Journal of the Korea Academia-Industrial cooperation Society
Vol. 13, No. 1 pp. 333-342, 2012

http://dx.doi.org/10.5762/KAIS.2012.13.1.333

청각 주파수 응답에 기반한 자동 모음 개시 지점 탐지

장 한[1], 김학태[1], 정길도[1*]
[1]전북대학교 전자정보공학부

Automatic Vowel Onset Point Detection Based on Auditory Frequency Response

Xian Zang[1], Hagtae Kim[1] and Kil To Chong[1*]
[1]**Electronics and Information Department, Chonbuk National University**

요 약 이 논문에서는 인간 청각 시스템에 기반한 모음 개시 지점 (VOP) 탐지 방법을 제시하였다. 이 방법을 통해 '지각의' 주파수 범위, 즉 선형 음향 주파수에서의 Mel Scale을 보여준 후 일련의 삼각 Mel-weighted Filter Bank를 만들어 인간의 청각 시스템에서 대역 필터링 기능을 시뮬레이션하였다. 이러한 비선형 임계 대역 Filter Bank는 데이터 차원수를 크게 감소시키고 비선형적으로 간격을 둔 Mel 스펙트럼에서 더욱 효과적으로 포먼트를 생성하기 위해 조파들의 영향을 제거해준다. Mel 스펙트럼의 첨두 에너지 합은 각 프레임의 특징으로 추출하고 에너지 진폭이 급격히 상승하기 시작할 때의 특성은 Gabor 윈도우를 사용하여 VOP로 탐지한다. 실험 결과를 통해서 다른 종류의 자음들과 연결된 12개의 모음들을 포함하는 한 단어 데이터베이스에 대한 제안된 방법의 평균 정확도는 단시간 에너지와 zero-crossing 비율에 기반을 둔 다른 모음 탐지 방법들보다 높은 72.73% 이상임을 확인하였다.

Abstract This paper presents a vowel onset point (VOP) detection method based on the human auditory system. This method maps the "perceptual" frequency scale, i.e. Mel scale onto a linear acoustic frequency, and then establishes a series of Triangular Mel-weighted Filter Bank simulate the function of band pass filtering in human ear. This nonlinear critical-band filter bank helps greatly reduce the data dimensionality, and eliminate the effect of harmonic waves to make the formants more prominent in the nonlinear spaced Mel spectrum. The sum of mel spectrum peaks energy is extracted as feature for each frame, and the instinct at which the energy amplitude starts rising sharply is detected as VOP, by convolving with Gabor window. For the single-word database which contains 12 vowels articulated with different kinds of consonants, the experimental results showed a good average detection rate of 72.73%, higher than other vowel detection methods based on short-time energy and zero-crossing rate.

Key Words : VOP detection, Formant, Human auditory system, Mel Scale, Triangular Mel-weighted filter bank, Gabor window

1. 서론

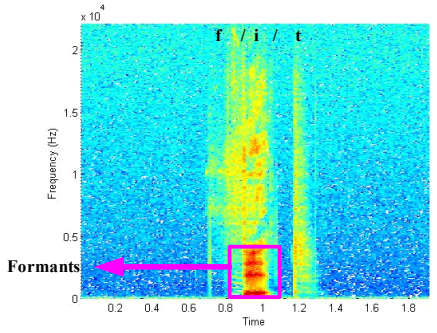
우리는 보통 교육 목적을 위해 모음만의 길이나 높낮이, 액센트를 변경할 필요가 있다. 그러므로 모음 개시 지점 (VOP)에 대한 정보를 필요로 한다. 수동 표시는 시간이 많이 소모되며 오류가 발생하기 쉽다. 인간의 시각과 청각 지각 능력 사이의 가변성 때문에 수동 표시 결과를 재생하는 것은 거의 불가능하다. 그러므로 수동 표시 방법은 본질적으로 일치하지 않는다. 이에 비해 자동 탐지 방법은 일정한 결과를 얻는데 비교적 시간이 적게 소요되며 일치하는 결과를 얻을 수 있다. 그러나 다양한 자음 모음 구성의 강인한 VOP 특성을 구하기가 어려우며 발성자 마다 사로 다른 방식으로 발음하기 때문에 오류가 존재 할 수 있다. 그러므로 자동 인식 시스템의 성능 향

*교신저자 : 정길도(kitchong@chonbuk.ac.kr)
접수일 11년 11월 04일 수정일 (1차 11년 12월 22일, 2차 12년 01월 03일) 게재확정일 12년 01월 05일

상을 위해서는 다양한 경우에서 VOP를 구분할 수 있는 특성을 구하는 것이다.

많은 언어들 사이에서 모든 모음들은 유성음에 속하고 대부분의 자음들은 무성음이다. 그러므로 우리는 모음의 두드러진 특징을 나타낼 수 있는 필요한 매개변수를 찾기 위해 이러한 관점에서 문제를 다루었다.

음성학에서 유성음은 성대의 진동으로 발생되는 반면에 무성음은 기류와 성도 사이의 충돌과 마찰에 의한 난류 효과이다. 이러한 차이는 서로 다른 스펙트럼 분포를 만든다. 유성음은 성도의 공명에서 발생하는 독특한 스펙트럼 정점을 가지는데 이것은 Fant에 의해 포먼트로 정의되었다[1]. 하지만 무성음은 이와 같은 규칙이 없다. 그림 1에 나타낸 단어 “fit”의 스펙트럼에서 이러한 차이를 발견할 수 있다. 그림에서 x-축은 시간을 y-축은 주파수를 나타낸다. 그리고 3차원은 특정한 시간에서 특정한 주파수의 크기를 색으로 나타내며, 색이 진할수록 에너지가 크다. 저 주파수 영역에서 단모음 /i/는 시간 축에 평행하는 명확한 고에너지 “bars”를 가진다. 하지만 자음 /f/와 /t/는 고주파수 대역에서 덩어리진 혼돈상태의 에너지 분포를 보여 준다. 이 “bars”들을 포먼트라고 부른다. 이 포먼트들은 높은 에너지를 운반하기 때문에, 우리는 이 포먼트 에너지를 에너지 진폭이 급격히 상승하기 시작할 때의 성질을 탐지하기 위한 특징으로 사용할 수 있다. 하지만 포먼트는 DFT 스펙트럼에서 사라지고 조파의 영향을 받는다. 이러한 특징으로 인해 더욱 효과적으로 포먼트 정보를 얻는 방법에 면밀한 탐구가 필요하다.



[그림 1] 단어 “fit”의 스펙트럼(모음은 단모음 /i/)
[Fig. 1] The spectrum of the word “fit” (vowel is a monophthong /i/)

음향심리학 분야에서의 인간 청각 지각에 대한 연구는 인간의 청각이 비선형 필터로 작용하고 확실한 주파수 요소들에만 집중한다는 것을 설명해 준다. 이것은 중간 주파수와 대역폭이 선형 청각 주파수 축에 비선형적으로 분포되는 Filter Bank를 구축함으로써 모델화될 수 있다. Filter Bank표시에는 두 가지 주된 목적이 있다:

1) 순음과 같은 자극에 대한 기저막에 따른 최대 변위 위치는 어조의 주파수 알고리즘에 비례한다. 이 가설은 장소 이론의 한 부분이다[2].
2) 인간 지각 실험은 아주 적은 주파수의 주파수대역 안에 복잡한 소리의 주파수는 개별적으로 확인될 수 없음을 보여왔다. 이 소리의 구성요소들 중 하나가 주파수대역의 밖에 있을 때, 우리는 이것을 개별적으로 구별해 낼 수 있다. 우리는 이 주파수대역을 임계 주파수대역으로 분류한다[3].

청각 주파수 축의 이러한 비선형 뒤틀림은 “지각의” 주파수 범위, 즉 stevens와 volkmen에 의해 개발된 이른바 Mel-scale에 의해 모델링 될 수 있다[4]. 임계 대역 filter bank는 선형적으로 정돈되는 선형위상 유한 임펄스 반응(FIR) 대역통과필터이다. 그러므로 변형 후에 이 filter bank 간격은 최대 1000Hz까지 대략 선형이고 청각 주파수축 상의 더 높은 주파수들은 대수이다.

본 논문에서는 청각 시스템에서 일어나는 대역통과 필터링을 시뮬레이션 하기 위해 법칙에 따라 일련의 삼각 Mel-weighted Filter Bank를 설계하였다. 구체적인 설계 방법은 본 논문의 2.2.3 절에 기술되어 있다.

삼각필터는 Mel 크기에 따라 선형적으로 연결되어있다. 그러나 청각 주파수 영역에서 보면, Mel 크기로 사상된 필터의 중앙 주파수는 비선형적으로 구성된다. 즉 삼각 필터는 각 가중된 Mel 크기의 중앙 주파수에 기초하여 같은 간격으로 나열 되어 있다. 그러므로 서로 인접한 필터는 50%가 중첩된다.

삼각 필터의 장점은 다음과 같다:

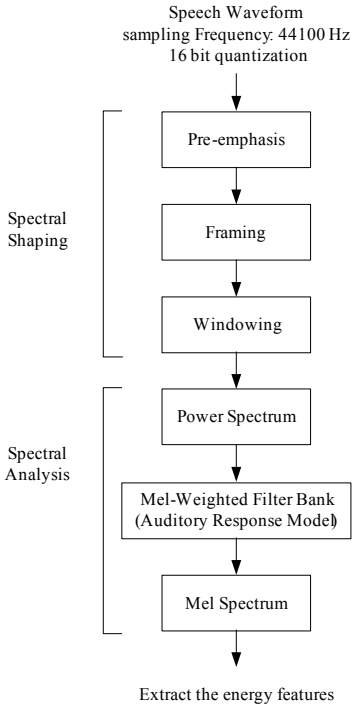
1) 삼각 필터는 스펙트럼을 매끄럽게 하고 formants prominent를 만들기 위해 조파를 제거한다. 결과적으로 입력 소리의 높낮이나 음색의 차이에서 발생하는 불필요한 외란들을 피할 수 있다.
2) 삼각 필터는 단순한 형태이기 때문에 데이터 차원수를 감소시킨다. 계산 난이도를 현저하게 감소시키는 동안 소리의 필수적인 정보들이 저장된다.

이러한 임계 대역통과 필터들에 근거해서 상응하는 필터 이득에 의해 각각의 FFT크기를 증가시킨 후 결과들의 합계를 구하는 “binning”에 의해 파워 스펙트럼(FFT)을 Mel-weighted 스펙트럼으로 변형시켰다. 동시에 에너지가 계산된다. 다음으로, 필터 채널의 차원수가 감소된 필수적인 정보들을 가지고 Mel 스펙트럼으로부터 포먼트 에너지의 합을 추출하고 이것을 각각의 말 신호 프레임에 대한 특징으로 사용한다. 첨두 에너지의 합과 Gabor 윈도우를 사용하여 에너지 진폭이 급격이 상승하기 시작

할 때의 성질을 VOP로 탐지한다.

2. 자동 탐지 과정

아래 그림은 모음 특징 추출 작업의 블록선도이다.



[그림 2] 특징 추출 흐름
[Fig. 2] Feature extraction flow

2.1 스펙트럼 성형

Pre-filtering, 샘플링, A/D 변환과 같은 약간의 프런트 엔드 처리 후에, 언어 파형 s(n)을 형성하기 위해 일련의 샘플들을 얻는다. 다음으로 스펙트럼 성형을 위한 다음 작업을 시작한다.

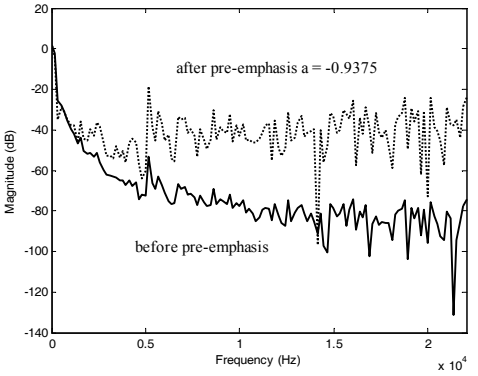
2.1.1 프리엠퍼시스

이전의 연구가 보여주는 것처럼, 음성 신호의 유성음 부분은 음성 생성 시스템의 생리적인 특징 때문에 자연스럽게 10년마다 대략 20dB(주파수 크기 증가 순서)의 음의 스펙트럼 기울기(감쇠)를 가지고 있다[5-6]. 그러므로 프리엠퍼시스 필터로 잘 알려진 하나의 계수 디지털 필터는 그림 3과 같이 음성 신호를 스펙트럼으로 평평하게 하기 위해 이러한 자연 경사를 벌충함으로써 분석의 효율성을 향상시킨다[6-7].

출력 $\hat{s}(n)$과 입력 $s(n)$ 사이의 관계는 다음과 같다.

$$\hat{s}(n) = s(n) - \alpha s(n-1) \tag{1}$$

여기서, 프리엠퍼시스 계수 α의 범위는 보통 0.9와 1 사이에 있다. 고정점 하드웨어에서 효율적으로 실행하기 위해 본 논문에서는 $\alpha = 1 - 1/16 = 0.9375$로 선택했다[8].



[그림 3] 프리엠퍼시스 전후의 말 신호의 주파수 응답(점선은 평평하게 된 음성 스펙트럼)
[Fig. 3] The frequency response of the speech signal before and after the pre-emphasis(the broken line shows the flattened speech spectrum)

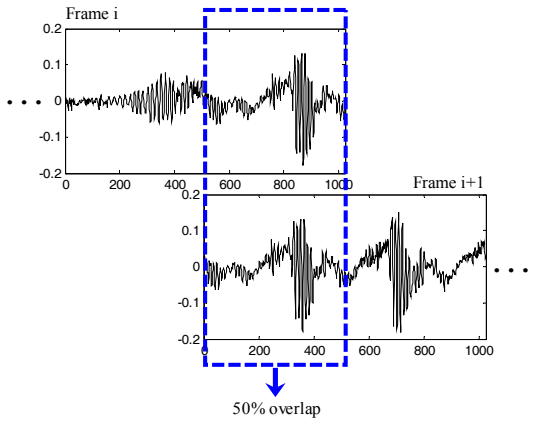
2.1.2 구성

음성은 일반적으로 시변 신호이다. 그러나 우리는 음성 신호의 특성이 상대적으로 천천히 짧은 기간을 가지고 변한다는 것을 추측할 수 있다.

이 단시간 안정성 가정은 프리엠퍼시스된 음성을 프레임이라 불리는 일련의 짧은 조각으로 만들어 전체 음성 신호 밖으로 분리하도록 한다. 이것들은 time sequence안에서 정렬되고 서로 중복된다.

각 프레임에 대한 시간 길이인 프레임 기간은 일반적으로 10~30ms정도이고 중복 길이는 인접한 프레임들 사이의 고정성을 보장하기 위해 보통 프레임 길이의 1/2~1/3로 정해진다.

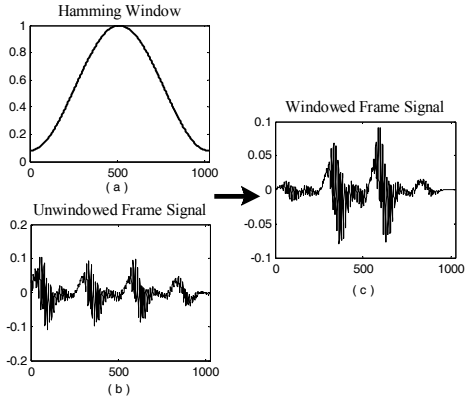
본 논문의 실험에서는 높은 샘플링 주파수 ($f_s = 44100Hz$)때문에 프레임 기간은 23.2ms(1024 samples)로 설정했고 그림 4와 같이 50%의 중복 부분을 가진다.



[그림 4] 음성 신호의 프레임 차단
[Fig. 4] Frame blocking of the speech signal

2.1.3 윈도잉

음성 파형의 결과가 직각 펄스와 곱해질 때 프레임이 보여질 수 있다. 주변효과를 감소시키기 위해 가중치나 윈도우의 중심으로 향하는 favor 샘플들에 대해 윈도잉 과정을 실시했다. 이러한 특성은 매개변수 평가에 원활하게 변화를 주는 것과 frame-by-frame기반에 대한 다음의 스펙트럼 분석을 얻는데 있어서 중요한 기능을 수행한다.



[그림 5] Hamming 윈도우와 윈도잉 전후의 프레임 신호의 파형(Hamming 윈도우의 목적은 각 프레임과 frame-by-frame에 원활한 변화를 실현하기 위해 윈도우의 중심으로 향하는 favor 샘플들의 주변 효과를 감소시키는 것임)

[Fig. 5] Hamming window and the waveform of frame signal before and after windowing(The motive of Hamming window is to reduce the edge effect of each frame and favor samples towards the center of the window for realizing smoothly varying frame-by-frame)

여기서, 우리는 hamming 윈도우(그림 5(a))를 각각의 프레임 신호(그림 5(b))에 적용했다. 그 수학적 표현은 식 (2)와 같다.

$$w(n)=\begin{cases}0.54-0.46\cos(2\pi n/(N-1)): & 0\le n\le N-1\\ 0 \quad : & else\end{cases} \tag{2}$$

N은 프레임 하나에서의 샘플 수이다. 우리는 시간 윈도우에 의해 프레임 신호를 증가시킴으로써 식 (3)으로부터 짧은 창이 있는 조각 $s_w(n)$을 얻는다. 시간영역 파형은 그림 5(c)와 같다.

$$s_w(n)=\begin{cases}\hat{s}(n)\cdot w(n-m) & : \quad n=m,m+1,...m+N-1\\ 0 & : \quad else\end{cases} \tag{3}$$

2.2 스펙트럼 분석

본 절에서는 인간 청각 시스템과 특징 추출을 위한 스펙트럼 분석의 모델링 과정을 상세히 기술할 것이다. 우선은 선형 청각 주파수 범위에 대한 파워 스펙트럼부터 시작한다.

2.2.1 파워 스펙트럼

주파수 영역에서 스펙트럼 분포를 관찰하기 위해, 식 4와 같이 신호의 크기 스펙트럼을 계산하기 위한 창이 있는 프레임에 푸리에 변환(FFT)이 적용된다[9].

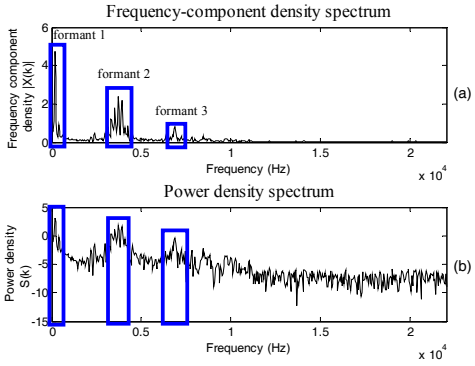
$$X(k)=\sum_{n=1}^{N}s_w(n)e^{-j2\pi kn/N}, \quad k=1,\ldots,K \tag{4}$$

여기서, K는 FFT의 길이이다.

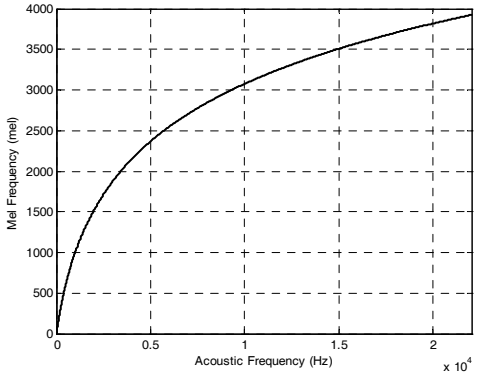
그리고 나서 이것의 크기를 계산하면, 파워 스펙트럼이 얻어진다.

$$S(k)=(real(X(k)))^2+(imag(X(k)))^2 \tag{5}$$

여기서, 각 프레임에서의 샘플의 숫자와 같은 K를 선정한다(i.e. 2^{10}=1024). 그림 6은 FFT후의 크기 스펙트럼과 파워 스펙트럼을 보여주고 있다. 분명하게 포먼트가 존재하는 곳을 나타내는 높은 스펙트럼 에너지를 가진 세 개의 영역이 있다. 하지만 포먼트 주변에는 많은 첨두치들이 있어 포먼트들을 분명하게 하기 위해서는 다음과 같은 공정을 필요로 한다.



[그림 6] 모음 신호의 FFT 결과 (a), (b)
[Fig. 6] FFT results (a), (b) of vowel signal



[그림 7] Mel 크기 곡선
[Fig. 7] The Mel Scale

2.2.2 Mel 범위

"Mel"은 음색의 지각 피치 또는 주파수 측정의 구성단위 중 하나이다. Mel 범위는 다음 절차를 사용한 인간 청각 지각 연구의 결과로 Stevens와 Volkman(1940)에 의해 개발되었다[4]:

1) 1,000Hz로 참조 주파수를 선택하고 이것을 "1,000Mels"라고 부른다.
2) 청자들에게 신호가 제공되고 그들이 인식했던 피치가 참조의 두 배가 되거나 참조가 10회 반복되거나, 참조가 절반이 되거나 1/10일 때까지 그 주파수를 변경하도록 요청된다.

이 데이터에서 Mel 범위가 만들어진다. Mel 범위는 주파수 범위에서 지각적으로 의미 있는 범위로의 변형으로 간주될 수 있다. 이것은 1,000Hz 이하에서 대략 선형이고 1,000Hz 이상에서 대수이다. 이것은 다음 식을 사용하여 근사화 되었다.

$$Mel(f) = 2595 \lg\left(1 + \frac{f}{700}\right) \tag{6}$$

또는

$$Mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \tag{7}$$

여기서, FFT의 대칭 때문에 우리는 샘플링 주파수의 1/2로 샘플 신호의 최대 주파수를 계산했다.

$$f_{\max} = \frac{f_s}{2} = 22050Hz \tag{8}$$

그림 7은 Warping 함수의 곡선을 나타내고 있다.

2.2.3 삼각 Mel-weighted Filter Bank

Mel-weighted filter bank는 청각 시스템에서 일어나는 대역통과 필터링을 시뮬레이션 하도록 설계된 일련의 삼각 임계 대역통과 필터이다. 모든 필터들을 위한 그룹 지연이 0과 같고 필터의 출력 신호들이 시간내에 동기화될 수 있도록 디지털 filter bank의 각 필터는 보통 선형 위상 필터로 구현된다. 이것은 mel 주파수 범위 상에 선형적으로 정렬된 대역통과 필터들의 연속에 상응한다. 처음에 필터들의 중간 주파수들은 mel 범위에 근거해 계산된다.

첫 번째로 Mel 범위에 최대한의 청각 주파수를 표시한다.

$$Mel(f_{\max}) = 1127 \ln(1 + \frac{f_{\max}}{700}) \tag{9}$$

이 필터들이 Mel 주파수 범위를 따라서 선형적으로 정렬된다고 가정되기 때문에 두 개의 인접한 필터들 사이의 간격은 다음과 같다.

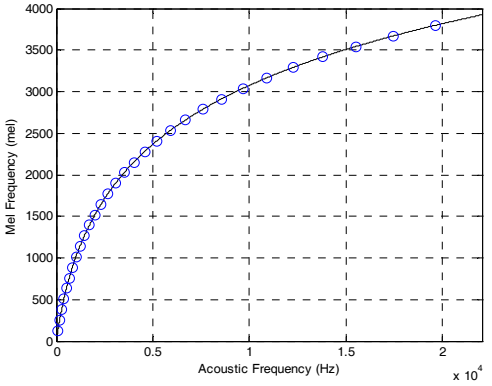
$$\Delta = \frac{Mel(f_{\max})}{M+1} \tag{10}$$

여기서, M은 삼각 Mel-weighted 필터들의 총 개수이고 보통 24~40개의 범위에 있다. 44,100Hz의 음성 신호에 대하여 30개의 필터를 고려해보자.

그리고 나서 식 (7)의 역을 이용해 이 중간 주파수들을 음향 주파수에 나타낸다.

$$Mid(m) = 700 \times \left[\exp\left(\Delta \times m / 1127\right) - 1\right]$$
$$m = 1, 2, \ldots, M \tag{11}$$

음향 주파수 범위에서의 중간 주파수들의 분포는 그림 8과 같다.

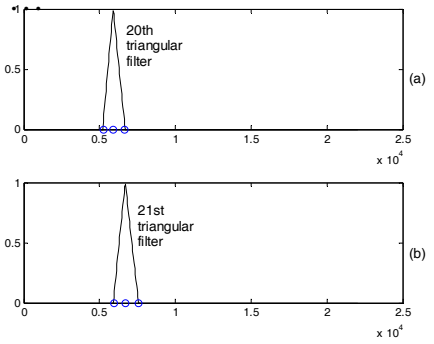


[그림 8] 삼각 Mel-weighted Filter Bank의 중간 주파수(원의 각 중심은 각 필터의 중간 주파수에 상응함)
[Fig. 8] Mid frequencies of the Triangular Mel-weighted Filter Bank (each center of circle corresponds to mid frequency of each filter)

본 논문에서는 청각 시스템에서 일어나는 대역통과 필터링을 시뮬레이션 하기 위해 음향 주파수 범위에 있는 중간 주파수에 기반을 두어 다음과 같은 과정에 따라 일련의 삼각 Mel-weighted Filter Bank를 설계하였다.

1) m번째 중간 주파수를 선택한다. 이 주파수는 음향 주파수 범위에 있는 m번째 필터의 중심을 결정하기 위해 사용된다.
2) m번째 중간 주파수에 중심을 두고 (m-1)번째와 (m+1)번째 중간 주파수를 선택한다. 두 중간 주파수 사이의 거리는 m번째 필터의 대역폭이다.
3) 대역폭을 삼각형의 밑변으로 취하고 상응 대역폭의 m번째 필터의 높이는 식(12)에 의해 구해진다. 최대 높이는 m번째 중간 주파수에 위치하고 대역폭 밖의 높이는 0이다.
4) 다음 삼각 필터를 만들기 위해 다음 중간 주파수를 선택하고 1) ~ 3) 단계를 반복한다.
5) 30개의 필터를 모두 설계한 후에 그것들을 삼각 Filter Bank의 형태로 중복시킨다. 음향 주파수에 있는 중간 주파수가 인간의 청각 응답을 시뮬레이션하기위해 Mel 범위로 한정되어 있기 때문에 우리는 이것을 Mel-weighted Filter Bank라고 명명했다.

20번째와 21번째 삼각 필터를 만드는 예가 그림 9에 설명되어 있다. 각 plot에 있는 세 개의 원은 각 삼각 필터의 대역폭을 형성하기 위해 사용되는 세 개의 중간 주파수를 나타낸다.



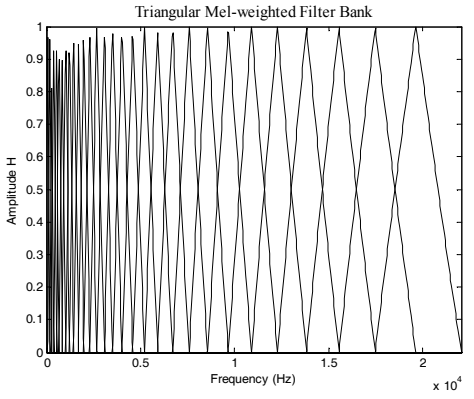
[그림 9] 각 삼각 필터의 제작
[Fig. 9] An illustration of establishing each triangular filter

$$H_m(f)=\begin{cases}0 & f<Mid(m-1)\\ \dfrac{(f-Mid(m-1))}{(Mid(m)-Mid(m-1))} & Mid(m-1)\le f\le Mid(m)\\ \dfrac{(Mid(m+1)-f)}{(Mid(m+1)-Mid(m))} & Mid(m)\le f\le Mid(m+1)\\ 0 & f>Mid(m+1)\end{cases}$$

$$f=(k-1)\times\frac{f_s}{K}\quad k=1,2,\ldots,\frac{K}{2}+1 \tag{12}$$

위 식은 [10]에 따라 $\sum_{m=1}^{M}H_m(f)=1$ 을 만족시킨다.

여기서, $f(m)$ 은 m번째 필터의 중간 주파수를 의미하며, K는 FFT의 길이이다.

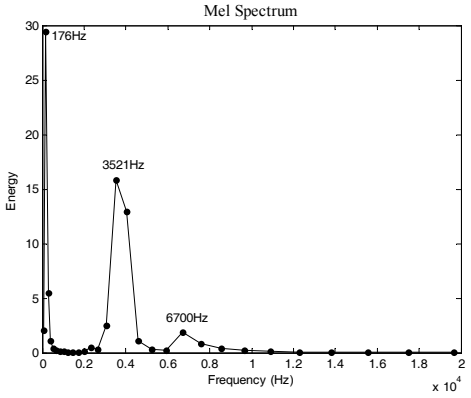


[그림 10] 정규화된 삼각 Mel-weighted Filter Bank
[Fig. 10] The normalized Triangular Mel-weighted Filter Bank

2.2.4 Mel 스펙트럼

청각 시스템에 의한 특별한 주파수 f의 인식이 f주변 주파수들의 임계 대역 에너지에 의해 영향을 받는다는 것을 발견했기 때문에 Mel 스펙트럼은 각각의 삼각 Mel-weighted 필터를 가진 파워 스펙트럼을 증가시키고 나서 각각의 임계 대역에서의 결과를 가산함으로써 계산된다[11].

$$\tilde{S}[m] = \sum_{k=1}^{\frac{K}{2}+1} S(k) H_m(k) \quad m = 1, 2, \ldots M \tag{13}$$



[그림 11] 30개의 필터 출력(모음 프레임 신호는 2.2.1절과 동일)

[Fig. 11] The outputs of 30 filters (the vowel frame signal is same as that in Section 2.2.1)

그림 11은 이 변환의 결과를 보여준다. 30개 필터들의 출력은 음향 주파수 축 상에 비선형적으로 간격을 둠에도 불구하고 각 프레임을 위한 음성의 필수 정보를 획득한다. 2.2.1절의 512-length 파워 스펙트럼에 비해 데이터 차원수가 현저하게 감소되었고 30개 출력의 더 작은 집합은 인간 청각 시스템에 가장 잘 부합한다. 에너지 추출을 위해 이 출력 시간 동안 세 개의 스펙트럼 정점이 분명해진다.

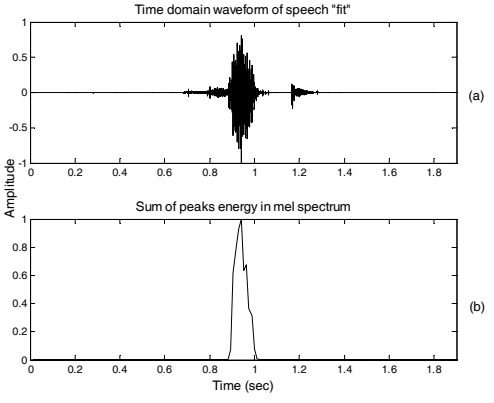
2.3 Mel 스펙트럼 첨두 에너지를 사용한 모음 개시 지점 탐지

필터를 통해 여과된 스펙트럼에서 포먼트 영역이 잘 분포되어 있음을 알 수 있다. 포먼트의 에너지는 Mel 스펙트럼에서 첨두치를 선별하여 쉽게 계산된다. 그림 12(a)와 (b)는 "fit"의 음과 이것의 첨두 에너지의 정규 합을 각각 보여준다. 모음의 개시 지점은 합계 plot에서 상당한 변화가 있는 순간으로 관찰된다. 이 순간을 탐지하기 위해 식 14와 같이 주어진 Gabor 윈도우를 사용한다[12].

$$g(n) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{(n-\mu)^2}{2\sigma^2}} \cos(\omega n) \tag{14}$$

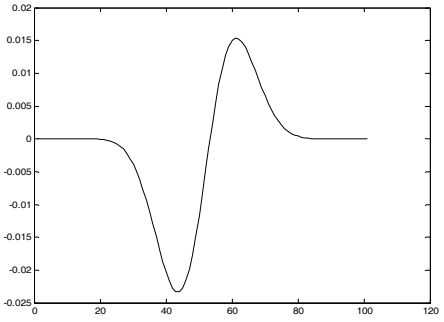
이 윈도우는 매개변수 $\sigma = 10$과 $\omega = 0.9$를 사용해 생성되는데, 여기서 σ는 Gabor 필터의 공간 넓이이고 ω는 필터 길이 n=100을 가진 사인 곡선 요소의 각 주파수이다. 억제 시간 코스가 CV유닛 자극 시간 코스 보다 길기 때문에 Gabor 필터의 매개변수들은 음의 부분이 양의 부분보다 크게 선택된다[13]. 그림 13은 Gabor 필터의 형태를 보여 준다.

그리고 나서 Gabor 윈도우 함수의 음의 부분을 가진 첨두 에너지의 합을 이용하면 VOP 증거 plot 이라 불리는 출력을 얻게 되는데, 여기서 최대값은 첨두 에너지의 합이 급격히 상승하기 시작하는 순간에 상응하도록 가정된 VOP를 나타낸다. 그림 14는 단어 "fit"에 대한 VOP 증거 plot을 보여 준다. (a)는 손으로 표시된 참조 VOP이고 시간 포인트는 0.889s이다. (b)는 탐지된 것으로 시간 포인트는 0.896s이고 오차는 7ms이다. 이러한 결과는 종점 탐지 및 유성음/무성음 범위 인식과 같은 응용을 위해 충분히 정확하다.

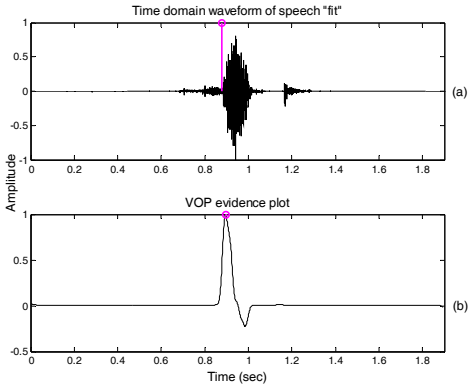


[그림 12] (a) "fit"의 말 신호, (b) Mel 스펙트럼에서의 첨두 에너지의 합

[Fig. 12] (a) Speech signal of "fit", (b) Sum of the peaks energy in the Mel spectrum



[그림 13] Gabor 윈도우 ($\sigma = 10, \omega = 0.9, n = 100$)
[Fig. 13] Gabor window ($\sigma = 10, \omega = 0.9, n = 100$)



[그림 14] VOP 증거 plot
[Fig. 14] VOP evidence plot

3. 시뮬레이션

본 논문의 실험에서는 C가 자음을 나타내고 V가 모음을 나타내는데 실험을 위해 모음의 위치가 다른 VC, CV, CVC와 같은 다음자 단어들을 사용하였다. 이 단어들은 남성 화자로부터 분리된 방법으로 기록된다. 데이터베이스는 표 1에 나타낸 9개의 단모음과 3개의 이중 모음 총 12종류의 모음을 포함한다. 각 모음은 표 2에 나타낸 다른 자음들과 이어진다.

[표 1] 모음들
[Table 1] Vowels

Monophthongs	/i:/, /i/, /e/, /u:/, /u/, /ɔ:/, /ɔ/, /ʌ/, /ə/
Diphthongs	/ei/, /æ/, /əu/

[표 2] 자음들
[Table 2] Consonants

Liquid	/l/
Glide	/j/
Nasal	/m/, /n/, /ŋ/, etc.
Plosive	/b/, /p/, /d/, /g/, /t/, /k/, etc.
Fricative	/f/, /h/, /s/, /ʃ/, etc.
Affricate	/z/, /ð/, /dʒ/, /ʒ/, etc.

표 3은 데이터 수집 및 분석의 세부사항들은 보여주고 있다.

[표 3] 데이터 수집 세부사항
[Table 3] Data collection details

매개변수	값
화자 성별	남성
총 발화	252 (각 모음당 20단어 + 12 모음)
샘플링 주파수	44,100Hz
프리엠퍼시스 계수	0.9375
프레임 길이	1024 samples ≈ 23.2ms
윈도우	1024-point Hamming Window
FFT 크기	1024
Filter Bank 종류	Mel Filter bank
Mel Filter Bank 개수	30

각 단어의 말 신호는 50%의 중복을 가진 23.2ms의 블록으로 처리된다. 2자의 자동 탐지 과정을 거치면 탐지된 VOP가 얻어진다. 만약 참조 VOP와 탐지된 VOP 사이의 오차가 +/- Tms의 허용 한계보다 작으면 VOP가 T의 시간 분해능을 가진 탐지 VOP로 간주된다. 부합되는 VOP의 백분율을 탐지율이라 한다. 표 4는 10ms의 시간 분해능에서 각 모음 범주에 대한 탐지율을 나타낸다.

[표 4] VOP 탐지율
[Table 4] VOP detection rate

모음	True VOP의 개수	탐지율(%) (within ≠ 10ms)
i:	22	68.19
i	22	63.63
ei	22	77.27
e	22	59.10
æ	22	77.27
u:	22	72.73
u	22	72.73
əu	22	81.82

ɔ:	22	72.73
ɔ	22	81.82
ʌ	22	72.73
ə	22	72.73
평균 정확도		72.73

단시간 에너지와 zero-crossing 비율에 근거한 다른 모음 탐지 방법들과 비교해 보면, 평균 정확도가 50%에서 72.73%로 상승했다. 탐지 결과를 통해 모음과 연결된 자음들이 "liquid", "glide", "affricate"일 때, VOPrk 정확하게 탐지될 수 없음을 확인할 수 있는데, 이것은 손실된 탐지율 때문이다. 위 자음들의 경우에서 나쁜 성능을 보이는 것은 VOP 전후의 신호 특징이 유사한 것에 기인한다. 그러므로 향후 이러한 문제들을 개선하기 위한 방법들에 초점을 맞춰 연구를 진행할 예정이다.

4. 결론

본 논문에서는 인간 청각 시스템 모델에 기반한 모음 개시 지점 탐지 방법을 살펴보았다. 자음과 모음의 발음에서 현저한 차이가 있으며, 포먼트 에너지는 VOP 위치를 파악하는 중요한 지표가 된다. 그러나 스팩트럼의 피치와 관련된 스팩트럼 구조, 즉 하모닉 파형은 포먼트를 파악할 때 장애가 되며, 또한 음성 특징에 변화를 유발한다. 본 논문에서는 이러한 문제를 해결하기 위해 인간의 청각시스템에 기반한 VOP 파악에 대한 방법을 제안한다. 인간의 청각 시스템에 일어나는 대역 통과 필터링 기능을 시뮬레이션하기 위해 Mel 범위에 근거한 30개의 삼각 Mel-weighted 필터들을 구성했다. 이 비선형 임계대역 filter bank는 데이터 차원수를 512에서 30까지 현저하게 감소시키고 피치와 관련된 세부 구조를 제거하므로 포먼트를 보다 우수하게 만들 수 있도록 해주었다. 비선형적으로 간격을 둔 Mel 스펙트럼에 근거하여 각 단어의 VOP를 탐지하기 위해 첨두 에너지의 합을 각 프레임에 대한 특징으로 사용하였다. 단시간 에너지와 zero-crossing 비율에 근거한 다른 모음 탐지 방법들과 비교해 보면, 평균 정확도가 50%에서 72.73%로 상승함을 알 수 있었다.

결론적으로 본 논문에서는 인간의 청각 시스템에 기반한 모음 식별 방법으로 VOP를 이용한 방법을 제안하였고, 실험을 통하여 제안한 방법이 기존의 방법에 비하여 우수한 인식률을 확보하였다.

향후 "liquid", "glide", "affricate" CV 유닛의 경우에 대한 탐지율을 향상시키기 위해 연구를 계속 진행할 예정이다.

References

[1] Fant, G. (1960). Acoustic Theory of Speech Production. Mouton & Co, The Hague, Netherlands.

[2] J. O. Pickles, “An introduction to the Physiology of Hearing”, New York: Academic press, 1988.

[3] A. R. Moller, “Auditory Physiology”, New York: Academic press, 1983.

[4] Stevens, SS, Volkman, J, “The relation of pitch to frequency”, American Journal of Psychology, Vol.53, pg. 329.

[5] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, “ Discrete Time Processing of Speech Signals”, New York: MacMillan,1993.

[6] J. Markel and A. H. Gray, Jr., “Linear Prediction of Speech”, New York: Springer-Verlag, 1980.

[7] L. R. Rabiner and R. W. Schafer, “Digital Processing of Speech Signals”, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[8] M. Sigmund, Voice Recognition by Computer, Tectum Verlag, Marburg, 2003.

[9] O. E. Brigham, “The Fast Fourier Transform”, Englewood Cliffs, NJ: Prentice-Hall, 1974.

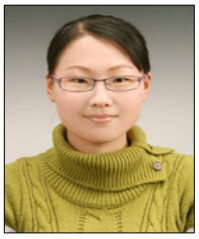
[10] X. Huang, A. Acero, and H.W. Hon, “Spoken Language Processing: A Guide to Theory, Algorithm and System Development”, Prentice Hall, 2001.

[11] Schroeder, MR, “Recognition of complex acoustic signals”, Life Science Research Reports, Vol.55, pp.323-328, 1977.

[12] D. Gabor, “Theory of communication”, Journal of IEE, vol. 93, pp. 429-457, 1946.

[13] R. L. Smith and J. J. Zwislocki, “Short-term adaptation and incremental response of single auditory-nerve fibers”, Biological Cybernetics, Vol.17, pp.169-182, 1975.

장 한(Xian Zang) **[정회원]**



- 2007년 7월 : 회혜대학교 공정장비제어학과 (공학사)
- 2009년 8월 : 전북대학교 대학원 전자공학 (공학석사)
- 2009년 9월 ~ 현재 : 전북대학교 대학원 전자정보공학부 박사과정

<관심분야>
Mechanical Engineering, Automotive Design, Electronic Control Technology, Speech Signal Processing

김 학 태(Hagtae Kim) **[정회원]**

- 2009년 2월 : 전북대학교 컴퓨터공학과 (공학사)
- 2009년 3월 ~ 현재 : 전북대학교 대학원 전자정보공학부 석사과정
- 2011년 1월 ~ 2012년 1월 : 미국 Texas A&M University 방문연구원

<관심분야>
선박전자장치, 원전제어시스템, 프로토콜 통합

정 길 도(Kil To Chong) **[정회원]**

- 1984년 2월 : Oregon State University 기계공학 (공학사)
- 1986년 2월 : Georgia Institute of Technology 기계공학 (공학석사)
- 1992년 2월 : Texas A&M University 기계공학 (공학박사)
- 2010년 3월 ~ 현재 : 전북대학교 전자정보공학부 교수

<관심분야>
Marine Navigation, Time-Delay, Robotics, 인공지능, 지능형 교통시스템