잡음환경의 ASR 성능개선을 위한 음성강조 파라미터

차영동, 김영섭, 허강인
동아대학교 전기전자컴퓨터공학부

Using speech enhancement parameter for ASR

Young-dong Cha, Young-sub Kim, Kang-in Hur
School of Electrical, Electronic, and Computer Eng., Dong-A University

요 약

음성인식시스템은 사람이 별도의 장비 없이 음성만으로 시스템의 사용이 가능한 편리한 장점을 지니고 있으나 여러 가지 기술적인 어려움과 실제 환경의 낮은 인식률로 폭넓게 사용되지 못한 상황이다. 그 중 배경잡음은 음성인식의 인식률을 저하시키는 원인으로 지적 받고 있다. 이러한 잡음환경에 있는 ASR(Automatic Speech Recognition)의 성능 향상을 위해 외측억제 기능 이 추가된 파라미터를 제안한다. ASR 에서 널리 사용되는 파라미터인 MFCC을 본 논문에서 제안한 파라미터와 HMM를 이용하여 인식률을 비교하여 성능을 비교하였다. 실험결과를 통해 제안된 파라미터의 사용을 통해 잡음환경에 있는 ASR의 성능 향상을 확인할 수 있었다.

I. 서 론

음성은 가장친숙한 의사표현 방법이며 별다른 도구 없이 사용하는 가장 편리한 매체 이다.그러나 음성은 화자간의 영향이나 주변 환경에 따라 특징변화가 심하기 때문에 다양하고 폭넓게 사용되지 못하는 상황에 있다. 음성인식시스템에서의 입력 신호는 순수한 음성신호 이외에 다양한 배경 잡음과 섞여 있어 인식 하기 때문에 인식률의 저하를 가져오게 된다. 이러한 잡음의 스펙트럼의 형상이 음성이나 음성의 피치성분과 유사 한 것이 많기 때문에 음성의 인식률을 현저하게 저하시키는 원인에 되고 있다.[5]

이러한 잡음환경에 있는 ASR(Automatic Speech Recognition)의 성능 향상을 위해 외측억제 기능의 파라미터를 제안한다. 신경상호간의 기능인 외측억제를 추가하여 본 실험의 파라미터를 생성하였다. ASR에서 널리 사용되는 파라미터인 MFCC(Mel-Frequency Cepstrum Coefficient)를 기본 파라미터로 설정하고 본 논문에서 제안한 파라미터와 실제적인 음성인식에 많이 사용하는 HMM를 이용하여 인식률을 비교하여 성능을 비교 분석하였다.

본 논문의 2장에서는 MFCC와 상호억제와 이론적인 배경과 제안된 파라미터를 설명하고 3장에는 다양한 잡음환경에서 HMM를 이용하여 학습한 실험결과를 나타내었고 4장에서 결론 및 향후 과제로 구성하였다.

II. 특징 파라미터

1. 특징파라미터

1) MFCC(Mel-Frequency Cepstral Coefficient)

음성인식에서 가장 중요한 과정은 음성의 특징을 잘 찾아 낼 수 있는 특징을 추출하는 것이다. 음성은 시간이나 문맥에 따라 변화가 심한 것이 특징이다.

Mel은 인간의 청각특성이 저주파영역에서는 민감하고, 고주파영역에서는 둔감한 반응을 log-sacle과 유사한 형태로 표현한 것이 Mel 단위이다. 그림 1에서 음성신호로부터 MFCC를 추출하는 과정을 나타내었다.

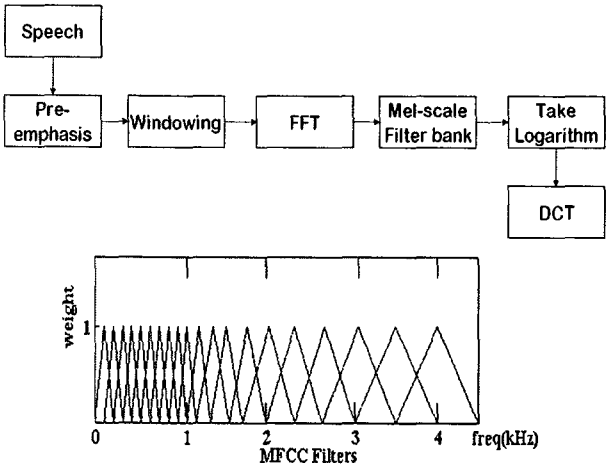


그림 1. MFCC 블럭도

음성신호는 A/D 변환을 거쳐 고역강조(Pre-emphasis)를 통해 입술의 방사에 의해 20dB/decade 로 감쇄되는 것을 보상하게 되어 음성의 성도 특징만을 취하게 된다. 음성의 특성이 고정되어 있다고 가정하는 20-30ms의 길이를 갖는 윈도우함수를 씌워 프레임단위로 나눈 다음 프레임별로 FFT를 이용해 주파수영역으로 변환한다. 변환된 주파수 대역을 인간의 청각 특성을 반영한 여러 개의 Mel-scale Filter bank 로 나누고 각각의 뱅크의 에너지를 계산한 후 계산된 에너지에 로그를 취한다음 DCT(Discrete Cosine Transform)를 하게 되면 최종적인 MFCC가 프레임 별로 얻어진다.

2. 외측억제 (Lateral Inhibition)

1) 외측억제

외측억제는 생물학적시스템에서 감각적 수용의 일반적인 현상에 깊게 관여되어있다. 외측억제의 근원적인 메커니즘은 뉴런들의 상호연결성이다. 사람의 시각에서 우리의 밝기 지각은 대상에서 반사되는 빛의 강도를 그대로 반영하지 않는다. 대신에 지각과정이 환경의 중요한 측면을 부각시키려 하기 때문에 대상의 모서리나 윤곽을 강조하는데 그 과정에 관여하는 것이 외측억제이다. 이 외측억제현상 때문에 그림 2 의 헤르만 격자에서 흰색 교차점에는 검은색 반점이 보인다.

이런 외측억제를 청각에 적용 시키는 연구가 진행되어왔었다.[3][4][6] 외측억제는 시간적이고 공간적인 영역에서 변화하는 부분을 날카롭게 만든다. 그래서 만약 이 외측억제 메커니즘을 스펙트럼에서 사용한다면, 외측억제는 피크부분은 더 높게 밸리 부분은 더 낮게 내려가게 되어 날카롭게 만든다.

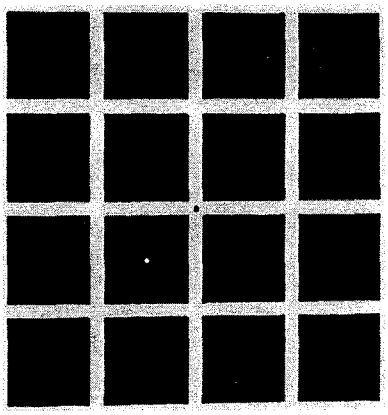


그림 2. 헤르만 격자(Hermann grid)

2) 잡음모델

이 장에서는 부가되는 백색잡음 같은 잡음들을 수식으로 나타내보았다. 음성신호와 잡음이 섞인 신호는 식(1)과 같이 표현된다.

$$x(k) = s(k) + n(k) \tag{1}$$

음성과 잡음이 둘 다 안정적 (stationary) 이라고 가정하고 음성 신호에서 잡음이 백색잡음이고 상관성이 없다면 식 (2) 처럼 표현 가능하다.

$$r_x(k) = r_s(k) + \sigma^2\delta(k) \tag{2}$$

식 (2) 에서 $r_x(k)$, $r_s(k)$, $\sigma^2\delta(k)$ 는 각각 잡음이 섞인 신호, 잡음이 섞이지 않은 깨끗한 신호, 잡음의 자기상관 수열 이다.

$$\begin{aligned} P_x(w) &= P_s(w) + P_n(w) \\ &= P_s(w) + \mu \end{aligned} \tag{3}$$

식 (3) 에서 $P_x(w)$, $P_s(w)$, $P_n(w)$ 은 각각 $x(k)$, $s(k)$, $n(k)$의 파워밀도스펙트럼들이고 μ 는 상수이다. 음성과 잡음은 시간 제한된(time-limited) 프레임에서만 안정(stationary)이기 때문에 우리는 단구간 파워스펙트럼을 사용한다.

$$\begin{aligned} P_x^{(i)}(w) &= P_s^{(i)}(w) + P_n^{(i)}(w) \\ &= P_s^{(i)}(w) + \mu^{(i)} \end{aligned} \tag{4}$$

식 (4) 에서 i 는 프레임의 인덱스이다. 이 경우의 음성의 비안정(non-stationary) 신호를 거의 안정(stationary)하게 만들 수 있다.

3) 공간적 외측억제 함수

공간적 외측억제 함수 (Function of Spatial Lateral Inhibition) FSLI 에서 $Z(w,\Omega)$로 표현되고 FLSI 가 모든 Ω에서 아래와 같은 적분식을 만족한다면 다음과 같이 표현가능하다.

$$\int_{-\infty}^{\infty} Z(w,\Omega)\,dw = 0 \quad \forall\ \Omega \tag{5}$$

식(6)은 공간적 외측억제 함수와 함께 잡음이 섞인 음성 입력의 파워스펙트럼을 컨벌루션 한 것이다.

$$\begin{aligned}\widehat{P}_x(\Omega) &= \int_{-\infty}^{\infty} P_x(w)Z(\Omega-w,\Omega)dw \\ &= \int_{-\infty}^{\infty} P_s(w)Z(\Omega-w,\Omega)dw \\ &\quad + \int_{-\infty}^{\infty} P_n(w)Z(\Omega-w,\Omega)dw \\ &= \widehat{P}_s(\Omega) + \mu\int_{-\infty}^{\infty} Z(\Omega-w,\Omega)dw \\ &= \widehat{P}_s(\Omega)\end{aligned} \tag{6}$$

식 (6)에서 잡음이 섞인 신호의 파워스펙트럼에서 입력신호와 잡음은 이론적으로 일치함을 알 수 있다.

3. 제안한 파라미터

1) 제안한 파라미터의 설계

먼저 음성신호는 16 kHz로 샘플링하고 고역강조(Pre-emphasis)를 거친다. 입력된 음성은 20ms 해밍윈도우를 거친 다음 FFT 파워 스펙트럼을 구하게 되고 공간적 외측억제 함수 (Function of Spatial Lateral Inhibition)의 성질을 같은 임펄스 응답을 구하여 시스템을 설계하고 자연로그에 절대 값을 취한다음 DCT(Discrete Cosine Transform) 연산을 취하면 최종적인 음성의 파라미터가 생성이 된다. 생성된 파라미터는 HMM을 통해 학습되고 음성별로 학습된 데이터에 임의의 음성을 입력하여 정확하게 인식 되었는지 확인한다

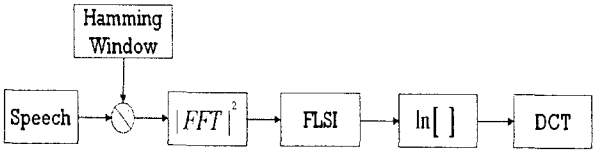


그림 3. 제안한 파라미터의 블럭도

Ⅲ. 실험 및 결과

1. 분석 조건

실험조건으로는 3명의 각각 다른 화자가 각각 동일 숫자를 두 번씩 발음하고 일반 마이크로 녹음한 10개의 숫자음(영, 일, 이, 삼, 사, 오, 육, 칠, 팔, 구)을 사용하였다. 특징은 표 와 같다.

A/D Convert	16kHz, 16bit
window	hamming
window length	20ms
shifting period	10ms
feature parameter	13th MFCC

표 . 1

A/D Convert	19.98 KHz 16bit
pre-filter	anti-aliasing filter
pre-emphasis	none
duration	235 sec
length (uncompressed)	approx 9 Mbyte

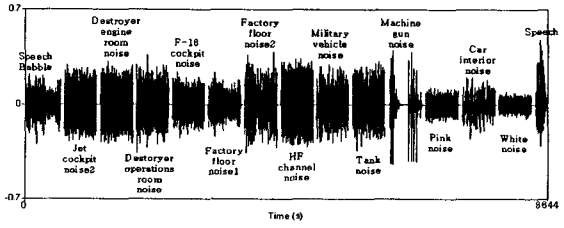
표 . 2

NOISEX92의 데이터베이스를 이용하였고 특징은 표과 같다. 본 데이터는 표 3 과 같은 다양한 종류의 잡음들로 구성되어 있으며 그 중에서 5개의 잡음 데이터를 음성신호에 추가하여 인식률 비교하였다.

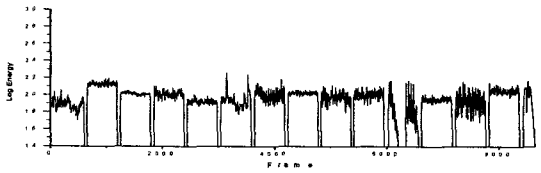
19.98 KHz - 16 bit으로 anti-aliasing 필터 처리된 것을 16KHz - 16비트로 재 샘플링하여 실험에 이용하였다.

	Noisex-92
1	Speech babble
2	Jet cockpit noise2
3	Destroyer engine room noise
4	Destroyer operations room noise
5	F-16 cockpit noise
6	Factory floor noise1
7	Factory floor noise2
8	HF channel noise
9	Military vehicle noise
10	Tank noise
11	Machine gun noise
12	Pink noise
13	Car interior noise
14	White noise

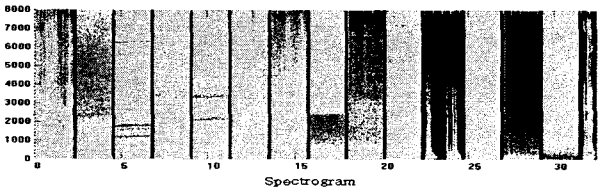
표 . 3 NOISEX-92



(a) Source Signal



(b) Log Energy



(c) Spectrogram

그림4. 잡음데이터의 파형, 로그에너지, 스펙토그램

그림 4에서 잡음데이터의 특징들을 살펴 볼 수 있다.

2. 특징 파라미터의 인식 결과

분석 조건에 해당하는 입력음성 데이터에 5개의 잡음을 각기 다른 SNR(5dB, 15dB, 25dB)로 첨가하여 특징 파라미터를 추출하고 총 150개의 데이터를 HMM을 이용하여 학습시킨 뒤 학습인식실험을 수행 하였다. 실험의 정확도를 위해 총 3번의 인식테스트를 한 후 그 값을 모두 더해서 평균값을 인식률로 결정하였다. 계산의 편의를 위해 소수점 2자리에서 반올림 한 값을 사용 했다.

	MFCC	제안한 파라미터
5dB	98.67%	99.11%
	(444/450)	(446/450)
15dB	94.89%	96.22%
	(427/450)	(433/450)
25dB	89.56	93.56%
	(403/450)	(421/450)

표 4. 인식결과

Ⅳ. 결 론

본 논문에서는 기존의 방식에 외측억제라는 개념을 도입하여 잡음 감소 효과가 있는 파라미터를 제안하였다. 또한 해당 프로그램을 시뮬레이션을 하여 기존의 파라미터와 성능비교를 하였다.

다양한 잡음 환경에서 기본적인 특징 파라미터는 잡음레벨이 강할수록 인식률의 저하가 두드러지는 반면에 특히 25dB 지점에서는 인식률이 90% 이하로 떨어지고 있으며 급격하게 인식률이 떨어짐을 볼 수 있다. 제안된 파라미터는 25dB 지점에서도 약 93%의 인식률을 지니고 있음을 알 수 있다. 이에 제안된 파라미터는 잡음에 비교 파라미터인 MFCC보다 효과적임을 알 수 있다.

그러나 고립단어의 인식이기 때문에 인식률이 상당히 높게 나온 것은 사실이다. 또한 파라미터의 잡음 제거효과가 상대적으로 크지 않았던 것 같다. 추후에 잡음에 대한 분석과 지속적인 연구를 하여 연속 단어에서도 높은 성능 목표로 할 것이며 앞으로 홈네트워킹의 시스템에 적용될 경우 다양한 소음에도 높은 인식률을 기대할 수 있을 것이다.

참 고 문 헌

[1] L.R.Rabiner, R.W.Schafer : "Digital Processing of Speech Signals", PRENTICE HALL

[2] Lu Xugang, "Lateral inhibition mechanism in computational auditory model and it's application in robust speech recognition," IEEE trans. Video Techno. vol. 6, pp.243-250,1996.

[3] Shihab A. Shamma, Speech processing in the auditory system I: The representation of speech sounds in the responses of auditory nerve,J, Acoust.Soc.Am.,78(5),P1612-1621,1985

[4] Shihab A. Shamma, Speech processing in the auditory system II: Lateral inhibition and central processing of evoked activity in the audio nerve ,J,Acoust.Soc.Am.,78(5),P1612-1621,1985

[5] Yadong Wu, Yan Li, "Robust speech/non-speech detection in adverse condition using the Fuzzy polarity correlation method", Systems, Man, and Cybernetics, 2000 IEEE International Conference on ,Vol. 4, 8-11 Oct.pp. 2935-2939, 2000.

[6] T.Houtgast "Psychophysical Evidence for Latral Inhibition in hearing"Acoust.Soc.Am., P1885-1894 ,1972