectory model)과 유사한 형태가 된다. 각 경우에 대해 다음과 같이 분석할 수 있다.

(경우 1) 분절 길이 $N=1$이고 회귀 차수 $R=1$이라면, 시간 t의 분절 $\boldsymbol{Z}\hat{\boldsymbol{B}}_t$는 단일 프레임 특징 $\boldsymbol{c}_t$가 된다. 따라서 내적 분절 변이의 확률 $P(\boldsymbol{C}_t|\ \boldsymbol{Z}\hat{\boldsymbol{B}}_t, s_i, \lambda)$는 1이 되고, 외적 분절 확률 $P(\boldsymbol{Z}\hat{\boldsymbol{B}}_t|s_i, \lambda)$은 추정된 궤적이 아닌 관측된

프레임 특징에 대한 가우시안 분포를 따르게 된다. 따라서 C_t는 단일 프레임 특징을 표현하게 되고, 추정된 궤적 $Z\hat{B}_t$와 같게 된다. 이 경우에 추정 오차 또는 적합도 χ^2는 0이되며, $Z\beta_i$는 상태 s_i에 대한 평균 특징을 표시하게 된다. 그러므로 분절 특징 HMM은 연속 HMM과 완전히 같게 된다.

(경우 2) 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 위하여 디자인 행렬 Z의 크기를 관측 열의 길이 T가 되도록 확장한다. 이 경우 디자인 행렬은 k번째 관측 분절에 종속적인 디자인 행렬 Z_k로써 표현되며 궤적 계수 행렬 $\hat{B}_k$는 다음과 같이 계산된다.

(11)

$$\hat{B}_k = [\ Z_k'\ Z_k]^{-1}\ Z_k\ C_k$$

각 모델은 단일 상태로 표현되기 때문에, 평균 궤적의 추적은 다항식에 의한 분절 모델과 동일하게 된다. 따라서 궤적 계수 행렬 $\hat{B}_k$와 평균 궤적 계수 행렬 $\overline{B}$이 분절 길이에 대응되도록 조정된다면, 분절 특징 HMM은 모수적 궤적 모델과 같이 가변 길이를 갖는 음성 분절을 모델링할 수 있다(윤영선 2002).

3.2. 분절 특징의 특성 비교

분절 특징은 여러 프레임 특징들로 구성되어 있으며 분절 길이와 회귀 차수 등에 따라 성능의 변화가 발생하기 때문에 다양한 조건의 특징 조합을 통하여 성능이 분석되었다(윤영선 2002). 분절 특징을 살펴보기 위하여 두 가지 조건의 특징 집합을 이용하여 성능 비교를 하였다. 일반 HMM에 사용되는 특징 벡터는 12차의 MFCC와 로그 에너지, 그리고 이들 특징 벡터의 1차 미분 계수를 이용하였으며, 분절 특징 HMM에는 13차의 기본 특징만 사용하였다. 이 연구에서 분절 길이가 5이고 회귀 차수가 3이상인 경우에는 SFHMM의 성능이 우수하다고 발표되었다. 또한, 분절 특징의 기본이 되는 특징 집합을 26차로 확장하여 비교한 결과 모든 특징 조합에서 HMM에 사용되는 특징 벡터 (13차의 2차 미분 계수를 추가하여 39차의 특징 벡터 사용)보다 성능이 우수함을 알 수 있었다. 이 연구에서 분절 특징과 프레임 특징의 성능 비교를 통하여 분절 특징이 동적 특징의 특성을 나타냄을 알

수 있다. 또한, 분절 길이와 회귀 차수의 상관 관계에서 분절 길이가 증가하고 회귀 차수가 높을수록 분절 특징 HMM의 성능이 향상됨을 알 수 있다. 특히 동일한 분절 길이에 대하여 회귀 차수가 증가하면 인식률이 증가하고 치환 오류가 감소한다. 이것으로부터 동일한 분절 길이에서 회귀 차수가 높아지면 변별력이 증가한다고 판단하였다. 반면에 동일한 회귀 차수에 대해 분절 길이를 증가시키면 삽입 오류와 치환 오류가 감소하지만 삭제 오류는 일정 부분 증가한다. 그러나 증가된 오류보다 감소된 오류가 크기 때문에 음성 인식의 성능 척도인 정확도(accuracy)에서는 증가하였다. 이것으로부터 치환 오류와 삽입 오류의 감소에서 동일한 회귀 차수인 경우 분절 길이의 증가는 전이 정보 (transitional information)가 증가한다고 판단하였다(윤영선 2002).

3.3. 가변 분산과 고정 분산

분산은 일반적으로 입력 벡터와 평균 벡터의 차이로써 구하므로, 평균 벡터가 분절 특징으로 표현되는 특징 표현 방법인 경우, 분산 표현 방법은 매우 중요하다. 특히 분절 특징 HMM인 경우 상태에서의 분산은 분절 특징의 분산을 의미하기 때문에, 입력 음성 분절에서 추정된 궤적과 주어진 상태의 평균 궤적의 차이로써 표현된다. 분절 길이가 N이라면 상태에서 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 궤적에서 프레임별로 복원된 점들의 거리를 이용한다. 이와 같이 분절의 분산이 분절을 구성하는 프레임 분산의 상대적인 순서로써 표현되는 방식을 시변 분산 또는 가변 분산(time-varying variance)이라 한다. 가변 분산 방식이 적용되면, 단일 혼합 밀도로 구성되는 한 상태를 표현하기 위해서는 궤적 계수 행렬 $\boldsymbol{B}$와 궤적의 분산을 나타내는 N개의 분산 행렬이 필요하다. 따라서 분절의 범위가 넓어지거나 혼합 밀도의 수가 증가하면 분절 특징 HMM을 표현하는 매개 변수의 수는 급격히 증가하게 되고, 더 많은 연산 시간을 필요로 한다. 이와 같은 문제점을 해결하기 위하여 매개 변수의 수를 줄이는 방법에 대한 고찰이 필요하다. 특히 분절 특징의 분산을 표현하기 위해서는 각 상태의 혼합 밀도에 N개의 분산 행렬이 필요하므로, 대표 분산을 이용하여 분절 분산을 표현한다면 매개 변수의 수를 줄일 수 있을 것이다. 이 경우, 대표 분산을 선택하는 것은 그 방법에 따라 음성 인식 시스템의 성능에 많은 영향을 미칠 수 있기 때문에 분절 내에서의 각 프레임 특징에 대한 분산을 평균하여 대표 분산으로 한다. 즉, 각 프레임 특징의 분산을 평균하여 분절 내의 모든 프레임에 공통적으로 사용한다. 분산 표현 방법에 따른 성능 평가를 비교하

기 위하여 분절 특징 HMM의 혼합 밀도 수와 분절 길이, 회귀 차수를 조정하며 실험하였다. 실험 결과에서 단일 혼합 밀도를 사용한 경우 가변 분산의 성능이 고정 분산보다 높은 성능을 보이나, 혼합 밀도 수를 증가시켜 비교한 결과 두 표현 방식의 성능 차이는 현저히 줄어든다. 이런 결과는 모수적 궤적 방식의 기존 연구와 다른 양상을 보이고 있는데, 분절 길이의 영향 때문으로 판단한다. 기존 연구에서는 완전 궤적(complete trajectory)을 이용하였기 때문에 각 분절은 음소와 같은 음향학적 모델에 대응된다. 이때의 분절 길이는 분절 특징 HMM에서 사용하는 분절 길이보다 상대적으로 크기 때문에 고정 분산은 분절의 길이를 충분히 표현하지 못하나, 분절 특징 HMM에서는 분절 길이가 작아 적절한 혼합 밀도인 경우, 고정 분산 방식이 가변 분산과 유사한 성능을 보임을 알 수 있다 (윤영선 2000b, 2002).

3.4. 경향 공유

분절 특징 HMM에서 각 분절은 고정된 길이를 갖으며, 다항식에 의한 궤적으로 모델링된다. 이 궤적은 모수적 방법에 의하여 음성 신호의 특징 열로부터 얻어지기 때문에, 궤적 계수로부터 쉽게 경향과 위치 정보를 분리할 수 있다. 경향정보는 음성의 변화 형태를 표현하며, 위치 정보는 분절 특징의 기준 위치를 나타낸다. 분절 특징 HMM에서 궤적 정보는 선형 회귀 방정식으로 표현될 수 있으므로 각 특징 차원은 궤적 계수와 디자인 행렬로부터 쉽게 복원될 수 있다.

(12)

$$y_{\tau,i} = b_{1,i} z_{\tau,1} + b_{2,i} z_{\tau,2} + \cdots + b_{R,i} z_{\tau,R}, 1 \le i \le D,$$

여기에서 $y_{\tau,i}$는 분절에서의 τ번째 프레임의 i차 켑스트럼(cepstrum) 벡터를 의미하며, $b_{r,i}$는 r번째 궤적 계수를 나타낸다. $z_{\tau,r}$은 디자인 행렬의 요소를 나타내며, $\left(\frac{\tau - t}{2M}\right)^{r-1}$로 표현된다. 위 식에서 디자인 행렬의 첫 번째 행 벡터는 1이므로, 즉 $z_{\tau,1} = 1$, $b_{1,i}$는 켑스트럼 특징 공간에서의 절편(intercept)을 의미하게 되고 나머지 부분은 분절 특징의 형태를 나타내는 경향 정보로 해석할 수 있다. 따라서, 궤적 표현에서 절편을 제외한 나머지 부분을 공유한다면, 다른 궤적 특징과 경향 정보를 공유한다고 할 수 있다. 이런 가정을 바탕으로 경향 부분의 양자화 과정을 통하여 매개 변수 수를 줄이고자 하였다.

분절 특징 HMM에서는 현재의 프레임 관측 벡터는 분절의 중앙에 존재한다. 따라서 $b_{1,i}$는 궤적 표현에 의해 평활화(smoothing)된 가운데 점의 위치를 나타내므

로, 궤적 행렬의 첫 번째 행 벡터 $\boldsymbol{b}_1$는 D차원 위치를 의미하고, 나머지 부분은 $(R-1)\times D$ 차원의 경향을 의미하게 된다. 경향 벡터를 공유하기 위해서 경향 양자화 방법을 이용한다. 경향 양자화 방법을 이용하여 각 분절 특징을 구성하는 경향 벡터는 가장 가까운 코드워드(codeword)로 교체된다. 경향 벡터가 이미 학습된 코드북(codebook)의 공유된 경향 $\widehat{\boldsymbol{T}}_t$으로 교체된 후, 기존의 행벡터 $\boldsymbol{b}_1$과 병합되어 최종 특징 벡터로 사용된다. 실험 결과, 제안된 시스템은 일반 HMM보다 우수한 성능을 보이나, 기존의 SFHMM에 비해 성능저하가 크지 않다고 보고되었다. 또한 동일 혼합 밀도를 사용하는 경우, 분절 길이가 확장될수록 더욱 더 많은 코드북이 필요하고, 혼합 밀도의 수가 증가할수록 마찬가지로 더 많은 코드북이 필요하다는 것을 성능의 변화를 통하여 알 수 있었다(Yun 2002).

3.5. 분절 프레임의 우도 계산

분절 특징 HMM은 일반 HMM보다 매개 변수의 수가 많기 때문에, 성능 향상과 더불어 매개 변수 수를 감소시키는 방향으로 연구가 필요하다. 그러나, 매개 변수 수를 급격히 감소시키게 되면 분절 특징 HMM의 성능은 일반 HMM의 성능과 유사하게 된다. 특히, 경향 공유와 같이 매개 변수 수를 줄이기 위하여 양자화 과정을 거치게되면, 양자화 과정에서 소비되는 시간 때문에 인식 시간은 더욱 늘어나게 된다. 따라서, 기존의 분절 특징 HMM을 크게 수정하지 않고 계산 시간을 빠르게는 연구가 필요하다고 할 수 있다. 분절 특징 HMM에서 한 분절은 여러 프레임 특징에 해당된다. 또한, 각 분절은 주위의 분절과 중첩되어 계산되므로, 평활화 작용에 의하여 그 분절을 대표하는 궤적이 달라진다고 할 지라도 정보의 중복이 발생된다고 할 수 있다. 이들 분절의 중첩에 의하여 발생될 수 있는 중복된 정보를 최소화하여 모델링한다면 계산 시간을 줄일 수 있을 것이다. 일반적으로 선형 시스템인 경우, 직선 형태의 궤적이 형성되기 때문에 두 점으로 표현할 수 있다. 그러나 두 점으로 직선을 표현하는 경우, 기울기에 의하여 발생되는 차이를 제대로 모델링할 수 없게 된다. 따라서 기울기에 의하여 발생되는 차이를 모델링하기 위해서는 적어도 세 점이상으로 모델링하여야 한다 (그림 1 참조).

<그림 1> 두 점(프레임) 비교에
의한 분절의 궤적 비교

비슷한 방법으로 2차 곡선으로 분절을 모델링하는 경우, 적어도 네 점이상으로 모델링하여야 하나, 2차 곡선 이상의 궤적 시스템인 경우 오차 계산에 의하여 세 점으로 모델링하더라도 큰 차이는 없을 것으로 보인다. 특히 분절 길이가 작은 경우에는 네 점이상으로 모델링하는 것은 과추정(over-estimation)으로 보일 수 있어, 본 연구에서는 분절 길이가 5인 경우에는 세 점 모델링을 제안한다. 만약 분절 길이가 7인 경우에 회귀 차수가 3보다 큰 경우 (즉, 3차 곡선으로 모델링하는 경우)에는 네 점 모델링이 적합하다고 생각한다. 이와 같이 분절 길이와 회귀 차수가 사전에 정의되면 적절한 모델링 방법을 선정하여 인식 시간을 빠르게 할 수 있을 것이다.

4. 실험 및 결과

분절 특징의 특성을 파악하고 3장에서 제안한 세 점 모델링의 유효성을 확인하기 위하여 모음 분류 실험을 하였다. 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 데이터베이스의 발음 기호로부터 추출되었다. 실험에 사용된 특징은 26차 또는 39차의 기본 특징 벡터로부터 추출한 분절 특징을 사용하였다. 문장에서 모음을 추출하여 학습에 41,429개의 모음을 사용하였으며, 평가에는 15,119개의 모음을 이용하였다.

첫 번째 실험은 분절 특징의 동적 특성을 살펴보기 위하여 26차의 특성에 기반한 분절 특징 HMM의 성능과 39차 특징에 기반한 분절 특징 HMM의 성능을 비교하였다. 실험 결과 26차 프레임 특징에 비해 26차 특징에 기반한 분절 특징 HMM이 더 낮은 성능을 보이고 있으나, 39차 프레임 특징을 이용한 HMM에 비해 분절 길이가 5이고 혼합 밀도가 2인 경우에 비슷하거나 높은 성능을 보이고 있다. 이것은 기존 연구에서는 가변 분산을 사용한 반면에 본 연구에서는 고정 분산을 이용하였기 때문에 나타난 현상이다.

<표 1> 26차 특징과 39차 특징에 기반을 둔 분절 특징을 이용한 경우의 분절 특징 HMM의 성능 비교 (M: 혼합 밀도의 수)

	M=1	M=2
HMM	52.88	55.86
N=3, R=2	53.56	56.69
N=3, R=3	53.66	56.74
N=5, R=2	54.62	57.22
N=5, R=3	54.63	57.25

(a) 26차 특징을 이용

	M=1	M=2
HMM	54.92	57.20
N=3, R=2	55.22	57.57
N=3, R=3	55.18	57.40
N=5, R=2	55.52	57.57
N=5, R=3	55.50	57.83

(b) 39차 특징을 이용

또한 39차 특징에 기반한 분절 특징 HMM의 경우, 동일한 차수의 프레임 특징을 이용한 일반 HMM에 비해 뚜렷한 성능 향상을 보이고 있지 않다. 이것은 음소 분류인 경우, 2차 미분 계수 이상의 특징은 일반적인 음소 길이 이상의 음성 특징을 모델링하고 있기 때문으로 판단된다.

두 번째 실험은 분절 특징에 의하여 분절간의 비교시 세 점(프레임) 비교를 통한 성능 차이를 비교한 것이다. 분절 길이가 3인 경우에는 분절 내의 프레임 수가 세 프레임이기 때문에, 분절 길이가 5인 경우에만 고려하였다. 세 프레임 비교에 의한 분절 비교 시 학습 단계에서는 5 프레임 비교를 한 경우와 학습 단계에서도 3 프레임 비교를 한 경우에 대해 성능을 평가하였다. 이 실험에서는 모두 26차 특징에 기반을 둔 분절 특징을 사용하였다.

<표 2> 세 프레임 비교에 의한 분절 특징 HMM의 성능 비교

	M=1	M=2
N=5, R=2	50.69	54.22
N=5, R=3	50.26	53.77

(a) 평가 단계에서만 세 프레임 비교

	M=1	M=2
N=5, R=2	54.95	57.77
N=5, R=3	54.87	58.17

(b) 학습, 평가 단계에서 모두 세 프레임 비교

위 실험에서 알 수 있듯이 세 프레임 비교에 의한 분절 비교는 26차 특징에 기반한 분절 특징을 사용하였음에도 불구하고 36차 특징에 기반을 둔 경우보다 더 좋은 성능을 보임을 알 수 있다. 이것은 궤적 시스템에 의하여 분절을 모델링하는 경우 세 프레임에 의한 비교는 궤적 표현에서 어느 정도 잡음 성분이 포함되어 평활화가 진행된 것으로 파악된다. 위의 실험에 기초하여 혼합 밀도의 수를 증가할 경우에 각 시스템의 성능을 파악하고자 일반 HMM과 N=5, R=3인 경우에서 26차, 39차 특징에 기반한 분절 특징 HMM의 성능 비교 실험을 하였다.

<표 3> 혼합 밀도의 수가 10인 경우의 26차 특징과 39차 특징에 기반한 시스템의 성능 비교

특징 차수	시스템	M=10
26차	HMM	62.37
	N=5, R=3	64.04
39차	HMM	62.57
	N=5, R=3	64.24

이 결과로부터 분절 특징 HMM을 음성 인식 시스템에 사용하고자 하는 경우, 분절 길이가 5이고 회귀 차수가 3인 분절 표현 시스템에서 26차 특징에 기반한 세 프레임 계산이 성능면에서나 계산 시간면에서 적합함을 알 수 있다. 또한 실험 결과로부터 음소 모델에 기반한 모음 분류 실험인 경우, 26차 특징 벡터를 이용하거나 39차 특징 벡터를 이용한 경우의 성능 차이가 크지 않음을 알 수 있었다. 그 이유는 음소 모델인 경우 음성의 길이가 짧기 때문에 정적 특징의 2차 미분(delta-delta) 값의 영향이 작은 것으로 판단된다.

5. 결 론

본 연구에서는 분절 특징 HMM의 특성을 정리하고, 지금까지의 연구 진행상황을 살펴보았다. 분절 특징 HMM이 비록 일반 HMM보다 성능이 우수하다고 할지라도 모델을 표현하는 매개 변수의 수가 많기 때문에 학습 시간이나 인식 시간이 많이 걸리는 단점이 있다. 따라서, 매개 변수의 수를 줄이려는 연구가 필요하다. 그렇지만 모델의 표현 방법에서 매개 변수를 줄이는 연구는 인식 시스템의 성능까지도 같이 저하시키는 효과를 가져와(Yun 2002), 본 연구에서는 매개 변수의 수보다도 계산 시간의 감소를 가져올 수 있는 세 점(프레임) 비교를 통한 분절 비교를 제안하였다. 제안된 시스템은 기존의 시스템보다도 동일한 환경에서 더 나은 성능을 보여, 계산 시간 면에서나 성능 면에서 우수함을 알 수 있었다.

참 고 문 헌

윤영선, 오영환(2000a), 모수적 궤적 기반의 분절 HMM을 이용한 연속 음성 인식, 「음향학회지」 **19**(3), pp. 35～44.

윤영선, 오영환(2000b), 분절 특징 HMM의 매개 변수 수의 감소에 관한 연구, 「음향학회지」 **19**(7), pp.48～52.

윤영선(2002), 분절 특징 HMM을 이용한 영어 음소 인식, 「한국정보과학회지」 **29**(3), pp.167～179.

Gales, M. J .F. and S. J. Young (1993), Segmental Hidden Markov Models, In *Proceedings of European Conference on Speech Communication and Technology*, pp.1579~1582.

Ostendorf, M., V. Digalakis and O. A. Kimball (1996), From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition, *IEEE Tr. on Speech and Audio Processing* **4**(5), pp. 360~378.

Gish, H. and K. Ng (1996), Parametric trajectory models for speech recognition, In *Proceedings of International Conference on Spoken Language Processing*, pp.I-466~469.

Holmes, W. J. and M. J. Russell (1999), Probabilistic trajectory segmental HMMs, *Computer Speech and Language* **13**, pp.3~37.

Furui, S. (1986), Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum, *IEEE Trans. on Acoustics, Speech and Signal Processing* **34**(1), pp.52~59.

Deng, L. (1992), A generalized hidden Markov model with state conditioned trend functions of time for the speech signal, *Signal Processing* **27**, pp.65~78.

Deng, L., M. Aksmanovic, D. Sun and J. Wu (1994), Speech recognition using hidden Markov models with polynomial regression functions as non-stationary states, *IEEE Trans. on Speech and Audio Processing* **2**(4), pp.507~520.

Gish, H. and K. Ng (1993), A segmental speech model with application to word spotting, In *Proc. of Int. Conf. on Acoustics, Speech and Signal Proc.*, pp.II-447~450.

Yun, Y.-S. (2002), Sharing trend information of trajectory in segmental-feature HMM, In *Proc. of Int. Conf. on Spoken Language Processing*, Sep., Denver, pp.2641~2644.

접수일자 : 2002년 5월 3일
게재결정 : 2002년 5월 24일

▶ 윤영선 (Young-Sun Yun)
주소: 대전시 대덕구 오정동 133번지
소속: 한남대학교 정보통신공학과
전화: 042) 629-7569
Fax: 042) 629-7843
E-mail: ysyun@mail.hannam.ac.kr

▶ 정호영 (Ho-Young Jung)
주소: 대전시 유성구 가정동 161번지
소속: 한국전자통신연구원 음성정보연구센터
전화: 042) 860-1328
Fax: 042) 861-1342
E-mail: hjung@etri.re.kr