순환 퍼지연상기억장치를 이용한 음성경계 추출
(Word Boundary Detection of Voice Signal Using Recurrent Fuzzy Associative Memory)

마 창 수 † 김 계 영 ††
(Chang-Su Ma) (Gye-Young, Kim)

요 약 본 논문에서는 음성인식의 전처리 단계로서 음성 영역과 비음성 영역 사이의 경계를 검출하는 음성경계 추출에 대하여 기술한다. 본 논문에서는 음성경계 추출을 위해 두 가지의 특징벡터를 사용한다. 첫 번째는 백색잡음(white noise)에 강건한 시간 영역의 정보인 정규화된 RMS이고, 두 번째는 주파수 영역의 정보인 정규화된 멜주파수 대역 최대 에너지(mel-frequency band maximum energy)이다. 본 논문에서 사용하는 음성경계 추출 알고리즘은 학습을 통해 규칙을 생성하고 음성의 시간 정보를 적용하기 위해 순환노드를 추가한 순환 퍼지연상기억장치이다. 퍼지부의 가중치 학습은 헤비안 학습 방법을 사용하고, 순환부의 가중치 학습을 위해서는 오류 역전파(error back-propagation) 알고리즘을 사용한다. 실험에서는 KAIST에서 제공한 연령과 성별로 구분된 음성 자료를 사용하였다.

키워드 : 음성경계 추출, 멜주파수, 순환 퍼지연상기억장치, 헤비안 학습

Abstract We describe word boundary detection that extracts the boundary between speech and non-speech. The proposed method uses two features. One is the normalized root mean square of speech signal, which is insensitive to white noises and represents temporal information. The other is the normalized mel-frequency band energy of voice signal, which is frequency information of the signal. Our method detects word boundaries using a recurrent fuzzy associative memory(RFAM) that extends FAM by adding recurrent nodes. Hebbian learning method is employed to establish the degree of association between an input and output. An error back-propagation algorithm is used for learning the weights between the consequent layer and the recurrent layer. To confirm the effectiveness, we applied the suggested system to voice data obtained from KAIST.

Key words : word boundary detection, mel-frequency, RFAM, hebbian learning

1. 서 론

컴퓨터가 생활화됨에 따라서 사람과 컴퓨터 사이의 인터페이스에 대한 많은 연구가 진행되어오고 있다. 특히, 음성인식은 사람에게 친숙하고 편리한 입력 방법이기 때문에 많은 관심의 대상이 되고 있다. 음성인식을 위해서는 전처리, 특징 추출, 인식, 결과 출력 등의 과정을 거쳐야 하는데, 이 중 첫 번째 단계인 전처리 과정은 음성인식을 위한 기본단계로서 중요한 의미를 갖고있다. 음성인식을 위한 전처리 단계로는 입력신호의 잡음을 제거하는 과정, 입력신호의 주파수를 인간의 청각에 민감한 신호로 여과시키는 과정, 그리고 음성과 비음성 부분을 구분하여 음성인식에 적용할 대상을 찾아내는 음성경계 추출 과정들로 구성된다. 이 중에서 랜덤잡음(random noise) 및 백색잡음 환경에서도 강인한 음성경계 추출은 정확한 특징을 추출하고 좋은 음성인식 결과를 얻기 위해서 필수적인 것이다.

음성경계 추출을 위해 사용되는 특징은 크게 시간 영역의 정보와 주파수 영역의 정보로 나눌 수 있다. 시간 영역의 정보로 많이 사용되는 특징벡터는 영교차율(zero crossing rate), RMS(root mean square), 레벨 교차율(level crossing rate), 봉우리와 골의 비율(peak valley rate), 피치 변화(pitch variance), 적응적 시간 주파수 파라미터(adaptive time-frequency parameter) 등이 있는데 이들은 주로 음성의 시간적 변화량을 측정하고 임계치에 의해 음성과 비음성을 구분하는 방법이다[1-4]. 그러나, 잡음에 민감한 시간 영역에서의 정보들을 사용

· 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음
† 비 회 원 : (주)핸디소프트 근무
magyver@dreamwiz.com
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
gykim@computing.ssu.ac.kr
논문접수 : 2003년 6월 30일
심사완료 : 2004년 7월 16일

하여 음성과 비음성을 구분할때의 단점은 시간 영역에서의 절대적인 임계치로 정확한 결과를 얻지 못한다는 것이다. 다시 말해, 학습에 사용된 자료와 인식에 사용된 자료의 잡음 환경이 서로 다를 경우 그 차이를 정확히 측정하지 못하기 때문에 생기는 현상이다. 이를 보완하기 위하여 본 논문에서는 정규화된 RMS를 이용하여 전체 시간에 걸쳐 추가된 백색잡음이나 서로 다른 대역의 특징벡터값을 가지므로 생기는 차이를 보정하는 효과를 얻으려고 한다.

주파수 영역의 정보로 사용되는 특징벡터는 멜주파수 켑스트럼 계수(mel-scaled frequency cepstral coefficient)[3], 정제된 시간 주파수(refined time frequency)[4], 필터 뱅크(filter-bank)[5], 엔빌롭 값(Envelope Value)[6], 웨이블릿 변환 계수(wavelet transform coefficient)[7] 등이 있다. 본 논문에서는 주파수 영역의 특징벡터를 얻기 위해 원시 주파수를 인간의 청각적 특성을 고려한 멜주파수로 변환한 후 필터뱅크를 이용하여 대역 에너지의 값을 얻는다. 그리고 이 중 최대값을 선택하고 그 값을 정규화시킨 멜주파수 대역 에너지를 사용하여 랜덤잡음에 강인한 인식 결과를 얻으려고 한다.

음성경계 추출을 위한 알고리즘으로는 퍼지-신경망(fuzzy neural network)을 이용한 방법[3], 인공신경망(artificial neural network)을 이용한 방법[8], 퍼지 규칙(fuzzy rule)을 이용한 방법[9,10], 임계화 모델(threshold model)을 이용한 방법 등이 있다. 퍼지-신경망을 이용한 방법[4]에서는 시간 잡음, 적응적 시간-주파수 파라미터, 영교차율 등을 입력 벡터로 사용하였고, 알고리즘으로는 학습을 통해 노드를 자동으로 생성하는 자동 구성 퍼지-신경망(SONFIN: Self Organized Neural-Fuzzy Inference Network)을 사용하였다. 여기에서 노드들은 퍼지 규칙으로 구성되고 시스템 전체는 신경망의 구조를 갖는 5단의 퍼지-신경 추론망으로 구성되어 있으며, 노드는 학습을 통해 생성되거나 제거되는 동적인 구조를 가진다. 인공신경망을 이용한 방법[5]에서는 음성경계 추출과 레이블링을 위해 신경망과 은닉 마르코프 모델, 퍼지 규칙이 혼합된 형태의 시스템을 이용하였다. 이 시스템은 단어 인식 기능과 단어 추출 기능이 결합된 형태이며 다소 복잡한 구조를 가지고 있다. 퍼지 규칙을 이용한 방법[9]에서는 음성경계 추출을 위해 퍼지 규칙을 사용한다. 이 방법에서는 퍼지 규칙을 효과적으로 생성하는 것과 적정한 노드의 개수를 결정하는 것이 어렵다. 또한, 몇 개의 노드가 가장 효율적인지 정하는 것도 어렵다. 또한 퍼지 규칙은 시간적인 정보를 표현하는데 한계가 있다. 임계화 모델을 사용한 방법은 특징벡터의 값들을 정해진 임계치에 따라 음성과 비음성으로 나누는 단순한 모델로서 구현하기 쉽고 단순하지만 잡음에 약하다는 단점이 있다.

본 논문에서는 퍼지 규칙의 단점인 규칙 생성의 비효율성을 개선하기 위해 학습을 통해 규칙을 자동 생성하는 퍼지연상기억장치를 사용하였고, 시간적 정보를 고려하기 위해 순환노드를 추가한 순환 퍼지연상기억장치(recurrent fuzzy associative memory)를 사용하였다. 또한 잡음에 민감한 기존의 특징들을 개선하기 위해 정규화된 RMS와 정규화된 멜주파수 대역 최대 에너지를 사용하였다.

본 논문에서 사용하는 순환 퍼지연상기억장치는 크게 다섯 부분으로 이루어져 있다. 입력부는 입력 벡터를 퍼지화층으로 전달하는 기능을 하고, 퍼지화층은 입력 벡터를 퍼지값으로 변환하여 조건부층으로 전달한다. 조건부층은 퍼지규칙을 구성하고, 순환부는 조건부층의 출력을 퍼지화층의 입력으로 순환시킨다. 마지막으로 결론부층은 조건부층과 가중치의 퍼지연산 결과를 결론부층 소속함수에 적용하여 최종 결론을 얻어낸다.

1장에서는 본 논문의 기본 방향과 기존의 방법에 대하여 기술하였다. 2장에서는 음성경계 추출에 사용할 특징벡터에 대하여 기술하고, 3장에서는 음성경계 추출에 사용할 순환 퍼지연상기억장치의 기본 구조와 학습방법 등에 대하여 기술한다. 4장에서는 실험 결과에 대하여 기술하고, 마지막으로 5장에서는 결론에 관하여 기술한다.

2. 특징 추출

음성경계 추출을 위해 본 논문에서는 시간 영역의 정보인 정규화된 RMS와 주파수 영역의 정보인 정규화된 멜주파수 대역 최대 에너지를 사용한다. RMS는 영교차율과 더불어 가장 많이 사용하는 시간 정보로서 영교차율에 비해 음성과 비음성이 잘 구분되고 계산하기도 쉽다. 정규화된 멜주파수 대역 최대 에너지는 잡음에 강하며 인간의 청각적 특성에 기인한 정보이다. 시간 영역 정보가 정보 값의 크기를 특징값으로 삼으므로 직관적인 분석이 가능한 반면 정지된 시간만을 나타낸다는 단점을 가지고 있고, 반면에 주파수 영역의 정보는 직관적으로 이해하기는 어렵지만 정지된 시간이 아닌 시간 대역의 특징을 잘 표현하는 장점이 있다. 본 논문에서 시간 영역의 특징인 RMS와 주파수 영역의 특징인 멜주파수 대역 에너지를 이용하는 이유는 앞에서 말한 각각 시간 영역과 주파수 영역의 대표적인 특징들을 혼합하여 사용함으로 서로의 단점을 상호 보완 할 수 있기 때문이다.

RMS는 주어진 구간 내에서 시간에 따른 정보의 변화량을 측정한 것이다. RMS는 식 (1)과 같이 음성신호의 심플값을 제곱하여 누적시키고 구간의 길이로 나눈

후 로그를 취하면 된다. 입력 신호가 0에 가까울수록 RMS는 작아지고 멀수록 커진다. RMS를 구하는 방법을 수식으로 나타내면 식 (1)과 같다. 여기에서 L은 프레임의 길이이며 본 논문에서는 128로 정하였고, m은 프레임의 인덱스, 그리고 n은 프레임 내에서의 신호 인덱스이다. 그리고 $x_{rms}(m)$은 m번째 프레임의 RMS 값이고, $x_{time}(m,n)$은 m번째 프레임 내에서 n번째 신호의 샘플값이다. [3]에서는 RMS를 식 (2)와 같이 스무딩을 수행한다.

$$x_{rms}(m) = \log\sqrt{\frac{\sum_{n=0}^{L-1} x_{time}^2(m,n)}{L}} \quad (1)$$

$$\hat{x}_{rms}(m) = \frac{x_{rms}(m-1) + x_{rms}(m) + x_{rms}(m+1)}{3} \quad (2)$$

퍼지소속함수를 정확하게 얻기 위해서는 특징벡터의 히스토그램이 종류별로 잘 구분되어 있어야 한다. 잡음에 오염된 영역의 RMS값은 음성 영역의 RMS값과 구분을 하기가 어려워 지는데, 이러한 문제를 해결하기 위해 잡음 영역에서 RMS값의 평균을 구하고, 그 값을 모든 신호의 RMS 값에서 빼주면 전체 프레임의 RMS 값에 대하여 잡음과 음성영역을 좀더 정확하게 구분할 수 있다. 그러나 아직도 잡음이 없는 신호의 음성과 비음성의 RMS보다 잡음이 섞인 음성신호의 음성과 비음성의 값이 크므로 상대적인 비교를 할 수 없다. 따라서 RMS의 최소값과 최대값을 제한하여 일반화시킨 정규화된 RMS(normalized RMS)를 식 (3)을 이용하여 구한다. 수식에서 $rms_{\max}$는 정규화 하기 위한 정의된 최대값이며 모든 데이타에 대하여 같은 값이다. 본 논문에서는 20의 상수값을 사용하였다. 하나의 프레임은 주파수 대역필터를 통해 대역 에너지 값을 얻고 각 대역마다 최대값인 $\hat{x}_{rms}$ 이 구해진다. 현재의 프레임에서 구해진 각 대역의 $\hat{x}_{rms}$ 값중 최대값인 $\hat{x}_{rms}^{\max}$ 이 현재의 프레임을 대표하는 값이 된다.

$$T(m) = \frac{\hat{x}_{rms}(m) \times rms_{\max}}{\hat{x}_{rms}^{\max}} \quad (3)$$

위의 식을 이용하여 구한 정규화된 RMS의 음성과 비음성 히스토그램의 한 예를 그림 1에서 보여주고 있다. 그림에서 가로축은 RMS의 값이고 세로축은 빈도수를 나타낸다. 그림에서 가는 실선은 비음성 부분으로서 정규화된 RMS의 값들이 작은 값들이므로 히스토그램도 0에 가까운 부분에서 봉우리(peak) 모양을 이루며, 굵은 실선은 음성부분의 히스토그램으로 RMS가 11에서 봉우리 모양을 이룬다. 그림 1은 본 논문에서 설명하는 정규화된 RMS의 값의 분포를 나타내며 음성과 비음성의 분포가 분명하므로 음성과 비음성의 구분을 명확히 할 수 있음을 보여주고 있다.

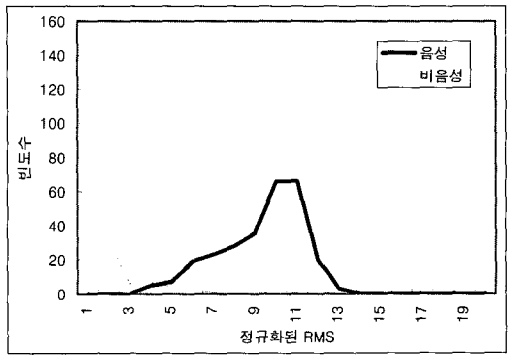

그림 1 음성과 비음성에 대한 정규화된 RMS 히스토그램

다음은 주파수 영역의 정보인데, 본 논문에서는 시간영역의 정보를 주파수 영역의 정보로 변환하기 위해 고속 퓨리어 변환을 수행한다. 주파수 영역의 정보에서는 포만트 분석을 통해 음성의 발성 특징을 알아낼 수 있다. 만약 신호 전체에 걸쳐서 백색잡음이 추가되더라도 음성 특유의 포만트 정보는 유지되므로 포만트 정보를 이용하여 랜덤잡음 혹은 백색잡음으로 이루어진 비음성 영역과 음성영역을 구분해 낼 수 있다. 본 논문에서는 멜주파수 대역 최대 에너지를 주파수 영역의 정보로 사용하는데, 주파수 영역에서 각 대역의 최대값은 포만트의 후보를 찾는 것이고 각 대역 최대 에너지중의 최대값을 선택하는 것은 한 시간을 대표하는 포만트 정보를 구하는 것이다. 이것을 구하기 위해 식 (4)와 같이 고속 퓨리어 변환된 주파수 정보를 인간의 청각적 특성에 맞는 멜주파수로 변환한다. 식 (4)에서 f는 원시 주파수이다. 멜주파수로 변환하는 이유는 인간의 청각적 특성이 고주파 보다는 저주파에 민감하기 때문에 저주파 영역에서 더욱 많은 정보를 얻어내기 위함이다.

$$f_{mel} = 2595 \times \ln\left(1 + \frac{f}{700}\right) \quad (4)$$

그림 2는 원시 주파수와 멜주파수의 축소 비율에 대한 곡선과 필터뱅크를 수행하기 위한 윈도우를 보여주고 있다. 멜주파수의 필터뱅크를 수행할 때, 가로축인 원시 주파수에서 크기가 다른 윈도우로 필터링을 하는 것 보다 세로축인 멜주파수에서 균일한 크기의 윈도우로 필터링을 하는 것이 구현하기 쉽기 때문에 필터링을 수행하기 전에 먼저 원시 주파수를 멜주파수로 변환한다. 원시 주파수를 멜주파수로 확대 및 축소한 후 비어있는 자료는 이웃 정보들의 선형 보간을 통해 채워넣는다. 식 (5)는 시간 영역의 자료를 주파수 영역으로 변환하는 식이고, 식 (6)은 필터뱅크를 수행하는 식이다.

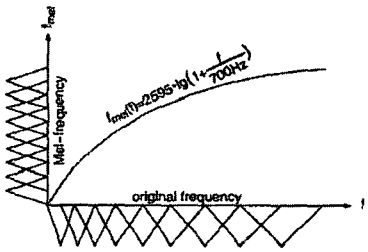

그림 2 멜스케일

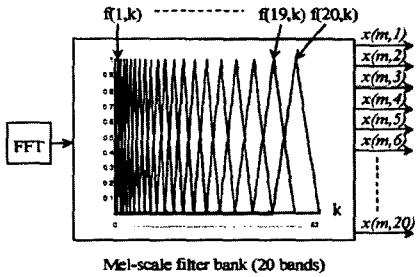

그림 3 멜스케일 필터뱅크

$$x_{freq}(m,k)=\sum_{n=0}^{N-1}x_{time}(m,n)W_N^{kn} \tag{5}$$

$$x(m,i)=\sum_{k=0}^{N-1}\left|x_{freq}(m,k)\right|f(i,k) \tag{6}$$

$x_{freq}(m,k)$ m 번째 프레임에서 k번째 스펙트럼의 크기
$x(m,i)$ m 번째 프레임에서 I번째 필터뱅크의 크기
W_N^{kn} 고속 퓨리어 변환 ($\exp(-j2\pi / N)$
N 프레임의 크기
k 스펙트럼 인덱스
i 필터 뱅크 인덱스
$f(i,k)$ 윈도우

이제 프레임마다 20개의 멜주파수 대역 에너지가 구해졌는데, 식 (7)을 이용하여 스무딩을 수행한다.

$$\hat{x}(m,i)=\frac{x(m-1,i)+x(m,i)+x(m+1,i)}{3} \tag{7}$$

그림 4에서는 멜주파수 대역 에너지의 히스토그램을 보여주고 있는데, 음성과 비음성의 히스토그램이 잘 분리되어 있기는 하지만 음성의 히스토그램이 넓게 분포되어 있는 것을 볼 수 있다. 퍼지소속함수를 정확하게 생성하기 위해서는 좀더 집중된 히스토그램을 갖는 특징벡터가 필요한데, 이를 위해 본 논문에서는 대역 에너지의 누적값을 사용하는 것이 아니라 프레임 내에서 가장 큰 대역의 값을 취하는 멜주파수 대역 최대 에너지를 사용한다. 수식으로 나타내면 식 (8)과 같다. 여기에서 i는 필터뱅크의 인덱스이다.

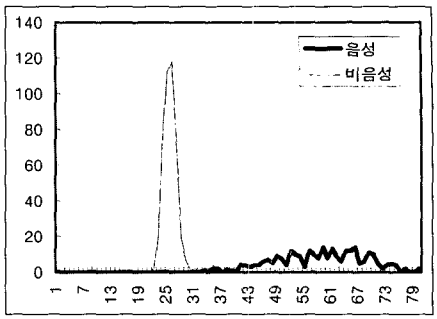

그림 4 음성과 비음성에 대한 멜주파수 대역 에너지의 히스토그램

$$F_{\max}(m)=\max[\hat{x}(m,i)]_{i=1,2...20} \tag{8}$$

식 (8)에서 구해진 값에 정규화를 수행하기 위해 식 (9)를 수행한다. 여기에서 F_{max}는 식 (8)에서 구해진 모든 프레임의 $F(m)$ 값 중 최대값이고 $mfbe_{max}$는 정규화를 위한 상수값으로 본 논문에서는 50을 사용한다.

$$F(m)=\frac{F_{\max}(m)\times mfbe_{max}}{F_{max}} \tag{9}$$

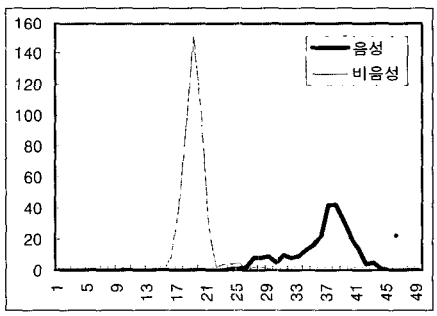

그림 5 음성과 비음성에 대한 멜주파수 대역 최대 에너지 히스토그램

그림 5에서는 멜주파수 대역 최대 에너지의 음성과 비음성에 대한 히스토그램을 보여주고 있는데, 음성의 히스토그램이 봉우리 모양의 형태를 갖는 것을 볼 수 있다. 이것은 본 논문에서 설명하는 멜주파수 대역 최대 에너지가 음성과 비음성을 구분하는데 적합한 특징이라는 것을 보여주는 것이다.

3. 음성경계 추출

순환 퍼지연상기억장치를 이용하여 음성경계 추출을 하기 위해서는 먼저 입력 벡터를 0에서 1까지의 퍼지값으로 변환시켜주는 퍼지 소속함수를 생성하여야 한다. 퍼지 소속함수는 입력 벡터의 히스토그램 분석을 통해

생성하고, 생성된 소속함수는 조건부층과 연결되어 퍼지 연상기억장치를 구성한다. 구성된 연상기억장치와 결론부층 사이의 가중치를 학습을 통하여 학습하고 학습된 가중치는 조건에 맞는 퍼지규칙을 생성한다.

음성신호를 구성하는 음성과 비음성은 시간적으로 연속성을 가지고 있다. 하지만 음성경계 추출에서 의미는 연속성은 음성과 비음성의 정보가 시간적으로 일정 기간 동안 지속적으로 발생하는 성질을 말하는 것인데, 그 특성은 보통 수십에서 수백 프레임 동안 지속된다. 본 논문에서는 이러한 연속적인 성질을 적용하기 위해 조건부층과 퍼지화층 사이에 순환 노드를 추가하여 현재 시점의 출력이 다음시점의 결과에 영향을 미치도록 구성한 순환 퍼지연상기억장치를 이용하였다.

그림 6에서는 본 논문에서 사용한 순환 퍼지연상기억장치의 시스템 구성도를 보여주고 있다. 시스템은 크게 입력부층, 퍼지화층, 조건부층, 결론부층, 순환부층의 다섯 부분으로 이루어져 있다.

입력부층은 입력된 자료를 별도의 처리 없이 퍼지화층으로 전달하는 역할을 한다. 입력은 두 개인데, 첫 번째는 시간 영역의 정보인 정규화된 RMS이고, 두 번째는 주파수 영역의 정보인 정규화된 멜주파수 대역 최대 에너지이다. 각 특징값은 해당하는 퍼지화층의 노드로 분할되어 입력된다. 퍼지화층은 6개의 퍼지소속함수로 구성되어 있는데, 서로 다른 입력을 0에서 1 사이의 퍼지값으로 변환한다. 그림 6에서 A_1과 A_2는 RMS에 대한 소속함수이며, 소속함수가 두 개인 이유는 그림 1과 같이 RMS의 히스토그램이 두 개의 피크를 갖기 때문이다. B_1과 B_2는 멜주파수 대역 최대 에너지에 대한 소속함수이며, 소속함수가 두 개인 이유는 그림 5와 같이 멜주파수 대역 최대 에너지의 히스토그램이 두 개의 피크를 갖기 때문이다. 또한 R_1과 R_2는 순환부층에서 들어온 입력을 위한 소속함수이며, 두 개의 소속함수를 갖는 이유는 음성경계 추출의 최종 결과가 음성과 비음성으로 나뉘기 때문이다. 그러므로 퍼지화층의 노드 개수는 6개가 이다. 소속함수를 생성하는 방법은 뒤에서 자세히 설명할 것이다. 조건부층에서는 퍼지화층에서으로부터 입력된 특징벡터의 퍼지값을 이용하여 퍼지곱의 형태로 퍼지규칙을 생성한다. 그림 7에서는 퍼지규칙의 예를 보여주고 있는데, 조건부층 노드의 개수는 8개이다. 조건부층은 결론부층에 연결되며, 동시에 순환부층과도 연결된다.

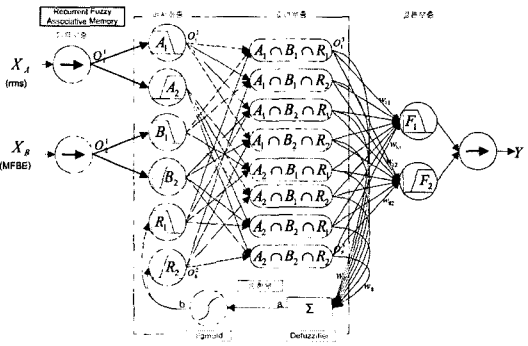

그림 6 순환 퍼지연상기억장치의 시스템 구성도

그림 8에서는 퍼지추론의 한 예를 보여주고 있다. 퍼지연산에서의 퍼지곱은 최소 연산을 의미하며, 최소 연산을 통해 계산된 값이 결론부층의 입력이 된다. 결론부층에서는 입력값과 소속함수와의 논리 곱을 통해 결과값을 얻게 된다. 그림 8에서 함수 F의 검게 칠해진 아래 부분이 결론부층의 입력과 소속함수와 논리곱을 나타내는 것인데, 결론부층을 이루는 두 소속함수의 결과를 가로축을 기준으로 누적시켜서 무게중심을 구한 것이 최종 결과값이 된다. 퍼지연상기억장치를 수식으로 나타내면 식 (10)과 같다. 여기에서 $\circ$는 퍼지 최대-최소 연산을 의미하고, x는 입력 퍼지함수로서 조건부층의 출력이다. 또한 M은 사상 행렬, 즉 가중치 행렬을 의미한다.

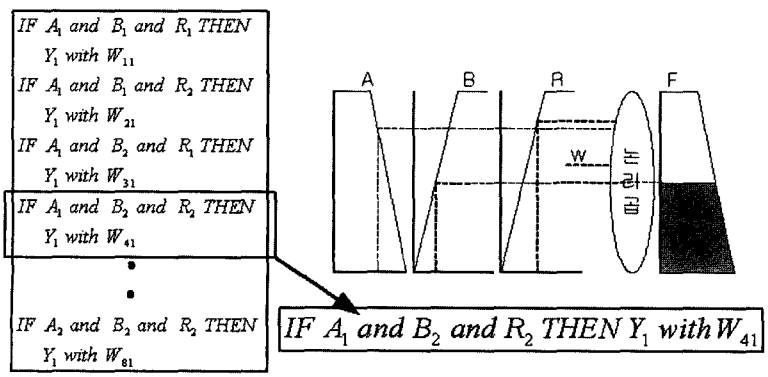

그림 7 퍼지 규칙

그림 8 퍼지 연상 기억장치를 이용한 추론 모델

$$y = M \circ x \tag{10}$$

식 (11)은 식 (10)을 퍼지연상기억장치에 적용하여 계산한 예이다.

$$y = M \circ x = \begin{bmatrix} 0.2 & 0.2 & 0.4 & 0.9 & 0.1 & 0.0 & 1.0 & 0.2 \\ 0.9 & 0.8 & 0.1 & 0.2 & 0.0 & 0.8 & 0.3 & 0.2 \end{bmatrix} \circ \begin{bmatrix} 0.0 \\ 0.1 \\ 0.1 \\ 0.8 \\ 0.7 \\ 0.3 \\ 0.2 \\ 0.4 \end{bmatrix}$$

$$= \begin{bmatrix} \max(\min(0.2,0.0), \min(0.2,0.1), \cdots, \min(1.0,0.2), \min(0.2,0.4)) \\ \max(\min(0.9,0.0), \min(0.8,0.1), \cdots, \min(0.3,0.2), \min(0.2,0.4)) \end{bmatrix} = \begin{bmatrix} 0.8 \\ 0.3 \end{bmatrix} \tag{11}$$

여기에서 x는 $\min(A_i, B_j, R_k)$ 연산을 통해 얻어진 조건부층의 출력을 말한다. x와 가중치 w의 퍼지최소연산을 수식으로 나타내면 $\min(\min(A_i, B_j, R_k), W_l)$와 같은데, 식을 단순화시키면 모든 입력과 가중치의 퍼지 최소연산을 동시에 수행하는 것과 같아지기 때문에 $\min(A_i, B_j, R_k, W_l)$와 같이 식을 단순화 시킬 수 있다.

순환부층은 역퍼지화 함수와 시그모이드 함수로 이루어진다. 역퍼지화 함수에서는 조건부의 출력과 가중치를 곱한 후 모두 더해서 하나의 출력값을 얻는다. 시그모이드 함수에서는 역퍼지화 함수의 출력을 0에서 1 사이의 퍼지값으로 변환하여 퍼지화층의 입력으로 전달한다. 역퍼지화 함수를 수식으로 나타내면 식 (12)와 같고, 시그모이드 함수를 수식으로 나타내면 식 (13)과 같다. 여기에서 w는 가중치로서 오류 역전파 알고리즘에 의해 학습된다.

$$a = \sum_{i=0}^{7} O_i^3 w_i \tag{12}$$

$$b = f(a) = \frac{1}{1+e^{-a}} \tag{13}$$

퍼지 소속함수의 생성은 그림 9와 같이 특징벡터의 히스토그램 분석을 통해 수행된다. 소속함수를 생성하기 위해서는 먼저 특징벡터의 히스토그램을 생성하고, 불규칙한 값을 제거를 위해 스무딩을 수행한 다음 지역 최대값을 찾아낸다. 그리고 선택된 최대값을 기준으로하는 삼각형 형태의 함수를 생성하고, 그림 9(b)와 같이 두 최대값 사이에서 최소값을 찾아 두 함수를 중첩시킨다. 이 단계가 끝나면 그림 9(c)와 같이 좌우 함수의 끝이 1이 되도록 조정한다. 본 논문에서는 그림 1과 그림 5의 히스토그램을 분석하여 각각 2개의 소속함수를 생성한다.

조건부층과 결론부층 사이에 있는 퍼지연상기억장치의 가중치 학습은 [8]에서 제안한 헤비안 학습방법을 사용한다. 식 (14)와 식 (15)는 헤비안 학습방법을 이용한 퍼지연상기억장치의 가중치 학습을 수식으로 나타낸 것

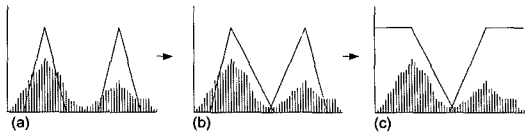

그림 9 퍼지소속함수 생성

이다. 여기에서 ⊕는 퍼지합 연산을 의미한다. 본 논문에서 헤비안 학습방법 에서의 곱과 합을 퍼지곱과 퍼지합으로 대체하여 사용한 것은 퍼지 연산에서의 곱과 합으로 적용하기 위함이고 [15]에서 보여주고 있다.

$$w_{ij}(new) = \Delta w_{ij} \oplus w_{ij}(old) \tag{14}$$

$$\Delta w_{ij} = \min(\min(A, B, R), y_i) \tag{15}$$

조건부층과 순환부층을 잇는 순환부의 가중치 학습을 위해서는 오류 역전파 알고리즘이 사용되는데, 학습시에 입력값 X_A와 X_B를 알고있고, 원하는 출력값 b를 알고 있으므로 지도학습이 가능하다. 초기의 X_R 값은 0으로 초기화 하고 학습을 시작한다. 수식으로 나타내면 식 (16), (17)과 같다.

$$\delta = b(1-b)(b-d) \tag{16}$$

$$w_i = w_i - c\delta O_i \tag{17}$$

여기에서 d는 목표 출력값이고, b는 실제 출력값, O는 이전 노드의 출력이다. δ는 목표값과 실제 출력간의 오류값이다. c는 학습률로서 본 논문에서는 0.2를 사용하였다. c가 클수록 수렴하는 속도는 빨라지지만 정확도는 떨어진다.

4. 실험 결과

본 장에서는 본 논문에서 사용한 순환 퍼지연상기억장치와 정규화된 두 개의 특징벡터를 이용하여 음성경계 추출을 수행한 실험 결과에 대하여 기술한다. 실험을 위해 펜티엄4, 1.8GHz의 CPU와 256MB의 메모리를 갖춘 컴퓨터가 사용되었고, 개발 언어는 MS Visual C++ 6.0을 사용하였다. 실험을 위한 음성자료는 KAIST에서 제공하는 로우 포맷의 자료를 사용하였는데, PC환경의 중가 마이크로 녹음된 낭독체의 음성자료이며 음소와 출신지방이 적절히 배분되어 있는 것이다. 또한, 자료는 나이에 따라 10대, 20대, 30대, 40대로 구분되어 있고, 성별에 따라 남성과 여성으로 구분되어 있다. 표 1에서는 학습에 사용한 자료에 대하여 설명하고 있다. 각 자료들은 8kHz로 샘플링 되어있는데 1회 학습을 수행할 때 표 1의 자료를 모두 한번씩 학습하게 되는데, 이 경우 전체 시간이 65.051초가 되므로 프레임 크기가 128이고 64 샘플을 중첩시킨 경우 약 8000 프레임을 학습에 사용한 것이 된다.

표 1 자료의 종류

파일 이름	성별	나이	시간
10Fm	여성	10대	6.575sec
10Mm	남성	10대	8.375 sec
20FFm	여성	20대	4.375 sec
20Fm	여성	20대	5.975 sec
20MMm	남성	20대	7.175 sec
20Mm	남성	20대	5.675 sec
30Fm	여성	30대	6.175 sec
30Mm	남성	30대	6.375 sec
40Fm	여성	40대	7.375 sec
40Mm	남성	40대	6.975 sec

그림 10은 학습이 진행됨에 따라 순환부의 가중치 학습 에러율이 변화하는 것을 그래프로 나타낸 것인데, 2회 학습 때부터 급격히 작아져서 10회 학습에서는 0.076의 매우 작은 값을 나타내고 있다.

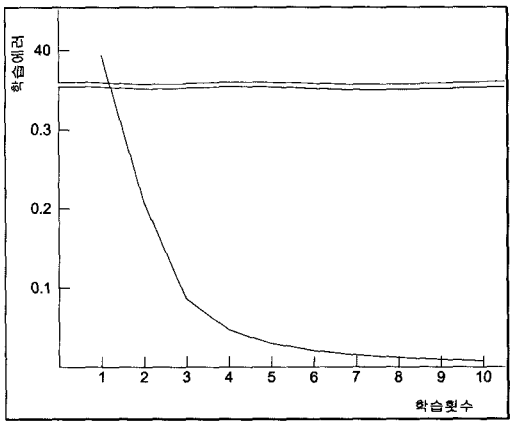

그림 10 순환부 가중치 학습 횟수에 따른 오류 변화

그림 11에서는 조건부에서 학습이 진행함에 따라 조건부의 가중치가 변화하는 것을 그래프로 보여주고 있는데, 약 9회 이후에는 그 값이 거의 변하지 않는 것을 볼 수 있다.

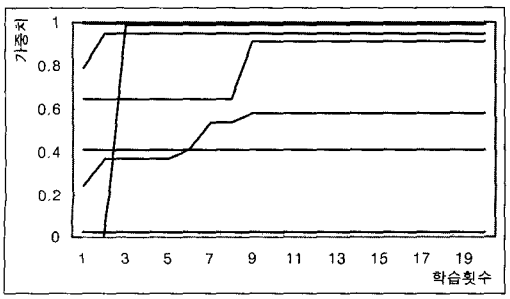

그림 11 조건부 학습에 따른 가중치 변화

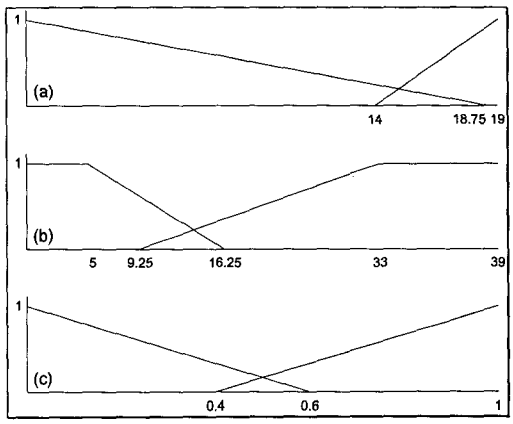

그림 12 생성된 소속 함수 (a) RMS에 대한 소속 함수 (b) 멜주파수 대역 최대 에너지에 대한 소속 함수 (c) 순환부에 대한 소속 함수

그림 12는 히스토그램 분석을 통해 생성된 퍼지소속함수를 나타낸다. (a)는 RMS에 대한 소속 함수이고, (b)는 멜주파수 대역 최대 에너지에 대한 소속 함수이며, (c)는 순환 노드에 대한 소속 함수이다. 가로축은 각 특징벡터의 값이고 세로축은 소속함수를 통한 퍼지값이다.

학습을 통해 구한 가중치와 특징벡터의 히스토그램 분석을 통해 구한 소속 함수를 이용하여 음성의 경계를 추출한 결과를 그림 13에서 보여주고 있다. 그림 13에서 볼 수 있듯이 음성 영역을 모두 포함하여 경계가 추출된 것을 볼 수

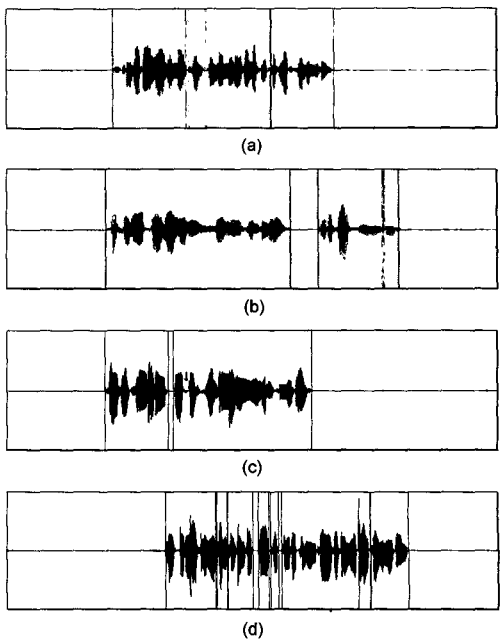

그림 13 경계 추출 결과의 예

있다. 이것은 음성인식의 특성상 비음성을 음성으로 인식하는 오인식률보다, 음성을 비음성으로 인식하는 오거부율이 낮아야 하는 특성을 충분히 반영하는 결과이다. 오인식률은 음성으로 분류된 결과중 실제로 비음성인 비율을 의미하고 $\frac{\text{음성으로 인식된 샘플중 비음성 샘플의 수}}{\text{음성으로 인식된 샘플의 수}}$ 로 구할 수 있다. 오거부율은 비음성으로 분류된 결과중 실제로는 음성인 비율을 의미하고 $\frac{\text{비음성으로 인식된 샘플중 음성 샘플의 수}}{\text{비음성으로 인식된 샘플의 수}}$ 로 구할 수 있다. 또한 그림 13(c)의 ↓부분은 자음 부분으로서 시간 영역의 정보인 RMS는 작은 값을 갖지만 멜주파수 대역 최대 에너지는 큰 값을 가지므로 비음성이 아니라 음성으로 추출된다. 만일 그러한 부분이 비음성으로 결정되어 음성추출 범위에서 제외된다면 자음이 없는 모음만을 갖는 음성이 될 것이다.

표 2 음성경계 추출 결과

샘플	오인식율	오거부율	정확도
샘플 1	1.21%	1.1%	97.69%
샘플 2	0.18%	1.28%	98.54%
샘플 3	0.45%	0.78%	98.77%
샘플 4	1.81%	0.52%	97.67%
샘플 5	0.57%	0.92%	98.51%
평균	0.844	0.92	98.236%

표 2는 음성의 경계를 추출한 실험 결과를 표로 보여주고 있는데 약 98%의 정확도를 갖는다.

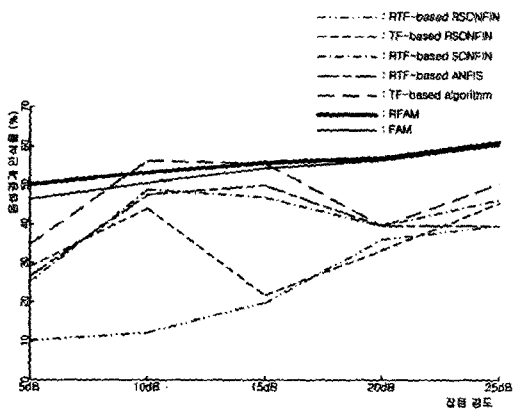

그림 14 음성경계 인식율

그림 14는 기존의 RSONFIN 방식과 비교한 결과를 보여주고 있다. 가로축은 추가된 신호 대 잡음비를 나타내고 세로축은 음성경계 인식율을 나타낸다. 음성 경계의 인식율을 계산하는 방법은 $\frac{\text{정확히 인식된 신호의 프레임 수}}{\text{전체 신호의 프레임 수}}$ 과 같다. 굵은 실선으로 표시된 부분이 본 논문에서 제안한 방법의 결과인데, 다른 방법과 비교해 신호대 잡음비가 작아짐에 따라 잡음의 강도가 강해져서 음성경계 인식율이 작아지지만 다른 방법들에 비해 안정적인 것을 볼 수 있다. 그림 14에서 비교되는 기존의 방법들[12]을 설명하면 앞의 약어(RTF)는 음성경계 추출을 위해 사용한 특징들을 의미하는데 T는 Time 의 약자로서 시간 영역에서 추출한 특징 정보를 의미하며 RMS와 ZCR을 나타낸다. 또한 F는 Frequency의 약자로서 주파수 공간에서 추출한 특징 정보를 의미하며 특정 시간에서의 주파수 최대 대역 에너지를 나타낸다. R은 Refine의 약자로서 각 시간영역과 주파수 영역의 정보를 추출함에 있어서 처음 20ms의 구간을 비음성 구간으로 규정하고 이때 얻어지는 잡음의 평균이 전체 시간대에 걸쳐서 분포한다고 가정하여 특징을 얻어내는 방법이다. 또한 그림 14에서 비교하는 기존 방법들 중에서 뒤에 붙는 수식어들을 설명하면 NFIS은 신경망과 퍼지 시스템이 결합된 형태를 의미하고 SO(Self-Organized)는 학습에 의하여 자동 생성되는 신경-퍼지망의 생성 특성을 의미하며, R(Recurrent)은 신경-퍼지망의 구성에 있어서 순환 노드를 포함한다는 의미이다. 그림 14에서 RFAM과 FAM의 실험 결과에 대하여 보면 신호대 잡음비가 작아질수록 RFAM 방식이 더 효과적인 것을 알 수 있다.

5. 결 론

본 논문에서는 음성인식을 위한 전처리 단계인 음성경계 추출을 위해 잡음에 강한 정규화된 특징벡터와 연속적인 성질을 이용한 순환 퍼지연상기억장치를 사용하였다. 기존의 RMS와 주파수 대역 에너지는 랜덤잡음과 백색잡음에 민감한 특성을 가지고 있으므로 학습 자료와 실험 자료의 환경이 달라질 경우에 잘못된 결과를 가져올 수 있다. 이러한 단점을 보완하기 위해 잡음의 영향을 적게 받는 정규화된 RMS와 정규화된 멜주파수 대역 최대 에너지를 사용하였다. 또한 음성의 연속성을 이용하여 랜덤잡음에 강인한 특성을 갖는 순환 퍼지연상기억장치를 이용하였다. 순환 노드는 퍼지 규칙이 갖는 시간적 자료 표현의 한계성을 해결하였다. 또한 퍼지연상기억장치는 시스템의 구성이 쉽고 직관적이어서 일반적인 퍼지 시스템이 갖는 구성상의 어려움을 해결할 수 있다.

참 고 문 헌

[1] Fabien Gouyon, Francois Pachet, Olivier Delerue, "On The Use of Zero-Crossing Rate for an

Application of Classification of Percussive Sounds," Conference on Digital Audio Effects, pp. 1-6, 2000.
[2] Ramana Rao G.V., Srichand J., "Word Boundary Detection Using Pitch Variations," Fourth International Conference on Spoken Language Processing, pp. 813-816, 1996.
[3] Gin-Der Wu, Chin-Teng Lin, "Word Boundary Detection with Mel-Scale Frequency Bank in Noisy Environment," IEEE Speech and Audio Processing, Vol. 8, No. 5, pp. 541-554, 2000.
[4] Sirko Molau, Michael Pitz, Ralf Schliiter, Hermann Ney, "Computing Mel-Frequency Cepstral Coefficients on The Spectrum," IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 73-76, 2001.
[5] Alain Biem, Shigeru Katagiri, Biing-Hwang Juang, "An Application of Discriminative Feature Extraction of Filter-Bank-Based Speech Recognition," IEEE Transaction on Speech and Audio Processing, pp. 96-110, 2001.
[6] Mark Marzinzik, Birger Kollmeier, "Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics," IEEE Speech and Audio Processing, pp. 109-118, 2002.
[7] 석종원, 배건성, "웨이블렛 변환을 이용한 음성신호의 끝점 검출", 한국음향학회지, 18권, 6호, pp. 57-64, 1999.
[8] D. O. Hebb, "The Organization of Behavior," John Wiley & Sons, New York, 1949.
[9] F. Beritelli, "Robust word boundary detection using fuzzy logic," Electronics Letters, Vol. 36, No. 9, pp. 846-848, 2000.
[10] Tong Zhao, Peng-Yung Woo, "Fuzzy Speech Recognition," International Joint Conference on Neural Networks, pp. 2959-2961, 1999.
[11] Vittorio Gorrini, Hugues Bersini, "Recurrent Fuzzy Systems," IEEE World Congress on Computational Intelligence, pp. 193-198, 1994.
[12] Gin-Der Wu, Chin-Teng Lin, "A Recurrent Neural Fuzzy Network for Word Boundary Detection in Variable Noise-Level Environments," IEEE Systems, Man and Cybernetics, Vol. 31, No. 1, pp. 84-97, 2001.
[13] Doroteo Torre Toledano, "Neural Network Boundary Refining for Automatic Speech Segmentation," IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 3438-3441, 2000.
[14] 배명진, 이상효, "디지털 음성분석", 동영출판사, 1998.
[15] 장대식, "퍼지연상기억장치에 기반한 퍼지 추론 시스템", 숭실대학교 석사청구논문, 1995.
[16] Martin T. Hagan, Howard B. Demuth, "Neural Network Design," PWS Publishing Company, 1995.

마 창 수
1999년 2월 수원대학교 정보통신공학과(공학사). 2003년 8월 숭실대학교 대학원 컴퓨터학과 졸업(공학석사). 2003년~현재 (주)핸디소프트 근무. 관심분야는 컴퓨터비전, 음성처리, 영상처리, 패턴인식 등

김 계 영
정보과학회논문지 : 소프트웨어 및 응용
제 31 권 제 8 호 참조