탠덤 구조를 이용한 강인한 음성 인식 시스템 설계

윤 영 선[1] 이 윤 근[2]
[1] 한남대학교 정보통신공학과
[2] 한국전자통신연구원 음성처리연구팀

Design of Robust Speech Recognition System Using Tandem Architecture

Young-Sun Yun[1], Yunkeun Lee[2]
[1] Depart. of Information and Communication Engineering, Hannam University
[2] Spoken Language Processing Team, ETRI
E-mail: ysyun@hannam.ac.kr, yklee@etri.re.kr

Abstract

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. The proposed method shows the better results than the baseline system on very low SNR environments.

I. 서론

은닉 마코프 모델 (HMM; Hidden Markov Model)은 구현의 용이성과 유연한 모델링 능력, 높은 성능으로 인하여 많은 연구 분야에서 오랫동안 널리 사용되어오고 있다. 그러나 HMM의 약점으로 지적되고 있는 시종속성에 대한 약한 모델링 방법을 극복하려는 연구가 꾸준히 진행되어 오고 있으며, 최근에는 모델링 방식을 비롯하여 시종속성을 효과적으로 표현하고자 하는 다양한 연구가 진행되고 있다.

또한 최근에는 HMM과 신경 회로망(NN; neural networks) 모델을 결합하여 두 시스템의 장점을 결합하고자 하는 연구가 진행되고 있다. 이런 영향을 받아 NN과 HMM을 직렬로 연결한 단순한 접근 방식이 제안되었으며, 탠덤 방식(tandem approach)이라 불린다. 탠덤 방식은 먼저 NN을 이용하여 음소 집합에 대한 사후 확률(posterior probability)을 계산하며, 그 계산된 사후 확률을 이용하여 전통적인 HMM을 이용하여 음성 인식에 적용하는 방식이다.

본 연구에서는 기존의 탠덤 구조에서 사용하던 음소 단위의 NN 출력 확률 대신, 음성 데이터에 기반한 분절 특징에 대한 NN 출력 확률을 사용하는 탠덤 구조를 제안하고, 그 가능성을 살펴본다. 기존의 연구에서 사용하는 유사 음소 유닛(PLU; phoneme like unit)에 대한 사후 확률(posterior probability)은 입력 음성의 아주 작은 부분만 관찰함으로써 PLU에 대한 사후 확률을 결정하기 때문에 출력 값에 대한 변별력이 떨어질 수 있다. 따라서 본 연구에서는 NN의 입력에 대한 음성 구간에 대해 분절 특징을 이용함으로써 PLU의 대체 가능성을 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 기존의 탠덤 구조의 소개와 제안된 탠덤 구조를 설명한다. 3장에서는 본 연구에서 사용하는 분절 특징에 기반한 경향 특징에 대해 간략하게 요약하고 직교 선형 변환법으로 많이 사용되는 KL 변환을 소개한다. 제안

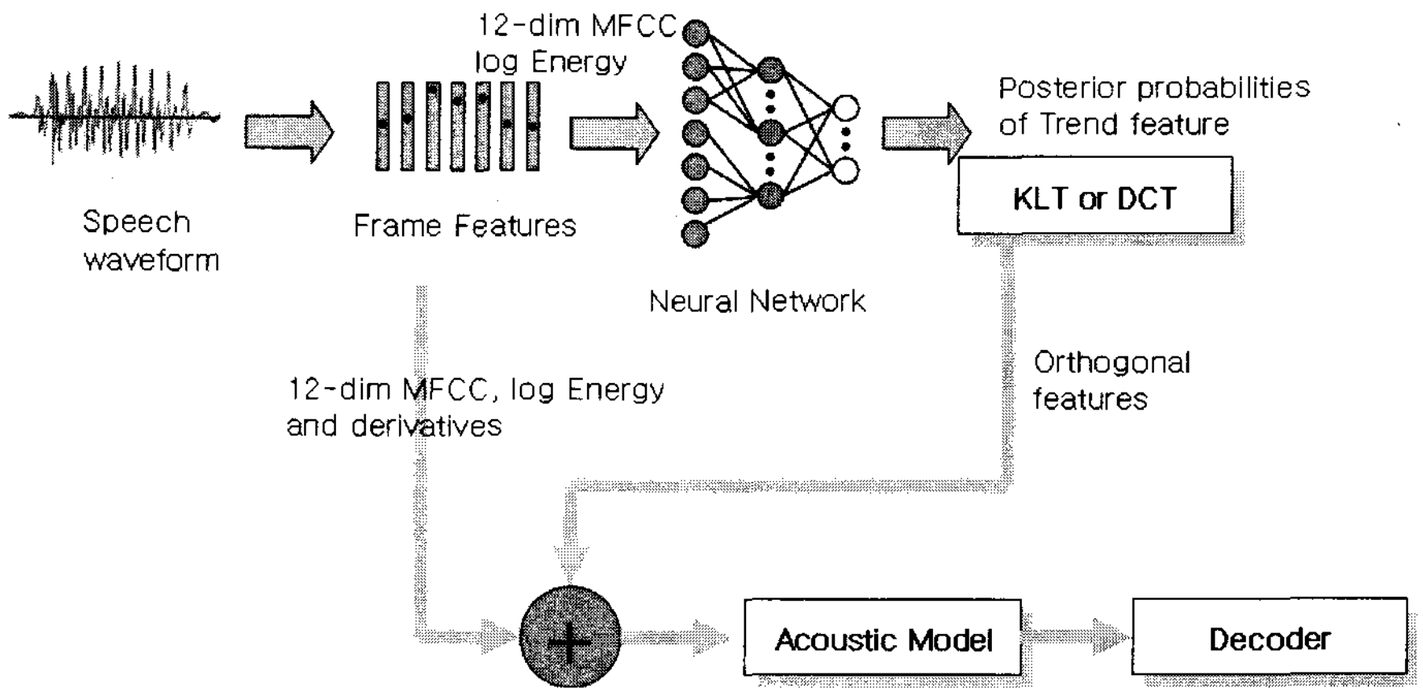

그림 1 탠덤 구조의 흐름도

된 방법의 적용 여부를 판단하기 위한 실험 및 결과를 4장에서 정리하며, 마지막으로 본 연구의 요약 및 결론을 맺는다.

II. 탠덤구조의 설계

일반적인 탠덤 구조는 Hermansky 등에 의해 잘 정의되어 있다[1]. 본 연구에서 제안한 방식은 기존의 시스템을 수정하여 PLU 대신 분절 정보를 포함하도록 하였다. 먼저 MFCC나 동적 특징에 기반을 두어 분절 정보를 추출한 후 신경회로망에 의하여 사후 확률 값으로 변환된다. 변환된 사후 확률 값은 미리 정의된 분절 정보에 대한 확률 또는 특징으로 나타나기 때문에 KLT(Karhunen Loéve Transform, 또는 Principal Component Analysis)에 의해 직교 특징 벡터(orthogonal feature vector)로 변환된다. <그림 1>에 보인바와 같이 변환된 특징 벡터는 HMM의 입력으로 전달되어 일반적인 음성 인식 시스템의 과정을 거친다.

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 NN의 입력에 따라 여러 가지 방법을 고려할 수 있으나, 본 연구에서는 프레임 특징을 경향 특징으로 변환하여 추정 오차를 제거한 다음, 변환된 경향 특징을 다시 프레임 특징으로 복원하여 NN의 입력으로 사용하는 방법을 사용하였다. 이 방식에서는 NN의 입력인 프레임 특징(복원된 프레임 특징)은 경향 특징을 추정하는 과정의 오차가 제거된 형태를 보이기 때문에 수렴속도가 빠르다는 장점이 있으나 분절 특징으로 변환하는 과정의 추정오차 때문에 왜곡된 정보가 전달될 수 있다는 단점이 존재할 수 있다.

III. 경향 특징 및 KL 변환

분절 특징은 여러 프레임 특징 또는 프레임 특징 집합에서 확률적인 방법 또는 수학적인 모델링 방식에 의해 표현되는 일련의 특징을 말한다. 최근에는 음성 인식의 성능 향상 뿐만 아니라 잡음의 영향을 받지 않거나 최소화하는 모델링 방식에 대한 연구가 진행되고 있으며, 그 대안으로 분절 특징에 대한 연구가 제시되고 있다. 분절 특징 시스템에서는 널리 사용되는 MFCC와 미분 값들을 프레임 특징들의 연속된 흐름을 나타내는 궤적 (trajectory)으로 변환하여, 분류 단계로 전달한다 [2,3,8]. 본 절에서는 분절 특징으로부터 경향 정보를 추출하는 과정과 신경회로망의 출력을 직교 특성으로 변환시키는 KL변환에 대해 요약 설명한다.

3.1 분절 특징

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. 모수적 방법이 여러 음성 단위에서 궤적의 평활화 효과가 있으며 잡음이나 환경 변화, 화자 변화에 강인한 성질을 보이고 있어, 여러 연구에서 분절 모델링에 모수적 방식을 이용하고 있다 [4,5].

일반적인 프레임 특징은 분석 구간에서 정 중앙의 단일 프레임을 나타내며, 1차 또는 2차 미분 계수를 구하는 과정에서는 여러 프레임 특징의 함축된 형식을 사용한다. 1,2차 미분계수가 프레임 집합의 특징 벡터를 평균화하여 기울기나 가속도의 형태로 축약적인 형태를 값으로 표현하는 방식에 비하여, 모수적 방식은 프레임 특징의 열에 의한 변화량을 열(sequence)로써 표현하는 방법이며 표현 방식에 따라 선형, 2차 곡선 등으로 나타내고 있다 [2,3].

3.2 경향 특징과 경향 양자화

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 [8].

이렇게 추출된 경향 정보는 벡터 양자화 알고리즘과 유사한 경향 양자화 알고리즘을 통하여 경향 정보를 공유하게 된다. 일반적으로 많이 사용되는 유클리드 거리(Euclidean distance)로 표현된 거리 척도는 분절 특징의 비교에서 첫 번째 행을 제거하여 두 경향을 비교하도록 다음과 같이 수정된다.

$$D(T_i, T_j) = \frac{1}{N}\sum_{\tau=1}^{N}\{\tilde{z}_\tau(T_i - T_j)\}\{\tilde{z}_\tau(T_i - T_j)\}',$$

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.

3.3 KL 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 정의하며, 널리 사용되는 PLU의 수와 비슷하게 64개정도를 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. 신경회로망의 출력은 각 클래스에 대한 확률 값을 가지기 때문에 음성인식 시스템에 사용하기 위해서는 특징을 직교 특성을 갖도록 변환하는 과정이 필요하며, 이때 사용하는 방법이 KL 변환이다. KL변환의 큰 특징은 우선 상관관계가 높은 특징들을 상관관계가 작은 특징으로 변환한다는 점과, 특징 중 중요 특징을 앞쪽으로 배치하는 에너지 압축의 효과가 있다. 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 계산 량이 늘어난다는 단점을 가지고 있다.

Ⅳ. 실험 및 검토

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 8kHz로 하향 샘플링하고 여러 가지 잡음과 선형 필터 (채널 잡음) 효과를 주었다. 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 인위적으로 첨가되었다. 본 연구에서는 조용한 환경에서 학습 시킨 후, Set C에 대한 인식 실험을 하고, 기존의 MFCC, Delta 특징 조합과의 비교 실험을 진행하였다. 실험결과 일반적으로 많이 사용하는 39차의 특징 벡터와 MFCC+Delta 26차와 탠덤 구조에서의 경향 벡터를 결합한 39차를 비교한 결과, 낮은 SNR에서 성능이 향상됨을 알 수 있어, 분절 특징을 이용한 탠덤 구조의 가능성을 엿보게 하였다.

V. 요약 및 결론

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, 새로운 분절 특징인 경향 특징을 이용하여 성능 향상을 꾀하였다. 기존의 탠덤 구조 방식의 신경회로망과 은닉 마코프 모델의 결합은 프레임 특징의 집합을 신경회로망의 입력으로 하고, 유사 음소 단위인 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 분절 특징으로 사용될 수

표 1 전형적인 특징 벡터 MFCC, Delta와 탠덤 구조에 의해 표현된 경향 특징 벡터 조합의 성능 비교

종류 \ 조합		MFCC	Delta (w=3)	Trend Feature	MFCC+ Delta	MFCC+ 13 TF	MFCC+ Delta+Delta2	MFCC+ Delta+13 TF
Subway	clean1	95.70	97.94	60.95	99.08	92.57	99.02	95.12
	SNR 20	59.13	93.83	51.82	93.31	82.13	94.29	89.65
	SNR 15	43.23	84.80	42.68	86.49	71.63	87.60	82.93
	SNR 10	27.82	61.34	32.02	71.08	53.95	73.23	67.15
	SNR 5	14.25	34.51	20.23	42.59	34.02	49.62	48.79
	SNR 0	7.86	19.16	13.02	17.90	16.49	23.89	26.68
	SNR -5	7.55	12.16	10.25	9.64	7.77	10.68	14.43
Street	clean2	95.28	98.00	60.10	98.73	93.38	99.00	96.28
	SNR 20	71.16	94.50	54.81	92.90	86.49	95.10	91.08
	SNR 15	58.98	87.94	46.64	85.97	77.78	88.72	84.95
	SNR 10	43.95	67.29	32.56	68.50	62.24	72.55	69.38
	SNR 5	29.63	39.33	21.52	44.65	44.65	46.86	49.88
	SNR 0	14.81	21.07	13.33	20.37	24.43	22.01	28.05
	SNR -5	9.67	11.52	9.40	10.43	10.34	10.97	14.63

있는 경향 특징을 적용하였으며, 기존 프레임 특징과 결합하여 잡음 환경 하에서 음성 인식의 성능을 높이고자 하였다.

실험결과 기존에 널리 사용되는 MFCC나 1차 미분 계수에 비해서 조용한 환경이나 높은 SNR에서 전반적으로 성능이 하락하나, 낮은 SNR에서는 성능이 향상됨을 알 수 있었다. 따라서 제안된 방식은 SNR이 낮은 잡음이 심한 환경에서 적용 가능성을 보여주었다. 그러나, recall 성능을 살펴볼 때 신경 회로망의 성능이 전체 특징의 변별력이나 모델링 능력을 좌우하며, 파라미터의 수를 줄이는 과정에서 인식 성능이 현저하게 하락됨을 알 수 있어 추가 연구 노력이 필요하다고 판단된다.

참고문헌

[1] H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," In Proc. of In. Conf. on Acoustics, Speech and Signal Processing, 2000, pp. 1635-1638, Instanbul, Turkey, 2000

[2] Y.-S. Yun and Y.-H. Oh, "A Segmental-Feature HMM for Speech Pattern Modeling," IEEE Signal Processing Letters, vol. 7, no. 6, pp. 135-137, 2000

[3] Y.-S. Yun and 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

[4] L.Deng and M.Aksmanovic and D.Sun and J.Wu, "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, 1994

[5] H.Gish and K.Ng, "A segmental speech model with application to word spotting," In Proc. of Int. Conf. on Acoustics, Speech and Signal Proc., 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, 27, pp. 65-78, 1992

[7] H.Gish and K.Ng, "Parametric trajectory models for speech recognition," In Proceedings of International Conference on Spoken Language Processing, pp. I-466-469, 1996

[8] Y.-S. Yun, "Sharing Trend Information of Trajectory in Segmental-Feature HMM," International Conference on Spoken Language Processing, pp. 2641-2644, Denver, 2002