분절 특징을 이용한 음성 신호의 모델링

윤영선 ° 오영환
한국과학기술원 전산학과
{ysyun, yhoh}@bulsai.kaist.ac.kr

Modeling of Speech Signals Using Segmental-Features

Young-Sun Yun ° Yung-Hwan Oh
Dept. of Computer Science, KAIST

요 약

본 논문에서는 분절 특징을 모수적 궤적 모델을 이용하여 표현하고, 이 특징을 분절 HMM (segmental HMM)의 입력으로 하는 음성 신호의 모델링 방식을 제안한다. 분절 특징은 음성의 경향을 나타내는 궤적으로 표현되고, 그 궤적은 연속되는 프레임 상에서 전이 정보를 포함하도록 디자인 행렬과 다항식의 회귀 함수를 이용하여 구해진다. 이 궤적을 분절 HMM에 적용하기 위하여, 외적 분절 변이와 내적 분절 변이에 대한 확률 분포 표현을 개선하였다. 제안된 방법의 효과를 살펴보기 위하여 TIMIT 데이터 베이스를 이용하여 실험한 결과, 제안된 분절 특징은 음성 신호의 인접한 프레임간의 상관관계를 표현하는 동적 특징과 같은 효과를 보였으며, 1차 미분계수를 포함하여 분절 특징을 구한 경우에는 기존의 특징 표현 보다 좋은 성능을 보였다.

1. 서론

HMM(hidden Markov model)은 구현하기 쉽고 유연한 모델링 능력을 가지고 있기 때문에 음성 처리를 비롯한 여러 분야에서 사용되어 왔으며, 그 성능을 인정받고 있다. 특히 HMM은 시간과 주파수 상에서의 변이를 고려하기 때문에 음성 패턴을 모델링하는데 적합한 구조를 제공하고 있다. 그러나, HMM의 기초가 되는 가정으로 인하여 실제 환경에서는 연속된 특징 벡터들의 시간적인 종속성을 제대로 표현하지 못한다고 알려져 있다. 이러한 약점을 보완하기 위하여 다양한 모델을 이용한 연구들이 제안되었다. 대표적인 연구로는 특징 추출 단계에서 인접한 특징들간의 관계를 표현하는 동적 특징 (dynamic feature) [1]과, 음향학적 모델을 개선하여 인식 성능을 높이려는 분절 모델 (segmental model) [2, 3, 4] 등이 있다. 그러나, 동적 특징은 켑스트럼 계수 (cepstral coefficients) 등과 같은 정적 특징 (static features)을 선형 회귀 함수를 이용하여 여러 프레임에서의 평균 변이를 표현하는 방법이기 때문에 좀 더 유연성 있는 특징의 표현이 필요하다. 또한, 분절 모델은 음성 문맥이 음성 신호의 궤적에 큰 영향을 끼친다는 사실에 입각하여 기본 단위를 주어진 음성 단위에서 관측된 음성 벡터 열로 보고 모델링하는 방식이나, 정확한 분절 정보를 얻지 못할 때에는 많은 계산을 요구하며, 가변길이의 신호를고정 길이의 분절 특징으로 표현하는데 문제점을 가지고 있다. 따라서 본 연구에서는 인접한 특징의 시간적인 종속성을 효율적으로 표현하는 특징의 표현 방법을 제안하고, 이 특징을 이용하여 인식 과정에 효율적으로 적용할 수 있도록 분절 HMM의 확률 분포를 개선한 분절 특징 HMM (SFHMM; segmental-feature HMM)을 제안한다.

2. 분절 특징 (Segmental feature)

연속된 음향학적 특징 벡터 열간의 관계는 특징 공간에서 궤적의 형태로 근사화될 수 있다. 이런 생각은 많은 분절 모델의 기본이 되었으며 모수적 (parametric) 또는 비모수적 (non-parametric)인 방법에 의하여 표현될 수 있다. 모수적 방법은 특정 영역으로부터 다항식의 궤적을 추정하며, 그 궤적으로부터 계산되는 점들로서 그 영역의 분포를 나타내는 방법이다. 반면에, 비모수적 방법은 각각의 영역에 대하여 확률 분포 함수로 표현하는 방법이다 [2]. 본 연구에서는 모수적 방법이 여러 음성 단위를 평활화시키는 경향이 있고 잡음 환경이나 환경 변화에 강점을 가지고 있기 때문에, 이 방법을 이용하여 분절의 궤적을 표현한다.

Gish 등이 제안한 모수적 궤적 모델에서는 N개의 프레임으로 구성된 음성 분절 $\mathbf{C}$를 디자인 행렬 $\mathbf{Z}$와 계수 행렬 $\mathbf{B}$, 그리고 잔차 행렬 $\mathbf{E}$를 이용하여 다음과 같이 표현하였다 [3].

$$\mathbf{C} = \mathbf{ZB} + \mathbf{E}, \tag{1}$$

여기에서 $\mathbf{C}$는 $N \times D$ 크기의 분절을 나타내는 행렬이며, $\mathbf{Z}$는 분절의 적용 범위를 결정하는 $N \times R$ 디자인 행렬, $\mathbf{B}$는 $R \times D$의

계수 행렬을 나타낸다.

음성 분절을 완전 궤적 (complete trajectory)으로 표현하는 기존의 연구는 궤적 특징을 $[0,1]$의 구간으로 정규화하여 표현하였다. 따라서, 가변의 길이를 갖는 음성 분절을 정규화하기 위해서는 음성 분절의 구간을 반드시 알아야 하며, 그렇지 않은 경우에는 분절 구간의 결정에 많은 시간을 필요로 한다. 따라서, 본 연구에서는 가변 길이의 분절 대신 고정 길이의 분절을 이용하여 분절 구간 결정의 문제점을 해소하며, 계산 시간을 줄이도록 하였다. 고정 길이의 분절을 이용하기 위해서는 음성 분절을 변화되는 시간에 따라 표현하여야 하므로, 관측 벡터 $\mathbf{Y}=\{\mathbf{y}_1,\ldots,\mathbf{y}_T\}$가 주어지면 음성 분절을 $\mathbf{C}_t=\mathbf{Y}_{t-M}^{t+M}=\{\mathbf{y}_{t-M},\cdots,\mathbf{y}_{t+M}\}$로 표현한다. 이 음성 분절을 효과적으로 표현하기 위해서는 분절의 범위를 결정하는 디자인 행렬을 수정하여야 한다. 분절의 중앙에는 현재 관측 벡터가 위치하므로, 디자인 행렬에서도 현재의 관측 벡터를 중앙에 위치하도록 수정하였다. 또한, 분절의 앞 부분과 뒷 부분이 시간의 흐름에 따라 이전 분절 또는 이후 분절과 중첩될 수 있기 때문에 상대적인 시간의 흐름을 표현하기 위하여 디자인 행렬의 전반부에는 음의 값을, 후반부에는 양의 값을 갖도록 하여, 전체 상대적인 시간을 $[-0.5,0.5]$로 정규화하여 분절의 길이에 분절 특징이 종속되지 않도록 하였다 [5].

분절의 범위가 결정되면, 분절의 특징을 표현하는 궤적 계수 행렬 $\mathbf{B}_t$는 행렬 연산이나 선형 회귀 방정식에 의하여 구해질 수 있으며, 추정된 궤적 $\mathbf{ZB}_t$와 원래의 음성 분절 $\mathbf{C}_t$의 차이에 의해 근사 적합도 (goodness-of-fit) χ_t^2를 구할 수 있다.

3. 분절 특징 HMM (Segmental-feature HMM)

주어진 음성 분절에 대해 궤적을 나타내는 궤적 계수 $\mathbf{B}_t$와 근사 적합도 χ_t^2가 구해지면, 이들 특징은 관측 확률과 분절간의 우도 (likelihood)를 계산하는데 사용된다. 분절 HMM은 음성 신호의 기본 궤적을 효과적으로 표현한다고 알려져 있어, 분절 특징을 분절 HMM에 적용하고자 하였다. 기존의 연구에서는 임의의 분절에 대한 관측 확률을 외적 분절 변이 (extra-segmental variation)와 내적 분절 변이 (intra-segmental variation)의 곱으로 표현한다.

$$P(\mathbf{C}_t|s_i,\lambda)=P(\mathbf{ZB}_t|s_i,\lambda)P(\mathbf{C}_t|\mathbf{ZB}_t,s_i,\lambda). \quad (2)$$

여기에서 외적 분절 변이 $P(\mathbf{ZB}_t|s_i,\lambda)$는 화자의 특성이나 발음상의 변이를 나타내며, 내적 분절 변이 $P(\mathbf{C}_t|\mathbf{ZB}_t,s_i,\lambda)$는 연속적인 조음 현상과 임의의 파동으로 발생되는 분절 내에서의 변이를 말한다. 그러나, 본 연구에서는 외적 분절 변이를 분절간의 관계를 나타내는 확률 분포로 표현하고, 내적 분절 변이를 분절과 궤적간의 추정 오차로 표현하고자 한다.

주어진 상태에서의 분절간의 관계는 그 상태에서의 평균 궤적과 분산이 주어지면, 관측 가능한 분절간의 우도에 의해서 계산된다. 반면, 내적 분절 변이는 주어진 분절과 그 분절에서 추정된 궤적간의 오차로서 표현되기 때문에 근사 적합도를 이용하여 정의할 수 있다. 다음은 외적 분절 변이를 나타내는 분절간의 우도와 그에 대응되는 내적 분절 변이를 표현하고 있다.

$$\begin{aligned}P(\mathbf{ZB}_t|s_i,\lambda)&=P(\mathbf{ZB}_t|\mathbf{Z}\mathcal{B}_i,\mathbf{\Sigma}_i)\\&=\prod_{\tau=t-M}^{t+M}\frac{1}{(2\pi)^{D/2}|\Sigma_{\tau-t,i}|^{1/2}}\cdot\\&\exp\left\{-\frac{1}{2}\left\{\mathbf{z}_\tau(\mathbf{B}_t-\mathcal{B}_i)\right\}\Sigma_{\tau-t,i}^{-1}\left\{\mathbf{z}_\tau(\mathbf{B}_t-\mathcal{B}_i)\right\}'\right\},\end{aligned} \quad (3)$$

$$P(\mathbf{C}_t|\mathbf{ZB}_t,s_i,\lambda)=P(\mathbf{C}_t|\mathbf{ZB}_t)=\exp\left\{-\frac{1}{2}\chi_t^2\right\}. \quad (4)$$

따라서 시간 t에 상태 j에서의 임의의 분절에 대한 관측 확률은 다음과 같이 계산된다.

$$\begin{aligned}b_j(\mathbf{C}_t|\mathbf{ZB}_t)&=P(\mathbf{C}_t|s_j,\lambda)\\&=P(\mathbf{ZB}_t|\mathbf{Z}\mathcal{B}_j,\mathbf{\Sigma}_j)P(\mathbf{C}_t|\mathbf{ZB}_t),\end{aligned} \quad (5)$$

여기에서 $\mathcal{B}_j$와 $\mathbf{\Sigma}_j$는 상태 j의 평균 궤적에 대한 궤적 계수 행렬과 분산을 나타내고 있다.

SFHMM의 한 상태에서의 평균 궤적은 다음과 같이 그 상태를 지나는 기대 값에 의한 평균치로 구할 수 있다.

$$\mathbf{Z}\bar{\mathcal{B}}_j=\frac{\sum_{t=1}^{T}\xi_t(j)\mathbf{ZB}_t}{\sum_{t=1}^{T}\xi_t(j)}. \quad (6)$$

이때 각 궤적은 공통적인 디자인 행렬 $\mathbf{Z}$를 사용하고 있기 때문에 양측에서 디자인 행렬을 소거하여 궤적 계수 행렬 $\bar{\mathcal{B}}_j$를 쉽게 구할 수 있다. 평균 궤적이 구해지면, 이 궤적으로부터 분산을 구할 수 있다. 본 연구에서는 한 상태에서 관측 가능한 분절에 대한 분산을 분절의 각 프레임에서 구한 시변 분산을 사용하고 있기 때문에 다음과 같은 식으로 구할 수 있다.

$$\bar{\Sigma}_{n,j}=\frac{\sum_{t=1}^{T}\xi_t(j)\left\{\mathbf{z}_n(\mathbf{B}_t-\bar{\mathcal{B}}_j)\right\}'\left\{\mathbf{z}_n(\mathbf{B}_t-\bar{\mathcal{B}}_j)\right\}}{\sum_{t=1}^{T}\xi_t(j)}, \quad (7)$$

여기에서 n은 분절에서의 상대적인 프레임 위치를 나타내며, $\mathbf{z}_n\mathbf{B}_t$와 $\mathbf{z}_n\bar{\mathcal{B}}_j$는 추정된 궤적과 상태 j의 평균 궤적으로부터 복원된 점들을 나타낸다.

4. 실험 및 검토

분절 특징의 유효성을 검증하기 위하여 정적 특징에 기반한 분절 특징을 사용하는 SFHMM과 정적 특징과 동적 특징을 동시

표 1: 정적 특징과 1차 미분 계수를 이용한 HMM과 정적 특징에 기반한 분절 특징을 이용한 SFHMM의 성능 평가

시스템	단일 혼합모델	두 혼합모델
기본시스템	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

표 2: 정적 특징과 1,2차 미분 계수를 이용한 HMM과 1차 미분 계수를 포함한 분절 특징을 이용한 SFHMM의 성능 평가

시스템	단일 혼합모델	두 혼합모델
기본시스템	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

에 반영한 일반 HMM의 성능 비교를 하였다. 인식 실험에 사용된 모델은 TIMIT 데이터베이스에 기반한 48 음소 모델을 사용하였으며, 인식 후 39개의 음소로 병합하여 결과를 계산하였다. 사용된 정적 특징은 12개의 MFCC와 로그 에너지를 사용하였으며, 동적 특징인 경우 이들 정적 특징의 1, 2차 미분 계수를 이용하였다. 학습에는 462명의 화자가 발성한 4,620 문장을 사용하였으며, 테스트에는 168명이 발성한 1,680 문장을 이용하였다.

먼저 1차 실험은 26차의 정적 특징과 1차 미분 계수를 사용하는 HMM과 13차의 정적 특징에 기반한 분절 특징을 사용하는 SFHMM의 성능을 평가하였다. SFHMM의 성능은 분절 길이 N과 회귀 차수 R의 변화에 따라 평가되었으며, 실험 결과는 표 1에 나타나 있다. 실험 결과 작은 분절 길이에서는 기본 시스템의 성능보다 저하되었지만, 분절 길이를 증가시키면서 회귀 차수를 높이면 성능이 향상됨을 알 수 있었다. 동일한 조건에서 정적특징에 대한 2차 미분 계수를 기본 시스템에 포함시키고, 1차 미분 계수를 포함하여 분절 특징을 구한 경우의 성능 평가는 표 2에 나타나 있다.

실험 결과, 정적 특징만을 이용하여 분절 특징을 표현한 경우에는 충분한 분절 길이와 회귀 차수가 주어지면 동적 특징의 경우보다 좋은 성능을 보였다. 또한, 1차 미분 계수를 포함한 경우 항상 기본 시스템보다 좋은 성능을 보여 제안된 분절 특징이 동적 특징보다 낳은 특성을 보임을 알 수 있었다.

5. 결론

본 연구에서는 인접한 프레임간의 시간적 상관관계를 표현하는 분절 특징을 제안하였다. 제안된 분절 특징은 현재의 관측 벡터를 분절의 중앙에 위치시켜, 인접한 분절간의 중첩을 허용하였다. 그 결과, 문맥 독립 모델로 인식 모델을 구성하더라도 음소의 경계 부분에서 인접한 음소의 정보를 가지게 되므로 부분적인 문맥 정보를 표현할 수 있다. 실험 결과, 충분한 분절 길이와 회귀 차수로 모델링된다면 정적 특징만을 이용하여도 동적 특징과 같은 특성을 갖게 됨을 알 수 있었다. 이 특징을 이용하여 분절의 관측 확률이나 우도 계산을 할 수 있도록, 기존의 분절 HMM 모델을 개선하여 분절 특징 HMM을 제안하였다. 개선된 모델은 외적 분절 변이를 분절간의 상관관계로 표현하고 평균 궤적과 분산을 이용하여 구하였으며, 내적 분절 변이는 추정된 궤적과 입력 음성 신호와의 추정 오차로 표현하였다.

6. 참고문헌

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

[2]. M. Ostendorf, V. Digalakis, O.A. Kimball, "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, 1996

[3]. H. Gish, K. Ng, "Parametric trajectory models for speech recognition," *In Proc. of Int. Conf. on Spoken Lang. Proc.*, pp. I-466–469, 1996

[4]. M. Russell, "A segmental HMM for speech pattern modeling," *In Proc. of the Inter. Conf. on Acosut., Speech and Signal Proc.*, pp. II-499–502, 1993

[5]. Y.-S. Yun, Y.-H. Oh, "A Segmental-Feature HMM For Speech Pattern Modeling," *IEEE Signal Processing Letters*, **7**(6), pp. 135–137, June 2000