고립단어 인식을 위한 빠른 전처리기의 구현

Implementation of A Fast Preprocessor for Isolated Word Recognition

안 영 목*
(Youngmok Ahn*)

※이 연구는 정보통신부의 지원으로 이루어진 결과물입니다.

요 약

본 논문에서는 고립단어 인식을 위한 빠른 전처리기를 소개한다. 제안하는 전처리기는 적은 계산량으로 후보 단어를 추출한다. 본 전처리기에서는 계산량을 줄이기 위해서 벡터 양자화 대신에 특징 정렬 알고리즘을 사용하였다. 이 전처리기의 유효성을 보이기 위해서 준연속 은닉 마코프 모델을 기반으로 한 음성 인식기와 벡터 양자화를 기반으로 한 전처리기에 대해서 화자독립 고립단어 인식에 대한 성능을 비교했다.

실험에 사용한 음성 데이터는 남성 화자 40명이 발성한 244 단어이며, 40명의 화자 중에서 20명은 전처리기의 훈련용으로 사용했으며 나머지 20명은 평가용으로 사용하였다. 실험의 결과, 음성 데이터에 대해서 90%의 감축률 조건에서 제안한 전처리기는 99.9%의 정확성을 보였다.

ABSTRACT

This paper proposes a very fast preprocessor for isolated word recognition. The proposed preprocessor has a small computational cost for extracting candidate words. In the preprocessor, we used a feature sorting algorithm instead of vector quantization to reduce the computational cost. In order to show the effectiveness of our preprocessor, we compared it to a speech recognition system based on semi-continuous hidden Markov model and a VQ-based preprocessor by computing their recognition performances of a speaker independent isolated word recognition.

For the experiments, we used the speech database consisting of 244 words which were uttered by 40 male speakers. The set of speech data uttered by 20 male speakers was used for training, and the other set for testing. As the results, the accuracy of the proposed preprocessor was 99.9% with 90% reduction rate for the speech database.

I. 서 론

DSP 기술의 발전으로 음성 처리 속도는 급격히 향상되고 있으나 음성 인식 제품의 구현에서 많은 계산량에 따른 어려움이 여전히 존재한다. 특히 기존의 음성 인식 시스템에서 은닉 마코프 모델(hidden Markov model) 혹은 동적 시간 휘어짐(dynamic time warping) 알고리즘을 기반으로 하였을 경우 대어휘 고립단어를 인식하기 위해서는 대단히 많은 계산이 요구된다. 또한 이러한 인식 알고리즘을 가전 제품 등에 적용할 경우 많은 계산량 때문에 여러 가지 문제점이 발생된다. 따라서 고립단어 음성 인식에 있어서 효율적인 계산량 감축 방법이 요구된다.

*한국전자통신연구소 음성언연구실
접수일자 : 1996년 11월 28일

계산량 감축 방법으로 은닉 마코프 모델을 기반으로 한 음성 인식기는 비터비 빔(Viterbi Beam) 탐색 기법을 이용하기도 한다[1]. 그리고 전처리기를 이용하여 적은 수의 인식 후보 단어를 미리 검출함으로써 세밀한 기준 패턴과의 비교가 이루어지는 본 탐색에서 그 계산량을 줄이는 방법도 사용된다[2]. 그런데 이러한 전처리기는 몇 가지 조건을 만족해야 한다. 첫째, 후보 단어 검출 단계에서 요구되는 계산량이 주 탐색기보다 매우 적어야 한다. 둘째, 전처리기를 통해서 추출한 후보 단어의 수는 전체 대상 어휘보다 훨씬 적어야 한다. 셋째, 축소된 후보 단어 때문에 원래의 음성 인식 시스템 성능이 떨어져서는 안된다.

본 논문에서 제안하는 전처리기는 이러한 조건을 바탕으로 설계되었다.

Ⅱ. 계산량 감축 방법의 개요

벡터 양자화 및 은닉 마코프 모델을 기반으로 한 음성 인식 시스템은 보통 그림 1과 같은 과정을 거치게 된다. 이러한 음성 인식 시스템의 전체 계산량은 벡터 양자화 과정 및 비터비 탐색이 대부분을 차지한다. 음성 인식 시스템에서 각 처리 단계별로 차지하는 계산량을 표 1에 나타냈다[3]. 표 1은 분석 구간, 분석 차수, 음성 특징 벡터 크기, 코드북의 크기, 인식 대상 어휘 등에 따라서 달라질 수 있다. 즉, 벡터의 크기 및 코드북의 크기가 커지면 벡터 양자화 과정에 소요되는 계산량이 증가하고, 인식 대상 어휘가 증가할 경우에는 비터비 탐색에 소요되는 계산량이 증가된다. 따라서 음성 인식 시스템의 인식 속도를 높이기 위해서는 위의 두 가지 측면을 함께 고려하는 것이 효율적임을 시사한다.

표 1. 각 단계별 계산량 비교

Preemphasis, Windowing	10 %
Autocorrelation	32.8 %
LPC	2.9 %
LPC to Cepstrum	2.2 %
Parameter Weighting	0.3 %
Differential	0.4 %
Vector Quantization	43.7 %
Viterbi Search	7.7 %

벡터 양자화의 가장 단순한 방법은 주어진 입력 벡터를 모든 코드워드와 비교하는 것이다. 그런데 이러한 방법은 벡터 양자화에 많은 시간을 소비하게 한다. 대부분의 경우 입력 벡터에 가장 가까운 코드워드를 찾는 것이 중요하지만 어느 정도의 에러가 허용될 경우에는 tree search 방법을 사용하여 벡터 양자화 시간을 줄인다. 그런데 이러한 방법의 단점은 코드북이 바뀔 경우 탐색 가지에 대한 정보를 또다시 계산해야 한다[4, 5]. 한편, 선형 탐색 알고리즘(linear search algorithm)은 현재의 입력 벡터와 코드워드 사이의 최소 거리를 이용하여 탐색 속도를 향상시킨다. 즉, 입력 벡터와 코드워드의 거리 계산에 있어서 지금까지 계산된 거리가 최소 거리 보다 작은 경우에만 다음 벡터 차원에 대해서 계산을 하고, 그렇지 않은 경우 다음 코드워드에 대한 비교가 이루어진다. 그리고 현재 저장되어 있는 코드워드의 최소 거리보다 작은 거리를 갖는 코드워드가 나올 경우 최소 거리를 갱신한다[6].

이러한 고속 벡터 양자화 방법은 음성 인식 시스템의 속도를 향상 시키지만 비터비 탐색과 같은 입력 패턴과 기준 패턴의 정밀한 비교 단계에서는 도움을 주지 못한다. 그리고 기준 패턴 비교 단계의 계산량에 대한 비중은 인식 대상 단어가 많아질 경우에 커진다. 이러한 경우에 은닉 마코프 모델을 근간으로 한 음성 인식 시스템에서는 탐색 노드의 제한을 통한 계산량 감축 방법을 사용하거나[7] 계산량이 적은 전처리기를 이용해서 인식 대상 어휘를 대폭 줄이는 방법이 사용되기도 한다[8].

이것은 다음의 수식을 바탕으로 한다[2].

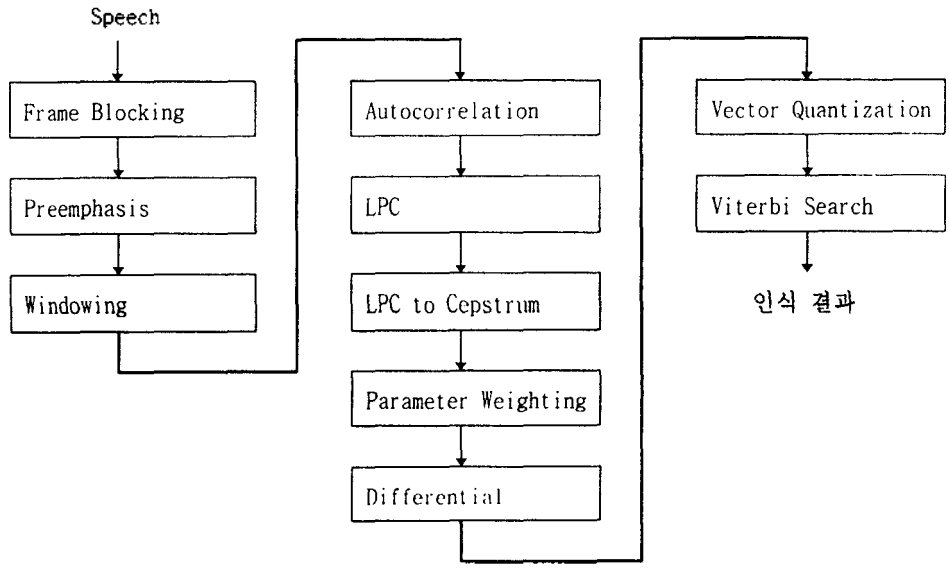

그림 1. 일반적인 음성 인식 시스템의 흐름도

$$S_w = \sum_{t=1}^{n} V(y_t, W) + i_w \qquad (w = 1, 2, \ldots, N) \tag{1}$$

식 (1)에서 S_w는 단어 w에 대한 발생값을 뜻하며, y_1, y_2, …, y_n은 벡터 양자화를 위한 코드워드 인덱스이다. $V(y_t, W)$는 y_1, y_2,…, y_n에 대한 단어 w의 모델에서 발생되는 값이며, N은 전체 인식 대상 단어이다. 여기에서 i_w는 S_w에 대한 초기값이며, 이것은 해당 단어에 대한 코드워드의 분포 및 단어의 길이에 대한 정보를 이용한 것이다. 여기에서 코드워드 발생의 순서 정보를 추가할 경우 전처리기의 성능은 개선되며, 또한 인식 대상 단어끼리 코드워드 발생 순서 및 분포가 비슷한 파라미터를 공유할 경우 보다 성능을 향상시킬 수 있다[9].

Ⅲ. 특징 정렬 기반의 전처리기

벡터 양자화를 기반으로 한 전처리기를 사용할 경우, 인식 후보 단어를 대폭 줄일 수 있다. 그러나 벡터 양자화에 필요한 계산량은 여전히 존재하므로 시스템의 보다 빠른 실행을 위해서는 이 계산량을 줄이는 방법이 요구된다. 또한 기존의 고속 벡터 양자화 방법을 사용함으로써 계산량을 줄이고 속도를 향상시킬 수 있으나, 이러한 경우 코드북이 꼭 필요하다. 만일 음성 인식 기술을 가전제품 등에 응용할 경우 이 코드북은 큰 기억 장치를 요구하게 되어 실용화에 장애를 준다. 따라서 코드북이 없이 벡터 양자화가 가능하다면 그것은 계산량 및 기억 장치의 관점에서 매우 쓰임새가 많을 것이다.

그런데 프레임 사이의 음성 특징 벡터끼리는 서로 상관 관계가 있으며, 음성 특징 벡터 안에서도 각 차원 사이에 상관 관계가 존재한다. 이러한 사실은 음성 특징 벡터 안에서 각 차원 사이의 상관 관계를 음성 인식에서 이용할 가치가 있음을 시사한다. 또한 이것은 하나의 벡터에 대한 특성을 코드북을 사용한 벡터 양자화 과정을 거치지 않아도 이와 유사한 정보를 얻을 수 있다는 가능성을 제공한다. 하나의 벡터에서 각 차원 사이의 상관 관계는 상대적인 위치로 나타낼 수 있다.

따라서 음성 특징 벡터를 각 차원의 값에 따라서 크기순으로 정렬한 결과는 해당 벡터에 대한 특성을 어느 정도 나타냈다고 할 수 있다. 이것은 벡터의 차원이 M일 경우, $M!$개의 코드워드 를 가진 코드북을 이용하여 벡터 양자화했다고 볼 수 있다. 따라서 이 방법은 코드북 및 많은 계산량이 필요하지 않지만 하나의 단어에 대한 기준 패턴을 M!개의 코드워드로 나타내기는 불가능하다. 그런데 이 문제는 벡터의 각 차원에 따른 결과만을 이용한다면 그 종류가 M개로 줄어들기 때문에 가능해진다.

이 결과를 이용하기 위해서 식 (1)을 다음의 식으로 변환하여 사용한다.

$$S_w = \sum_{t=1}^{n} \sum_{k=1}^{M} V(y_{t,k}, W) + i_w \qquad (w = 1, 2, \ldots, N) \tag{2}$$

식 (2)에서 $y_{t,k}$는 임의의 시간 t에 해당된 벡터에서 k번째 차원에 해당되는 값이 해당 벡터의 어느 위치에 속하는가를 나타낸다. 여기에서 M은 음성 특징 벡터의 크기를 나타내며, S_w 및 N 그리고 i_w 등은 식 (1)과 동일하다. 모든 인식 대상 단어에 대한 음성 특징은 $y_{t,k}$의 형태로 표시되며 그 발생 순서가 전처리기의 기준 패턴에 저장된다.

Ⅳ. 실험 및 결과

1. 시스템 개요

기본 시스템의 성능 평가에 사용된 음성 데이터는 244개의 고립단어로 이루어졌다. 이것은 40명의 남성 화자가 1회씩 발성한 것이다. 음성 데이터의 내용은 '10호'에서 '99호', '1월'에서 '12월', '1일'에서 '31일' 및 영어 알파벳, 기타 호텔 예약과 관련된 어절들로 구성되어 있다.

본 실험에서는 제안하는 전처리기의 성능 평가를 위해서 준연속 은닉 마코프 모델(Semi Continuous Hidden Markov Model)을 이용한 단어 인식기 및 벡터 양자화 기반의 전처리기의 244단어에 대한 인식 결과를 비교하였다.

제안하는 전처리기 및 가변 어휘 인식기에 사용한 음성 특징 벡터 추출 과정은 다음과 같다. 먼저, 10 msec (160 samples)마다 256 point FFT를 수행하고, 이로부터 PLP(perceptually linear prediction) 특징 벡터를 구한다. 구해진 특징 벡터로부터 dynamic feature를 구하기 위해 FIR filter를 사용하여 first-order dynamic feature를 얻고, 이 두 가지 벡터를 연결한 26차 벡터에 mean-substraction을 이용한 정규화를 거쳐 최종적인 26차 특징 벡터를 구한다. 한편 벡터 양자화 기반의 전처리기는 분석 구간이 20 msec 이고, 분석 주기가 10msec인 12차 가중 켑스트럼을 음성 특징으로 사용하였다.

SCHMM 기반의 음성 인식기는 가변 어휘 단어 인식기로써 발음 사전 생성기를 이용하여 새로운 어휘에 대한 모델을 생성한다. 모델의 훈련에는 POW 3848 DB[10]를 사용하였으며, 40개의 기본 음소 및 1,548개의변이음[11] 모델을 사용하였다[12].

2. 실험 결과 및 분석

전처리기의 기준 모델 생성을 위해서 40명의 화자 중에서 20명의 화자에 대한 음성을 사용하였으며, 나머지 20명은 실험에 사용하였다. 244 단어에 대한 실험 결과, 벡터 양자화 기반의 전처리기는 92.54%이었으며, 가변 어휘 인식기의 인식 성능은 71.86%이였고 제안하는 전처리기는 68.31%이였다. 그리고 전처리기의 상위 Top 4에 대한 성능도 추가적으로 표 2에 나타냈다.

전처리기의 후보 단어 감축률에 대한 실험 결과는 표 3에 나타냈다. 여기에서 감축 조건이란 각 후보 단어의 발생 값과 첫 번째 후보 단어의 발생 값에 대한 비율을 뜻한다. 후보 단어수는 감축 조건을 만족하는 단어의 수를 의미한다. 그리고 감축률이란 전체 인식 대상 단어를 어

느 정도 줄였는가에 대한 값이다. 다시 말하면 첫 번째 후보의 발생 값과의 비율이 5%인 단어인 단어는 평균적으로 3.6 단어이며, 이것은 후보 단어를 98.5% 감축시킨 것이며, 3.6 단어로 인식할 경우 91.8%의 인식 성능을 보인다는 뜻이다.

표 2. 제안된 전처리기의 Top 4 인식 성능

Top N	인식 성능
Top 1	68.3 %
Top 2	81.7 %
Top 3	87.7 %
Top 4	91.0 %

표 3. 제안된 전처리기의 후보 단어 감축 성능

감축 조건	인식 성능	후보 단어수	감축률
5 %	91.8 %	3.6	98.5 %
10 %	98.9 %	10.6	95.7 %
15 %	99.9 %	24.3	90.0 %
20 %	99.9 %	45.5	86.4 %
25 %	99.9 %	72.7	70.2 %
30 %	100.0 %	102.6	58.0 %

V. 결 론

본 논문에서는 고립단어 인식에서의 계산량 감축 방법을 살펴보았다. 그 방법으로써 계산량이 매우 적은 전처리기를 사용하였으며, 음성 특징의 정렬 방법을 통해서 기존의 벡터 양자화에 필요한 많은 계산을 피하였다. 제안한 전처리기는 전체 인식 대상 후보 단어를 90%줄일 경우 99.9%의 정도의 성능을 보였다. 또한 본 전처리기는 인식 대상 단어의 유사도가 크지 않은 수십 단어의 어휘에 대해서는 97% 이상의 성능을 얻을 수 있었다. 따라서 제안한 전처리기는 소규모 고립단어 음성 인식 기술을 이용한 제품 등에 효과적으로 사용될 수 있으며, 대어휘 고립 단어 음성 인식에서의 계산량을 대폭 감소시킬 수 있다.

앞으로 해야 할 연구는 본 전처리기를 연속어 음성 및 대화체 음성에 적용하는 일이다.

참 고 문 헌

1. N.Y. Han, H.R. Kim, K.W. Hwang, Y.M. Ahn, J.H. Ryoo, "A Continuous Speech Recognition System Using Finite State Network and Viterbi Beam Search for the Automatic Interpretation," Proc. ICASSP 1995, Vol. 1, pp 117-120.
2. Lalit R. Bahl, Raimo Bakis, Peter V.de Souza and Robert L. Mercer, "Obtaining candidate words by polling in a large vocabulary speech recognition system," Proc. ICASSP 1988, pp. 489-493.
3. Eloi Batlle, Jose A.R. Fonollosa, "Computational cost and memory requirements for real-time Speech Recognition systems," Proc. ICSPAT 1996, Vol. 2, 1730-1734.
4. Nader Moayeri, David L. Neuhoff, "Time-Memory Tradeoffs in Vector Quantizer Codebooks Searching Based on Decision Trees," IEEE Trans. Speech and Audio Processing, Vol.2, pp. 490-506, Oct. 1994.
5. V. Ramasubramanium, K.K. Paliwal, "Fast k-dimensional tree algorithms for nearest neighbor search with application to vector quantization encoding," IEEE Trans. Signal Processing, Vol. 40, pp. 656-659, June 1989.
6. L. Fissore, P. Laface, P. Massafra, F. Ravera, "Analysis and Improvement of the Partial Distance Search Algorithm," Proc.ICASSP 1993, Vol. 2, pp. 315-318.
7. P. Laface, C. Vair, L. Fissore, "A FAST SEGMENTAL VITERBI ALGORITHM FOR LARGE VOCABULARY RECOGNITION," Proc. ICASSP 1995, Vol. 1, pp. 560-563.
8. 김회린, "중규모 어휘 단어인식시스템의 인식속도 및 성능 개선," 제6회 신호처리합동학술대회 논문집, pp. 735-738, 1993.
9. Y.M. Ahn, H.R. Kim, K.W. Hwang, "Development of a Simple VQ-Based Preprocessor," Proc. ICSPAT 1994, Vol. 2, pp. 1697-1699.
10. Yeonja Lim, Youngjik Lee, "Implementation of the POW (Phonetically Optimized Words) algorithm for speech database," Proc. ICASSP 1995, pp. 89-91, 1995.
11. 서영주, 성철재, 이정철, 한민수, 이영직, "한국어 대화체 인식 시스템의 구현," 제13회 음성 통신 및 신호처리 워크샵 논문집(KSCSP'96), 13권, 1호, pp. 344-347, 1996.
12. 김회린, 이항섭, "POW 3848 단어 인식기 구현 및 어휘 독립 실험," 제13회 음성통신 및 신호처리 워크샵 논문집 (KSCSP'96), 13권, 1호, pp. 127-130, 1996.

▲안 영 목(Youngmok Ahn)

1991년 1월 : 홍익대학교 전자공학과 졸업(학사)

1991년 1월~현재 : 한국전자통신연구소 음성언어연구실 연구원

※주관심분야 : 음성인식, 화자인식, 음성신호처리