음소경계검출과 신경망을 이용한 음소인식 연구

임유두 · 강민구 · 최영호

호남대학교

Phoneme-Boundary-Detection and Phoneme Recognition Research using Neural Network

You-doo Lim · Min-goo Kang · Young-ho Choi

Honam University

E-mail : kangmg@honam.honam.ac.kr

요 약

음성 인식 연구는 유사음소 단위의 인식시스템을 구축하는 방향과 단어 단위의 인식시스템에서의 효율을 최대화하는 방향으로 이루어지고 있다. 이중 유용한 유사음소 단위의 인식시스템 구현을 위해서는 음소의 경계 검출 문제와 검출된 음소에 대한 인식률 향상 문제가 해결되어야 한다.

기존의 LPC(Linear Predictive Coefficient) 방법들은 기준 음소데이터의 LPC와 입력 음성프레임의 LPC 사이의 거리를 Itakura-Saito 방법으로 구하여 음소의 경계를 검출하였으며, 근래에는 MFCC(Mel-Frequency-Cepstrum Coefficient)를 이용하여 스펙트럼의 천이부분을 음소의 경계로 검출하는 방법들이 제안되어왔으나 이러한 방법들은 공통적으로 적응성이 미비하다는 단점이 있다.

본 논문에서는 이러한 단점을 극복하기 위해 음소경계검출을 위해서는 auto-correlation을 이용하고 음소인식을 위해서는 적응성이 뛰어난 다층 Feed-Forward 신경망을 사용하는 새로운 인식시스템을 제안하였다. 제안하는 시스템은 기존의 방법들보다 적응성이 뛰어나고 특징추출부분과 인식 부분의 알고리듬이 독립적이라는 장점을 가지며 프레임단위의 음소인식시스템의 구현 가능성을 확인해 주었다.

ABSTRACT

In the field of speech recognition, the research area can be classified into the following two categories: one which is concerned with the development of phoneme-level recognition system, the other with the efficiency of word-level recognition system. The resonable phoneme-level recognition system should detect the phonemic boundaries appropriately and have the improved recognition abilities all the more. The traditional LPC methods detect the phonemic boundaries using Itakura-Saito method which measures the distance between LPC of the standard phoneme data and that of the target speech frame. The MFCC methods which treat spectral transitions as the phonemic boundaries show the lack of adaptability.

In this paper, we present new speech recognition system which uses auto-correlation method in the phonemic boundary detection process and the multi-layered Feed-Forward neural network in the recognition process respectively. The proposed system outperforms the traditional methods in the sense of adaptability and another advantage of the proposed system is that feature-extraction part is independent of the recognition process.

The results show that frame-unit phonemic recognition system should be possibly implemented.

Ⅰ. 서 론

음성인식의 주제는 10년 이상 연구를 거듭하면서 수 많은 알고리즘이 있다. 초기의 패턴 매칭 방식부터 시작한 여러 시도들은 이제 신호처리와 인공지능의 모든 가능한 방법들을 섭취하고 더 나아가 그들을 혼합하여 최적의 새로운 방법에까지 이르고 있다. HCI (Human Computer Interaction)의 측면에서 음성인식의 성취는 필수불가결하며 이미 Command & Control 수준을

넘어 키보드를 대신하는 Dictation시스템까지 그 활용을 넓혀가고 있다. 이러한 음성인식 기술의 발전을 이루기 위하여 가장 중요한 세가지 분야를 들자면, 전자공학 분야, 인공지능 분야, 실험음성학 분야이다.

음성인식의 주제는 이러한 학술적 분류외에도 다양한 분류가 존재하며, 본 논문에서는 인식단위에 의한 분류에 집중하여 수행하고자 한다.

인식단위에 있어 현재 제시되고 있는 방법은 음소, 유사음소, 음절, 단어가 대표적이다. 음소나 유사음소는 시스템 효율이 최대이나 실제 구현에 있어 난해한 점이 많으며, 음절이나 단어에 의한 인식시스템은 알고리즘적 구현이 용이하나 방대한 데이터 요구 때문에 비효율적 시스템을 유발하게 된다. 현재 음성인식 연구는 유사음소 단위의 인식시스템을 구축하는 방법과, 단어 단위의 인식시스템이 최대의 효율을 내는 방향으로 진행되고 있다. 앞서 언급되었듯이 유사음소를 이용한 인식시스템 구현에는 난해한 부분이 많은데, 특히 음소의 경계를 어떻게 검출할 것이며, 검출된 음소를 어떠한 알고리즘을 사용하여 정확히 인식하는가에 맞추어져 있다.

음소의 특징을 추출하는 방법에는 크게 인간의 성도를 모델링하는 방법, 청각을 모델링하는 방법으로 나뉘는데, 현재 가장 많이 사용되고 있는 방법으로는 청각을 모델링한 MEL-FREQUENCY-CEPSTRUM이 있으며, 최근에는 신호처리 기술의 발달에 힘입어 잡음환경에 강한 RelAtive-SpecTrAl Perceptual-Linear-Prediction 방식이 사용되고 있다.

음소경계검출에 있어 기존의 LPC (Linear Predictive Coding)방법에서는 기준 음소 데이터의 LPC와 입력 음성프레임의 LPC의 거리를 Itakura-Saito방법으로 구하여 경계를 검출하는 방식이 사용되었으며, MFCC의 사용에 들어서면서 스펙트럼의 천이 부분을 음소의 경계로 삼는 방법등이 시도되었다. 그러나 현재 이러한 방법은 적응성이 미비하여, 새로운 음소 경계검출 알고리즘이 필요하며, 특히 이러한 알고리즘의 개발은 특징추출과 인식알고리즘에 독립적이어야 한다.

본 논문에서는 이러한 음소경계검출 알고리즘으로 Autocorrelation을 이용한 방법을 제안하였다. 인식알고리즘으로는 Command & Control방식에서 최대의 효과를 보이는 Dynamic Time Warping이 있으며, 인식시스템에서는 Hidden Markov Model과 Neural Network이 사용되고 있다. 최근에는 이러한 세 가지의 알고리즘의 장점만을 취해 혼합 사용하는 추세를 보이고 있다.

본 논문에서는 적응성이 뛰어난 다층 Feed-Foreward 신경망을 사용하여 프레임단위의 음소인식시스템의 가능성을 확인 하였다.

본 논문의 구성은 다음과 같다. 본론에서는 음성신호의 전처리 및 특징 추출, Short-term Energy를 이용한 음성구간검출, 음소경계검출, 음소인식 시스템에 대하여 설명하며, 결론에서는 실험결과 및 추후 연구 과제로 끝을 맺는다.

II. 본 론

1. 음성신호의 전처리 및 특징 추출

음성에서 얻어지는 신호에는 여러 가지 원인에 의해 여러 주파수 성분이 포함되어 있으나 원치 않은 잡음등이 포함될 수 있다. 이러한 경우 전처리 필터링에 의해 주파수 대역을 한정한 후, 신호처리를 하는 것이 유효하다.

1. Preemphasis

음성신호의 주파수 스펙트럼은 일정하지 않고, 주파수가 높을수록 그 성분이 작아지게 되어 주파수가 2배로 되면 약 6dB의 기울기로 그 파워의 진폭 특성이 작아진다. 그러므로, 음성신호 분석전에 6dB/Oct 기울기를 갖는 고역 강조 필터를 통과시켜, 음성신호의 스펙트럼이 저역부터 고역까지 같은 S/N ratio를 갖게 하는데 이것이 preemphasis이다.[1]

이러한 preemphasis의 전달함수는 다음과 같다.

$$H(z) = 1 - a z^{-1}$$

여기서, a는 1이나 1에 근사한 값을 갖는다. 본 논문에서는 위와 같은 preemphasis를 거친후에 음성인식을 하였다.

2. Hamming Windowing

음성신호의 대부분은 비주기적인 정현파가 되므로 누설 현상이 발생한다. 음성선호의 비주기 문제는 대부분 양 끝에 존재한다. 만일 Fourier Transform을 중앙부분에만 집중할 수 있게 된다면 더 정확한 스펙트럼을 구할 수 있다. 만일 시간 기록을 양 끝은 0이고 가운데가 큰값을 가진 함수와 곱셈을 한다면 시간 기록 중앙부에 집중할 수 있을 것이다. 즉, 시간영역 신호에 Windowing을 함으로써 스펙트럼상에 크나큰 개선을 얻을 수 있게 된다.[1]

본 논문에서는 고역강조된 새로운 음성을 분석에 필요한 프레임별로 분리하게 되는데, 분리할 때 생기는 프레임 가장자리의 갑작스런 변화의 영향을 적게 받기 위해 Hamming Window를 사용하였다.

$$w(n) = 0.54 - 0.46 \cdot \cos\left(\frac{2n\pi}{N-1}\right)$$

$$n = 0, 1, 2, \cdots, N-1$$

3. Mel-Cepstrum

음성인식을 위해 음성신호 전체를 신경망의 입력으로 사용할 수는 없다. 본 논문에서는 음성신호에서 Mel-Cepstrum 계수를 구하여 MFCC 12차만을 신경망의 입력 특징 벡터로 사용하였다.

이 방법은 귀가 음성을 분석하는 방식을 이용하는 청각분석방식이다. 이는 저주파 영역에서는 상세히 분석하고 고주파수 영역에서는 상대적으로 개략적인 분석을 한다.[1] Mel-Cepstrum 계수를 구하는 과정은 입력음성을 이산 코사인 변환하여 주파수 영역내의 weighting 함수를 구하여 Mel-scale로 저장하고 이 Mel 단위로 생성된 filter bank를 거쳐 통과한 신호 중 각각의 필터를 통과한 신호의 크기를 더하여 원하는 갯수 만큼의 계수를 뽑아낸다.

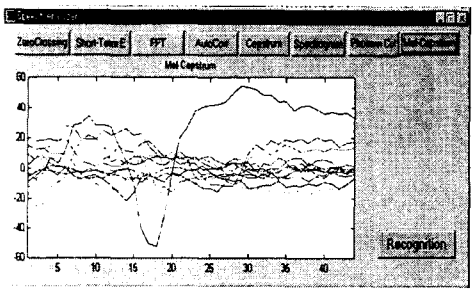

그림 1. 입력음성의 12차 MFCC

II. Short-term Energy를 이용한 음성구간검출

잡음 환경하에서 음성의 시작과 끝의 위치를 찾는 문제는 음성 처리의 중요한 부분이다. 음성신호의 시작과 끝의 위치를 찾게 되면 의미없는 부분을 제거하여 실제적인 음성 부분만 처리하게 된다.

본 논문에서는 time domain에서 측정한 energy와 zero-crossing rate를 사용하는 알고리즘을 이용하였다.

보통 음성의 시작시에는 zero-crossing rate가 뚜렷하게 증가하고 energy도 비교가 된다. 그러나 다음과 같은 경우에는 구별하기가 매우 난해하게 된다.

- 시작과 끝에 약한 마찰음이 있을 경우
- 시작과 끝에 약한 파열음이 있을 경우
- 끝에 비음이 올 경우
- 단어의 끝에 무성음이 오는 유성 마찰음
- 발성의 끝에 모음이 끌리면서 끝나는 경우

위와 같은 어려움에도 zero-crossing rate와 short-term energy를 사용하여 쓸만한 음성검출 알고리즘을 구현할 수 있다.

기본적으로 zero-crossing rate와 energy는 10msec마다 계산한다. 즉 초당 100번씩 계산되는 것이다. 우선 100msec동안은 음성이 포함되어 있지 않다고 가정한다. 그러면 평균 magnitude와 zeor-crossing rate는 통계적인 잡음을 나타내게 된다.

이러한 통계학적 특성과 간격 사이의 최대 평균 magnitude를 사용하여 영교차율과 에너지의 경계값(ITL, IZCT)이 계산되어 진다. 평균 magnitude보다 낮게 적당히 ITU를 잡고, 시작점과 끝점이 그 간격밖에 존재한다고 가정한다. 우선 ITU보다 큰 임의의 한점을 잡고 그 점이 ITL보다 작아지는 첫 번째 점을 시작점이라 추정한다. 마찬가지로 끝점을 검출한다.[1] 이 이중 경계치 처리는 평균 magnitude안에 끝점이 잘못하여 포함되는 것을 막아준다. 다음 단계는 N_1과 N_2를 바깥쪽으로 이동하며 영교차율을 경계값과 비교하는 것이다. 최대 25 frame까지만 진행시킨다. 만약 영교차율이 경계치보다 3번 이상 초과하면, 시작점 N_1은 처음 초과했던 지점이 되고, 그렇지 않으면 초기 N_1을 시작점으로 정의한다. 끝점도 마찬가지이다.

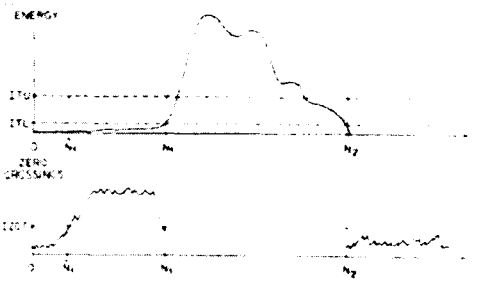

그림 2. 평균 magnitude와 zero-crossing

III. 음소경계검출

1. Autocorrelation을 이용한 Pitch 검출

피치는 유성의 음성신호에서 관찰되는 주기성분에서 봉우리와 봉우리 또는 유사한 모양의 골짜기와 골짜기 간격으로 정의될 수 있으며, 피치주파수는 피치주기의 역수가 되어 보통 기본주파수라고 부른다. [4]

일반적으로 주기적인 신호에서 주기성을 검출하기 위하여는 Autocorrelation function이 사용된다. 이 함수는 어떤 신호의 한 부분과 다른 부분과의 상관관계를 연속적으로 나타내 주는 함수로써 연산결과로 나오는 신호는 신호의 주기적인 부분을 강조해 주어 피치를 특정할 수 있게 해준다. 이 때 측정할 수 있는 피치는 계산에 포함되는 신호의 평균피치값이 된다.

아래의 그림에서 왼쪽에서부터 첫 번째 가장 높은 최대치 봉우리까지의 거리가 피치구간을 나타낸다.

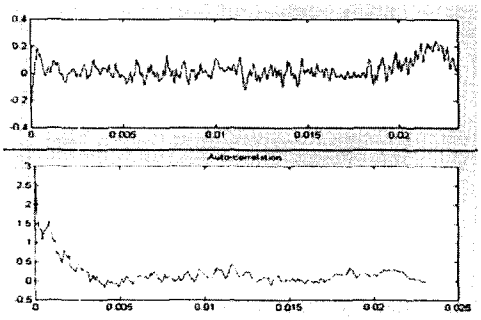

그림 3. Autocorrelation을 이용한 Pitch 검출

2. Cepstrum을 이용한 Pitch 검출

Cepstrum 분석은 DFT의 크기에 로그함수를 취하여 IDFT한 것이다.[1]

speech signal : s_n , excitation : e_n

linear filter : $H(e^{j\theta})$

$S(e^{j\theta}) = H(e^{j\theta})E(e^{j\theta})$

: the frequency domain

Calculated using Matlab

: abs(fft(hamming(512) .* sig))

위 식을 복소수로그함수

$$\log z = \log|z| + i \arg\{z\}$$

정의에 적용시키면

$$\log(S(e^{j\theta})) = \log(H(e^{j\theta})) + \log(E(e^{j\theta}))$$

Calculated using Matlab

: 10 log10(abs(fft(hamming(512) .* sig)))

대부분의 음성처리응용에서는 amplitude spectra만을 요구하므로[1]

$$\log(S(e^{j\theta})) = \log(|H(e^{j\theta})|) + \log(|E(e^{j\theta})|)$$

이와 같이 다시 쓰여진다.

Calculated using Matlab

: ifft(log(abs(fft(hamming(512) .* sig))))

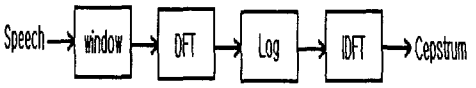

그림 4. Cepstral Analysis

Cepstrum의 결과는 Autocorrelation function의 경우에서와 유사하게 원점에서부터 최초의 최대 봉우리까지의 거리가 피치값이 된다. 이 봉우리의 안쪽 부분은 따로 성도의 스펙트럼 특성을 구하는데 사용된다.

3. 음소경계검출

Correlation이란 말 그대로 신호의 상관도를 나타내는 파라미터이다. 음성신호의 피치를 검출하고자 할 때 이러한 Autocorrelation, 즉 n-1 프레임 신호와 n 프레임의 신호의 상관도를 체크하여 피치를 검출한다.

그때 사용되는 값은 첫 번째 피크가 나타나는 시간이 피치가 되는데 그때의 magnitude를 아래와 같이 플로팅 해보면 과도부분이나 경계부분에서 상관도가 떨어지기 때문에 낮은 값이 나타나는 것을 알 수가 있다.

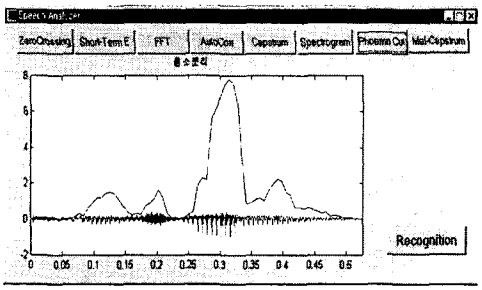

그림 5. 음소분리

한 음소는 여러 주파수가 혼합되어 나타나지만 가장 강한 주파수를 기본주파수라 한다. 이때 잡음이 음성 기본주파수 마스킹 레벨을 넘지 않았다 하면, 최고 magnitude를 가지는 주파수의 변화는 음소의 경계를 검출할 것이다.

그러나, 유성음소의 경우 비슷한 기본주파수를 가지는 것이 많아 난점이 있다.[5] 이러한 문제는 여러 음소분리 알고리즘의 복합 사용으로 해결할 수 있을 것으로 보인다.

IV. 음소인식 시스템

1. 데이터 베이스와 신경망 훈련

인식 실험에 사용된 데이터 베이스는 단일 화자의 음성을 사용하였다. 모두 33단어로 구성되었으며, 샘플링 주파수 16KHz, 16Bit로 녹음하였다.

학습 데이터로 사용된 각 음소의 데이터는 모두 17개로 자음이 11개, 모음이 6개로 구성되어 있다.

표 1. 음소 데이터

자 음	ㄱ	ㄴ	ㄷ	ㅁ	ㅂ	ㅅ
	ㅇ	ㅈ	ㅊ	ㅌ	ㅎ	
모 음	ㅏ	ㅓ	ㅗ	ㅜ	ㅣ	ㅐ

한 음소당 512 프레임씩 5개의 데이터를 가지고 '%d-%d.wav'로 번호를 매긴다음 텍스트 코드화 하여 저장한 후, 이것을 가지고 12차 MFCC를 추출하였고, 이 음소 특징계수들은 신경망의 훈련을 위한 입력 데이터가 된다.

이렇게 구해진 입력 데이터는 Normalizing 기법을 이용하여 정규화 시키는데,[7] 평균을 뺀 뒤, 표준편차로 나누어 그 편차가 1이 되게 정규화 시켰다. 이 값을 신경망에 가하면 역전파 알고리즘에 따라 학습을 하게 되고, 평균 제곱 오차가 최소가 되는 방향으로 입력층과 은닉층, 은닉층과 출력층 사이의 가중치의 값을 지속적으로 수정한다.

2. 신경망을 이용한 음소인식

본 논문에서는 적응성이 뛰어난 다층 Feed-Forward 신경망을 이용하여 대표 음소 데이터를 훈련시킨후 앞서 데이터 베이스로 녹음해 두었던 음성 데이터를 프레임의 분석을 통하여 표 1.의 음소 중 어느 것과 입력 음성이 유사한지를 찾아내는 인식방법을 사용하였다.

이때, 각 음성 프레임은 음성정보의 손실을 줄이기 위해 512 포인트 50%씩 중첩하였으며[2] 특징 추출에 있어서는 Mel-Cepstrum을 이용하여 12차 MFCC를 추출 신경망의 입력 데이터로 사용하였다.[3] 신경망의 출력결과는 그림 6과 같이 나타난다.

그림 6에서 보이는 바와 같이 인식결과는 프레임별 음소인식결과이다. 그러므로 이 결과는 음소조합 알고리즘을 이용하여 필터링함으로써 최종 인식결과를 이끌어 낸다. 즉, 음소인식결과를 첫 번째 필터링을 통해 겹치는 문자를 제거한다. 다음으로 두 번째 필터링을 통해 한번 이상 출력된 음소를 선택한다. 바로, 이것이 최종 인식 결과가 되는 것이다.

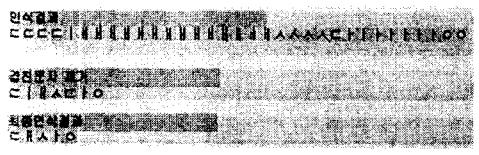

그림 6. '대상.wav'의 인식결과

V. 결 론

본 논문에서는 음소경계검출 알고리즘으로 Autocorrelation을 이용한 방법을 제안하였으며 음소단위 인식시스템의 가능성을 확인하였다.

그러나, 실험에 사용하였던 33단어 중에서 인식결과가 저조한 몇 개 단어의 인식결과를 분석해 본 결과 그림 7, 그림 8과 같이 'ㅂ', 'ㅎ', 'ㅅ' 이 한번만 출력된 경우 음소 조합 알고리즘을 거치면서 2차 필터 통과시 필터링이 되고, 필요치 않은 'ㄴ'는 2번이상 출력되어 1, 2차 필터를 통과하고도 필터링이 되지 않아 올바르지 못한 최종 인식결과가 발생하였다.

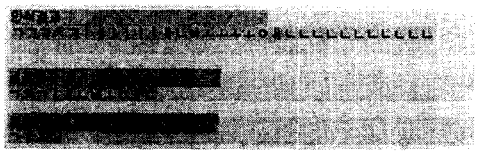

그림 7. '기분.wav'의 인식결과

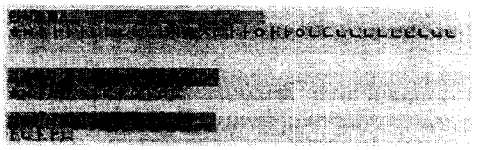

그림 8. '한산.wav'의 인식결과

이처럼 인식결과를 두고 음소 조합 알고리즘을 이용하여 완전한 최종인식결과를 유도할 때 꼭 필요한 음소와 전혀 불필요한 음소를 구별할 수 있는 효과적인 음소 조합 알고리즘의 연구가 요구 된다.

추후 연구과제로는 효율적인 인식 시스템 구현을 위해 RASTA-PLP 적용, Parameter Adaptation, 잡음제거, 영역에 대한 지원과 연구등이 필요하다.

참고문헌

[1] Lawrence R. Rabiner / Ronald W. Schafer, "Digital Processing of Speech Signals", PRENTICE HALL, pp. 116-445, 1981.

[2] 박인정, 이천우, 남상엽, 김형배, "음성-영상 정보의 통합처리에 의한 음성인식", 정보처리학회지, Vol.6, No.4, pp-701-713 1999년 7월.

[3] 이태한, 양태영, 박상택, 이충용, 윤대희, 차일환, "차량 항법용 음성인식 시스템의

구현", 전자공학회논문지, Vol.36, No.9, pp-1105-1114, 1999년 9월.

[4] Rabiner, L. R., "A comparative study of several pitch detection algorithm", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-24, 5, 1976.

[5] Atal, B. S. and Rabiner, L. R., "A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition", IEEE Trans. Acoust., Speech, Signal Processing, ASSP-24, 3, 1976.

[6] J. D. Markel and A. H. Gray, "Linear Prediction of speech", springer verlag, New York, 1976.

[7] Ben Yuhas, "NEURAL NETWORKS IN TELECOMMUNICATIONS", KLUWER ACADEMIC PUBLISHERS, pp. 271-283, 1994.