순환 퍼지연상기억장치를 이용한 음성경계 추출

마창수○ 김계영 최형일
숭실대학교
magyver@nate.com○ {gykim, hic}@computing.ssu.ac.kr

Word Boundary Detection of Voice Signal Using Recurrent Fuzzy Associative Memory

Chang-Su Ma○ Gye-Yong Kim, Hyung-Il Choi
Department of Computing, Soongsil University

요 약

본 논문에서는 음성인식을 위한 전처리 단계로 음성인식의 대상을 찾아내는 음성경계 추출에 대하여 기술한다. 음성경계 추출을 위한 특징 벡터로는 시간 정보인 RMS와 주파수 정보인 MFBE를 사용한다. 사용하는 알고리즘은 학습을 통해 규칙을 생성하는 퍼지연상기억장치에 음성의 시간 정보를 적용하기 위해 순환노드를 추가한 새로운 형태의 순환 퍼지연상기억장치를 제안한다.

1. 서론

음성인식을 위한 전처리 단계로는 입력 신호의 잡음을 제거하는 과정, 음성 강화 과정, 그리고 음성과 비음성 부분을 나누어 실제 음성인식에 적용할 대상을 찾아내는 음성경계 추출 과정이 있다. 이중에서 잡음 환경에도 강인한 음성경계추출과정은 음성인식의 성능 향상을 위해 중요한 부분이다.

음성경계추출은 음성인식 뿐만 아니라 통신망에도 사용할 수 있는데, 전화와 같은 일반적인 대화의 경우 실제 말을 하는 voice 부분이 silent 부분에 비해 더 작은 비율을 가지게 되는 것이 보통이다. 그러므로 전체 신호에서 음성경계를 추출하여 추출된 음성 부분만을 전송한다면 별도의 알고리즘을 사용하지 않더라도 통신 채널을 절약하는 효과를 얻을 수 있다.

음성처리를 위해 사용되는 특징 벡터들을 보면 ZCR(Zero Crossing Rate), LPC(Linear Prediction Code), RMS(Root Mean Square) Energy, LCR(Level Crossing Rate), PVR(Peak Valley Rate), Pitch Variations[1] 등의 시간 정보와 MFCC (Mel-scaled Frequency Cepstral Coefficient), Filter-Bank, Envelope Value, RTF (Refined Time Frequency) parameter[2], Wavelet Transform Coefficient[3]등의 주파수 정보들이 있다.

보통 시간 정보들이 잡음에 약하다는 특성이 있는데 정규화 RMS는 신호 전반에 걸쳐 추가된 잡음이나 에너지 레벨을 보정하는 효과를 얻을 수 있다. 또한 주파수 정보를 위해 논문에서는 mel-scale에 의해 인간 청각 특성을 추가한 mel-scaled frequency에서 band에너지를 얻고 그 중 최대값을 취하는 MFBE(Mel-scaled Frequency Band Energy)를 사용하여 랜덤 잡음에 강한 특징을 추출해낸다.

음경 경계 추출을 위한 알고리즘으로는 기존의 패턴인식에서 사용하는 대부분의 방법들이 사용될 수 있는데 HMM, Neural Network, Fuzzy, Neural Fuzzy, DTW, Threshold Model 등의 방법 등을 들 수 있다. Gin-Der Wu와 Chin-Teng Lin이 제안한 방법[4]에서는 특징 벡터로 Noise time, ATF, ZCR을 사용하였고 신경망과 퍼지를 결합한 self-organized Neural Fuzzy Network을 제안하였다. 각 노드는 퍼지 규칙으로 구성되며 시스템 전체로는 신경망의 구조를 따르는 5단의 퍼지-신경망의 형태를 갖는다. 또 다른 방법은 퍼지의 로직을 이용하는 방법[5]이 있다. 퍼지의 패턴 매칭을 위해서 퍼지 규칙이 생성되어야 하고 각 규칙에 맞는 소속 함수가 구성되어야 한다. 그러나 규칙을 효과적으로 생성하는 것이 쉽지 않고 몇 개의 노드를 사용할 것인지 정하기 위해 전문가적 지식을 필요로 하게 된다. 또한 음성은 시간적 정보를 갖는데 퍼지 로직은 이를 잘 표현하지 못한다.

본 논문에서는 퍼지 로직의 단점인 규칙생성의 비효율성을 보완하기 위하여 학습을 통해 규칙을 자동 생성하는 퍼지 연상 장치를 음성경계 추출에 적용하였다. 또한 시간 정보를 추가하기 위해 최근 신경망에서 많이 사용하는 순환 노드를 추가 하여 시간 정보를 on-line시에 적용하는 순환 퍼지연상기억장치 (RFAM : Recurrent Fuzzy Associative Memory)을 제안한다.

2. 특징 벡터 RMS와 MFBE

본 논문에서 음성경계 추출을 위해 사용하는 특징은 RMS와 MFBE이다.

RMS(Root Mean Square)는 주어진 구간 내에서 시간 에너지의 변화량을 측정한 것인데, 부호에 상관 없는 값을 얻기 위해 제곱을 해서 누적한 다음 구간 크기로 나누어 표준화 하고 로그함수를 씌워 제곱으로 인해 커지는 값을 완충시킨다. 신호의 값이 0값에 가까울수록 RMS값은 작아지고 멀수록

RMS값이 커진다. 이를 수식으로 나타내면

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

이다. 여기에서 L은 프레임의 크기이고 m은 프레임 인덱스, n은 프레임 내에서 신호의 인덱스이다. 구해진 RMS값을 스무딩하기 위해 수식(2)를 적용하고

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

Noise_Time을 제거하여 정규화 한다.

$$R(m) = \hat{x}_{rms} = Noise\_time_{rms} \quad (3)$$

여기에서 Noise_time은 초기 5 프레임의 스무딩 RMS값의 평균이다.

$$Noise\_time_{rms} = \frac{\sum_{m=0}^{4} \hat{x}_{rms}(m)}{5} \quad (4)$$

위의 식들을 이용하여 구한 정규화 RMS의 음성과 비음성 히스토그램을 그림.1에서 보여주고 있다. 그림에서 점선은 비음성 부분을 나타내고 실선은 음성 부분을 나타내는데 비음성은 대부분 0에 가까운 값들이므로 히스토그램도 0에 가까운 부분에 분포하는 것을 볼 수 있고 음성 부분은 11값을 기준으로 산모양을 나타내는 것을 볼 수 있다.

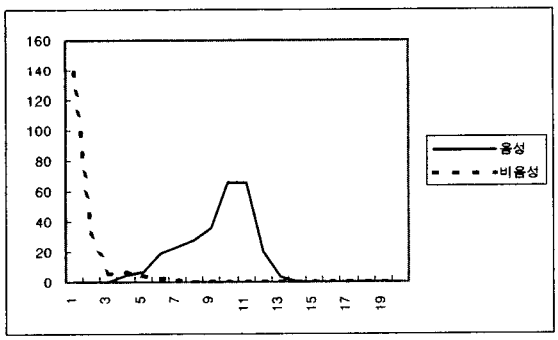

그림1. RMS 히스토그램

특징 벡터가 높은 구분도를 나타내어야 하는 이유는 퍼지 시스템에 적용하기 위해서는 각 규칙에 해당하는 퍼지 소속함수로 나타내기가 용이하기 때문이다.

여기서 구해진 RMS가 음성경계 추출을 위한 첫 번째 특징 벡터로 사용된다.

두번째 특징으로 MFBE(Mel-scaled Frequency Band Energy)를 사용하는데 입력 신호를 FFT변환을 통해 주파수 공간으로 변환 시키면 백색 잡음 등의 추가에도 변하지 않는 포만트 정보를 얻어낼 수 있다. 특히 청각 특성에 입각한 정보를 얻어내기 위해 주파수를 Mel-scale을 이용해 비선형 변환하는데 식으로 나타내면

$$x_{freq}(m,i) = \sum_{k=0}^{N-1} \left| x_{freq}(m,k) \right| f(i,k) \quad (5)$$

이다. 여기에서 i는 filter bank 인덱스이고, k는 스펙트럼 인덱스이다. $f(i,k)$는 mel-scale filter bank의 weight function을 나타낸다. 이렇게 해서 각 frame마다 20개의 누적 밴드 에너지가 구해졌는데 RMS와 마찬가지로 스무딩과 정규화 과정을 거치게 된다. 식은 아래와 같다.

$$\hat{x}_{freq}(m,i) = \frac{x_{freq}(m-1,i) + x_{freq}(m,i) + x_{freq}(m+1,i)}{3} \quad (6)$$

$$F(m,i) = \hat{x}_{freq} - Noise\_freq \quad (7)$$

$$Noise\_freq = \frac{\sum_{m=0}^{4} \hat{x}_{freq}(m,i)}{5} \quad (8)$$

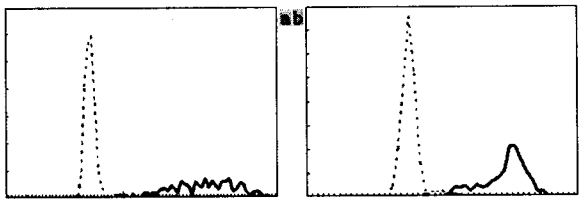

그림2. a. mel-scale 에너지 히스토그램 b. mfbe 히스토그램

그림.2에서 정규화 mel-scaled frequency 누적 에너지 히스토그램을 보여주고 있다. 그림2.a에서 보듯이 음성과 비음성의 구분도가 상당히 높게 나타나고 있지만 음성 부분 히스토그램이 넓게 퍼져 분산이 큰 것을 알 수 있다. 음성경계 추출에 있어서 비음성 부분을 음성으로 인식하는 FAR(False Acceptance Rate)은 허용하지만 음성을 비음성으로 인식하는 FRR(False Rejection Rate)은 낮아야 한다. 또한 퍼지 함수를 구성하기에도 부적합하다. 이를 개선하기 위해 frame내의 각 밴드 에너지 중 가장 높은 값을 선택하는 mel-scaled maximum band energy(MFBE:Mel-scaled Frequency Band Energy)를 사용 한다. 이를 구하는 수식은 다음과 같다.

$$M(m) = \max[F(m,i)]_{i=1,2,\ldots,20} \quad (9)$$

이를 이용한 결과가 그림2.b이다. 그림2.a 보다 분산이 작고 집중성 있는 향상된 결과를 나타낸다.

3. RFAM (Recurrent Fuzzy Associative Memory)

자동 생성 신경망과 같이 퍼지규칙을 자동생성 하기 위해 본 논문에서는 순환 퍼지연상기억장치(RFAM:Recurrent Fuzzy Associative Memory)를 사용하도록 한다. 퍼지연상기억장치의 각 입력 벡터는 히스토그램 분석을 통해 퍼지 소속 함수로 구성 되고 모든 소속함수는 조건부와 결합되는 연상 장치를 구성한다. 구성된 연상 장치는 학습을 통해 가중치가 조절되고 각 가중치는 조건에 맞는 규칙을 생성하는 역할을 한다. 또한 음성경계 데이터의 시간적 연속성을 반영하기 위해 순환 노드를 추가한다.

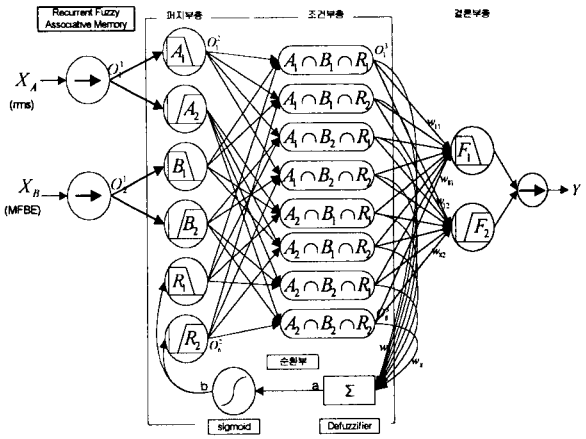

그림3. RFAM 구성도

본 논문에서 제안하는 순환 퍼지연상기억장치는 그림.3과 같이 입력부층, 퍼지화층, 조건부층, 결론부층, 순환부층으로 구성된다.

입력부층은 전송층이라고도 하는데 입력 데이터를 별도의 처리 없이 퍼지화층으로 전송하는 역할을 한다. 입력은 모두 두 개인데 하나는 시간 영역의 특징인 RMS이고 다른 하나는 주파수 영역의 특징인 MFBE이다. 각 입력은 해당하는 퍼지화층의 노드로 분할되어 입력으로 들어가게 되는데 노드의 개수는 특징벡터의 히스토그램 분석을 통해 구해진다.

퍼지화층은 총 6개의 퍼지 소속 함수로 구성되어 있는데 서로 다른 크기의 입력된 특징 값들을 [0,1] 범위의 퍼지값으로 변환하는 역할을 한다. A는 RMS에 대한 소속 함수로 구성된 노드이며 소속함수를 두개로 나눈 이유는 그림.1에서 보는 것과 같이 히스토그램이 두개로 크게 나뉘어 분포되어 있기 때문이다. B는 MFBE에 대한 소속 함수로 구성된 노드이며 두개의 노드인데 그림2.b에서 확인할 수 있다. 또한 R은 순환부층에서 온 입력의 소속함수 노드로 두개이다. 이유는 시스템의 최종 결과가 음성과 비음성의 두개이기 때문이다. 이렇게 해서 총 노드의 개수는 2(A) + 2(B) + 2(R) 로 6개가 된다.

조건부층은 퍼지화층에서 생성된 각 소속 함수값과 조건부와 결론부를 잇는 가중치의 퍼지곱에 의해 생성된다. 조건부층의 개수는 2(A) × 2(B) × 2(R)로 8개이다. 8개 노드의 출력은 결론부층에 완전연결 되고 또한 순환부층으로 입력된다.

결론부층은 조건부 층과 함께 퍼지 연산을 하게 되고 결과를 출력하게 된다. 출력 Y의 값에 따라 최종 결론이 구해지게 된다. 조건부층과 가중치 퍼지곱의 최대값이 결론부의 입력이 되는데 두 출력부의 누적 면적의 무게중심이 최종 결론이 된다.

순환부층은 조건부층의 각 노드 출력과 가중치를 곱해서 얻어지는데 신경망에서 net를 구하는 것과 같다. 조건부 층의 출력은 가중치를 제외한 3개 조건의 퍼지 곱에 의해서 얻어지고 얻어진 출력이 해당 가중치와 산술 곱의 누적을 계산해서 역퍼지화 된다. 수식으로 나타내면

$$a = \sum_{i=0}^{7} O_i^3 w_i \qquad (10)$$

이고 얻어진 a 값이 시그모이드 함수를 거쳐 b값이 얻어진다.

4. 실험결과 및 결론

실험을 위한 데이터는 8kHz로 샘플링하고 프레임 크기는 16ms로 128샘플이다. 프레임 중첩은 8ms로 하였고 샘플마다 16비트이다. 초기 프레임 크기는 5프레임으로 384샘플을 사용하였고 데이터베이스는 KAIST에서 제공한 로우 포맷 나레이션 데이터와 자체 녹음한 숫자음을 사용하였다.

본 논문은 음성경계 추출을 위해 순환 노드를 추가한 퍼지 순환 연상 장치를 제안하였다. FAM의 특성으로 인해 규칙은 학습을 통해 자동 생성되므로 시스템 구성을 위해 학습 데이터 이외의 전문가적 지식은 필요하지 않았고 학습을 통해 가중치를 생성하였다. 또한 순환 노드가 연속성을 증가시켜 주므로 불규칙하게 발생하는 잡음을 제거하는 효과를 거두게 된다. 특징 벡터를 선택하기 위해 ZCR, RMS, Frequency energy, MFBE등을 실험하였고 이중 RMS와 MFBE를 선택하였다. 선택된 특징 벡터에 스무딩과 정규화를 통해 랜덤 잡음과 백색 잡음을 제거하는 효과를 거두었다.

향후에는 퍼지 시스템에서 연결도가 낮은 불필요한 노드를 제거하여 속도 성능이 향상된 시스템을 구성하려고 한다. 또한 제안한 RFAM 시스템을 음성인식 등의 좀더 복잡한 시스템에 적용하기 위해 연상 장치를 동적으로 구성하는 연구가 요구된다.

Acknowledgement

본 논문은 첨단정보기술연구센터를 통하여 과학재단의 일부 지원을 받았음.

참고문헌

[1] Ramana Rao G.V. and Srichand J., "Word boundary detection using pitch variations", ICSLP'96, pp813-816, 1996

[2] Alain Biem and Shigeru Katagiri, "An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition", IEEE, SAP. Vol 9., No. 2., pp96-110, 2001

[3] 석종원 and 배건성, "웨이블렛 변환을 이용한 음성신호의 끝점 검출", 한국음향학회지, 18권, 6호, pp. 57-64, 1999

[4] Gin-Der Wu and Chin-Teng Lin, "Word Boundary Detection with Mel-Scale Frequency Bank in Noisy Environment", IEEE, Vol. 8., No. 5., pp. 541-554, 2000

[5] Gin-Der Wu and Chin-Teng Lin, "A Recurrent Neural Fuzzy Network for Word Boundary Detection in Variable Noise-Level Environments", IEEE Vol. 31., No. 1., pp. 84-97, 2001