세그멘트 통계량을 이용한 HMM의 한국어 음절 인식

Syllable Reconition by HMM Using Segmental Statistics

°박창호*, 이영재**, 허강인*

°Chang Ho Park*, Young Jae Lee**, Kang In Hur*

* 동아대학교 전자공학과
** 창신전문대학 전산정보처리과

요 약

기존의 연속 출력 분포형 HMM은 시계열의 과도적 변화에 대하여 표현 능력이 부족하다는 단점이 있다. 이것을 보완하기 위해 본 논문에서는 음성의 동적 변화를 반영하기 위한 특징 파라메타로서 여러 개의 프레임을 결합하여 세그멘트를 구성하여 각각에 대해 한개의 벡터를 만들었다. 이것을 그대로 이용하면 세그멘트의 프레임수에 대응하는 파라메타의 차원수가 증가하기 때문에 학습 데이터가 불충분한 경우 모델의 파라메타를 잘 추정할 수 없으므로 K-L전개로서 파라메타의 차원을 압축하여 파라메타수를 감소시켰다.

인식실험은 한국어 단음절에 대하여 멜캡스트럼을 K-L전개로 압축한 벡터를 이용한 결과와 멜캡스트럼, 멜캡스트럼 + 선형회귀계수를 파라메타로 이용한 경우를 비교하였다.

실험결과 K-L전개로 압축한 벡터만을 이용한 경우는 멜캡스트럼 + 선형회귀계수를 파라메타로 이용한 경우보다 인식율이 낮았으나 멜캡스트럼 + K-L전개로 압축한 경우와 거의 동등한 결과를 얻을 수 있었다.

Ⅰ. 서론

기존의 HMM(hidden markov model)에서 멜캡스트럼만을 특징 파라메타로 이용할 경우 시계열의 과도적인 변화에 대한 표현 능력이 부족하다는 단점이 있다. 그래서 음성파라메타에 대해 동적 변화를 고려하기 위한 방법으로 상대수를 늘이는 방법, 시간축 방향의 회귀계수를 파라메타로 이용하는 방법, 선형변화나 비선형예측기를 이용하는 방법등이 연구되고 있다.[1] [2]

특히, 음성 특징량의 동적변화를 HMM에 도입하기 위하여 멜캡스트럼에 대해 여러개의 프레임을 결합해서 고정장의 세그멘트를 구성하여 이 세그멘트를 벡터로 표현하는 경우가 있다. 이 경우 세그멘트내의 프레임수에 대응해서 추정 파라메타의 차원수가 증가하기 때문에 학습데이터가 불충분할 경우 모델의 파라메타를 충분히 추정할 수 없게 된다. 따라서 파라메타의 차원을 압축해서 추정할 파라메타의 차원수를 감소시킬 필요가 있다. 압축방법으로는 K-L(Karhanen-Loeve)전개에 의한 방법과 NN에 의한 방법등이 있다.[3]

본 논문에서는 4개 혹은 7개의 프레임을 결합하여 이 세그멘트를 한개의 벡터로 표현하여 특징량의 동적변화로 이용하였으며 차원 압축은 K-L전개를 이용하였다. 인식 실험에서는 한국어 단음절 100개에 대해 멜캡스트럼, 회귀계수와 K-L전개에 의한 압축데이터를 이용하여 비교 분석하였다.

Ⅱ. 차원 압축법

2.1 K-L전개

몇개의 프레임을 결합하여 이 세그멘트를 1개 벡터로 하여 음성 특징량으로 취급하면 추정해야 할 벡터의 차원수가 증가한다. 그러므로 세그멘트를 표현하는 벡터에서 분산이 적은 1차 결합성분을 소거하고 큰 분산을 갖는 성분을 취하기 위해 K-L전개를 다음과 같은 순서로 한다.

N : 파라메타의 차원

x_i : 샘플 ($i=1,\cdots,I$)

$x_i=[\ x_i^1, x_i^2, \cdots, x_i^N\ ]^T$

단, 기호 T는 전치를 나타낸다.

x_i^k : 벡터 x_i의 k번째 요소

x_M : 벡터의 평균 벡터

z_i : $x_i - x_M$

p : 차원 압축 후의 특징 파라메타의 차원

y_i : 차원 압축 후의 특징 파라메타

$y_i=[\ y_i^1, y_i^2, \cdots, y_i^p\ ]^T$

① 샘플에 의한 공분산 행렬 $A=[\ a_{lm}\ ]$ 의 추정.

$$a_{lm}=-\sum_{i=1}^{I} z_i^l z_i^m \qquad (1)$$

② 고유치 $\{\ \lambda_i\ \}$와 고유벡터 $\{\ \phi_i\ \}$를 계산한다.

$$A\phi_i=\lambda_i\phi_i \qquad (2)$$

③ $\lambda_1\geq\lambda_2\geq\lambda_3\geq\cdots\geq\lambda_p$로 되도록 고유치와 그것에 대응하는 고유벡터를 정렬한다.

④ 압축후의 파라메타를 계산한다.

변환행렬을 $B=[\ \phi_1\phi_2\cdots\phi_p\ ]^T$로 하면

$$y_i=Bx_i \qquad (3)$$

2.2 K-L전개에 의한 차원 압축

그림 1(a)와 (b)에 각각 4프레임복과 7프레임복을 세그멘트로 하였다. 1세그멘트는 40차원이다. 이 세그멘트를 1프레임씩 이동시키면서 전음절 구간에 대하여 K-L전개를 한다.

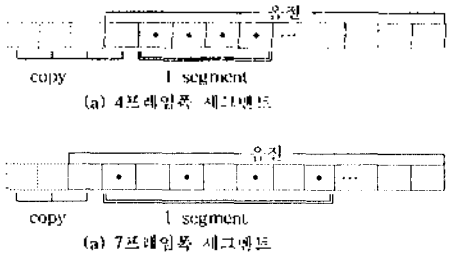

그림 1. 세그멘트 구간

K-L전개에 의한 데이터 압축정도는 학습데이터에 대해서 공분산행렬의 고유치의 누적기여율로 평가했다. 누적기여율 Θ_i는 다음과 같이 정의한다.

$$\Theta_i = \lambda_i / \sum_k \lambda_k \quad (\lambda_k \geq \lambda_{k+1}) \tag{4}$$

표 1의 누적기여율에서 4프레임복보다 7프레임복의 세그멘트를 압축한 경우가 파라메타의 변화가 크며 높은 차원으로 갈수록 오차가 적어짐을 알 수 있다.

그리고, K-L전개에 의한 세그멘트내의 스펙트럼의 동적특징량이 잘 보존되었는지를 보기 위해 K-L전개로서 압축된 데이터를 원래의 세그멘트(4프레임)로 복원하여 프레임간 자승오차와 인접하는 2프레임간의 평균 자승거리를 비교하였다.

여기서 복원오차가 인접 2프레임간 평균거리보다 적으면 복원된 벡터계열이 원래의 인접 프레임보다 동적특징량을 보존하고 있다고 할 수 있다.

또 회귀계수를 파라메타에 부가한 경우, 세그멘트내의 프레임은 직선에 근사화할 수 있으므로 이 1차 직선과 실제 프레임과의 거리(오차)를 구하였다.

프레임 i에서 제 n차원의 최소자승 오차직선(산형회귀계수)의 기울기와 오차직선의 절단 및 직선으로 근사화시킨 경우의 평균 자승오차를 각각 식(5), (6), (7)으로 나타낸다.

$$a_i^n = \sum_{k=-w}^{w} \cdot y_{i+k}^n \Big/ \sum_{k=-w}^{w} k^2 \tag{5}$$

$$b_i^n = \sum_{k=-w}^{w} y_{i+k}^n \Big/ 2w+1 \tag{6}$$

$$\text{평균오차} = \frac{1}{2w+1} \sum_{k=-w}^{w} \sum_n (y_{i+k}^n - k \cdot a_i^k - b_i^n)^2 \tag{7}$$

단 y_i^n은 프레임 i의 제 n차원 요소이고, 2w+1은 회귀계수의 계산시간 프레임복이 된다. 본 실험에서는 w=5로 하였다.

표 2에 인접 2프레임간 평균거리, 최소자승 직선에 근사화시킨 경우의 오차, K-L전개에서의 압축오차를 나타내었다. 자승오차 직선의 근사화시킨 경우는 인접 2프레임간의 평균거리보다 적었다. 그리고 세그멘트의 복이 다르므로 단순하게 비교할 수 없지만 K-L전개에 따른 압축오차(복원오차)는 학습데이터와 평가데이터에서 인접 2프레임간 평균거리보다 적으므로 양호한 스펙트럼의 동적특징량을 보존하고 있었다. 그리고 4프레임복과 7프레임복의 세그멘트에서는 누적기여율과 비례해서 압축오차도 감소함을 알 수 있었다.

표 1. 누적기여율 (20차원까지 나타내었다)

(a) 4프레임복

i	θ_i	i	θ_i
1	0.4251	11	0.9116
2	0.6992	12	0.9651
3	0.7974	13	0.9682
4	0.8420	14	0.9711
5	0.8814	15	0.9736
6	0.9042	16	0.9759
7	0.9218	17	0.9781
8	0.9354	18	0.9800
9	0.9466	19	0.9820
10	0.9556	20	0.9837

(b) 7프레임복

i	θ_i	i	θ_i
1	0.4065	11	0.9453
2	0.6762	12	0.9509
3	0.7725	13	0.9561
4	0.8163	14	0.9600
5	0.8546	15	0.9632
6	0.8774	16	0.9663
7	0.8987	17	0.9694
8	0.9149	18	0.9721
9	0.9275	19	0.9746
10	0.9367	20	0.9769

표 2. K-L전개에서의 압축오차

방법		학습데이타	평가데이타
인접 프레임간 평균거리		0.1040	0.1060
회귀계수(11프레임복)		0.0349	0.0352
K-L (4프레임복)	10차원	0.0212	0.0211
	14차원	0.0154	0.0158
	20차원	0.0096	0.0096
K-L (7프레임복)	10차원	0.0260	0.0254
	14차원	0.0187	0.0183
	20차원	0.0103	0.0102

III. 인식실험 및 고찰

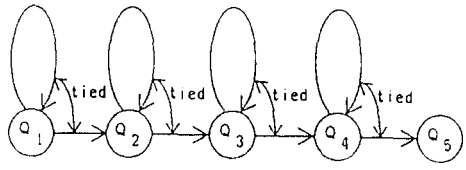

그림 2. 음절인식용 HMM의 구조

그림 2는 연만적인 HMM의 구조를 나타내며 본실험의 경우, 모델은 5상태 4출력분포의 인속출력 분포형으로 하였다. 학습은 Bawm-Welch 알고리즘을 이용하였고, 출력 분포는 다차원 정규분포로 가정하였다.[8]

모델의 초기치 추정은 학습용 단음절 데이터를 균등하게 4개의 구간으로 나누어서 각 구간에 할당된 백디로부터 평균벡터와 공분산행렬의 초기치를 구했다. 최대 반복학습의 회수는 10회로 하였다.

또 스펙트럼의 동적 특징량에 있어서 회귀 계수를 이용한 경우 11프레임복에 대해서 구하였다.

3.1 실험 데이터 및 분석 방법

음성데이터는 남성화자 5명이 5회 발성한 한국어 단음절을 사용하였다.(표 4. 100음절×5명×5회=2500개) 음성의 분석 조건은 표 3과 같다.

표 3. 음성의 분석 조건

샘플링 주파수	10 KHz
창함수(해밍창)	20 ms
프레임주기	5 ms
분석	14차 LPC 분석
특징 파라메타	10차 LPC mel cepstrum 계수

표 4. 한국어 단음절(100개*5명*5회)

가 간 갈 감 거 게 고 구 기
나 난 날 남 너 노 누 니
다 단 달 담 대 도 동 두 디
라 란 람 로 롱 루 리
마 만 말 에 모 무 미
바 반 발 보 봉 부 비
사 산 살 삼 새 소 수 시
아 안 알 암 애 우 이
자 잔 잘 잠 재 조 종 주 지
차 찬 찰 참 채 초 추 치
카 칸
타 탄 탈 탐 토 투 티
파 판 팔 포 피
하 한 할 함 후 히

3.2 실험 결과 및 고찰

본 실험에서는 5명의 화자가 5회 발성한 음절중 3회분을 학습용, 나머지 2회분은 평가용 데이터로 하여 다음의 6가지 방법으로 비교하였다. 각 방법에 대한 평균 음절 인식율을 표 5에 나타내었다.

① CEP
특징파라메타로서 프레임마다 멜켑스트럼을 이용.

② CEP + ΔCEP
①에 동적특징량으로서 회귀계수를 부가.

③ K-L(4프레임폭 세그멘트)
1세그멘트 40차원을 10차원, 14차원, 20차원으로 압축한 것을 1프레임씩 이동하여 이용.

④ K-L(7프레임폭 세그멘트)
1프레임 건너서 4프레임을 1세그멘트(40차원)로 하여 10차원, 14차원, 20차원으로 압축한 것을 1프레임씩 이동해서 이용.

⑤ K-L(4프레임폭 세그멘트) + CEP
③의 압축법에 멜켑스트럼을 부가.

⑥ K-L(4프레임폭 세그멘트) + ΔCEP
③의 압축법에 회귀계수를 부가.

표 5 에서 실험 ②의 경우는 실험 ①의 평균 음절인식율 85.59% 보다 1.57% 인식율이 향상되었다.

표 5. 음절 인식율(단위:%)

방법	세그멘트폭	차원수	평균인식율
①CEP		10차원	85.59
②CEP+ΔCEP		10+10차원	87.16
③④K-L	4프레임폭	10차원	78.72
		14차원	82.27
		20차원	86.46
	7프레임폭	10차원	76.63
		14차원	82.02
		20차원	83.52
⑤K-L+CEP	4프레임폭	10+10차원	82.79
		14+10차원	85.48
		20+10차원	88.75
⑥K-L+ΔCEP	4프레임폭	10+10차원	80.20
		14+10차원	85.18
		20+10차원	88.29

그리고 실험 ③의 경우, 10차원, 14차원 압축인 경우에는 인식률이 각각 78.72 %, 82.27%로 실험 ①,②의 경우보다 안 좋았지만 20차원 압축인 경우에서는 86.45%로 실험 ①의 경우보다 0.86% 향상되었다. 그러나 실험 ②의 경우보다는 인식율이 떨어졌다.[HH]

실험 ④에서는 실험 ①,②,③의 경우보다 낮은 인식율이 나타났다. 이것은 7프레임폭 세그멘트의 경우 1프레임 건너 4프레임을 1세그멘트로 하기 때문에 떨어진 프레임간의 상관관계가 없어졌기 때문이라고 생각된다.

위 결과에서 K-L전개에 의한 압축 데이터만을 이용한 경우는 모음이 같은 다른 음절로의 오인식률이 높았다. 이것은 자음(초성 또는 종성)의 구간(평균 20-30 ms)이 세그멘트폭(4프레임폭일 경우 35ms, 7프레임폭일 경우 50ms)에 긴구간을 갖는 모음(중성)과 함께 포함이 되어 자음이 모음에 영향을 받아 무시되기 때문이다. 실험 ③과 ④의 비교에서 너무 넓은 폭의 세그멘트 역시 이러한 근거로 인식율의 저하를 가져온다는 것을 알 수 있었다.

그리고 실험 ⑤는 10차원 압축인 경우에는 실험 ①의 경우보다 2.81% 인식률이 떨어졌지만 14차원 압축인 경우에는 거의 비슷했고 20차원 압축인 경우에는 실험 ①,②의 경우보다 각각 3.16%, 1.59% 향상되었다.

실험 ⑥에서는 실험 ①의 경우보다 10차원 압축인 경우에는 5.38%, 14차원 압축인 경우에 0.41% 떨어졌으나 20차원 압축인 경우에는 2.7% 향상되었다. 또, 20차원 압축인 경우 실험 ⑤의 경우보다 인식율이 0.46% 떨어졌다.

IV. 결론

본 논문에서는 음성의 동적 변화를 반영하기 위한 특징 파라메타로서 몇개의 프레임을 결합하여 세그멘트를 구성하고, 각각에 대해 한개의 벡터를 만들어 K-L전개로 파라메타의 차원을 압축하여 추정한 파라메타의 차원수를 감소시켰다. 학습 및 인식 실험을 행해서 멜켑스트럼만을 이용한 경우, 멜켑스트럼 + 회귀계수를 이용한 경우, K-L전개에 의한 압축 벡터만을 이용한 경우와 멜켑스트럼을 부가한 경우 및 회귀계수를 부가한 경우에 대하여 한국어 단음절 인식을 하여 비교 분석하였다.

K-L전개에 의한 압축 벡터만을 이용한 경우에는 자음의 구간(평균 20-30ms)이 세그멘트폭(4프레임폭일 경우 35ms,

7프레임폭인 경우 50ms)에 포함이 되기 때문에 자음(초성과 종성)이 모음(중성)에 영향을 받아 모음이 같은 다른 음절로 오인식이 많이 발생함을 알게 되었다. 특히 너무 넓은 폭의 세그멘트 역시 잔음 성분으로 인식율의 저하를 가지온다는 사실도 함께 되었다. 그래서 이것을 보완하기 위해 파라메타로서 델타스트림과 최귀계수를 부가한 경우 향상된 인식율을 얻을 수 있었다.

그리고 실험을 하지 않았으나 델타스트림과 최귀계수를 함께 부가한 경우는 인식율의 향상을 가지올 것이지만 추정 파라메타의 증가로 계산량의 관점에서 문제가 있다고 하겠다.

참고논문

[1] L.R. Rabiner, J.G. Wilpon, and F.K. Soong, "High performance connected digit recognition using hidden Markov models," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-37, pp. 1214-1225, Aug. 1989.

[2] F. Jelinek, " Continuous speech recognition by statistical methods", Proc. IEEE, 64, pp.532-556 1976.

[3] L. R. Bahl, et al., "Acoustic Markov Models used in the TANGORA speech recognition system", Proc. ICASSP, pp.467-500, 1988.

[4] 中川聖一, "確率モデルによる音聲認識", 電子情報通信學會編, 1988.

[5] 中川聖一, "連續出力分布型HMMによる日本音韻認識", 論文志, Vol. 46, pp.486-496,1990.

[6] 김상범, 박창호, 허강인, "CHMM을 이용한 음소와 단어 음성 인식에 관한 연구", 한국통신학회 부산·경남지부 학술논문 발표회 논문집, 제 1 집, pp.67-71, 1994 .

[7] 박창호, 허강인, "음성인식을 위한 파라메타 확장에 관한 연구", 제11회 음성통신 및 신호처리 워크샵 논문집, 제 SCAS-11권 1호, pp.153-156, 1994.