Communications of the Korean Statistical Society Vol. 15, No. 2, 2008, pp. 293-302

숨은마코프모형을 이용하는 음성구간 추출을 위한 특징벡터

홍정우[1], 오창혁[2]

요 약

본 논문에서는 숨은마코프모형을 사용하여 음성구간을 추출하는 경우에 사용되는 새로운 특징벡터인 평균파워를 제안하고, 이를 멜주파수 켑스트럴 계수(mel frequency cepstral coefficients, MFCC)와 파워계수와 비교한다. 이들 세 가지 특징벡터의 수행력을 비교하기 위하여 일반적으로 추출이 상대적으로 어렵다고 알려진 파열음을 가진 단어에 대한 음성 데이터를 수집하여 실험한다. 다양한 수준의 잡음이 있는 환경에서 음성구간을 추출하는 경우 MFCC나 파워계수에 비해 평균파워가 더 정확하고 효율적임을 실험을 통해 보인다.

주요용어: 멜주파수 켑스트럴 계수; 숨은마코프모형; 파워계수; 평균파워.

1. 머리말

컴퓨터와 같은 기계를 이용한 음성인식 방법에 대하여 다양한 연구가 이루어져 왔다. 이러한 음성인식의 첫 단계는 마이크와 같은 기기를 통해 입력되는 음압 신호에서 음성구간을 추출해내는 것이다. 대체로 음성신호는 통상 잡음이라고 일컬어지는 비음성 신호와 섞여 있으며 순수 잡음구간을 제외한 음성구간 만을 추출하는 것은 중요한 일로 알려져 있다. 자동화 기법에 의해 추출된 음성구간은 잡음구간을 일부 포함하고 있거나 혹은 음성구간이 일부 잘려나간 형태를 띠는 것이 일반적이다. 따라서 추출된 음성구간의 정확성 정도에 따라 음성인식시스템의 효율성이 큰 영향을 받게 되므로 정확한 음성구간 추출을 위한 많은 연구가 이루어져 왔다.

음성구간 추출에 흔히 사용되는 특징벡터는 음성에너지 수준과 영교차율이다 (Rabiner와 Juang, 1993). 음성에너지와 영교차율을 사용하여 음성구간을 추출하는 방법은 잡음이 없는 환경에서는 효율적이지만, 작은 잡음에도 민감하게 반응함으로 인해 잡음환경에서는 비효율적임이 알려져 있다 (Abdulla, 2002). 또한, 음성구간 추출에 있어 시작이나 끝부분에 파열음이나 마찰음이 존재하는 경우 유성음 구간에 비해 신호의 에너지의 크기가 작아서 잡음환경 하에서 추출하기가 용이하지 않으며 정확한 음성구간의 추출에 실패하는 주요한 이유 중 하나가 됨이 알려져 있다 (Seok과 Bae, 1999).

1) (712-749) 경상북도 경산시 대동 214-1, 영남대학교 통계학과, 석사과정.
E-mail: brbravo9915@yu.ac.kr
2) (712-749) 경상북도 경산시 대동 214-1, 영남대학교 통계학과, 교수. 교신저자: choh@yu.ac.kr

잡음환경 하에서 음성구간을 효율적으로 추출하기 위한 다양한 방법이 시도되었다. Acero 등 (1993)는 신호의 델타 로그에너지와 표준화된 로그에너지를 사용하여 숨은마코프모형에 적용하였고, Abdulla (2002)는 웨이브렛 변환을 통한 잡음제거와 멜켑스트럼을 특징벡터로 이용한 숨은마코프모형을 사용하였다. 그러나 기존의 방법과는 달리 숨은마코프모형을 이용한 방식은 음성구간 추출과정에 있어 음성구간 시작 전과 끝부분에 존재하는 무음 구간에서도 계속적인 연쇄과정을 거침으로 인해서 계산비용을 증가시킨다. 특히 긴 무음구간을 가지는 경우에는 비효율적임이 알려져 있다. 숨은마코프모형에서 계산비용에 영향을 미치는 요인은 상태의 개수, 관측값의 차원, 관측치에 대한 분포를 혼합분포로 하는 경우 성분분포의 개수 등을 들 수 있다. 따라서 Abdulla (2002)에서의 13차 특징벡터를 저차원으로 낮추는 것은 비슷하거나 나은 구간추출의 효율성이 보장되는 경우 그 유용성이 기대된다. 본 논문에서는 음성구간 추출을 위한 1차원 특징벡터를 제시한다. 제안된 특징벡터는 인간의 인지에 바탕을 둔 멜켑스트럼에서 유도되었다. 또한, 제시된 특징벡터와 숨은마코프모형을 이용한 음성구간 추출의 효율성을 기존의 특징벡터인 멜켑스트럼과 파워에 대하여 비교 조사한다. 한편, 잡음환경 하에서, 제시된 특징벡터와 기존의 웨이브렛을 이용한 소음제거 방식의 음성구간 추출기능을 비교하여 보였다. 본 논문의 구성을 보면, 2절에서는 특징벡터의 유도와 숨은마코프모형에 대한 소개, 3절에서는 실험 및 결과, 그리고 마지막 절에서는 결론을 다루었다.

2. 특징벡터와 숨은마코브모형

2.1. 멜켑스트럼과 특징벡터

음성구간을 추출하는 것은 끝점검출이라고 불리워지기도 한다. 끝점에는 음성구간의 시작을 나타내는 시점, 그리고 마지막을 나타내는 종점이 있다. 흔히 사용되어지는 끝점검출 방법은 음성신호 에너지와 영교차율이다. 이 방법은 잡음수준이 낮은 환경에서는 효율적이나 잡음수준이 높은 환경 하에서는 비효율적임이 알려져 있다. 이러한 문제점을 해결하기 위하여 Teager 에너지, 웨이브렛을 이용한 방법과 로그에너지, 멜켑스트럼 특징벡터 등이 제안되었다.

이들 중 멜켑스트럼은 사람의 인지적 특성을 고려하는 멜주파수 척도를 사용하는 특징벡터이며, Abdulla (2002)는 멜켑스트럼을 이용하는 구간추출법을 제안하였다. 멜주파수 척도는 소리의 주파수에 대한 사람의 인지는 선형 척도가 아닌 로그 척도를 따른다는 사실을 고려한 것이다 (Ganchev 등, 2005). 멜주파수 척도는 1 KHz 이하의 주파수에 대하여는 선형적이며, 1 KHz 이상의 주파수에 대하여는 로그형태를 따르도록 정의된다. 즉, 1 KHz 이상인 주파수 f에 대하여 멜주파수는 흔히

$$\eta(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.1}$$

으로 주어지며 식 (2.1)은 인지실험으로부터 얻은 음성데이터를 로그 회귀선으로 적합하여 얻은 결과이다.

특징벡터를 얻는 절차는 입력된 음성 신호를 세부 구간으로 나누는 것부터 시작한다. 이때 세분화된 각 구간을 프레임이라고 한다. 음성신호의 특징을 위하여 프레임은 적당한 크기로 겹치도록 하는 것이 일반적이다. 입력된 음성신호 $s(1),\ldots,s(T)$에 대하여 프레임 $F(n),\ n=1,\ldots,N$은 다음과 같이 나타낼 수 있다.

$$F(n)=\{s(t):(n-1)\times(u-v)+1\le t\le(n-1)\times(u-v)+u\}. \tag{2.2}$$

여기서 u는 프레임 크기, v는 중첩된 구간의 크기이며 v는 u보다 작은 값이다. 또한 N은 전체 프레임의 수이다. 각 프레임 $n=1,2,\ldots,N$에 대하여 지정되는 주파수에 대응되는 크기는 퓨리에변환

$$X^k(n)=\sum_{s(t)\in F(n)} s(t)e^{\frac{-j2\pi tk}{u}},\quad k=0,1,\ldots,u-1 \tag{2.3}$$

으로 얻어진다. 여기서 패스트 퓨리에변환을 가정하므로 u는 2의 거듭제곱수로 두며, j는 허수 단위수 $\sqrt{-1}$이다. 식 (2.3)의 퓨리에변환을 거쳐 얻어진 주파수 영역의 값은 멜필터링을 통하여 멜주파수 영역으로 변환된다. 여기서는 $B=20$개의 삼각필터를 사용하며, $i=1,2,\ldots,B$에 대하여 다음으로 주어진다 (Ganchev 등, 2005).

$$H_i(k)=\begin{cases}\dfrac{k-b_{i-1}}{b_i-b_{i-1}}, & b_{i-1}\le k\le b_i,\\ \dfrac{b_{i+1}-k}{b_{i+1}-b_i}, & b_i\le k\le b_{i+1},\\ 0, & \text{그 이외.}\end{cases} \tag{2.4}$$

단, 샘플링 주파수 F_s에 대하여 필터 경계값 b_i는 다음으로 주어진다.

$$b_i=\left(\frac{u}{F_s}\right)\eta^{-1}\left(\eta(f_{low})+i\frac{\eta(f_{high})-\eta(f_{low})}{B+1}\right), \tag{2.5}$$

식 (2.5)에서 $\eta^{-1}(\cdot)$는 식 (2.1)에 주어지는 $\eta(\cdot)$의 역함수이며, f_{low}와 f_{high}는 고려 대상이 되는 주파수의 최저 및 최고 값이다. 이때 B개의 멜주파수 크기는 주파수 크기와 삼각필터의 합의 로그변환으로 주어진다.

$$Y^i(n)=\log_{10}\left(\sum_{k=0}^{u-1}|X^k(n)|H_i(k)\right),\quad i=1,2,\ldots,B. \tag{2.6}$$

Abdulla (2002)는 프레임 $n=1,2,\ldots,N$에 대하여 멜주파수 크기의 합으로 파워를 정의하였다.

$$P(n)=\sum_{i=1}^{B}Y^i(n). \tag{2.7}$$

프레임 n에 대하여 멜주파수 켑스트럴 계수 MFCC는 코사인변환을 이용하여 얻어진다.

$$D^l(n)=\sqrt{\frac{2}{B}}\sum_{i=1}^{B}Y^i(n)\cos\left(\frac{l\pi(i-0.5)}{B}\right),\quad l=0,1,\ldots,L. \tag{2.8}$$

Abdulla (2002)는 $L = 12$일 때의 멜주파수 켑스트럴 계수와 파워로 구성된 13차 벡터 $D(n) = (D^1(n), D^2(n), \ldots, D^{12}(n), P(n))$을 음성구간 추출을 위한 특징벡터로 제안하였다. 또한, 신호의 변동성을 보정하기 위하여 $D(n)$의 각 성분 별로 표준화하여 사용하였다. 표준화에 관해서는 Haeb-Umbach (1999)를 참조하기 바란다. 여기서는 파워가 추가된 $D(n)$을 멜켑스트럼이라고 부르기로 한다.

멜켑스트럼 $D(n)$은 12개의 주파수 영역에 대한 에너지의 크기 $D^l(n)$와 신호 전체의 에너지 크기 $P(n)$으로 구성되어 있다. 멜켑스트럼 $D(n)$을 음성구간 추출을 위한 특징벡터로 사용하는 경우에 저잡음의 환경에서는 파워 $P(n)$이 음성구간 추출의 주요 정보원으로 사용되며, 잡음환경에서는 영역별 에너지 $D^l(n)$이 음성구간 추출에서 주요한 역할을 하는 것으로 판단된다. 잡음환경에서 파워만을 사용하는 경우 잡음의 영향으로 음성구간 추출에서 성능이 저하되는 것으로 보여진다. 따라서 파워에 추가된 잡음을 제거하는 하나의 방법으로 파워의 이동평균을 고려할 수 있다. 각 프레임, $n = 1, 2, \ldots, N$에 대하여 이동평균은 다음과 같이 정의한다.

$$\mathrm{AP}(n) = \frac{P(n-a) + \cdots + P(n+a)}{2a+1}, \tag{2.9}$$

여기서 a은 음이 아닌 정수이며, $P(N+1) = P(N+2) = \cdots = 0$, $P(0) = P(-1) = \cdots = 0$으로 정한다. 식 (2.9)에 의해 주어지는 이동평균을 평균파워라고 부르기로 한다.

2.2. 신호의 잡음제거

Abdulla (2002)는 소음환경 하에서의 음성구간 추출에서 고주파 잡음효과를 줄이는 방안으로 웨이브렛을 적용하였다. 웨이브렛을 이용한 잡음 제거의 방법은 소리신호의 데이터에 대한 웨이브렛 계수 중 고주파 영역에 해당하는 웨이브렛 계수를 제거한 나머지 계수만을 사용하여 소리신호를 표현하는 것이며 Abdulla (2002)는 이 방법의 우수성을 실증적으로 보였다. 한편 식 (2.9)에서 정의된 평균파워는 이동평균의 기본적 성질에 의해 평활의 성질을 가지고 있고 이는 곧 신호에 포함된 고주파 부분을 제거하는 효과를 가지고 있음을 나타낸다. 평활의 정도는 이동평균의 크기 $2a+1$에 의해 결정된다. 이동평균은 몇 개의 주변 값들의 평균만으로 얻어지므로 웨이브렛의 계수를 구하는 것보다 계산이 쉽고 빠르며 웨이브렛 변환 적용시 계산을 위한 추가적 메모리 할당과 같은 절차가 필요없는 장점이 있다. 그림 2.1은 신호대잡음 비율(signal-to-noise ratio: SNR)이 5dB인 신호의 예에서 파워와 크기 $2a+1 = 5$인 평균파워를 나타낸 것이며 평균파워가 파워에 비해 평활되어 있는 모습을 볼 수 있다.

본 논문에서는 다양한 크기의 백색잡음 환경에서 웨이블렛을 이용한 방법과 평균파워를 이용한 방법의 음성구간 추출의 성능을 비교조사한다.

2.3. 숨은마코프모형의 소개

숨은마코프모형(hidden Markov model, HMM)은 음성인식뿐만 아니라 영상인식, 유전학 등 여러 분야에서 널리 사용되어지는 확률모형이다. 숨은마코프모형은 관측할 수 없

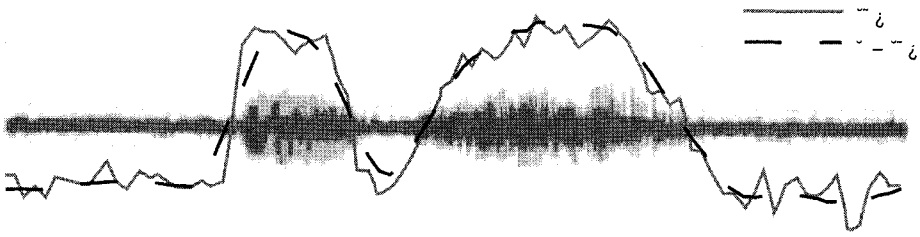

그림 2.1: SNR = 5dB인 음성신호에 대한 파워와 $2a+1=5$인 평균파워

는 확률과정인 마코프연쇄 $\{Q_n\}$과 관측할 수 있는 확률과정 $\{Y_n\}$으로 구성된 이중확률과정이며 관측 가능한 값은 주어진 시각에서 마코프연쇄의 상태에 의존한다고 가정한다. 본 논문에서는 관측가능한 확률과정은 평균파워 AP(n)과 같은 특징벡터 열에 해당하며, 관측되지않는 마코프연쇄는 각 프레임의 상태 즉 음성구간 혹은 잡음구간 등과 같은 상태에 관한 것이다. 숨은마코프모형은 상태와 상태 간의 상태변이 확률, 주어진 상태에 대한 기호관측확률로 구성된다. 숨은마코프모형에서 관측값에 대한 분포의 형태에 따라 이산형 혹은 연속형으로 구분하며 연속형인 경우에는 흔히 혼합정규분포를 가정한다.

마코프연쇄의 상태의 개수는 $\{1,2,\ldots,I\}$로 주어진다고 하고 상태변이 확률 행렬은 $A=(a_{ij})_{I\times I}$으로 나타내자. 그리고 상태 $Q_n=i$가 주어진 경우 관측치 Q_n에 대한 분포는 M개의 성분을 가지는 혼합정규분포를 따른다고 하자. 즉, 확률밀도함수

$$b_i(o_n)=\sum_{k=1}^{M} c_{ik}N(o_n,\mu_{ik},\Sigma_{ik}) \tag{2.10}$$

를 가진다고 하자. 단, $N(\cdot,\mu,\Sigma)$는 평균벡터가 μ, 분산공분산행렬이 Σ인 다변량 정규분포의 확률밀도함수이며 c_{ik}는 혼합가중치로서 $c_{i1}+\cdots+c_{iM}=1$, $c_{ik}\geq 0$ 이다. 이와 같은 주어진 상태에 따른 확률밀도함수의 집합을 B라고 나타낸다. 초기상태 확률집합은 $\Pi=\{\pi_i\}$으로 나타내자. 단, $\pi=\Pr\{Q_1=i\}$, $1\leq i\leq I$이다. 숨은마코프모형을 $\lambda=(A,B,\Pi)$로 나타내기로 하자. 모형 λ가 주어진 경우에 관측열 $o=(o_1,\ldots,o_N)$의 확률은 전진절차 혹은 후진절차로 구할 수 있다. 또한 모형 λ와 관측열 $o=(o_1,\ldots,o_N)$이 주어진 경우 상태변이 확률 행렬과 관측치 분포는 Baum-Welch 알고리즘에 의해 개정될 수 있다. 또한 관측열 $o=(o_1,\ldots,o_N)$이 주어진 경우에 마코프연쇄의 상태열 $Q=(Q_1,\ldots,Q_N)$에 대한 추정은 Viterbi 알고리즘에 의해 이루어질 수 있다. 숨은마코프모형에 관한 자세한 내용은 Rabiner와 Juang (1993)을 참조하기 바란다.

3. 실험 및 결과

실험을 위한 음성데이터는 분리형 에어콘 잡음과 같은 저수준의 생활 잡음이 있는 조용한 환경 (SNR> 30)에서 녹음된 음성과 이들 음성에 백색잡음을 포함시킨 (SNR =

표 3.1: 평균파워의 프레임 개수 에 따른 시점/종점 검출률(SNR = 5dB)

오차범위	2a+1=3		2a+1=5		2a+1=7	
(ms)	시점	종점	시점	종점	시점	종점
30	86.67	55.56	87.78	58.89	86.67	46.67
50	88.89	62.22	91.11	65.56	94.44	63.33
70	96.67	72.22	96.67	75.56	97.78	68.89

5, 10, 15) 음성으로 이루어졌다. 실험 데이터는 총 90개의 단어를 녹음한 데이터로, 여성화자 5명과 남성화자 4명의 총 9명이 각각 10개의 단어를 녹음하여 얻었다. 모든 단어는 시작 부분에 파열음 'ㅍ'을 가지고 있으며, 2음절을 가지는 80개의 단어와 3음절을 가지는 10개의 단어로 구성되었다. 디지털 녹음의 형식은 채널수 1, 샘플링주파수 11,025 Hz, 샘플크기 16bit 인 PCM 형식이다. 한편, 프레임 크기는 20ms, 프레임의 중복크기는 10ms로 설정하였다.

HMM의 모형 설정은 마코프연쇄과정의 설정과 상태가 주어졌을 때의 관측분포의 설정으로 구성된다. 먼저 마코프연쇄과정의 상태의 개수는 $I = 5$로 하였으며, 좌에서 우로만 상태이동을 허용하는 Bakis모형을 가정하였다. 영어단어에 대한 음성구간 추출에서 Abdulla와 Kasabov (1999)에서는 $I = 7$개의 상태를, Abdulla (2002)에서는 $I = 3$개의 상태를 사용하였으나 본 논문에서는 5개의 상태를 사용하였다. 이는 강세중심언어인 영어와는 달리 한국어가 음절중심 언어임을 고려한 것이다. 한편, 주어진 상태에 대한 관측치의 분포는 혼합성분의 개수 $M = 1$인 혼합정규분포를 가정하였다. 혼합성분의 개수를 $M = 2$ 이상인 경우도 고려할 수 있으나, 본 연구의 실험에서 추출율의 향상을 가져오지 않았다. 즉, 성분의 개수를 두 개 이상으로 하더라도 음성구간추출에서 개선효과가 나타나지 않는 반면 계산비용만 증가하였다. 이러한 결과는 선행연구인 Abdulla (2002)의 실험결과와도 일치한다. 상태변이 확률 행렬 와 식 (2.10)의 각 관측분포에 대한 평균 μ_{ik}와 분산공분산 Σ_{ik}행렬 의 추정은 Baum-Welch 알고리즘에 의한다. 다만, 모형 $\lambda = (A, B, \Pi)$의 초기값 설정은 Abdulla (2002)의 방법을 사용하였다. 한편 음성구간을 추출하기 위하여 주어진 HMM에 대하여 Viterbi 알고리즘을 이용하여 각 프레임에 대응되는 상태를 추정한다.

식 (2.9)의 평균파워를 특징벡터로 사용하는 경우에 이동평균을 위한 프레임의 개수를 먼저 정해야한다. 이를 위하여 이동평균의 프레임의 개수가 각각 $2a + 1 = 3, 5, 7$인 평균파워를 사용하여 음성구간을 추출할 때의 검출율을 조사하여 보았다. 표 3.1은 이에 대한 결과이다. 오차의 범위는 각 단어의 기준 시점 (또는 종점)에 대하여 추정된 시점 (또는 종점)과의 시간 거리를 나타내며 단위는 밀리초(ms)이다. 단어의 기준시점과 기준종점은 각 단어의 음성파형을 시각 및 청각으로 확인하여 정해진다. 프레임의 개수가 $2a + 1 = 5$일 때 다른 경우에 비해 전반적으로 검출율이 더 높은 것으로 나타났다. 따라서 평균파워를 위한 프레임의 개수는 $2a + 1 = 5$를 사용한다. 표 3.1에서 사용한 음성데이터는 비교적 잡음이 큰 SNR = 5dB의 백색잡음이 추가된 데이터이다.

평균파워와 멜켑스트럼, 파워에 대한 음성구간 검출율을 비교하기 위하여, 신호대잡음비 SNR = 15, 10, 5dB인 상태와 무잡음 상태에서 각 단어에 대하여 추정된 시점 또는 종

표 3.2: 특징벡터와 SNR에 따른 시점/종점 검출백분률

특징벡터	오차범위	무잡음		15dB		5dB			
	(ms)	시점	종점	시점	종점	시점	종점	시점	종점
멜켑스트럼	30	97.78	78.89	91.11	72.22	84.44	65.56	80.00	51.11
	50	98.89	91.11	93.33	82.22	87.78	80.00	84.44	64.44
	70	100.00	95.56	97.78	92.22	94.44	86.67	88.89	72.22
파워	30	97.78	88.89	91.11	70.00	86.67	57.78	86.67	46.67
	50	100.00	93.33	92.22	76.67	88.89	68.89	88.89	53.33
	70	100.00	97.78	96.67	86.67	95.56	81.11	95.56	62.22
평균파워	30	97.78	86.67	91.11	72.22	86.67	66.67	86.67	60.00
	50	100.00	93.33	95.56	80.00	94.44	73.33	91.11	67.78
	70	100.00	97.78	97.78	93.33	96.67	85.56	96.67	77.78

표 3.3: 특징벡터별 음성구간검출에 걸린 시간

검출시간	멜켑스트럼	파워	평균파워
시간(ms)	0.090	0.082	0.083

점이 기준 시점 또는 기준 종점과의 시간거리가 30, 50, 70ms 이내인지를 조사하였다. 결과비교 방법은 Seok과 Bae (1999)을 참고하였다. Seok과 Bae (1999)는 음소를 골고루 포함하는 40개의 단어를 사용하여, 신호대 잡음비 SNR = 20, 10dB인 상태와 잡음을 추가하지 않은 무잡음 상태에서 오차범위를 25ms부터 75ms 사이에서 12.5ms 간격으로 변화시키면서 검출된 시점과 끝점이 오차범위에 들어가는 경우를 백분율로 표시하였다. 표 3.2는 이 실험에 대한 검출백분률이다. 표 3.2에서 전반적으로, 평균파워가 멜켑스럼이나 파워에 비해 시점과 종점에서 모두 음성 검출률이 높게 나타났다. 멜켑스럼은 파워를 포함하는 특징벡터임에도 불구하고 파워에 비해 검출율이 크게 높지 않고 SNR의 값이 작아지는 경우 즉 잡음이 커지는 경우 시점에서 파워 보다 검출률이 낮아지는 경우도 있다. 이에 비해 평균파워는 주어진 모든 SNR에 대하여 시점과 종점의 검출률이 높게 나타났다. 다만, SNR = 10dB, 오차범위 50ms에서의 종점 검출에서만 멜켑스트럼이 평균파워에 비해 약간 높은 검출율을 나타내고 있다.

한편, 멜켑스트럼, 파워, 평균파워에 대하여 음성구간검출에 필요한 평균계산시간을 조사하였다. 총 90개 단어 각각에 대하여 계산시간을 구한 후 이를 평균하여 결과를 비교해 보았다. 표 3.3은 위의 세 가지 특징 벡터의 사용에 따른 음성구간추출에 걸린 계산시간을 나타낸 것이다. 표 3.3에서 볼 수 있듯이 멜켑스트럼은 13차 벡터를 사용함으로 인해서 1차 벡터만을 사용하는 파워 및 평균파워의 두 방법에 비해 더 많은 계산시간이 소요됨을 알 수 있다. 파워와 평균파워를 사용한 방법은 서로 계산속도가 비슷함을 알 수 있다. 이는 이동평균에 사용된 계산시간이 아주 작음을 나타낸다

잡음환경 하에 불규칙성분을 제거하기 위한 다른 방법으로 웨이브렛을 사용하는 방법이 있으며, Abdulla (2002)는 웨이브렛을 이용한 잡음제거 방식의 효율을 조사하였다. 본

표 3.4: 평균파워와 웨이브렛 방법과의 시점/종점 검출백분율

SNR (dB)	오차범위 (ms)	평균파워		웨이브렛	
		시점	종점	시점	종점
15	30	91.11	72.22	91.11	65.56
	50	95.56	80.00	95.56	73.33
	70	97.78	93.33	97.78	84.44
10	30	86.67	66.67	88.89	60.00
	50	94.44	73.33	92.22	67.78
	70	96.67	85.56	96.67	80.00
5	30	86.67	60.00	86.67	50.00
	50	91.11	67.78	90.00	58.89
	70	96.67	77.78	96.67	66.67

실험에서는 잡음제거를 위하여 Abdulla (2002)의 고주파 웨이브렛 계수 제거 방법을 사용하여 파워평균을 이용하는 방법과 비교하여 보았다. 표 3.4는 웨이브렛과 파워평균을 사용하였을 때의 음성구간 검출백분율이다. 표 3.4에서 평균파워을 사용하는 경우가 웨이브렛을 사용하는 것보다 대체적으로 높은 구간 검출율을 보여준다.

4. 결론

본 논문에서는 음성구간 검출을 위해 멜켑스트럼에 기초한 1차원 특징벡터인 평균파워를 제시하였다. 평균파워와 기존의 특징벡터인 멜켑스트럼과 파워에 대하여 음성구간추출율을 실제 음성데이터를 이용하여 비교하여 보았다. 다양한 백색잡음 수준에 대하여 평균파워가 비교대상이 되는 특징벡터에 비해 음성구간 검출 능력이 뛰어남을 파열음 'ㅍ'을 시점으로 가지는 단어의 음성데이터에 대한 실험으로 보였다. 또한 평균파워를 사용하는 방법이 멜켑스트럼을 사용하는 방법과 비교했을 때 시점과 종점 검출에서 계산 비용이 적음을 보였다.

한편, 평균파워는 신호에 섞인 고주파 잡음을 제거하는 기능이 있으므로 이를 웨이브렛을 이용한 잡음 제거 방법과 비교하여 보았다. 다양한 수준의 백색잡음의 음성 데이터에 대한 실험에서 평균파워가 전체적으로 웨이브렛에 의한 방법보다 음성구간 검출율이 높음을 보였다. 즉, 상대적으로 복잡한 기법인 웨이브렛을 사용하는 것보다 단순 계산식에 의해 표현되는 이동평균을 이용하는 것이 더 효율적임을 보였다.

본 논문에서 제시된 특징벡터 평균파워는 다양한 백색잡음 수준에서, 끝점 검출에 어려움이 많다고 알려진 파열음을 끝점으로 가진 단어에 대한 음성구간 검출에 효율적임을 보였다. 따라서 음성구간 검출이 상대적으로 잘된다고 알려진 다른 음들의 끝점을 가진 단어의 구간 추출에서도 같은 결과가 기대되며 이는 후속연구에서 이루어질 수 있을 것이다.

참고문헌

Abdulla, W. H. (2002). HMM-based techiques for speech segments extraction. *Scientific Programming*, **10**, 221–239.

Abdulla, W. H. and Kasabov, N. K. (1999). Two pass hidden Markov model for speech recognition systems. In *Proceeding of the ICICS'99.*

Acero, A., Crespo, C., Torre, C. de la and Torrecilla, J. C. (1993). Robust HMM-based endpoint detector, In *Proceeding of the EuroSpeech*, **3**, 1551–1554.

Ganchev T., Fakotakis N. and Kokkinakis G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. In *Proceeding of the* 10^{th} *International Conference on Speech and Computer, SPECOM 2005,* **1**, 191–194.

Haeb-Umbach, R. (1999). Investigations on inter-speaker variability in the feature space. In *Proceeding of the IEEE ICASSP'99*, **1**, 397–400.

Rabiner, L. R. and Juang, B. H. (1993). *Fundamentals of Speech Recognition.* Prentice Hall PTR, New Jersey.

Seok, J. W. and Bae, K. S. (1999). Endpoint detection of speech signal using wavelet transform. *The Journal of the Acoustical Society of Korea*, **18**, 57–63.

[2007년 12월 접수, 2008년 3월 채택]

A New Feature for Speech Segments Extraction with Hidden Markov Models

Jeong Woo Hong[1], Chang Hyuck Oh[2]

Abstract

In this paper we propose a new feature, average power, for speech segments extraction with hidden Markov models, which is based on mel frequencies of speech signals. The average power is compared with the mel frequency cepstral coefficients, MFCC, and the power coefficient. To compare performances of three types of features, speech data are collected for words with explosives which are generally known hard to be detected. Experiments show that the average power is more accurate and efficient than MFCC and the power coefficient for speech segments extraction in environments with various levels of noise.

Keywords: Average power; hidden Markov model; Mel frequency cepstral coefficients; power coefficient.

1) Graduate Student, Department of Statistics, Yeungnam University, Gyungsan 712-749, Korea. E-mail: brbravo9915@yu.ac.kr

2) Professor, Department of Statistics, Yeungnam University, Gyungsan 712-749, Korea. Correspondence: choh@yu.ac.kr