신경망 기반의 동적 파라미터들을 이용한 음성 경계 추출

마창수[0] 김계영 최형일
숭실대학교
csma@vision.ssu.ac.kr {gykim, hic}@computing.ssul.ac.kr

A Voice Boundary Detection Method Using Dynamic Parameters Based On Neural Network

Chang-Su Ma[0] Gye-Yong Kim Hyung-Il Choi
Department of Computing, SoongSil University

요 약

본 논문에서는 음성인식 성능을 높이기 위한 기본적 단계인 음성과 비음성 부분의 경계를 추출하는 음성 경계 추출 방법을 제안한다. 음성경계 추출을 위한 특징들로는 시간영역 분할 파라미터인 ZCR, MA를 사용하고 주파수 영역 분할 파라미터로 주파수 대역 파워 에너지 (Frequency band power energy), 포만트 계수 (Formant coefficient)를 사용하였고 각 파라미터들을 이용하여 음성 경계를 결정할 때 경험에 의해 임계치를 결정하는 단점을 보안하기 위해서 신경망을 이용한다. 신경망의 가중치와 임계치들은 지도학습을 통해 최적화 되고, 학습을 통해 구성된 망을 음성과 비음성의 경계치 구분에 사용한다.

1. 서 론

음성 신호를 음성 인식, 인증, 합성 등의 분야에 이용할 때 입력된 신호에서 음성과 비음성을 구분하여 그 경계를 정확하게 추출하는 것은 매우 중요하다. 음성 분석분야 이외에 전화와 같은 통신의 경우에도 비음성 부분을 찾아내어 불필요한 데이터를 제외시키고 전송하는 것은 데이터 양을 줄여 전송시 네트웍의 효율을 향상을 시킬 수 있다. 특히 고립단어 인식의 경우 단어를 정확히 분리해 내는 것은 최종 인식 결과에도 지대한 영향을 미치게 된다.

음성 인식 기법으로 사용하는 HMM(Hidden Markov Model), DTW(Dynamic Time Warping), NN(Neural Network) 등의 구현을 위해서는 단어, 혹은 음절을 구분해내는 분절(segmentation) 과정을 기본으로 하고 있다. 그림 1에서 보이는 바와 같이 입력된 음성 데이터는 경계추출 알고리즘에 따라 음성 부분과 비음성 부분의 경계 추출(Boundary Detection)을 하게 된다. 기존에 사용되고 있는 경계추출 방법으로는 에너지 레벨과 주기를 이용한 EPD-ATA(End Point Detection - Automatic Threshold Adjustment), 피치 정보와 에너지 변화를 이용한 EPD-PCH(Use of Pitch Information), rms 신호 에너지, ZCR, 주기 정보 등을 이용하는 EPD-NAA(Noise Adaptive Algorithm), 에너지, ZCR, 결정 규칙(decision rule), 임계치 결정(*threshold setting*)을 이용한 EPD-VAA(A *Voice* Activation Algorithm)[1]등이 있다. 위의 경계 추출 방법들에서 사용되는 특징들을 보면 에너지 레벨, ZCR(Zero Crossing Rate)[2], LCR(Level Crossing Rate), PVR(Peak Valley Rate), LPC(Linear Prediction Coding), LPER(Linear Prediction Error Rate), MFCC(Mel-Frequency Cepstral Coefficient), Pitch information, Filter-Bank[3], Envelope Values, RTF(Recurrent Time Frequency Parameter)[4], Wavelet Transform coefficient[5] 등이 있다. 이 중에서 ZCR, LCR, PVR 등은 초기 연구단계부터 사용되던 특징들로 시간 축 위에서 얻어지는 데이터들이며 잡음 등에 민감한 특성을 가지고 있다. 본 논문에서도 ZCR을 사용하지만 잡음에 민감한 단점을 감소시키기 위해 이동평균(Moving Average)를 사용하여 시간축 상에서의 음성 데이터의 추이 정보를 사용한다. 또한 Envelope을 이용한 음성 경계 추출 방법[6]의 경우 특징들의 임계값을 경험적으로 결정하고 조건 구성을 사람이 직접 해야한다는 단점이 있다.

본 논문에서 사용하는 또 다른 특징은 포만트 계수이다. 비음성 부분에서 포만트가 형성되지 않는 특징을 이용한 것이다.

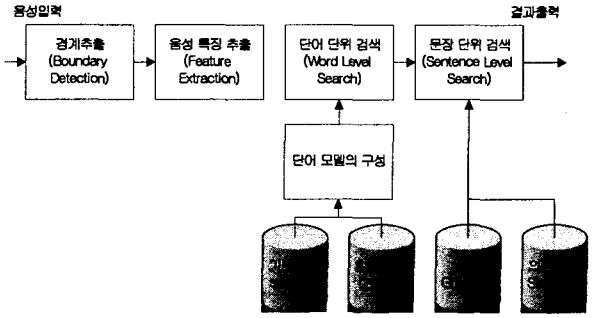

그림 1. 음성인식 시스템에서의 경계 추출

본 논문에서는 특징 파라미터들을 사용하는데 있어서 임계값을 경험적으로 정하는 불편함과 부정확성을 보완하기 위해 신경망을 이용한다. 신경망의 지도학습을 통해 각 노드에서의 임계값과 연결 가중치를 자동적으로 학습하여 오류를 최소화한다.

본 논문의 구성은 2장에서 사용하게될 특징들에 대해 살펴보고 3장에서는 2장에서 설명한 특징들을 이용하여 경계추출을 하게될 신경망에 대해 살펴본다. 4장에서는 제안한 방법의 실험 및 고찰을 하고 5장에서 본 논문의 결론을 맺는다.

2. 경계 추출 파라미터

2.1 시간적 계수

■ 영 교차율 (ZCR : Zero Crossing Rate)

영 교차율은 단어 경계 추출 방법에서 쓰이는 가장 오래된 파라미터로서 정해진 시간 프레임 안에서 시간 축 상의 신호 파형이 x축을 몇 번 교차하는지에 대한 비율을 가지고 결정하는 것이다. 비 음성의 경우 0점 부근에 대한 가능성이 높기 때문에 이를 이용한 것이다.

■ 이동 평균 (MA : Moving Average)

ZCR이 랜덤 잡음에 약하다는 단점을 보완하기 위해 이동평균을 이용하는데 이동평균을 이용할 경우 신호의 에너지 이동 추이를 볼 수 있고 에너지가 0에 가까이 있을 경우 잡음이라고 생각할 수 있으므로 비음성 부분을 찾아낼 수 있다.

2.2 주파수 대역 파워 에너지

Mark, Birger이 제안한 Envelope을 이용한 음성 경계 추출 방법[2]의 경우 두 가지의 문제점을 가지고 있는데, 첫째는 얻어진 High pass, Low Pass 최대/최소 값, 차이 값의 특징들을 경계 추출에 이용하기 위해 각 경우에 해당하는 조건들을 사람이 일일이 구성하여야 한다는 것이다. 이 경우 성능은 조건문의 구성에 의존적이기 때문에, 효율적이지 못한 구성은 곧바로 오류로 이어진다. 두 번째 문제점은 조건문에서 사용하는 임계값과 상수 값을 사용자가 임의로 결정해 주어야 한다는 것이다. 이때 임계값의 결정에 따라 전체 성능에 영향을 미칠 수가 있다. 또한 최대/최소값 들이 에너지 감소 방향으로 수렴하는 음량의 범위(dynamic)를 감소시키는 envelope의 경사도의 정의도 경험적이기 때문에 문제가 될 수 있다. 따라서 본 논문에서는 고주파, 저주파 영역에서의 envelope 대신 에너지 제곱 누적값을 특징으로 사용한다.

주파수 대역 파워 에너지는 2kHz를 기준으로 저주파 영역과 고주파 영역으로 나누고 각 대역의 에너지 총합의 제곱 누적값으로 정의한다. 각 대역에서의 음성 에너지 총합이 음성과 비음성을 구분하는 정보를 제공하기 때문이다. 이 파라미터들은 식 (1),(2)과 같이 구할 수 있다.

$$E_{LP}(p) = \sum_{l} |X(p, w_k)|^2 \quad (1)$$

$$E_{HP}(p) = \sum_{m} |X(p, w_k)|^2 \quad (2)$$

식(1)에서, l은 2kHz 이하의 주파수 대역을 의미하고 식(2)에서 m은 나머지 고주파 대역을 의미한다. 위에서 얻어진 $E_{LP}(p)$는 시간 프레임 p에서의 저주파 에너지 제곱 누적 값을 의미하고 $E_{HP}(p)$는 고주파에서의 에너지 제곱 누적 값을 의미한다. 이 두 값들이 경계 추출을 위한 신경망의 세 번째와 네 번째 입력 파라미터로 쓰이게 된다.

2.3 포만트 계수

음성은 모음, 자음 혹은 마찰음, 파찰음 등 그 특성에 따라 포만트를 형성하고 형성된 포만트의 주파수 상에서의 위치와 크기를 정보를 가지고 있다. 이와 반대로 비음성 즉, 잡음의 경우 특별한 포만트적 특성을 나타내지 않는다. 이런 성질을 기반으로 다음과 같이 포만트 계수를 구할 수 있다.

$$F(t) = \sum_{k=1}^{n} S(t,k)^2 \quad (3)$$

$$S(t,k) = \begin{cases} M_{t,k} - 100 & if\, M_{t,k}\ is\ Formant \\ 0 & else \end{cases} \quad (4)$$

식(4)에서 $M_{t,k}$는 특정 시간 t에서의 주파수k 가지는 크기(magnitude)인데, 보통 포만트는 주파수 스펙트럼 공간에서 주파수의 값이 100을 넘을 경우를 의미하므로 주파수의 값이 100을 넘을 경우 $S(t,k)$의 값은 포만트의 값을 뺀 나머지 값이 된다. 만일 주파수의 크기가 100보다 작을 경우 $S(t,k)$의 값은 0이 된다. 이렇게 해서 얻어진 $S(t,k)$의 값을 식(4)와 같이 전 주파수 영역 0~4kHz에서의 포만트 잉여값을 제곱하여 누적시키면 시간 t에서 포만트 계수를 얻을 수 있다. 얻어진 포만트 계수는 포만트가 존재할 때만 0보다 큰 값이 되고 포만트 값이 없는 시간의 경우 $F(t)$의 값은 0이 된다. 따라서 $F(t)$의 값이 작을수록 비음성일 가능성이 높다. 이 포만트 계수를 신경망의 다섯 번째 입력 파라미터로 사용한다.

그림 2.b에서는 입력 신호의 스펙트럼을 보여주고 있는데 검게 표시된 부분이 포만트를 나타낸다. 그림 2.a에서 x축의 중간을 지날 때부터 waveform이 음성신호를 나타내기 시작하는데 그림 2.b에서 보면 이때부터 포만트가 형성되기 시작하는 것을 볼 수 있다. 이때 주파수의 높고 낮음은 의미가 없고 다만 주파수에서의 크기만이 의미가 있다.

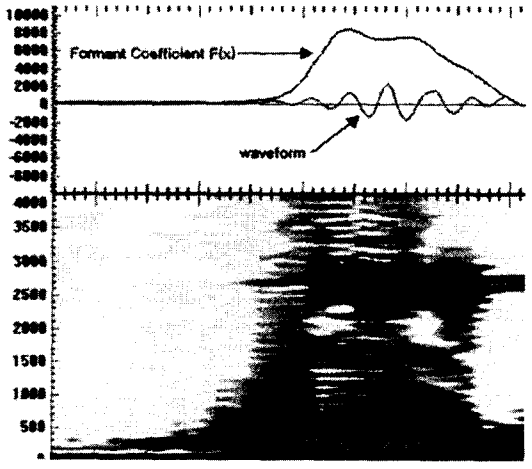

그림 2. a.wave form곡선/포만트 계수 곡선
b.스펙트럼

3. 신경망을 이용한 경계 추출 시스템

실제 경계 추출을 위한 방법으로 신경망을 사용하게 되는데 이때 2장에서 소개한 다섯 개의 특징들을 입력 파라미터로 이용하게 된다. 그림 3에서 제안하는 신경망의 구성을 보여주고 있다. 신경망은 하나의 은닉층을 포함하고 있고 다섯 개의 입력 파라미터를 사용한다. 처음 두 개의 파라미터를 시간적 특징이고 나머지 세 개는 주파수 공간에서의 특징들이다. 입력층

으로 입력된 데이터들은 시그모이드 함수에 의해 입력층의 출력이 결정이 되고 가중치와 곱해져서 다음 은닉층의 입력으로 들어가게 된다. 시그모이드 함수와 은닉층의 입력값은 식 (5),(6)와 같다.

$$f(x) = \frac{2}{1+e^{-x}} - 1 \quad (-\infty < x < \infty) \tag{5}$$

$$net_{pj} = \sum_{i=0}^{N-1} W_{ji} X_{pi} - \theta_j \tag{6}$$

은닉층의 입력은 시그모이드 함수를 통해 은닉층의 출력이 된다.

$$O_{pj} = f_j(net_{pj}) \tag{7}$$

결정 함수로는 비선형 함수인 시그모이드를 사용하고 지도 학습을 통해 각 노드의 가중치와 임계값을 학습하게 된다. 학습이 끝나면 각 연결가중치와 은닉층, 출력층의 각 노드 임계치는 최적화 되고 실제 경계 추출에 사용한다. 그림에서 2,3 계층에서의 임계치 θ는 해당 노드가 활성화 될 것인지 비활성화 될 것인지를 결정하고 활성화되어야 상위 계층의 입력으로 사용되게 된다.

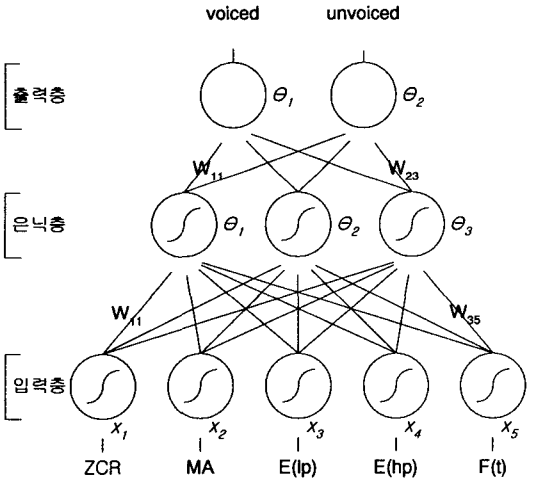

그림 3. 경계 추출을 위한 신경망

4. 실험

본 논문에서는 자체 수집한 20개의 5초~10 이내의 연속음성 문장을 사용하여 실험하였다. 16kHz PCM 형식으로 녹음되었다. 신경망의 지도 학습을 위해서 10개의 문장이 사용되었고 음성의 경계 부분은 사람이 직접 표시를 하였다. 실험을 위해서는 학습에 사용되지 않은 10개의 문장이 사용되었다. 10 포인트 이동평균을 사용하였고 초기 9개의 데이터는 잡음으로 간주하여 0으로 초기화하였다. 주파수 변환을 위해서 DCT를 이용하였다.

	음절수	비음성->음성	음성->비음성	정확도
문장 1	12	1253	1326	96%
문장 2	9	1244	987	96.4%
문장 3	13	1402	1322	96.9%
문장 4	8	887	980	96.6%
문장 5	9	2112	1225	92.3%
문장 6	10	2521	1023	96.2%
문장 7	8	1534	889	95.6%
문장 8	9	1148	1533	95.6%
문장 9	10	908	1017	97.1%
문장 10	16	2103	1730	96.5%
총계	104	15112	12032	95.9%

표 1. 신경망을 이용한 경계 추출 결과

표 1은 실험 결과를 보여주고 있다. 두 번째 줄은 입력 데이터의 음절수를 나타내고 세 번째 줄은 비음성을 음성으로 오인식한 경우를, 네 번째는 음성을 비음성으로 오인식한 결과를 나타낸다. 최종 결과는 95.9%로 높은 정확도를 나타내고 있다.

5. 결론

실험 결과에서와 같이 본 논문에서 제안한 방법은 신경망의 학습을 통한 음성의 경계 추출을 적응적으로 수행하는 것을 알 수 있고, 경험에 의해 임계치를 결정하는 모호성을 제거할 수 있었다.

Acknowledgement

본 논문은 첨단정보기술연구센터를 통하여 과학재단의 일부 지원을 받았음.

6. 참고 문헌

[1] Jean-Claude Junqua, A Robust Algorithm for Word Boundary Detection in the Presence of Noise, IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 3, 1994

[2] Thippur V. Sreenivas, Zero-Crossing Based Spectral Analysis and SVD Spectral Analysis for Formant Frequency Estimation in Noise, IEEE Transactions on Signal Processing, Vol. 40, No. 2, 282-293, 1992

[3] Alain Biem, Shigeru Katagiri, An Application of Discriminative Feature Extraction to Filter-Bank-Based Recognition, IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 2, 96-110, 2001

[4] Gin-Der Wu and Chin-Teng Lin, A Recurrent Neural Fuzzy Network for Word Boundary Detection in Variable Noise-Level Environments, IEEE Transactions on Systems,Man and Cybernetics-Part B : Cybernetics, Vol. 31, No. 1, 84-97, 2001

[5] 석종원, 배건성, 웨이블렛 변환을 이용한 음성신호의 끝점 검출, 한국음향학회지, 18권, 6호, 57-64, 1999

[6] Mark Marzinzik and Birger Kollmeier, Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics, IEEE Transaction on Speech and Audio Processing, Vol. 10, No. 2, 109-118, 2002