DHMM을 이용한 숫자음 인식의 Data Fusion에 관한 연구

유 강 주[1], 신 옥 근[2]

A Study on Data Fusion of DHMM-Based Korean Digits Recognition

Gang-Ju You, Ok-Keun Shin

Department of Control & Instrumentation, Korea Maritime University, Pusan Korea

Abstract

We choose in general DHMM(Discrete Hidden Markov Model) speech recognition method, as the recognized word, the index of the word model whose output probability is maximum among those of all the words in the vocabulary. In this case, the decisions are made by comparing the similarities of the input feature vector with respect to those of reference features, which might result in the misrecognition when there exists more than two reference features similar to each other.

In this paper, we present two methods based on Data Fusion in Korean digit recognition. The first is to combine linearly Multiple Discrete Hidden Markov Models(MHMM). This method is based on the fact that, given an utterance, the recognition rate of a given DHMM varies as a function of feature vectors taken. We consider LPC(Linear Prediction Coefficient), CEP(Cepstrum Coefficient), WCEP(Weighted Cepstrum Coefficient) and MEL(Mel Coefficient)

1) 한국해양대학교 제어계측공학과 석사과정
2) 한국해양대학교 자동화 정보공학부 교수

as the basic feature vectors and generate DHMMs for each of these coefficients. Then the weights for each model are estimated for each word, so that the elementary DHMM models with high recognition rate are weighted heavily and those with lower recognition rate lightly. We need two sets of training data, one for the training of DHMMs and the other for the weight training. To demonstrate the effectiveness of the proposed combination method, the four DHMMs of a vocabulary consisting of 13 isolated digits are generated and the weights of models are estimated by Linear Programming(LP). Experimental results show the moderate improvement of recognition rate by 3.2%.

The second is a method of comparing, for a given utterance, the distribution of the output probabilities from all the word models. In fact, we adopt four DHMMs whose feature vectors are LPC, CEP, Weighted CEP and MEL, and linearly combine all the output distributions of these four DHMMs. This method is based on the assumption that every word has a unique output probability distribution for a given DHMM with a specific feature vector, and that the output probability distribution is not the same for DHMMs employing different feature vectors. We need two training data sets : the first for DHMM training, and the second for the generation of reference distribution pattern. To test the effectiveness of the proposed method, the four DHMMs and a vocabulary consisting of 13 isolated digits are generated, and the reference distribution pattern corresponding to the four feature vectors of each word is estimated. Experimental result shows an improvement of recognition rate by 6.3%.

1. 서 론

음성 인식기를 구현하려는 연구는 최근 몇 년 동안 많은 발전을 하여 현재 수십에서 수백 단어의 어휘에 대해 신뢰성 있는 인식을 할 수 있는 단계에 이르렀으며, 이미 상용화된 것들도 많이 있다. 이러한 음성 인식기들에 널리 사용되고 있는 알고리즘으로는 DTW(Dynamic Time Warping)[1], 신경망(Neural Network)[2]에 의한 방법, HMM(Hidden Markov Model)[1, 3, 4, 5, 6]을 이용한 방법 등을 들 수 있다.

DTW는 패턴을 비교하는 알고리즘으로서 임의의 단어에 대한 표준적인 특징을 가지고 있는 시계열 패턴을 기준 패턴으로 설정하고 입력된 시계열 패턴을 비선형적으로 신축해가며 비교하는 방법이다. 이 방법은 시계열 패턴의 시간적 구조의 변화를 잘 처리할 수 있으나, 개인차에 의해서 발생하는 스펙트럼의 변화를 처리하는 것에 어려움이 있다.

HMM은 화자의 개인차에 따른 음성 패턴의 변동을 통계적으로 처리한 다음, 그 통계량을 확률적인 형태의 모델에 적용하여 음성을 인식하는 방법이다. 이 방법은 확률모델을 사용하기 때문에 개인차나 조음결합등의 영향으로 나타나는 음성 패턴의 변동을 보다 정확하게 반영할 수 있다. 그러나 모델의 구조를 결정할 때 시행착오나 경험에 의하는 경우가 많고, 학습시에는 다량의 샘플데이터와 계산능력이 필요하다. HMM의 이러한 결점에 대한 대책과 인식률을 향상시키기 위하여 다양한 음성 특징 벡터를 병용하는 방법 등이 연구되고 있다.

신경망은 HMM과 같이 화자의 개인차에 따른 스펙트럼의 변화를 망의 가중치로 변환해서 음성을 인식하는 방법이다. 이 방법은 음성과 같이 개인차에 의한 입력 패턴의 시계열 길이가 다를 경우 그 구조나 가중치로 이들을 표현하는 것에 어려움이 있다.

본 논문에서는 제한된 개수의 숫자음 인식을 위하여 Data Fusion을 이용한 두 가지 종류의 인식기를 구현하였다.

첫 번째 인식기는 DHMM(Discrete HMM)의 특정 단어에 대한 인식률이 사용한 특징 벡터에 따라 차이가 난다는 점을 이용하여 4개의 특징 벡터를 바탕으로 하는 4개의 DHMM을 구현한 다음, 이들의 인식 결과를 조합하여 최종 결정을 내린다. 제안한 인식기는 각 단어의 각 특징 벡터에 해당하는 DHMM에 가중치를 부여하여, DHMM의 출력을 선형 조합하여 인식률을 개선하는 방법이다.

두 번째는 주어진 특정 발화의 특징 벡터를 모든 인식 대상 단어의 모델에 인가하여 얻은 출력값의 분포를 유사도 분포 패턴(Similarity Distribution Pattern)이라 할 때, 특정 발화를 모든 인식 대상 단어의 모델에 인가하여 얻은 유사도 분포 패턴과 각 인식 대상 단어에 대한 기준 유사도 분포 패턴의 거리를 이용한 인식기이다. 이 인식기는 각 단어의 특징 벡터마다 모든 인식 대상 단어의 모델에 대한 고유의 유사도 분포 패턴을 가지고 있다는 점을 이용하여 4개의 특징 벡터를 바탕으로 하는 4개의 DHMM을 구현한 다음, 특정 발화의 각 특징 벡터에 대한 모든 인식 대상 단어 모델에서의 유사도 분포 패턴과 각 단어의 특징 벡터에

해당하는 기준 패턴의 거리를 조합하여 최종 결정을 내리는 방법이다.

제안된 방법들의 타당성을 입증하기 위해 13개의 숫자를 단어로 하는 DHMM을 구성하였으며, 각 단어의 DHMM에 대한 가중치는 LP(Linear Programming)[7, 8]에 의해서 추정하였고, 각 단어의 기준 패턴은 각 단어의 특징 벡터마다 30개의 발화를 모든 인식 대상 단어의 모델에 입력하여 얻은 출력값의 기하 평균으로부터 추정하였다. 제안된 방법들에 의해서 구현된 인식기들이 기존의 DHMM보다 인식률이 나아질 수 있다는 것을 확인하였다.

2. Data Fusion을 이용한 음성인식 시스템

그림 2.1은 데이터 퓨전(Data Fusion)[15]시스템의 일반적인 형태이다. 데이터 퓨전 시스템에 임의의 입력이 인가되었을 때 디시전 메이커(Decision Maker)가 각 부시스템(Subsystem)의 출력을 조합하여, 이 시스템의 최종 출력을 결정하게 된다.

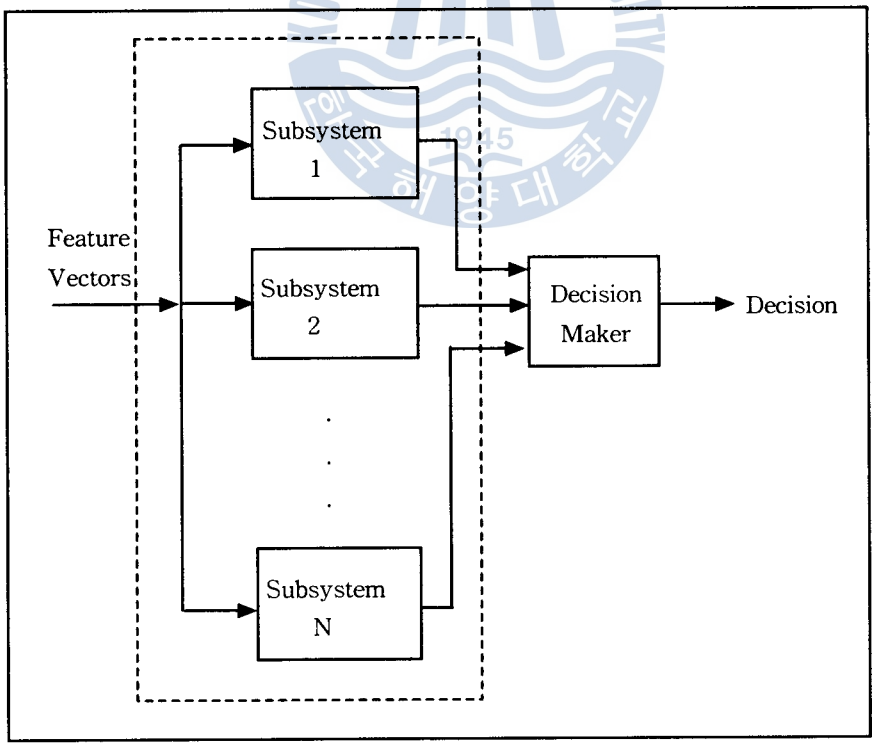

그림 2.1 데이터 퓨전 시스
Fig 2.1 Data fusion system

각 부시스템의 출력을 조합하는 방법으로는 리니어 오피니언 풀(Linear Opinion Pool)과 로그 오피니언 풀(Log opinion Pool)이 있다. 리니어 오피니언 풀은 각 부시스템의 출력과 가중치를 곱해서 합한 것을 데이터 퓨전 시스템의 출력으로 하는 방법이고, 로그 오피니언 풀은 각 부시스템의 출력의 대수에 가중치를 곱한 것을 데이터 퓨전 시스템의 출력으로 하는 방법이다. 이들을 수식으로 표현하면 각각 식 (2.1)과 (2.2)와 같다.

$$P_{Linear}(x) = \sum_{i=1}^{N} w_i P_i(x) \tag{2.1}$$

$$P_{Log}(x) = \sum_{i=1}^{N} w_i \log(P_i(x)) \tag{2.2}$$

여기서 $P_i(x)$는 i번째 부시스템의 출력이고, w_i는 i번째 부시스템의 가중치이며 N은 부시스템의 총개수이다. 그리고 $P_{Linear}(x)$는 리니어 오피니언 풀을 사용하는 데이터 퓨전 시스템의 출력이고, $P_{Log}(x)$는 로그 오피니언 풀을 사용하는 데이터 퓨전 시스템의 출력이다.

3. 실험 및 결과 고찰

3.1 음성 데이터

인식 실험에 사용된 데이터는 "영", "일", ..., "십", "백", "천"과 같이 13개의 숫자로 구성되어 있으며, 남성화자 25명이 3번씩 발음한 975개의 음성 데이터와 여성화자 25명이 3번씩 발음한 975개의 음성 데이터 중에서 남성화자 10명과 여성화자 10명의 데이터(단어별 60개의 발화)로 크기가 512인 양자화 데이블[11, 12, 13, 14]과 DHMM모델을 훈련시켰다. 그리고 DHMM모델의 훈련에 참여하지 않은 데이터 중에서 남성화자 5명과 여성화자 5명의 데이터(단어별 각 30개의 발화)를 가지고 기준 패턴의 학습과 가중치의 학습에 사용하였으며, 나머지 남성화자 10명과 여성화자 10명의 데이터(단어별 60개의 발화)를 가지고 인식 실험에 이용하였다. 11Khz로 샘플링된 모든 음성신호를 30msec의 길이로 프레임을 분할하고, 프레임과 프레임 사이의 겹침 구간을 20msec로 하였다.

그리고 각 프레임들을 $1-0.97\,Z^{-1}$의 디지털 필터로 고주파 성분을 강조한 다음, 해밍 윈도우를 적용했다. 또한 각 프레임에 대해서 자기상관 계수와 Levinson-Durbin 알고리즘을 이용하여 12차의 선형예측 계수[1, 9, 10]를 구한 후에 이를 이용하여 켑스트럼 계수[1, 9, 10]를 구하고, 이들로부터 가중 켑스트럼 계수[1, 9, 10]와 멜 켑스트럼 계수[1, 9, 10]를 구하였다. 이들을 이용하여 각 특징 벡터에 해당하는 양자화 테이블을 만들고, 모든 음성 데이터를 양자화하였다.

3.2 인식 실험 및 결과

그림 3.1은 각 특징 벡터의 DHMM, 다중 DHMM 그리고 유사도를 이용한 인식 시스템의 전체적인 인식률을 나타낸 것이다. 이 그림에서는 인식률이 가장 높은 특징 벡터는 멜 켑스트럼 계수이고, 인식률이 가장 낮은 특징 벡터는 선형예측 계수라는 것을 알 수 있다. 그리고 멜 켑스트럼 계수의 경우에만 기준 패턴과의 유사도를 이용한 시스템은 멜 켑스트럼 계수의 DHMM모델보다 0.9%의 인식률 향상을 보이고, 다중 DHMM은 멜 켑스트럼 계수의 DHMM보다 3.2%의 인식률이 개선됨을 보인다. 그리고 모든 특징 벡터에 대해서 기준 패턴과의 유사도를 이용한 인식 시스템은 멜 켑스트럼 계수의 DHMM모델 보다 6.3%의 인식률이 향상됨을 보인다.

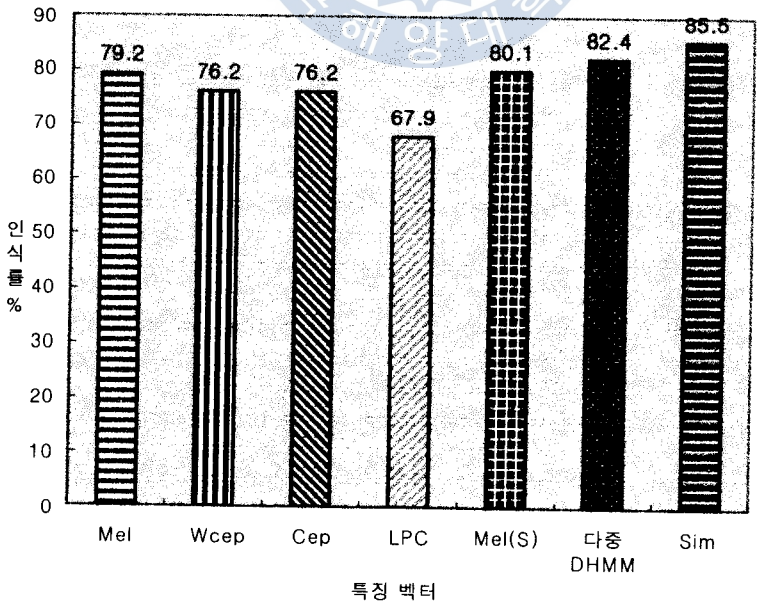

그림 3.1 인식률의 비교
Fig 3.1 Comparison of recognition rates

4. 결 론

본 논문에서는 데이터 퓨전을 이용한, "다중 DHMM 방법"과 "DHMM의 유사도 분포 패턴과 각 단어의 기준 패턴의 유사도 방법"을 제안하였다.

첫 번째 방법은 동일 단어일지라도 특징 벡터의 종류에 따라 인식률이 다르다는 것을 기초로 하여 인식률이 높은 특징 벡터의 모델에는 큰 가중치를, 인식률이 낮은 특징 벡터의 모델에는 작은 가중치를 부여해서, 모델의 출력값을 선형조합으로서 전체적인 인식률을 개선하는 것이다.

두 번째 방법은 각 단어의 특징 벡터마다 모든 인식 대상 단어의 모델에 대한 유사도 분포 패턴이 다르다는 점을 기초로 하여, 각 인식 대상 단어의 각 특징 벡터에 대해서 기준 패턴을 만든 다음, 이 기준 패턴과 모든 인식 대상 단어의 모델에 대한 특정 발화의 유사도 분포 패턴의 거리를 이용해서 인식률을 개선하는 것이다.

제안한 방법들의 타당성을 검증하기 위하여 13개의 숫자 음성을 인식대상 단어로 하는 DHMM을 선형예측 계수, 캡스트럼 계수, 가중 캡스트럼 계수, 델 캡스트럼 계수 각각의 특징 벡터에 대하여 구성하였다. 각 음성 모델에 대한 가중치는 LP을 이용해서 추정하였으며, 각 단어의 각 특징 벡터에 해당하는 기준 패턴은 각 단어의 특징 벡터마다 30개의 발화를 모든 인식 대상 단어의 모델에 인가하여 얻은 출력값의 기하 평균으로부터 추정하였다. 그리고 남성화자 10명과 여성화자 10명이 각각 세 번씩 이야기한 13개의 숫자 음성으로 인식 실험을 하였다. 그 결과, (1) 첫 번째 방법은 가장 인식률이 높은 델 캡스트럼 계수의 DHMM 보다 약 3.2%정도의 인식률이 향상 되었으며, (2) 두 번째 방법은 가장 인식률이 높은 델 캡스트럼 계수의 DHMM 보다 약 6.3%정도의 인식률이 향상되었으므로, 인식률 향상에 효과가 있음을 알 수 있었다. 그러나 특징 벡터의 종류가 많아지면 모델의 학습 시간이나 인식 시간이 길어지고 계산량이 증가되는 단점이 있다.

제안한 방법들의 인식률을 더 향상시키기 위해서는 음성의 시작점과 끝점을 좀 더 효과적으로 찾는 알고리즘과 벡터 양자화에서 발생하는 오차를 줄일 수 있는 양자화 대이블 생성 알고리즘이 필요하다고 생각되며, 지속 시간을 고려한 CHMM(Continuous HMM)과 DHMM, 혹은 DHMM과 DTW를 이용한 방법 등, 성격이 서로 다른 특징 벡터들 사이에 데이터 퓨전을 적용하면 더 효과적일 수 있으리라 기대된다.

참고 문헌

[1] Lawrence Rabiner and Biing Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] Rav P. Ramachandran and Richard J. Mammone, *Modern Methods of Speech Recognition*, Kluwer Academic, pp.159-183, 1995.

[3] Lawrence Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, Vol.77, No. 2, FEBRUARY 1989.

[4] L.R. Rabiner, S.E. Levinson and M.M.Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent Isolated Word Recognition", Bell System Technical Journal, Vol. 62, No. 4, APRIL 1983.

[5] S.E. Levinson, L.R. Rabiner and M.M. Sondhi, "An introduction to the Application of the Theory of Probabilistic Function of a Markov Process to Automatic Speech Recognition", Bell System Technical Journal, Vol. 62, No. 4, APRIL 1983.

[6] S.E. Levinson, "Structural Method in Automatic Speech Recognition", Proc. IEEE, Vol. 73, No. 11, NOVEMEBER 1985.

[7] William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery, *Numerical Recipes in C (The art of scientific Computing)*, Cambridge University Press, pp.430-444, 1992.

[8] James P. Ignizio and Tom M. Cavalier, *Liner Programming*, Prentice Hall, 1994.

[9] John Makhol, "Linear Prediction: A Tutorial Review", Proc. IEEE, Vol. 63, No. 4, APRIL 1975.

[10] Joseph W. Picone, "Signal Modeling Techniques in Speech Recognition", Proc. IEEE, Vol. 81, No. 9, SEPTEMBER 1993.

[11] John Makhoul, Salim Roucos and Herrert Gish, "Vector Quantization in Speech Coding", Proc. IEEE, Vol. 73, No. 11, pp.1551-1588, NOVEMBER 1985.

[12] Alex Waibel and Kai Fu Lee, *Reading in Speech Recognition*, Morgan Kaufmann, pp.75-100, 1990.

[13] Yoseph Linde, Andres Buzo and Robert M. Gray, "An Algorithm for Quantizer Design", IEEE Trans. Communications, Vol. COM-28, No. 1, pp.84-95, JANUARY 1980.

[14] Ravi P. Ramachandran and Richard J. Mammone, *Modern Methods of Speech Recognition*, Kluwer Academic, pp.23-50, 1995.

[15] Ravi P. Ramachandran and Richard J. Mammone, *Modern Methods of Speech Recognition*, Kluwer Academic, pp.279-297, 1995.