어휘 그룹화를 이용한 음성인식시스템의 성능향상에 관한연구

우 상 욱, 권 승 호, 한 수 영, 이 동 규, 이 두 수
한양대학교 전자통신전파공학과
전화 : 02-2290-0358

A Study on the Efficient Speech Recognition System using Database Grouping

Sang-Wook Woo, Seung-Ho Kwon, Soo-Young Han, Dong-Gyu Lee, Doo-Soo Lee
Dept. of Division of Electrical and Computer Engineering
E-mail : wsw77@ihanyang.ac.kr

Abstract

In this paper, the Classification of Energy Labeling has been proposed. Energy parameters of input signal which is extracted from each phoneme is labelled. And groups of labelling according to detected energies of input signals are detected. Next, DTW processes in a selected group of labeling. This leads to DTW processing faster than a previous algorithm.

In this Method, because an accurate detection of parameters is necessary on the assumption in steps of a detection of speeching duration and a detection of energy parameters, variable windows which are decided by pitch period is used. Extract algorithms don't search for exact frame energy, because 256 frame window-sizes is fixed. For this reason, a new energy extraction method has been proposed. A pitch period is detected firstly; next window scale is decided between 200 frames and 300 frames. The proposed method make it possible to cancel an influence of windows.

Ⅰ. 서론

음성인식기술은 산업 전반에 걸친 지식과 기술의 밀접한 결합을 요하는 첨단 기술로써 컴퓨터, 휴대폰, 녹음기, 전자사전, 게임기, 장난감, 멀티미디어 가판대, 자동차, 로봇, 통신기기, 건물 자동화, 가정용 전자기기 등의 제품에 채택된 음성을 통한 다이얼링, 정보검색, 회의록 작성, 게임 학습, 보안 시스템(출입문, 전자상거래, 금융거래), 무인전화번호 안내, 음성구동 주문형 비디오, 각종 음성안내시스템 등을 가능하게 한다. 이와 같이 생활 전반에 음성인식 및 합성기술에 관한 중요성이 점점 높아져가고 있다.

이중 고립단어인식(Isolated Word Recognition) 분야에서는 동적패턴정합(DTW) 방법이 많이 이용되고 있는데 이는 알고리즘의 복잡성과 하드웨어로 구현이 비교적 간단하기 때문이다. 하지만 패턴동적정합은 반복적인 정합방식에 의존함으로 간단한 알고리즘에 비해 처리시간이 길다는 단점을 가지고 있다[3].

본 논문에서는 음소 단위의 에너지 파라미터를 이용한 기준패턴 그룹화를 이용하여 기존의 동적시간정합법(DTW)을 이용한 고립단어 인식시스템에 처리시간을 단축시켜 시스템의 성능을 향상 시키고자 하였다. 연산량이 줄어든다면, 저가의 프로세서를 이용하더라도 원하는 속도를 얻을 수 있으므로 보다 저렴한 비용으로 음성인식을 할 수 있는 분야가 확대 될 것이다.

Ⅱ. 특징 매개변수 검출 및 패턴정합

2.1 음성구간 검출

정확한 끝점검출에 따라 음성구간의 검출의 정확성이 인식의 정확도에 큰 영향을 미치게 된다. 끝점 검출의 방법으로는 단구간 에너지(short-time energy)를 이용하여 에너지 값이 작은 부분은 묵음구간으로 하고 큰 부분은 음성구간으로 결정하는 방법이 간단한 방법이다[2].

Fig1은 영교차율과 에너지를 이용하여 음성끝점 검출하는 블록 다이어그램을 나타내었다.

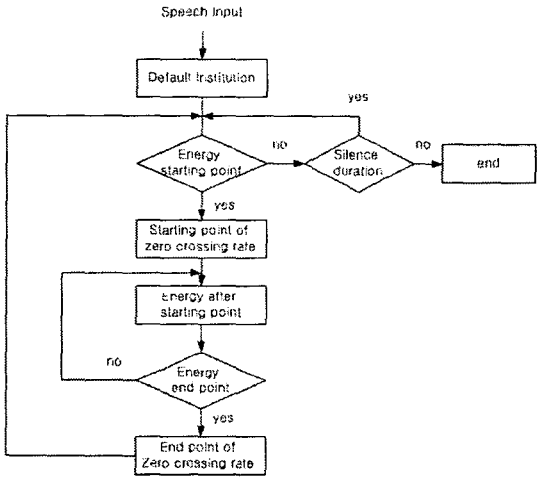

Fig 1. End-point Detection

2.2 LPC(선형 예측 계수) 검출

음성생성모델은 Fig2와 같다. 먼저 음원 성분인 유성음과 무성음을 생성한 뒤 여기신호의 크기를 조절하고 성도성분을 나타내는 시변환 필터를 거치게 되어 음성신호가 생성된다. 이때 성도를 나타내는 필터는 시변환 특성을 가지고 전극(all pole) 구조이기 때문에 다음과 같은 구조를 갖게 된다.

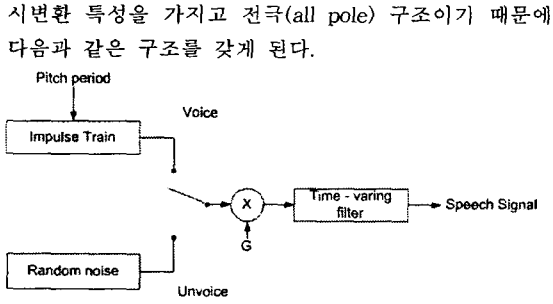

그림 2 Speech Generation Model

LPC 방법은 음성을 전극(all pole)모델로 가정하고 그에 따른 필터의 계수를 이용하여 음성 신호를 모델링하는 방법으로 음성 신호 처리에서 가장 널리 쓰이고 있는 알고리즘의 하나이고, 또한 실제 구현 시 쉽게 적용될 수 있기 때문에 많이 사용되고 있는 알고리즘이다[3]. LPC 계수에 의해 구성되는 필터는 전극특성으로 가정하여 음성이 어떻게 생성되는가하는 것을 분석해서, 성도의 특성을 모델링하게 된다.

식(1)에서 성도모델에 관한 필터계수 $a_i\,(i=1,\ldots,p)$ 가 LPC계수이다.

$$H(z)=\frac{1}{1-\sum_{i=1}^{p}a_i z^{-i}}=\frac{1}{A(z)} \quad (1)$$

2.3 동적패턴정합(DTW)

패턴 정합 방법인 동적정합법(DTW: Dynamic Time-Warping)은 길이가 서로 다른 두 개의 자료에서 최적의 정합 경로를 서로 비교할 수 있는 방법으로, 비교적 간단한 알고리즘과 최소의 하드웨어를 요구하므로 여러 분야에 효율적으로 응용할 수 있다. 이 기술은 고립단어 인식에서 개발되었으나 연속음성 인식에 역시 적용할 수 있다. 그러나 동적 프로그래밍(DP: Dynamic Programing)으로 인해 계산량이 많고, 수많은 화자내의 변위를 수용할 수 있는 기준패턴의 작성이 어려워 사용어휘가 제한되는 단점이 있다.

입력 단어 음성 패턴 T 에 대해서, 모든 단어 음성의 기준 패턴 R 과 DP 매칭을 행한 다음, 거리 $D(T,R)$ 을 최소로 하는 R 의 패턴을 T 의 패턴으로 선택한다. 즉 단어 음성 패턴 R 에 의해서 표시되는 단어에 대응하는 것으로 판단한다.

$$\begin{aligned} D(T.R) &= \min_{F(k)} D(T,R) \\ &= \frac{1}{N}\min_{F(k)}\sum_{k=1}^{K} w(k)d(a^{i(k)},b^{j(k)}) \end{aligned} \quad (2)$$

식(2)에서 특징벡터 a^i 와 b^j 의 거리 $d(a^i,b^j)$ 는 식(3)과 같은 유클리드 거리로서 주어질 수 있다.

$$d(a^i,b^j)=\|a^i-b^j\|_2=\left\{\sum_{m=1}^{M}(a_m^i-b_m^j)^2\right\}^{\frac{1}{2}} \quad (3)$$

거리를 생각하는 2개의 벡터의 시계열 a^i 와 b^j 를 대응시키는 것은 i 축과 j 축 상에서 평면상의 쌍을 식 으로 표시하고, 여기서 $F(k)$ 를 시간변환 함수라고 부른다[1].

Ⅲ. 제안한 음성인식 알고리즘

DTW를 사용한 고립단어 인식시스템의 연산량을 감소시키기 위하여 기준패턴의 음성의 최초 음절을 그 에너지에 따라 유무성음을 분리하고 그 구분된 유성음 구간을 라벨링하여 그 라벨링된 값에 따라 기준패턴을 네 개의 그룹으로 분리하였다.

단어인식 수행단계에서는 입력되어진 테스트 패턴 음성의 최초 음절의 에너지를 추출하여 그 레벨에 맞는 그룹 내에서 DTW를 수행함으로서 DTW 수행 시 소요되던 처리시간을 감소시켰다.

본 논문의 음성 라벨링 단계는 음성의 구간 검출 및 에너지 파라미터의 추출의 단계에서 정확한 파라미터의 검출을 전제로 하기 때문에 이를 보안하기 위해 피치의 주기에 따른 가변윈도우를 사용하였다.

기존의 방법은 윈도우 사이즈 256프레임으로 고정 시켜 에너지를 구함으로써 정확한 프레임 에너지를 구할 수 없었다. 따라서 피치주기를 먼저 구하고 그 주기에 200 프레임에서 300프레임 사이에서 윈도우의 크기를 결정함으로써 윈도우의 영향이 제거된 에너지를 구하는 방법을 제안하였다.

$$E_n = \frac{\sum_{i=0}^{PitchPeroid} E(i)}{PitchPeriod} \qquad (4)$$

제안된 가변 윈도우를 사용하여 에너지를 구하는 방법은 식(4) 와 같이 표현할 수 있다. 제안한 가변 윈도우에 대한 흐름도는 Fig3에서 보여 주고 있다.

Fig 3. Energy Extraction using Variable Window

본 논문에서 제안한 음성인식 알고리즘을 정리하면, 기준패턴의 음성을 그 에너지에 따라 유무성음을 분리하고 그 구분된 유성음 구간을 라벨링 함으로서 전체 인식기의 연산량 감소를 구현하였다. 하지만 음성 라벨링 단계는 음성의 구간 검출 및 에너지 파라미터의 추출의 단계에서 정확한 파라미터의 검출을 전제로 하기 때문에 이를 보안하기 위해 피치의 주기에 따른 가변윈도우를 사용하여 보다 정확한 파라미터 검출을 구현하였다.

Fig4는 본 논문에서 제안된 에너지 라벨링을 이용한 음성인식 시스템의 전체적인 구조를 보여주고 있다.

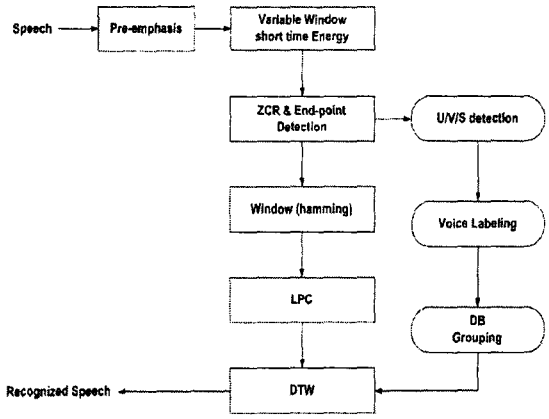

Fig 4. Proposed Recohnition System

Ⅳ. 모의실험 및 결과 분석

알고리즘을 실험하기 위해 IBM PC와 마이크가 장치된 16비트 A/D변환기를 인터페이스 하였다. 알고리즘을 구현하기위한 도구로는 Matlab6.0을 사용하여 파형분석을 하였고, 전체 인식률과 속도를 측정하기 위한 도구로 visual C++을 사용하여 인식기를 구현하였다.

4.1 에너지를 이용한 기준패턴 그룹화

가변윈도우를 사용하여 에너지 파라미터를 구하고 구해진 에너지 파라미터를 이용하여 음성의 U/V/S를 구별한다. 구별된 유성음 구간에서 에너지를 라벨링 함으로써 기준패턴을 구룹화 하였다. Fig5는 유/무성음 분리와 유성음 구간에서의 에너지 라벨링 결과를 보여주고 있다. 음성시료는 8㎑로 샘플링되고 16bit로 양자화된 raw화일을 사용하였다.

입력된 음성의 첫 음절의 에너지 값이 2.0×10^7 미만인 경우는 무성음(Unvoice)을 나타내고 에너지 값이 2.0×10^7 이상인 경우는 유성음을 나타낸다. 또 에너지 값이 1.0×10^7 미만인 경우는 무음(Silence)구간으로 나타났다. 구해진 유성음 구간을 3.0×10^7, 4.0×10^7, 5.0×10^7 으로 라벨링 하였다.

4.2 검색된 그룹 내에서의 음성인식 결과

실험은 밀폐된 일반 실험실 환경에서 20명의 화자가 각각 20개의 지정된 단어를 발성한 후 화자가 다시 테스트 음성을 발성하였다. 이 과정을 지속적으로 수행하여 인식률과 처리속도를 계산하였다.

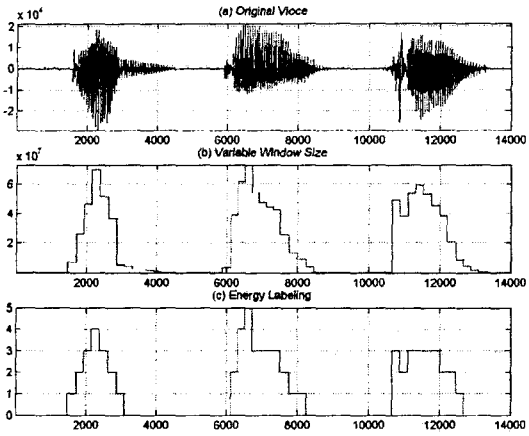

Fig 5. Energy Labeling

실험결과를 보면, 인식률의 면에서는 기준음성의 화자와 실험음성의 화자가 동일한 경우에는 기존의 방법과 제안된 방법이 약 92% 정도로 비슷한 성능을 나타내었고, 기준음성의 화자와 실험음성의 화자가 다른 경우에는 기존의 방법이 약 85% 정도이고 제안된 방법이 81%정도로 제안한 방법이 4%정도의 성능 저하를 나타내었다. 하지만 처리시간의 관점에서는 동일화자와 복수화자의 경우 모두 제안된 방법에서 76%정도로 기존의 방법보다 약 25%정도의 빠른 연산 시간을 보여 연산량의 감소를 확인할 수 있었다.

표 1. 제안된 인식기의 인식률과 처리시간

	기존의 방법	제안한 방법
동일화자 경우의 인식률	93%	92%
복수화자 경우의 인식률	85%	81%
처리시간(sec)	1.92	1.48

V. 결론

현재의 음성인식시스템은 그 응용분야에 따라 동적패턴정합(DTW), 벡터양자화(VQ), 은닉마코프 모델(HMM), 신경망(NN) 등의 다양한 방법으로 개발되어 있다. 본 논문에서는 패턴동적정합의 단점인 반복적인 정합방식에 의존함으로 알고리즘의 복잡성에 비해 처리시간이 길다는 점을 극복하고자, 음소단위로 추출된 에너지 파라미터를 이용하여 에너지를 라벨링하고 라벨링된 값에 따라 입력음성을 그룹화 하였다. 그리고 동적패턴정합 수행 시 입력된 실험음성에서 검출된 에너지의 크기에 따라 선택되어진 라벨의 그룹 내에서 DTW를 수행시켜 처리시간을 단축시켰다.

결과적으로 기존의 DTW를 사용한 고립단어 인식시스템에 연산량을 감소시킴으로서 저가형 프로세서에서도 고속으로 음성인식을 구현할 수 있을 것이다. 저가형 프로세서를 사용하여 음성인식을 수행할 수 있다면, 음성인식 알고리즘을 사용한 제품에 저가형 프로세서로 대체하여 전체 제품의 가격을 절감시키는 효과를 기대할 수 있어 음성인식기술의 적용분야가 많이 넓어 질 것 이다.

참고문헌

[1] L. R. Rabiner & R.W.Schafer, "Digital Processing of Speech Signal" , Prentice-Hall, Englewood Cliffs, N.J., U.S.A., 1978

[2] L. R. Rabiner & Biing-Hwang Juang, "Fundamentals Of Speech Recognition" , Prentice-Hall AT&T, U.S.A, 1993

[3] S. Funui, "Digital Speech Processing, Synthesis and Recognition" , Marcel Dedder, Inc., 1992

[4] Guoqing Chen , "Discovering similar time series patterns with fuzzy clustering and DTW methods", IEEE, 2001

[5] Vlandal, "Dynamic Time-Warping Method for Isolated Speech Sequence Recognition" , IEEE, 2001

[6] 조태수, "윈도우 영향이 제거된 에너지 파라미터에 관한 연구", 전자공학회 하계종합학술대회,2001