피치 동기된 에너지 유사도에 의한 음성신호의 전이구간 검출

On A Detecting the Transition Segments of Speech Siganl by Energy Approximation degree of the Synchronized Pitch

김 종득, 박 형빈, 김 대호, 배 명진
숭실대학교 정보통신공학과

Jongdeuk Kim, Hyungbin Park, Daeho Kim, Myungjin Bae
Dept. of Inf. & telecomm. engineering soongsil university
jdkim@assp.soongsil.ac.kr, mjbae@saint.soongsil.ac.kr

Abstract

In a large number of words and the continued speech recognition system using a phoneme as the recognition unit, it is necessary to segment proocessing. In this paper, a normalized AMDF detects the segments in time doamin, we propose a new method. The suggested parameter represents a degree of sharpness at valley point. This method can detect the speech segment between the steady state and transient region to the continued speech without a prior informaion of speech signal.

Ⅰ.서론

대용량의 단어나 연속된 음성의 인식을 위해서는 음성신호를 음성학적 단위인 단어, 음절, 음소등의 단위로 분할하면, 고성능 시스템의 구현이 가능하다. 또한 음성신호의 분석시, 분석의 반복을 줄일 수 있고 음성인식 과정에서 고립단어의 인식기법을 연속 음성 인식에 쉽게 연장시켜 적용할 수 있게 된다. 음성신호의 전이구간(혼합음)은 유성음에서 무성음으로 또는 무성음에서 유성음으로 연결되는 혼합영역이며 유·무성음의 성질이 동시에 나타나게 된다.

연결음이나 연속음에서, 유·무성음들은 시간에 따라 변화하게 되며, 이것은 프레임당 평균진폭의 변화 형태로 나타나게 된다. 이때 평균진폭의 변화도는 음소나 음절의 변화를 근사적으로 대별하게 된다. 음성신호의 전이구간을 검출하는 연구는 특징파라미터를 추출한 영역에 따라 크게 시간영역법, 스펙트럼영역법, 혼성영역법으로 나눌 수 있다. 시간영역법은 시간영역에서 계산의 간편성을 취할 수 있으며, VOT(Voice Onset Time)의 연속성이나 진폭의 증감을 이용하는 방법들이 제안되어 졌다[4-5,9]. 스펙트럼영역법은 음성신호의 음소의 변화에 따른 포만트의 전이특성이나 주파수성분별 에너지를 이용하는 것 등이 제안 되어져 있다[2,4]. 또한 혼합영역법은 변환영역에서의 특징 파라미터들을 이용하는 것[3,6]으로 LPC계수의 전이특성, LPC에러의 변화특성 등을 이용하고 있다. 시간영역법에서 파라미터의 검출은 비교적 쉬우나 그 변화정도를 정확히 파악하기 위한 결정논리가 상대적으로 어렵다. 반면 스펙트럼영역법이나 혼성법은 비교적 정확하지만 계산의 정밀도나 변환차수의 영향을 받게 되고, 전처리과정의 계산량이 시간영역법에 비해 큰 편이다.

따라서 우리는 시간영역법에서 전이구간 검출용 파라미터들 중 에너지 파라미터가 갖는 부정확성과 결정논리의 복잡성에 대해 알아보고 이러한 문제점을 제거할 수 있는 새로운 파라미터를 제안하고자 한다.

Ⅱ.음성신호의 구간 검출

음성신호는 생성모델에 따라 유성음, 무성음, 혼합음, 묵음으로 구분지울 수 있다. 유성음은 폐에서 올라온 공기를 성문을 통하여 배출시킴으로서 생성되므로 성대의 진동을 수반하게 된다. 그리고 성도에서의 공명으로 energy가 크고 준주기적인 파형이 된다. 무성음은 성도에 있는 어떤 점에서 협착을 형성하여 공기의 흐름이 고속으로 협착점을 통과하여 생성 되고, 준색잡음의 낮은 energy를 갖는 파형이 된다. 혼합음은 유성음에서 무성음으로 또는 무성음에서 유성음으로 연결되는 혼합영역이며 유·무성음의 성질이 동시에 나타나게 된다. 연속음에서 구간검출을 사전에 판별하면 음성인식이나 피치검색시에, 정확한 음성인식과 피치검색시에 소요되는 시간, 그리고 정확한 피치검출을 위해서는 각 음절에서 음소의 안정·전이 구간검출은 매우 중요한 요소로 작용하게 된다. 또한 시간영역에서 주기를 찾는 경우에 AMDF, ACM, parallel processing법 등을 이용하여 파형의 주기성을 강조한 뒤 결정논리에 의해 피치를 찾지만, 음소가 천이구간에 걸쳐있는 경우에는 프레임내의 음소변화가 심하고 기본주파수의 주기가 일정치 않으므로 검출에 어려움이 따른다. 이때에 음소의 전이 구간을 미리 검출하여 피치검출구간에서 제외시키면 피치 검출에 따르는 시간 소요를 줄일 수 있으며, 피치구간검출 오류증가를 막을 수 있다.

Ⅲ. 한 프레임내에서의 규준화된 AMDF

음소변화는 음성파형에 비해 서서히 변화하기 때문에 프레임 단위로 분석하는 것이 보통이다. 현재의 프레임이 전이구간이나 정상상태구간에 속하는지를 판정하는 방법은 현재의 프레임을 반분하여 평균진폭비로 판정할 수 있다. 평균진폭비는 다음과 같다.

$$MR(fr)=\frac{\sum_{k=N/2}^{N-1}|s(n-k)|}{\sum_{k=0}^{N/2}|s(n-k)|} \qquad (3\text{-}1)$$

MR(fr)은 평균진폭비이고 s(n)은 음성신호이다. n은 이 프레임이 시작되는 첫 시컨스의 위치이며, N은 프레임의 길이를 나타낸다. 이 평균진폭비는 프레임길이를 1/2로 했을 때의 인근 프레임의 평균진폭비를 나타내기 때문에 윈도우의 영향을 받게 된다. 윈도우의 영향에 무관하게 현재의 프레임이 어떤 상태에 존재하는지를 측정하는 새로운 방법으로는 다음과 같이 규준화된 AMDF를 정의해서 사용할 수 있다. N은 AMDF를 구하려는 윈도우 구간이다.

$$NAMDF(i)=\frac{\sum_{n=1}^{N-1}|s(n)-s(n-i)|}{\sum_{n=1}^{N-1}\{|s(n)|+|s(n-i)|\}} \qquad (3\text{-}2)$$

지연인자를 점차 증가시키면서 이 NAMDF를 구해보면, 지연인자가 프레임내 음성피치에 정수배가 될 때마다 NAMDF는 거의 영이 된다. 그림3-1에서 보면 $Y=X^2$ 그래프와 $Y=|X|$그래프는 Auto-correlation과 AMDF를 취했을 경우 주기에서의 피크 값의 그래프를 나타내고 있다. 0점 위치에 나타난 것처럼 자기상관함수의 정확한 피크 값을 찾는 것이 AMDF의 피크 값을 찾는 것 보다 더 어렵다는 것을 알 수 있다. 따라서 피치검색시에 잘못된 피크값을 얻게 됨으로써 피치를 찾는데 오차를 발생시킬 수 있는 문제를 내포하고 있기 때문에 AMDF가 Auto-correlation 대신에 주기성을 강조하는데 오랫동안 사용되어 왔다.

식 (3-2)의 NAMDF값은 지연인자 i간격을 갖는 N개 샘플간의 진폭에 대한 평균 차이 값이 되기 때문에 음성 파형의 i구간사이의 유사도를 나타내는 규준화된 거리 값으로 적용할 수 있다. 식 3-2의 NAMDF는 i간격의 두 음성파형 블록에 대해, 평균진폭 차이 값을 나타내지만 음성신호가 갖는 피치주기의 변화는 배제하지 않았다. 따라서 두 음성파형 블록에 대한 피치주기의 영향을 제거하려면 시간을 고려하는 지연인자 i의 값이 음성피치에 일치하였을 때의 NAMDF값을 두 파형블럭의 유사도 값으로 사용할 수 있게된다. 여기서 i는 NAMDF를 측정하려는 지연인자이며, i의 값이 음성피치와 일치하였을 때는 규준화된 AMDF값이 최소치를 이룰 때이므로, 각 프레임내에서 i를 변화시키면서 NAMDF의 최소값을 다음과 같이 구할 수 있다.

$$NAMDF(fr)=\underset{NAMDF(101),\ldots,NAMDF(199)]}{MIN[NAMDF(100),} \qquad (3\text{-}3)$$

여기서 MIN[·]함수는 주어진 변수영역에서 최소치를 선택하는 함수이고, fr은 현재 프레임의 위치를 나타낸다. 지연인자를 100샘플(8kHz표본율에서 12.5ms) 부터 구한 이유는 AMDF값은 지연인자가 0 일 때와 음성피치의 정수배일 때 최소치가 되기 때문이다. 따라서

NAMDF지연인자 i를 실존하는 음성피치(2.5ms에서 25ms까지)의 최장길이의 1/2에서 부터 증가시켰을 때 만 두 파형블럭의 유사도를 나타내는 거리 값이 된다. NAMDF값이 영에 근접하면 i간격을 유지하는 두 음성파형 N개 블록간에는 유사성이 최대가 되고 이 점에서 양과음의 경사도간의 간격정보를 측정하여 유사성이 최대 일 때와 비교하여 보면 큰 간격정보를 갖게 된다.

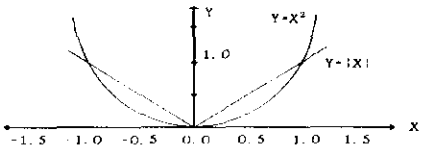

그림. 3-1 AMDF와 AUTOCORRELATION

따라서 각 프레임에서 간격정보가 갑자기 큰 값이 나타나게 되면 전이구간에 놓여있는 상태가 되고 간격정보가 작은 값이 유지되거나 나타나게 되면 안정구간에 놓여 있게 됨을 알 수 있다.

Ⅳ. 규준화된 AMDF법을 이용한 구간상태의 결정

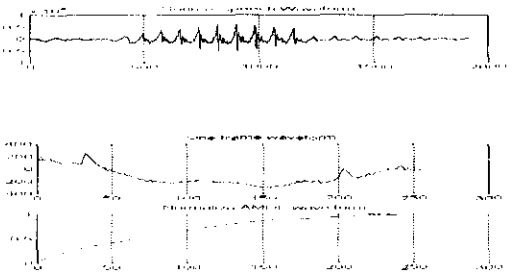

그림. 4(A) /삼/에서의 무성음 구간에 대한 NAMDF

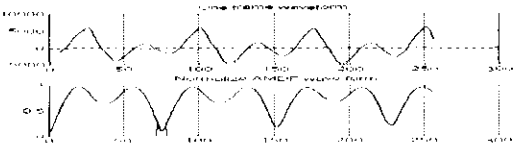

그림. 4(B) /삼/에서의 유성음 구간에 대한 NAMDF

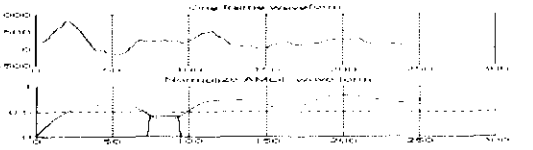

그림. 4(C) /삼/에서의 혼합음 구간에 대한 NAMDF

23세의 남성화자가 발음한 고립단어 /삼/의 무성음, 유성음, 전이구간의 음성신호에 대해 규준화된 AMDF를 구한 것을 그림. A), 그림. B), 그림. C)에 각각 나타내었다. 여기서 그림. 4(A)는 무성음의 파형을 나타내며, 그림. 4(B)는 유성음의 파형을 나타내고, 그림. 4(C)는 전이구간에서의 파형과 NAMDF파형을 나타내었다. 파형의 구조가 단순한 비음 /ㅁ/에 비해 파형의 구조가 복잡한 /아/음의 프레임에서 NAMDF는 상대적으로 높아지는 특징이 있다. 그림. 4(D)에서 NAMDF의 피치 점에서의 간격정보의 변화도를 나타내고 있다. 평균진폭은 각 프레임을 256샘플 단위로 하고 128프레임씩 overlap하여 구한 것이다. 그림. 4(A)를 살펴보면, 일반적인 피치존재 구간인 20-200샘플 사이에서 무성음구간일 경우 피치가 존재하지 않기 때문에 간격정보는 존재하지 않는다. 그림. 4(B)는 유성음구간이므로 피치가 약 80샘플 정도에 존재하고 있는 것을 볼 수 있고 피치가 존재하므로 피치 점에서 양·음의 경사도 간격을 구할 수 있다.

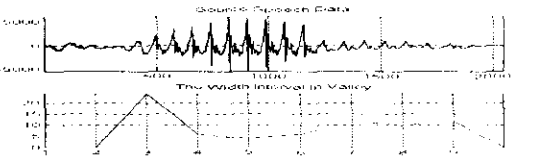

그림. 4(D) /삼/에서의 전체 구간에 대한 NAMDF

마지막으로 혼합음에 대해 전이되는 과정을 그림4-3에 나타내었다. 그림. 4(C)에서 보듯이 간격정보를 살펴보면 값이 커지는 것을 알 수 있다. 그림4-4에서는 각 프레임에서의 전체적인 간격정보의 변화를 나타내었다. 음소가 시작되는 영역에서 부터, 피치 점에서 NAMDF의 간격정보가 점점 커지는 것을 볼 수 있고 음소가 끝나 가는 부분에서 간격정보가 다시 감소하는 것을 볼 수 있다. 간격정보가 0이라는 것은 피치가 존재하지 않는 부분으로 무성음구간이라고 할 수 있다. 간격정보의 값이 갑자기 증가할 경우에는 음소가 전이되는 구간으로 볼 수 있고, 유성음구간에서는

피치가 존재하므로 간격정보의 값은 유성음구간에서 가장 작다. 음소의 변화가 빠른 자음연결이나 음소가 끝나는 프레임구간에서는 모음구간에 비해 NAMDF의 골점에서의 sharpness 정도가 완만하다는 것을 알 수 있다.

이상을 정리하여 음소의 구간을 선택하기 위한 결정논리를 만들 수 있다. NAMDF로 구한 골점에서의 sharpness 정도에 따른 간격정보는 음성 파형의 전반적인 변화도를 나타내기 때문에 각 프레임마다 규준화된 AMDF를 이용하여 골점으로부터 간격정보를 구함으로서 그 프레임에서의 안정상태·전이구간을 검출할 수 있다.

V. 실험 및 결과

시뮬레이션을 위해 IBM PC/Pentium -150MHz 시스템에 마이크로폰의 입력이 가능하도록 16-비트 Analog to Digital 변환기를 인터페이스 시켰다. 화자는 남성화자 2명과 여성화자1명을 통해 다음의 연속음을 발성케하고 8kHz로 표본화하여 저장시켰다:

발성 1): /인수네 꼬마는 천재소년을 좋아한다./

발성 2): /숭실대학교 정보통신공학과 음성통신 연구실 입니다./

발성 3): / 공일이삼사오육칠팔구./

각 음성시료에 대해 한 프레임의 길이를 256샘플로 하여 128샘플씩 겹치게 처리하였다. 각 프레임에 대해 NAMDF를 이용하여 골점을 찾고 그 점에서의 양과음의 경사도에 따른 간격정보를 측정하였다. 결국 피치 존재구간인 20-200샘플사이에 존재하는 골점(valley)에서의 골짐의 sharpness 정도에 따른 간격정보만을 고려하였다. 발성 1), 발성2), 발성3) 에 대한 처리 결과중 발성1)에 대한 규준화된 AMDF의 간격정보 값을 그림5-1에 나타내었다.

VI.결론

연속음 인식을 위해서는 음성신호의 분할 과정이 필요하다. 음소단위의 분리가 잘 이루어지면 음성분석시나 인식시에 고립단어의 분석과 인식에 적용했던 많은 기법들을 쉽게 적용할 수 있게 된다. 지금까지 음성 파형의 구간검출법들이 많이 제안되어 왔지만 평균진폭의 변화도에서 전이구간을 검출하는 것이 쉽고 우수한 편이다. 그렇지만 적용 과정에서 윈도우의 영향을 많이 받게 되어 전이구간 검출에 대한 결정논리가 복잡해진다.

따라서 본 논문에서는 시간영역에서 음소의 구간 검출시에 평균진폭이 갖는 제반 문제점들을 제거하기 위해 NAMDF를 이용한 파라미터를 제안하였다. 제안된 파라미터를 음성 파형에 적용하면, 간단한 비교논리에 의해 쉽게 그 구간을 찾을 수 있다. 또한 규준화된 AMDF에 의해 혼합음 구간의 간격이나 유성음 구간의 성질도 근사적으로 파악할 수 있다.

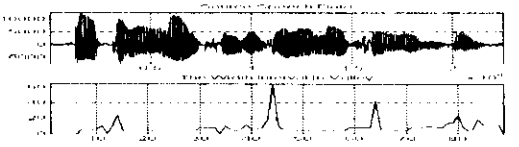

그림. 5-1 /발성 1)/에 대한 전체 구간의 NAMDF

본 연구는 정보통신부의 대학기초연구지원사업 연구지원비로 이루어졌다.

VII. 참고 문헌

[1] C.J. Weinstein, S.S. McCandless, L.F. Mondshein, and V.W. Zue, "A System for Acoustic Phonetic Analysis of Continuous Speech", IEEE Trans. on ASSP, Vol. ASSP-23, No.1, pp.54-67, Feb. 1975.

[2] W.F. Ganong, and R.J. Zatorre, "Measuring Phoneme Boundaries Four Ways", J. Acoust. Soc. Am. Vol. 68, No.2, pp.431-439, Aug. 1980.

[3] 김수중 외 2인, "A Segmentation Algorithm of the Connected Word Speech by Statistical Method", 대한전자공학회지, Vol.26, No.4, pp.151-162, Apr., 1989.

[4] R. Mori, P. Laface, and E. Piccoo, "Automatic Detection and Description of Syllabic Features in Continuous Speech", IEEE Trans. On ASSP, Vol.ASSP-24, No.2, pp.880-883,Oct., 1976.

[5] L.R. Rabiner, and M.R. Sambur, "Some Preliminary Experments in the Recognition of Connected Digits", IEEE Trans. on ASSP, Vol.ASSP-24, No.2, pp.170-182, Apr., 1976.

[6] 배명진 외 1인, "On Detecting the Steady State Segments of Speech Waveform by using the Normalized AMDF", 대한전자공학회지, Vol.14, No.1, pp.600-603, Jun., 1991.

[7] A.M. Kondoz *"Digital Speech"*, John Wiley & Sons Ltd, Chichester, England, 1994

[8] L.R. Rabiner, and R.W. Schafer *"Digital Processing of Speech Signals"*, Prentice-Hall, Englewood Cliffs, New Jersey, 1978.

[9] 배명진, 안수길, "음성파형의 대칭율을 이용한 천이구간 검출", 음성통신 및 신호처리WORKSHOP논문집, 한국음향학회, pp.79-83, 1990년 8월.