강인한 음성 인식을 위한 탠덤 구조와 분절 특징의 결합*

윤영선(한남대학교), 이윤근(한국전자통신연구원)

<차 례>

1. 서론
2. 탠덤 구조의 설계
3. 분절 특징(Segmental Feature)
4. 경향 특징(Trend Feature)
4.1 궤적 정보의 분리
4.2 경향 양자화
5. KL(Karhunen-Loéve) 변환
6. 실험 및 결과
7. 요약 및 검토

<Abstract>

Combination Tandem Architecture with Segmental Features for Robust Speech Recognition

Young-Sun Yun, YunKeun Lee

It is reported that the segmental feature based recognition system shows better results than conventional feature based system in the previous studies. On the other hand, the various studies of combining neural network and hidden Markov models within a single system are done with expectations that it may potentially combine the advantages of both systems. With the influence of these studies, tandem approach was presented to use neural network as the classifier and hidden Markov models as the decoder. In this paper, we applied the trend information of segmental features to tandem architecture and used posterior probabilities, which are the output of neural network, as inputs of recognition system. The experiments are performed on Aurora2 database to examine the potentiality of the trend feature based tandem architecture. From the results, the proposed system outperforms on very low SNR environments. Consequently, we argue that the trend information on tandem architecture can be additionally used for traditional MFCC features.

* Keywords: Speech recognition, Tandem architecture, Segmental feature, Neural network.

* 본 연구는 정보통신부 및 정보통신연구진흥원의 IT신성장동력핵심기술개발사업의 일환으로 수행하였음. [2006-S-036-01, 신성장동력산업용 대용량/대화형분산/내장처리음성인터페이스기술개발]

1. 서 론

은닉 마코프 모델(HMM; Hidden Markov Model)은 구현의 용이성과 유연한 모델링 능력, 높은 성능으로 인하여 많은 연구 분야에서 오랫동안 널리 사용되어오고 있다. 그러나 많은 선행 연구에서 HMM은 독립관측(independence observation) 가정과 프레임 특징의 이용으로 인해 음성 신호의 시종속성(temporal dependency)을 효과적으로 표현하지 못한다고 보고하고 있다. 시종속성이란 음성 신호가 시간에 따라 변화되는 특성을 나타내는 것으로써, 일반적으로 특징 표현 단계에서는 동적특성(dynamic feature)을 통하여 시종속성을 표현하여 인식 성능을 향상시키고 있다. 일반적으로 특징 표현 단계에서 동적특성을 이용함으로써 시종속성을 표현하고자 하였으나, 다른 연구 방식으로는 음향 모델링 단계에서 시종속성을 표현하고자 하는 연구가 진행되고 있다. 대표적으로는 확률 분포에 의한 확률 분절 모델(probabilistic segmental model)[1]-[3]이나 궤적 방식(trajectory approach)을 이용한 모수적 분절 모델(parametric segmental model)[4]-[7] 등의 연구가 있다. 확률 방식의 분절 모델은 모델링 과정에서 정확한 해를 구할 수 없어 특징 변수의 분포에 제약을 두거나 고려하지 않는다는 단점이 존재하며, 모수적 방식의 분절 특징은 모델링의 일반화와 쉽게 기존 HMM을 확장할 수 있다는 장점이 있으나 기존 음향 모델링보다 분석 구간의 크기 배수만큼 계산량이 늘어난다는 단점이 있다.

또한 최근에는 은닉 마코프 모델과 신경 회로망(NN; neural networks) 모델을 결합하여 두 시스템의 장점을 결합하고자 하는 연구가 진행되고 있다. 이 방법은 HMM의 뛰어난 시간 정규화(time normalization) 특성과 NN의 분별 능력(discriminative ability)을 결합하여, 음성 신호를 모델링한다. 이런 영향을 받아 NN과 HMM을 직렬로 연결한 탠덤 방식(TA; tandem approach or tandem architecture)이 제안되었다. 탠덤 방식은 먼저 NN을 이용하여 음소 집합에 대한 사후 확률(posterior probability)을 계산하며, 그 값을 HMM의 입력으로 전달하는 음성 인식 방식이다.

본 연구에서는 모수적 방식의 분절 방식의 적용 시 증가하는 연산량을 최소화하고 성능을 유지하기 위하여 기존의 탠덤 구조와의 결합 방식을 고려하였다. 분절 특징을 인식기에서 사용하는 경우 성능은 향상되나, 계산량이 많이 증가되어 성능 향상에 따른 이득이 줄어든다. 따라서 본 연구에서는 분절 특징의 특성을 살리면서 계산량을 줄이기 위한 방법으로 경향 정보만을 이용하여 직교 특징을 표현하고 인식시스템에 적용하는 방법을 제안한다. 따라서 기존의 탠덤 구조에서 사용하던 음소 단위의 NN 출력 확률 대신, 음성 데이터에 기반을 둔 분절 특징에 대한 NN 출력 확률을 사용하는 탠덤 구조를 제안하여 성능 변화를 살피고 적용 가능성을 타진한다. 즉, 기존의 연구에서 사용하는 언어 특징인 유사 음소 유닛(PLU; phoneme like unit)에 대한 사후 확률(posterior probability)은 입력 음성의 아

주 작은 부분만 관찰함으로써 PLU에 대한 사후 확률을 결정하기 때문에 출력 값에 대한 변별력이 떨어질 수 있기 때문에, 본 연구에서는 NN의 입력으로 음향 데이터에 바탕을 둔 분절 특징을 이용하여 세밀한 음성 특징을 표현하고자 한다.

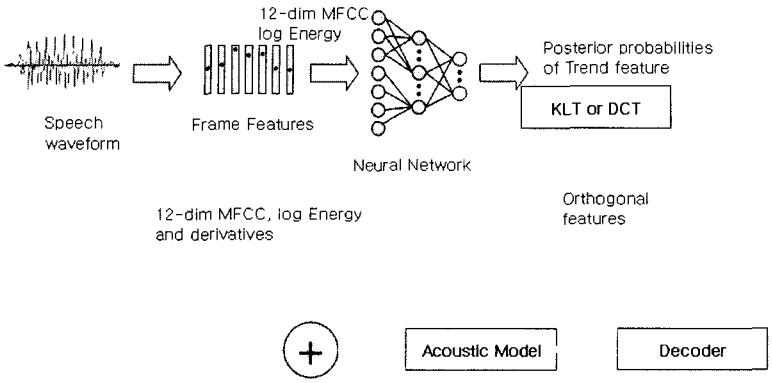
본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존의 탠덤 구조의 소개와 제안된 방법을 설명하고, 3장에서는 본 연구에서 사용하는 분절 특징을, 4장에서는 경향 특징에 대해 간략하게 요약하고 5장에서는 직교 선형 변환법으로 많이 사용되는 KL 변환을 소개한다. 제안된 방법의 적용 여부를 판단하기 위한 다양한 실험 및 결과를 6장에서 정리하며, 마지막으로 본 연구의 요약 및 향후 연구에 대해 7장에서 기술한다.

2. 탠덤 구조의 설계

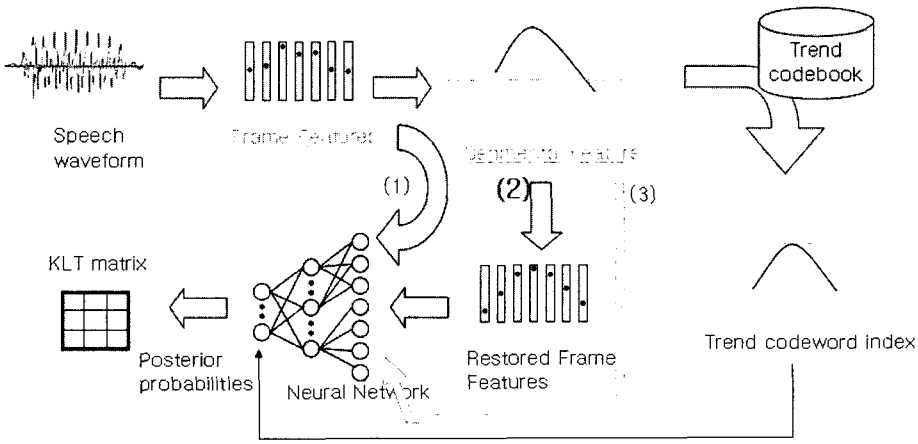
탠덤 구조는 Hermansky 등에 의해 제안되었으며[8], PLU에 대한 사후 확률 값을 의미하는 신경회로망의 출력을 KLT(Karhunen Loéve Transform) 또는 PCA (Principal Component Analysis)를 통해 직교 특징으로 변환한 후 HMM을 학습시켜 음성 인식 시스템을 구성한다. 본 연구에서 제안한 방식은 기존의 시스템을 수정하여 PLU 대신 분절 정보를 이용하도록 하였다. 분절 특징은 일반적으로 많이 사용되는 MFCC나 동적 특징을 대체하지 않고 부가적인 특징으로도 사용될 수 있다. 부가적인 특징으로 분절 정보를 이용하는 경우 다음과 같은 단계를 거친다. 먼저 MFCC나 동적 특징에 기반을 두어 분절 정보를 추출한 후 신경회로망에 의하여 사후 확률 값으로 변환한다. 변환된 사후 확률 값은 미리 정의된 분절 정보에 대한 확률 또는 특징으로 나타나기 때문에 KLT에 의해 직교 특징 벡터(orthogonal feature vector)로 변환된다. <그림 1>에 보인 바와 같이 변환된 특징 벡터는 기존의 음성 특징과 결합하거나 또는 단독으로 HMM의 입력으로 전달되어 일반적인 음성 인식 시스템을 구성한다.

제안된 구조에서는 NN의 입력으로 12차의 MFCC와 로그 에너지를 이용하며, 미리 정의된 경향 특징(trend feature)의 클래스에 대한 사후 확률 값을 출력으로 한다. 경향 특징의 클래스는 학습에 사용되는 전체 음성 특징에 대해 벡터 양자화(Vector Quantization) 과정을 거쳐 결정된다. 따라서 신경회로망의 출력으로 사용되는 경향 특징의 수는 경향 양자화(Trend Quantization)에 사용된 코드북 크기와 일치한다.

NN과 KLT를 이용하여 직교 특징을 구하는 방법으로 세 가지를 고려할 수 있다(<그림 2>참조). 첫째는 음성 신호의 프레임 특징(예; MFCC나 동적 특징)을 직접 NN의 입력으로 주고, NN은 경향 특징의 인덱스에 대한 확률을 출력 값으로 하는 방법이다. 이 방법은 입력 음성에 대해 별도의 변환 과정을 거치지 않는다는

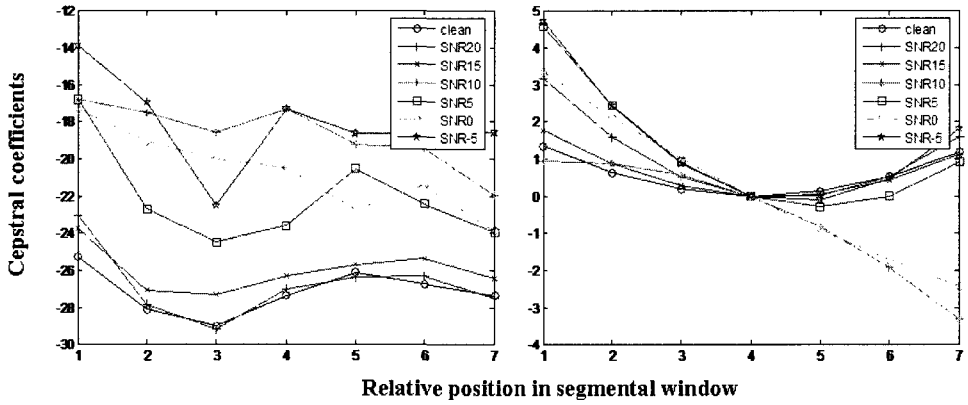


<그림 1> 일반적인 탠덤 구조(Tandem architecture)



<그림 2> 직교 특징을 구하는 과정

장점이 있으나, 표현된 경향 특징의 추정 오류가 같이 포함되어 NN의 수렴 속도 및 가능성이 저하될 수 있다. 두 번째 방법으로는 프레임 특징을 경향 특징으로 변환하여 추정 오차를 제거한 다음, 변환된 경향 특징을 다시 프레임 특징으로 복원하여 NN의 입력으로 사용하는 방안이다. 이 방법은 여러 과정을 거치나, 첫 번째 방법과 달리 NN의 입력인 프레임 특징(복원된 프레임 특징)은 경향 특징의 추정 오차가 제거된 형태를 보인다. 즉, 프레임 특징이 경향 특징으로 변환되는 과정의 추정 오차를 제거한 완전한 경향 특징으로 NN을 학습시킨다는 것이다. 마지막 방법은 두 번째 방법과 유사하나 프레임 특징으로 복원하지 않고 경향 특징의 계수들을 직접 NN의 입력으로 사용하는 방법이다. 분절 특징의 계수 정보는 분석 대상의 길이(프레임 수)에 독립적이기 때문에 분석 대상의 범위가 넓어지더라도 입력의 수는 변하지 않는다는 장점이 있지만, 계수 간의 거리와 프레임 특징 간의 거리는 다른 의미와 변화를 가져올 수 있기 때문에 NN의 성능이 안정적이지 않을



<그림 3> 원 프레임 특징과 경향 특징에서 복원된 프레임 특징 비교(1차 켑스트럼 특징)

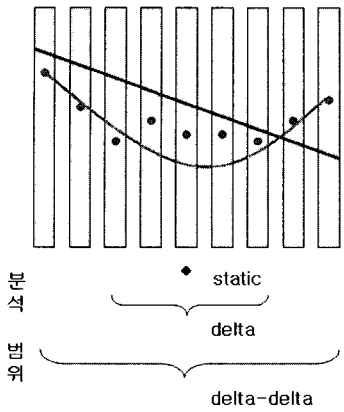
수 있다. 이상의 이유로 인해 본 연구에서는 두 번째 학습 방법을 선택하여 경향 정보를 추출하여 사용하였다.

원 프레임 특징과 경향 특징을 이용한 복원된 프레임의 비교는 <그림 3>에 보인다. 그림에서 보는 바와 같이 원 프레임 특징의 경우 포함된 잡음의 양에 따라 프레임 특징의 변화가 심하나, 복원된 프레임 특징의 경우 경향만을 나타내고 있어 음성 특징 비교에 사용될 경우 잡음의 영향을 줄일 수 있을 것으로 보인다. 그러나, <그림 3>에서 보인 바와 같이 프레임 특징의 상대적 위치가 제거되기 때문에 분석 구간이 작을 경우 원 프레임 특징의 왜곡을 가져 올 수 있다는 점이 지적될 수 있다.

프레임 특징에서 분절 특징으로 변환하는 과정과 경향 특징에 관한 내용은 3장과 4장에서 자세히 다루겠다.

3. 분절 특징(Segmental Feature)

분절 특징은 여러 프레임 특징 또는 프레임 특징 집합에서 확률적인 방법 또는 수학적인 모델링 방식에 의해 표현되는 일련의 특징을 말한다 [9][10]. 최근에는 음성 인식의 성능 향상뿐만 아니라 잡음의 영향을 받지 않거나 최소화하는 모델링 방식에 대한 연구가 진행되고 있으며, 그 대안으로 분절 특징에 대한 연구가 제시되고 있다. 선행 연구에서 분절 특징 표현에 계산 성능의 향상과 모델링 방법의 유연성을 꾀하기 위해 모수적 방법(parametric approach)을 채택하였고, 고정된 크기의 분석 구간을 이용하였다. 이 방법은 여러 프레임 특징에서 보일 수 있는



<그림 4> 분절 특징의 표현

비정상적인 잡음의 영향을 줄일 수 있고, 쉽게 확장할 수 있다는 장점이 있다. 분절 특징 시스템에서는 널리 사용되는 MFCC와 미분 값들을 프레임 특징들의 연속된 흐름을 나타내는 궤적(trajectory)으로 변환하여, 분류 단계로 전달한다. 본 절에서는 프레임 특징 열을 궤적으로 변환하는 과정을 소개한다.

음성 신호의 연속적인 음향 특징 벡터들 간의 관계는 특징 공간에서 궤적의 형태로 근사될 수 있다는 기본적인 생각에서 출발한 분절 모델링은 구현 방법에 따라 모수적(parametric) 또는 비모수적(non-parametric) 방식으로 분류된다. 본 연구에서는 모수적 방식이 여러 음성 단위에서 궤적의 평활화 효과를 보이기 때문에 [4][5] 잡음이나 환경 변화, 화자 변화에 강인할 것으로 생각되어 분절 모델링에 모수적 방식을 사용하였다.

모수적 방식에 의한 단순화된 형태는 <그림 4>와 같이 표현된다. 일반적인 프레임 특징은 분석 구간에서 정 중앙의 단일 프레임을 나타내며, 1차 또는 2차 미분 계수를 구하는 과정에서는 여러 프레임 특징의 축약된 형태로 표현된다. 1,2차 미분계수가 프레임 집합의 특징 벡터를 평균화하여 기울기나 가속도의 형태로 축약적인 형태를 값으로 표현하는 방식에 비하여, 모수적 방식은 여러 프레임 특징의 변화량을 열(sequence)로써 표현하는 방법이며 표현 방식에 따라 선형, 2차 곡선 등으로 나타내고 있다.

본 연구에서 채택한 분절 특징 표현 방법은 연속 음성 인식의 사용에 용이하도록 음성 패턴을 고정된 길이의 분절들의 열로써 표현하였으며 각 분절은 중첩이 가능하도록 하였다. 또한 각 분절은 인접한 분절들과 중첩될 수 있으므로 기준점을 분절의 중앙에 두었다. 이것을 고려하여 고정길이를 갖는 분절을 표현하면 다음과 같이 나타낼 수 있다.

$$C_t = ZB_t + E \tag{1}$$

위 식에서 C_t와 B_t는 각각 시간 t에서의 음성 분절과 궤적 계수를 나타내며, E는 잔차 오차(residual error)를 의미한다. 궤적으로 표현되는 분절 특징은 주어진 분절 안에서 적용할 프레임의 범위와 표현 형태를 나타내는 디자인 행렬(design matrix) Z와 궤적 계수 B_t의 곱으로 표현된다.

Yun 등이 제안한 분절 특징 HMM은 Gish 등이 제안한 모수적 궤적 모델(parametric trajectory model)의 약점으로 파악되었던 경계 문제(boundary problem)를 해결하기 위하여 다음과 같은 디자인 행렬을 이용하여 현재 시간에서의 관측 벡터가 분절의 중앙에 오도록 조정하였다.

분절 표현에서 알 수 있듯이 현재 시간 t의 분절 C_t의 기준점은 분절 중앙에 있는 프레임 특징이 되기 때문에, $t-1$ 또는 $t+1$시간의 분절과 중첩될 수 있다. 이와 같은 음성 분절을 표현하기 위하여 $\tau - M$가 현재 중앙 프레임에서의 상대적 위치를 표현한다고 할 때, 음성 신호의 프레임 특징 열의 범위를 조절하는 디자인 행렬 Z는 다음과 같이 정의될 수 있다.

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

여기에서 M은 현재 프레임의 분석 구간의 크기 $N = 2M+1$을 만족하며 R은 회귀 차수를 나타낸다. 또한 z_τ는 디자인 행렬 Z의 τ번째 행 벡터(row vector)를 나타낸다. 이와 비슷한 방법으로 D가 특징 벡터의 차수를 의미할 때 궤적 계수 행렬 B_t는 다음과 같이 정의된다.

$$\begin{aligned} B_t &= \begin{bmatrix} b_{1'} \\ \vdots \\ b_{R'} \end{bmatrix} \\ b_{i'} &= \left[b^t_{i,1} \cdots b^t_{i,D} \right], \quad 1 \le i \le R. \end{aligned} \tag{3}$$

음성 분절 C_t와 디자인 행렬 Z가 주어지면 추정되는 궤적 계수 행렬 $\hat{B}_t$는 선형 회귀(linear regression) 방정식이나 다음과 같은 행렬 연산에 의하여 계산될 수 있다.

$$\hat{B}_t = [Z^T Z]^{-1} Z^T C_t, \tag{4}$$

여기에서 T는 행렬의 전치(transpose)를 의미한다. 궤적 계수 행렬 $\hat{B}_t$가 추정되면, 추정 오차를 나타내는 최적 적합도(goodness-of-fit)는 분절을 구성하는 모든 프레임 특징에 대한 잔차 오차를 더하여 계산된다.

4. 경향 특징(Trend Feature)

분절 특징 표현에서 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. 경향 정보[11]는 음성의 변화 형태를 표현하며, 위치 정보는 분절 특징의 기준 위치를 나타낸다.

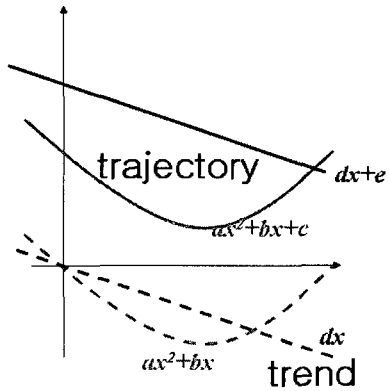
4.1 궤적 정보의 분리

궤적 정보는 선형 회귀 방정식으로 표현될 수 있으며, 각 특징 차원은 궤적 계수와 디자인 행렬로부터 다음과 같이 복원된다.

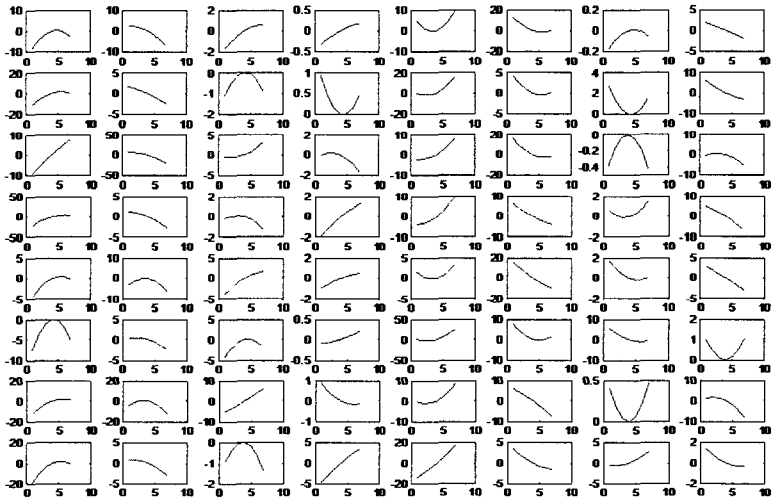
$$c_{\tau,i} = b_{1,i} z_{\tau,1} + b_{2,i} z_{\tau,2} + \cdots + b_{R,i} z_{\tau,R}, \quad 1 \le i \le D, \tag{5}$$

여기에서 D는 켑스트럼 차수를 나타내고, $c_{\tau,i}$는 분절 내에서 τ번째 프레임의 i차 켑스트럼(cepstrum) 벡터를 의미하며, $b_{r,i}$는 r번째 궤적 계수를 나타낸다. $z_{\tau,r}$은 디자인 행렬의 요소를 나타내며, $\left(\frac{\tau - M}{2M}\right)^{r-1}$로 표현된다.

위 식에서 디자인 행렬의 첫 번째 행 벡터는 1임을 알 수 있다, 즉 $z_{\tau,1} = 1$. 따라서 $b_{1,i}$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 이용한다면, 궤적 특징에서 경향 정보를 표현할 수 있다.



<그림 5> 분절 특징과 경향 특징의 비교



<그림 6> 경향 특징 코드북의 예(크기: 64, 1차 켑 스트럼 특징)

분절 특징 표현에서는 현재의 프레임 관측 벡터는 분절의 중앙에 존재한다. 따라서 $b_{1,i}$는 궤적 표현에 의해 평활화(smoothing)된 가운데 점의 위치를 나타낸다 (<그림 4> 참조). 만약 식 (5)가 행렬 연산으로 변환되면, 궤적 행렬의 첫 번째 행벡터 b_1는 위치를 의미하고, 나머지 부분은 경향을 의미한다. 경향을 공유하기 위해서는 궤적 표현으로부터 경향과 위치를 분리하여야 하는데, 행렬의 처음 행벡터를 제거하면 경향 벡터가 되며, 다음과 같이 표현된다.

$$\boldsymbol{T}_t = \left[b_2^t \cdots b_R^t \right]^T \tag{6}$$

이 경향 벡터를 공유하기 위해서 경향 양자화 방법을 이용한다. 경향 양자화 방법을 이용하여 각 분절 특징을 구성하는 경향 벡터는 가장 가까운 코드워드(codeword)로 교체된다.

4.2 경향 양자화

경향 양자화 알고리즘은 널리 알려진 벡터 양자화 알고리즘과 유사하다. 그러나, 유클리드 거리(Euclidean distance)로 표현된 거리 척도는 두 경향을 비교하도록 수정되어야 한다. 경향 특성을 반영하기 위하여 유클리드 거리는 다음과 같이 수정된다.

$$D(T_i, T_j) = \frac{1}{N}\sum_{\tau=1}^{N}\{\tilde{\mathbf{z}}_\tau(\mathbf{T}_\mathrm{i} - \mathbf{T}_\mathrm{j})\}\{\tilde{\mathbf{z}}_\tau(\mathbf{T}_\mathrm{i} - \mathbf{T}_\mathrm{j})\}^\mathrm{T}, \quad (7)$$

여기에서 $\tilde{z}_\tau$는 경향 벡터에 대응하도록 디자인 행렬에서 첫 번째 열(column)을 제외한 행 벡터를 나타내고, T_i와 T_j는 경향 벡터를 나타낸다.

<그림 5>는 궤적으로 표현되는 분절 특징과 위치 정보가 제거된 경향 특징의 의미적인 비교를 보이며, <그림 6>은 1차 켑스트럼 벡터에 대한 경향 특징 벡터를 코드북 크기 64로 양자화한 결과를 보이고 있다. 그림에서 보는 바와 같이 각 양자화된 코드북 벡터는 전체 학습 벡터를 분할한 경향 특징 벡터를 잘 표현하고 있으며, 각각의 코드워드가 신경회로망의 지정 출력(desired output)이 된다.

5. KL(Karhunen-Loéve) 변환

프레임 특징이 신경회로망의 입력으로 전달된 후, 각 경향 특징에 대한 사후 확률 값(또는 특징 벡터)을 얻게 된다. 경향 특징의 클래스 수는 데이터의 양에 따라 미리 정의되며, 널리 사용되는 PLU의 수와 비슷하게 64개 정도를 사용한다. 따라서 경향 특징의 클래스 수가 결정되면 벡터 양자화에 의하여 전체 훈련 데이터를 지정된 수만큼 분류하고, 그 분류된 인덱스에 대한 확률 값이 신경 회로망의 출력으로 나온다. 경향 특징은 특징의 성격상 위치(offset, location) 정보가 생략되어 있기 때문에, 음성 인식에 단독으로 사용할 때 성능이 저하될 수 있어 기존의 MFCC, 동적 특징과 함께 사용될 수 있다. 이 경우 기존의 특징 정보와 결합하여 새로운 특징 조합을 만들거나 특징 차수를 줄이기 위하여 사용하는 방법이 KL 변환이다. KL변환의 큰 특징은 우선 상관관계가 높은 특징들을 상관관계가 작은 특징으로 변환한다는 점과, 특징 벡터 중 중요 특징을 낮은 차수에 배치하는 에너지

압축의 효과를 가져온다는 것이다. 이 KL변환은 이산 환경에서 PCA, 또는 호텔링 변환이라고도 불리며, 큰 분산을 갖는 부분 공간을 유지하는 최적의 선형 변환 법으로 알려지고 있다 그러나 다른 선형 변환법과 비교해서 자료 집합에 종속된 basis 특징 집합을 가지며, 자료 집합에 따라 계산량이 늘어난다는 단점을 가지고 있다.

KL 변환은 자료 집합으로부터 KL 변환 행렬을 구하여 입력 자료에 곱함으로써 선형변환된 새로운 자료 집합을 구할 수 있다. 만약 신경회로망의 출력 특징 벡터의 차수가 n이라면 각 차원의 상관관계에 대한 n개의 고유벡터(eigenvector)로 구성된 $n \times n$ KL 변환 행렬 M을 구할 수 있다. $V(i)$가 현재 분절 특징에 대한 신경회로망의 출력 값이라면,

$$V(i) = [x_1, x_2, \cdots, x_n]^T \tag{8}$$

이며, 평균 벡터와 분산벡터는 다음과 같이 얻어진다.

$$\mu_v = E[V] = \frac{1}{K}\sum_{l=1}^{K} V(l) \tag{9}$$

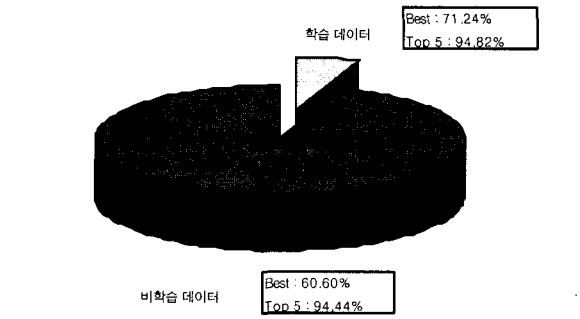
$$C_v = E[(V-\mu_v)(V-\mu_v)^T] = \frac{1}{K}\sum_{l=1}^{K}[(V(i)-\mu(i))(V(i)-\mu(i))^T] \tag{10}$$

$m_1, m_2, \ldots, m_n$이 C_v의 고유벡터라 할 때 KL 변환 행렬 M은 고유벡터를 행벡터(row vector)로 갖는다. 즉,

$$M = [m_1, m_2, \ldots, m_n]^T. \tag{11}$$

6. 실험 및 결과

제안된 방법의 가능성을 검토하기 위하여 ETSI의 Aurora2 DB를 이용하여, 채널 잡음과 가산 잡음이 함께 존재하는 경우에 제안된 방식의 성능 변화를 조사하였다. Aurora2 DB는 ETSI에서 DSR(distributed speech recognition) 시스템의 전단계(front-end) 알고리즘의 성능을 객관적으로 평가하기 위한 표준 데이터로 TI-DIGITS DB를 8 kHz로 하향 샘플링하고 여러 가지 잡음과 선형 필터(채널 잡음) 효과를 적용하였다. 제공되는 DB 중 Set A와 B는 가산 잡음만을 고려한 것이며 Set C는 일반 전화 채널의 효과(MIRS 특성)를 필터링하고 "subway"와 "street"의 잡음을 첨가한 것이다. 부가된 잡음은 잡음 레벨(SNR 20, 15, 10, 5, 0, -5 dB)에 따라 인위



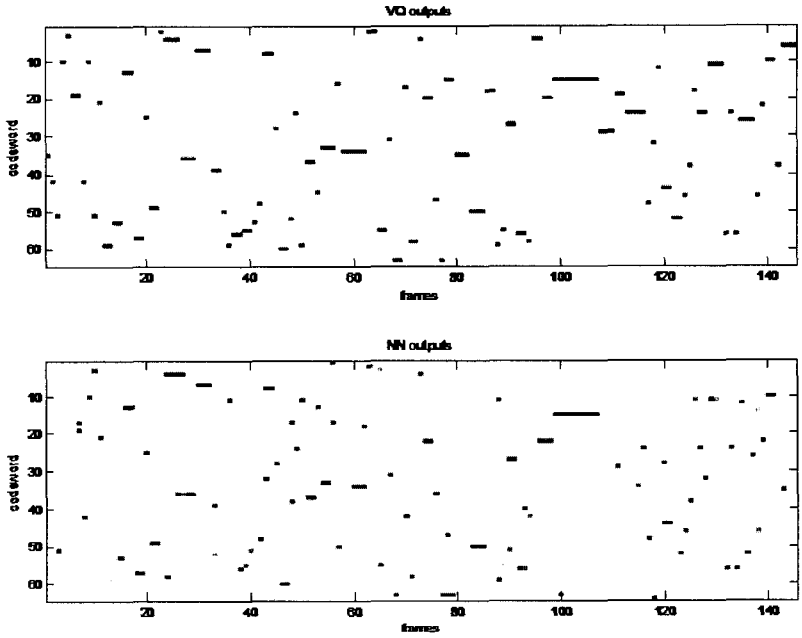
<그림 7> 신경회로망의 분류 성능

적으로 첨가되었다. 본 연구에서는 조용한 환경에서 학습 시킨 후, Set C에 대한 인식 실험을 하고, 기존의 MFCC, Delta 특징 조합과의 비교 실험을 진행하였다.

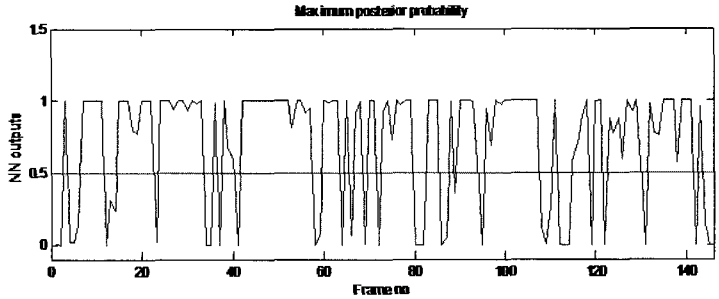
인식 실험을 진행하기 전에 신경회로망의 성능을 검증하기 위하여, 학습된 신경회로망의 최대 출력 값과 분절 특징의 코드북 인덱스를 비교하는 평가 실험을 진행하였다. 신경회로망은 HMM과 달리 모델을 학습시키기 위해서는 전체 자료를 이용하기 때문에 수렴 및 성능 검증에 많은 시간이 걸린다. 특히 NN의 구조에 따라 성능의 변화도 수반되기 때문에 입력과 출력이 결정되면 NN의 성능은 은닉층의 노드 수에 따라 결정된다. 그러나 모든 경우의 노드 수에 대하여 실험을 할 수 없기 때문에 출력 노드의 약 2배인 160개로 은닉 층의 노드 수를 결정하였다(입력 노드: 7개의 프레임 중 현재 프레임을 제거한 6개의 프레임 수 × 13차 정적 특징, 출력 노드 : 64개의 코드북 인덱스). 또한 전체 데이터를 이용하는 경우 많은 계산량과 시간을 필요로 하기 때문에 약 5%의 자료만을 이용하여 학습시킨 후, 나머지 95%에 대해 평가를 하였다.

<그림 7>에서 보는 바와 같이 5%정도의 학습 데이터를 이용한 경우 분류 오류는 약 29%를 나타냈으며, 전체 자료에 대해서는 약 40%의 오류율을 나타냈다. 그러나 5위 후보나 10위 후보까지 고려한 경우, 학습 자료나 비 학습자료에 대한 분류율이 큰 차이를 보이지 않아 신경회로망의 출력을 특징으로 사용할 수 있는 가능성을 보여준다. 테스트에 사용된 Aurora2 DB의 Set C 데이터에 대해서 VQ 코드북 인덱스와 신경회로망의 출력 노드의 결과 값을 살펴본 결과, 비교적 유사한 출력 분포를 보임을 알 수 있다(<그림 8> 참조).

그러나 각 프레임별 신경 회로망의 최대 출력 값을 비교하면 신경회로망의 성능이 아직 미비함을 파악할 수 있다. <그림 9>에서 보는 바와 같이 신경회로망의 최대 출력 노드 값 중에서 0.5 미만에 해당하는 값들이 약 27%대에 머물고 있다. 이것은 입력 특징벡터에 대한 신경회로망의 반응이 미약하다는 것을 의미하며, 하



<그림 8> Aurora2 DB의 set c/clean1/MAH_1390A 문장의 코드북 인덱스와 신경회로망의 출력 비교



<그림 9> 신경회로망의 최대 출력 노드의 값 분포

위 값(<0.5)들의 기여도가 낮다는 것을 의미한다. 따라서, 신경회로망의 출력을 정규화하거나 KL 변환과정에서 정규화가 필요하다는 것을 의미한다.

다음으로 KL 변환을 통한 신경회로망의 출력을 이용한 인식 실험들을 진행하였다. <표 1>은 일반적으로 음성인식에 많이 사용되는 정적 특징인 12차 MFCC와 로그 에너지, 그 미분 계수(분석 윈도우 크기를 2, 3 프레임 구분), 8차(1/8), 13차, 64차(전체)를 비교 실험한 내용이다. 실험 결과 1차 미분 계수를 사용한 경우에서 전반적으로 성능이 가장 우수함을 보였으며, MFCC와 로그 에너지를 사용한 경우, 마지막으로 경향 특징을 사용한 경우 순으로 성능을 보였다. 특히 clean1, clean2를

비롯하여 높은 SNR에서 경향 특징은 미분 계수에 비해 저조한 성능을 보이고 있는데, 이것은 위치 정보가 생략되었기 때문으로 보이며, 미분 계수는 현재 프레임의 정보를 이용하지는 않으나 주위 프레임에 의해 현재 분석 구간의 위치 정보를 포함하기 때문에 잡음 정보에 강인한 특성을 보이는 것으로 보인다. 그러나 낮은 SNR에서는 기존의 MFCC에 비해 높은 성능을 보이고 있어 경향 특징을 이용한 탠덤 구조의 경우 성능 개선에 대한 잠재적 능력을 가지고 있다고 파악된다.

<표 1> 전형적인 특징 벡터 MFCC, Delta와 탠덤 구조에 의해 표현된 경향 특징 벡터의 성능 비교(accuracy)

종류 \ 조합		13차 MFCC	13차 Delta (w=3)	8차 Trend Feature	13차 Trend Feature	64차 Trend Feature
Subway	clean1	95.70	97.94	57,41	60.95	76.54
	SNR 20	59.13	93.83	45.84	51.82	62.17
	SNR 15	43.23	84.80	36.78	42.68	50.51
	SNR 10	27.82	61.34	26.47	32.02	33.37
	SNR 5	14.25	34.51	17.87	20.23	20.48
	SNR 0	7.86	19.16	12.93	13.02	14.31
	SNR -5	7.55	12.16	9.43	10.25	9.82
Street	clean2	95.28	98.00	56.59	60.10	77.48
	SNR 20	71.16	94.50	46.10	54.81	58.74
	SNR 15	58.98	87.94	37.79	46.64	45.71
	SNR 10	43.95	67.29	26.57	32.56	31.23
	SNR 5	29.63	39.33	17.68	21.52	20.65
	SNR 0	14.81	21.07	11.70	13.33	13.48
	SNR -5	9.67	11.52	9.31	9.40	9.04

다음은 기존의 12차 MFCC와 로그에너지에 1차 미분 계수 또는 경향 특징을 추가하여 비교 실험한 결과이다. 이 실험 또한 기존의 MFCC와 로그에너지만을 사용했을 경우보다는 성능이 향상됨을 알 수 있으나, 미분 계수를 포함한 경우보다는 성능이 저하됨을 알 수 있다. 또한, SNR 10 dB에서는 급격하게 성능이 저하되고 있는데 이는 MFCC와 로그 에너지의 특징 표현력 감소에 경향 특징이 충분히 보상하지 못한데서 이유를 찾을 수 있을 것이다. 그러나 아주 낮은 SNR에서는 비슷한 성능을 보이고 있어 제안하는 경향 특징은 낮은 SNR에서 적용 가능성을 보인다(표 2 참조).

<표 2> 기존의 특징과 결합된 특징 벡터의 성능 비교(accuracy)

종류 \ 조합		13차 MFCC+ Delta	13차 MFCC+ 8 TF	13차 MFCC+ 13 TF	13차 MFCC+ 64 TF
Subway	clean1	99.08	95.18	92.57	83.97
	SNR 20	93.31	74.21	82.13	70.49
	SNR 15	86.49	56.49	71.63	57.63
	SNR 10	71.08	37.21	53.95	37.67
	SNR 5	42.59	20.76	34.02	23.21
	SNR 0	17.90	12.31	16.49	15.54
	SNR -5	9.64	7.98	7.77	10.41
Street	clean2	98.73	95.16	93.38	83.92
	SNR 20	92.90	86.37	86.49	67.32
	SNR 15	85.97	74.37	77.78	50.57
	SNR 10	68.50	57.80	62.24	34.16
	SNR 5	44.65	36.46	44.65	22.34
	SNR 0	20.37	17.62	24.43	14.00
	SNR -5	10.43	9.55	10.34	9.31

비슷한 방법으로 기존의 특징 벡터 조합(MFCC와 로그에너지, 미분계수)에 2차 미분 계수(가속도) 또는 경향 특징을 추가한 경우의 비교 실험이다. 실험 결과(표 3) 앞선 비교에서와 비슷하게 매우 낮은 SNR에서 성능이 향상됨을 알 수 있다.

<표 3> 기존의 MFCC, Delta와 결합된 특징 벡터의 성능 비교(accuracy)

종류 \ 조합		13차 MFCC+ Delta+Delta2	13차 MFCC+ Delta+8 TF	13차 MFCC+ Delta+13 TF	13차 MFCC+ Delta+64 TF
Subway	clean1	99.02	97.48	95.12	87.41
	SNR 20	94.29	91.16	89.65	72.92
	SNR 15	87.60	82.71	82.93	60.33
	SNR 10	73.23	66.26	67.15	38.66
	SNR 5	49.62	43.26	48.79	24.07
	SNR 0	23.89	23.79	26.68	15.63
	SNR -5	10.68	11.33	14.43	10.84
Street	clean2	99.00	97.37	96.28	87.21
	SNR 20	95.10	93.17	91.08	71.28
	SNR 15	88.72	86.97	84.95	54.72
	SNR 10	72.55	73.07	69.38	36.43
	SNR 5	46.86	51.84	49.88	24.09
	SNR 0	22.01	27.51	28.05	15.30
	SNR -5	10.97	11.58	14.63	10.13

이와 같은 결과에서 제안된 경향 특징 벡터는 낮은 SNR에서 적용 가능함을 보여주고 있으며, 경향 특징의 추출 단계에서 MFCC와 로그에너지만을 사용할 것이 아니라, 1차 미분 계수까지 포함한다면 높은 SNR에서도 적용 가능할 것으로 생각된다. 특히 SNR 0이나 SNR-5 dB에서는 3~40%의 높은 성능 향상률을 보이고 있어, 윈도의 크기의 확장이나 경향 특징의 추출 조절에 따른 성능 향상을 기대케 한다.

7. 요약 및 검토

본 연구에서는 패턴 인식 분야에서 분류 성능이 뛰어난 신경회로망과 시간 정규화 기능과 구현의 용이성, 높은 성능으로 여러 분야에서 활용되고 있는 은닉 마코프 모델을 순차적으로 결합한 탠덤 구조의 적용 여부를 파악하고, 분절 특징인 경향 특징을 이용하여 성능 향상을 꾀하였다. 기존의 탠덤 구조 방식의 신경회로망과 은닉 마코프 모델의 결합은 프레임 특징의 집합을 신경회로망의 입력으로 하고, PLU의 인덱스에 대한 사후 확률 값을 출력하여 KLT를 통한 직교 특성을 구하였다. 이렇게 구해진 직교 특징 벡터는 일반적인 HMM의 입력으로 전달되어 음성인식에 사용되었다. 본 연구에서는 기존의 방식을 수정하여 PLU 대신 경향 특징을 적용하였으며, 기존 프레임 특징과 결합하여 잡음 환경 하에서 음성 인식의 성능을 높이고자 하였다.

실험결과 기존에 널리 사용되는 MFCC나 1차 미분계수에 비해서 조용한 환경이나 높은 SNR에서 전반적으로 성능이 하락되나, 낮은 SNR에서는 성능이 향상됨을 알 수 있었다. 따라서 제안된 방식은 SNR이 낮은(잡음이 심한) 환경에서 적용 가능성을 보여주었다. 그러나, 신경 회로망의 성능이 전체 특징의 변별력이나 모델링 능력을 좌우하며, KL 변환을 통한 파라미터의 수를 줄이는 과정에서 인식 성능이 현저하게 하락됨을 알 수 있어 추가 연구 노력이 필요하다고 판단된다.

고려중인 향후 연구 방향으로는 다음과 같은 항목을 들 수 있다.

1) 경향 특징의 표현 방법

현재 제안된 경향 특징 벡터는 기존의 모수적 분절 특징(궤적 모델)에서 위치 정보를 제거한 것이기 때문에 경향만으로는 잡음 환경에서의 모델 표현력에 한계가 있다고 판단한다. 따라서 경향 정보외에도 분절 평균 위치를 차감한 위치 보상 방법에 대해 고려해 볼 수 있다.

2) 신경회로망의 출력 정규화

신경회로망의 출력을 정규화하는 방법에는 sigmoid 방식이나 softmax, simplemax 등 여러 방법이 존재한다. Sigmoid 방식은 특정 범위를 벗어나는 경우 모두 0과 1로 한정시키는 특징이 있으며, softmax는 지수 값(exponential

value)의 합에 의하여 전체를 정규화한다. 일반적으로 사후 확률에는 softmax 방식을 많이 사용하나 본 연구에서 적용하였을 경우, 신경회로망의 수렴 시간이 오래 걸리고 수렴이 잘되지 않았다. 이것은 1)항에서 지적한 바와 같이 경향 특징에 의해 음성 특징을 잘 표현하지 못했거나 학습 데이터의 변화가 심해 발생할 수 있다고 본다. 또한 simplemax는 일반적으로 은닉층에서 널리 사용되는 방식으로 전체 출력을 합에 의해 정규화하는 방법이다. 신경 회로망의 학습에서는 이미 sigmoid 함수를 사용하여 학습하였기 때문에 softmax나 simplemax 방식에 의해 전체 출력 값을 정규화하는 방법도 고려할 만한다고 본다.

3) 신경회로망의 구조 변경

현재 본 연구에서 사용하는 신경회로망은 1개의 은닉층으로 구성된 다층구조 신경회로망(MLP; Mulit-Layer Perceptron)이다. MLP는 단순한 구조에 성능이 우수하다는 장점이 있으나 시간 정규화 기능이 약하여 여러 특징 집합을 한꺼번에 입력으로 받아 처리하고 있다. 음성 신호는 시간에 따라 진행되는 시계열 특징이기 때문에 지연 개념을 도입한다면 입력 수를 줄일 수 있으며, 음성 모델링도 강화될 수 있을 것으로 판단한다. 고려 대상인 지연 개념을 도입한 신경회로망으로는 RNN(Recurrent Neural Net)이나 RBF(Radial Basis Funcion) 등이 있다.

4) KL 변환 단계에서 정규화 적용

KL 변환은 특징 벡터를 직교 좌표로 변환시키는 선형 변환법이며, 5장에서 설명한바와 같이 특징 벡터의 공분산에 대한 고유 값과 고유 벡터를 이용하여 변환 행렬 W을 구한다. 변환 행렬에 변환하고자 하는 특징 벡터 X를 곱하여 직교 좌표를 구하는데, 변환하고자 하는 특징 벡터를 정규화하여 KL 변환을 수행하는 방법을 고려 중이다.

원 KL 변환:

$$Y = W^T \cdot X$$

수정 KL 변환:

$$Y = W^T \cdot \{\frac{X - \mu}{\sigma}\},$$

여기서 μ와 σ는 학습 벡터들의 평균과 표준 편차를 나타낸다.

5) 분석 길이의 변화에 따른 성능 평가

현재 본 연구에서는 1차 미분계수의 분석 길이와 동일한 7 프레임(윈도우 크기 3)을 사용하고 있으나 잡음의 영향을 줄이면서 안정적인 경향 벡터를 추출하기 위해서는 9 프레임, 11 프레임 등 다양한 분석 길이에 대한 성능 변화를 관찰하여야 할 것이다.

6) 연속 음성 인식에 적용

실험에 사용한 Aurora2 DB는 TIDIGIT 자료에 다양한 잡음을 추가하거나 변형하여 주로 잡음 환경에서의 음성 모델링 또는 전처리 과정을 주 대상으로 한다. 또한 인식 단위가 단어 단위이기 때문에 모노폰 단위 환경이나 트라이폰 단위에서의 성능 변화에 대한 확인 작업이 필요할 것으로 보인다. 숫자음 자료와 연속 음성 자료는 자료의 분포나 특성이 달라질 수밖에 없으며, MFCC나 1차 미분 계수의 성능 변화도 달라지기 때문이다.

참 고 문 헌

[1] M. Gales, S. Young, *The theory of segmental hidden Markov models*, CUED/F-INFENG/TR-133, Cambridge University Engineering Department, England, 1993.

[2] W. Holmes, M. Russell, "Experimental evaluation of segmental HMMs", *Proc. ICASSP*, pp. 536-539, 1995.

[3] W. Holmes, M. Russell, "Probabilistic-trajectory segmental HMMs", *Computer Speech and Language*, Vol. 13, No. 1, pp. 3-37, 1999.

[4] L. Deng, M. Aksmanovic, D. Sun, J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as non-stationary states", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 507-520, 1994.

[5] H. Gish, K. Ng, "A segmental speech model with application to word spotting", *Proc. ICASSP*, pp. II-447-450, 1993.

[6] L. Deng, "A generalized hidden Markov model with state conditioned trend functions of time for the speech signal", *Signal Processing*, Vol. 27, No. 1, pp. 65-78, 1992.

[7] H. Gish, K. Ng, "Parametric trajectory models for speech recognition", *Proc. ICSLP*, pp. I-466-469, 1996.

[8] H. Hermansky, D. Ellis, S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems", *Proc. ICASSP*, pp. 1635-1638, 2000.

[9] Y.-S. Yun, Y.-H. Oh, "A segmental-feature HMM for speech pattern modeling", *IEEE Signal Processing Letters*, Vol. 7, No. 6, pp. 135-137, 2000.

[10] Y.-S. Yun, Y.-H. Oh, "A segmental-feature HMM for continuous speech recognition based on a parametric trajectory model", *Speech Communication*, Vol. 38, No. 1, pp. 115-130, 2002.

[11] Y.-S. Yun, "Sharing trend information of trajectory in segmental-feature HMM", *Proc. ICSLP*, pp. 2641-2644, 2002.

접수일자: 2007년 5월 11일
게재결정: 2007년 6월 9일

▶ 윤영선(Young-Sun Yun) : 교신저자
주소: 306-791 대전광역시 대덕구 오정동 133번지
소속: 한남대학교 정보통신공학과
전화: 042) 629-7569
E-mail: ysyun@hannam.ac.kr

▶ 이윤근(YunKeun Lee)
주소: 305-700 대전광역시 유성구 가정로 138 한국전자통신연구원
소속: 한국전자통신연구원 음성처리연구팀
전화: 042) 860-1869
E-mail: yklee@etri.re.kr