잡음환경에서의 음성인식을 위한 변이특성을 고려한 파라메터

박진영* · 이광석** · 고시영*** · 허강인*

*동아대학교 전자공학과 · **진주산업대학교 전자공학과 · ***경일대학교

Parameter Considering Variance Property for Speech Recognition in Noisy Environment

Jin-Young Park* · Kwang-Seok Lee** · Si-Young Koh*** · Kang-In Hur*

*Dept. of Electronic Engineering, Dong-A University

**Dept. of Electronics Engineering, JinJu National University

***Dept. of Electronic & Information Engineering, Kyung-il University

E-mail : erajjang@donga.ac.kr

요 약

본 논문에서는 음석인식 시스템을 구현함에 있어서 잡음의 영향에 강인한 특성을 가지는 효과적인 음성특징 파라미터에 대해 제안한다. ASR(Automatic Speech Recognition)에 사용되는 가장 기본적인 파라미터인 MFCC와 DCT를 이용한 DCTCs를 기본적인 파라미터로 설정하였다. 또한, 음성의 변이구간에 대한 정보를 가지도록 Cepstrum을 재구성한 delta-Cepstrum, delta-delta-Cepstrum 파라미터를 제안하고, HMM을 이용하여 인식성능을 비교하였다. 그리고 각각의 파라미터의 차원을 축소하기위해 LDA 알고리즘을 적용하고 이에 대한 인식성능을 비교하였다. 실험결과 다양한 조건의 잡음 환경에서 기존의 파라미터보다 LDA를 이용하여 차원 축소된 delta-delta-Cepstrum 파라미터가 향상된 인식성능을 나타내었다.

ABSTRACT

This paper propose about effective speech feature parameter that have robust character in effect of noise in realizing speech recognition system. Established MFCC that is the basic parameter used to ASR(Automatic Speech Recognition) and DCTCs that use DCT in basic parameter. Also, proposed delta-Cepstrum and delta-delta-Cepstrum parameter that reconstruct Cepstrum to have information for variation of speech. And compared recognition performance in using HMM. For dimension reduction of each parameter LDA algorithm apply and compared recognition. Results are presented reduced dimension delta-delta-Cepstrum parameter in using LDA recognition performance that improve more than existent parameter in noise environment of various condition.

키워드

MFCC, DCT, LDA, HMM, DCTCs, DCSCs

Ⅰ. 서 론

음성은 가장 편리한 의사소통수단이지만 음성데이터의 특성이 화자간의 영향이나 잡음환경에서 특징변화가 상당히 심하기 때문에 특징 파라미터 추출에 대한 다양한 방법이 제시되고 있다. 본 논문에서는 기존의 ASR에 많이 사용되는 MFCC(Mel-Frequency Cepstrum Coefficient)와 DCT(Discrete Cosine Transform)를 이용한 DCTCs(Discrete Cosine Transform Coefficients)를 기본적인 파라미터로 설정하고 이에 대해 음성의 변이특성을 고려한 파라미터로 delta-Cepstrum과 delta-delta-Cepstrum을 사용하였다 또한, 음성인식 시스템의 성능향상을 위해 특징 파라미터의 차원을 축소하기 위해 LDA(Linear Discriminant Analysis)를 사용하여 잡음 환경에서 단계별로 노출된 음성데이터에 대하여 제안한 파라미터들을 HMM을 사용하여 인식 실험을 수행하고 인식결과를 비교, 분석하였다.

본 논문은 2장에서는 본 논문에서 제안한 특징

파라미터들에 대해 설명하고 3장에서는 특징 파라미터의 차원을 축소하기 위해 사용된 LDA 알고리즘에 대해 설명하였다. 4장에서는 단계별 잡음 환경에서 각각의 파라미터를 HMM을 이용하여 학습한 실험결과를 나타내었고, 5장에서 결론 및 향후과제로 구성하였다.

II. 특징 파라미터

II-1. MFCC(Mel-Frequency Cepstral Coefficient)

Mel은 인간의 청각 특성을 고려한 주파수 단위로 주파수에 대한 주관적인 척도라 할 수 있다. 인간의 청각특성은 저주파수 대역으로 갈수록 민감한 반응을 나타내고, 고주파수 대역으로 갈수록 둔감한 반응을 나타낸다. 이러한 인간의 청각특성을 log-scale과 유사한 형태로 표현한 것이 Mel 단위이다. 그림 1은 음성 신호로부터 MFCC를 추출하는 과정을 나타내었다.

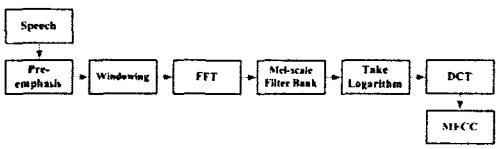

Fig 1. MFCC Extraction

10차의 MFCC는 ASR에서 가장 기본적인 특징 파라미터로 사용되고 있다.

II-2. DCTCs(Discrete Cosine Transform Coefficient)

이산 코사인 변환 계수(DCTCs)는 Mel 단위의 cepstrum 정보를 가진 MFCC을 보완하기 위해 cepstrum과 시간의 정보를 포함한다. 특히 본 논문에서 사용하는 인식알고리즘인 HMM은 시간정보가 중요한 요소로 작용한다.

본 논문에서 사용되는 DCTCs는 1차원 DCT로써 스펙트럼-시간 matrix를 2차원 DCT를 이용하여 구할 수 있다. 2차원 DCT는 2개의 1차원 DCT로 표현가능하기 때문이다. cpestral-time matrix $C_t(m,n)$은 식(1)과 같이 연속하는 MFCC의 특징 벡터 $C_t(n)$을 누적하여 1차원 DCT를 통해 얻어질 수 있다.[1]

$$C_t(m,n) = \sum_{k=0}^{M-1} C_{t+k}(n)\cos\frac{(2k+1)m\pi}{2M} \quad (1)$$

DCTCs는 MFCC에서 표현하는 spectral envelope, pitch 정보를 시간변이에 따라 표현할 수 있다.

II-3. delta-Cepstrum, delta-delta-Cepstrum

음성신호를 사람의 성문신호(glottal signal)와 성대(vocal tract)특성을 이용하여 나타낼 수 있는 특징파라미터가 cepstrum이다.

Fig 2. Speech Model

이는 식(2)과 같이 표현할 수 있다.

$$s(t) = x(t)*h(t) \quad (2)$$

식(2)의 magnitude spectrum은 다음과 같이 표현할 수 있으며,

$$|S(w)| = |X(w)||H(w)| \quad (3)$$

식(3)에 log를 취하고 fourier 변환을 하면 아래의 식(4)과 같이 cepstrum을 구할 수 있다.

$$c(n) = \sum \log|S(w)|e^{-jwn} \quad (4)$$

cepstrum은 낮은 차수에서 spectral envelope가 나타나고 높은 차수에서 pitch 정보를 나타낸다.

이런 cepstrum의 특징은 음성을 모델링하기위해 여러 개의 프레임을 사용하여 보다나은 성능을 얻을 수 있는데 이런 경우 전체적으로 계수의 수가 증가하게 된다. 따라서 이를 최소한의 파라미터로 표현할 필요가 있고, 그 방법 중 하나가 미분계수(delta)와 가속계수(delta-delta)를 사용하는 것이다.

III. 특징 파라미터의 차원 축소

II장에서 추출된 특징들은 인식성능에 결정적인 영향을 미친다. 더불어 특징벡터의 차원을 형성하는 특징의 수 역시 인식률에 영향을 준다. 특징의 수가 많으면 인식기의 속도가 저하되고, 모델링에 필요한 학습 집합의 크기가 증가한다. 또한 인식 대상 신호의 잡음성분을 포함할 수 있기 때문에 인식률의 저하를 초래한다. 따라서 추출된 특징벡터의 차원을 축소시켜 분류기를 설계하면 적은 양의 표본으로 충분히 정확한 데이터의 분포를 나타내고 분류할 수 있다.

본 논문에서 사용한 특징 파라미터의 차원 축소 알고리즘은 PCA(Principle Component Analysis)와 더불어 대표적인 특징 벡터 차원 축소기법으로 사용되는 LDA(Linear Discriminant Analysis)이다. LDA는 클래스간의 분산과 클래스내의 분산 비율을 최대화하는 방식으로 특징 공간속에서 클래스 분리를 최대화하는 주축으로의 사상을 통해 데이터의 특징 벡터의 차원을 축소한다. 또한 LDA는 가능한 클래스간의 분별정보를 최대한 유

기시키면서 차원을 축소한다.

LDA에서 사용하는 Fisher의 선형판별식은 최적화된 변환행렬 W^*을 식(5)과 같이 표현한다.

$$W^* = \arg\max_W \left\{ \frac{W^T S_B W}{W^T S_W W} \right\} = S_W^{-1}(\mu_1 - \mu_2) \quad (5)$$

W : 변환 행렬
S_W : 클래스내의 분산행렬
μ_1, μ_2 : 두 클래스의 평균 벡터

식(5)을 일반화하여 C개의 클래스에 대한 경우는 하나의 사영(y)을 구하는 것이 아니라 $C-1$개의 사영 $[y_1, y_2, \cdots, y_{N-1}]$을 구해야 한다. 사영벡터 W_i을 사용하여 사영 행렬 $W = [W_1 | W_2 | \cdots | W_{N-1}]$을 정의한다.

각 클래스내의 데이터에 대한 분산행렬의 일반식은 다음과 같다.

$$S_W = \sum_{i=1}^{C} S_i \quad (6)$$

여기서, $S_i = \sum_{x \in w_i} (x - \mu_i)(x - \mu_i)^T$,

$\mu_i = \frac{1}{N} \sum_{x \in w_i} x$ 이다.

각 클래스간의 분산행렬일반식은 다음과 같다.

$$S_B = \sum_{i=1}^{C} N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (7)$$

여기서, $\mu = \frac{1}{N} \sum_{\forall x}^{C} x = \frac{1}{N} \sum_{x \in w_i} N_i \mu_i$ 이고, 전체분산행렬은 $S_T = S_B + S_W$가 된다.

사영된 표본들에 대한 평균벡터($\tilde{\mu_i}$)와 분산행렬($\tilde{S_B}$, $\tilde{S_W}$)은 다음과 같이 정의한다.

$$\tilde{\mu_i} = \frac{1}{N} \sum_{y \in w_i} y \quad (8)$$

$$\tilde{\mu} = \frac{1}{N} \sum_{\forall y} y \quad (9)$$

$$\tilde{S_W} = \sum_{i=1}^{C} \sum_{y \in w_i} (y - \tilde{\mu_i})(y - \tilde{\mu_i})^T \quad (10)$$

$$\tilde{S_B} = \sum_{i=1}^{C} N_i (\tilde{\mu_i} - \tilde{\mu})(\tilde{\mu_i} - \tilde{\mu})^T \quad (11)$$

$$\tilde{S_W} = W^T S_W W \quad (12)$$

$$\tilde{S_B} = W^T S_B W \quad (13)$$

클래스내의 분산에 대한 클래스간의 분산의 비를 최대화하는 사영을 구하는 것이 LDA의 목적이므로 사영은 $C-1$차원을 가지며 스칼라가 아니다. 따라서 목적함수를 구하면 다음과 같다.

$$J(W) = \frac{\tilde{S_B}}{\tilde{S_W}} = \frac{|W^T S_B W|}{|W^T S_W W|} \quad (14)$$

이 목적함수를 최대로 하는 사영행렬 W^*을 구한다.

$$W^* = [W_1^* \mid W_2^* \mid \cdots \mid W_{C-1}^*]$$
$$= \arg\max \left\{ \frac{W^T S_B W}{W^T S_W W} \right\}$$
$$\Rightarrow (S_B - \lambda_i S_W) W_i^* = 0 \quad (15)$$

식(15)에서 S_B는 랭크가 1이하인 C개의 행렬의 합이고, 평균벡터는 다음과 같다.

$$\frac{1}{C} \sum_{i=1}^{C} \mu_i = \mu \quad (16)$$

이는 S_B는 랭크가 $(C-1)$이하가 될 것이고, 고유값 λ_i의 $(C-1)$개만이 영이 아님을 의미한다. 또한 최대 클래스간의 분류를 최대로 하는 사영은 $S_W^{-1} S_B$의 가장 큰 고유값과 관련된 고유벡터임을 알 수 있다.

Fig 3과 Fig4는 3가지의 클래스를 갖는 데이터를 LDA를 이용하여 차원을 축소한 예를 보여준다.

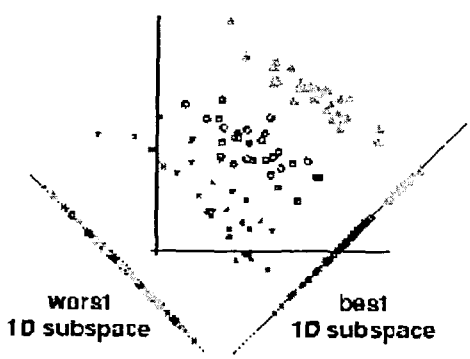

Fig 3. 3가지 클래스를 갖는 특징 데이터

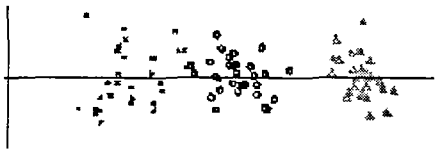

Fig 4. LDA를 사용한 차원 축소결과

Ⅳ. 실험 및 결과

Ⅳ-1. 분석조건

실험데이터로는 ETRI의 음성정보연구센터에서 녹음한 한국어 중가 마이크 음성인식용 숫자 데이터를 사용하였고, 10개의 숫자음(영, 일, 이, 삼, 사 오, 육, 칠, 팔, 구)을 사용하였다. 실험데이터의 특징은 Table 1과 같다.

Table 1. Extraction of feature parameter

A/D Convert	16㎑, 16bit
window	hamming
window length	24ms(384 samples)
shifting period	8ms(128 samples)
feature parameter	10th MFCC, DCTCs delta-Cepstrum, delta-delta-Cepstrum

Ⅳ-2. 특징파라미터의 인식 결과

분석조건에 해당하는 입력음성데이터에 백색잡음을 각기 다른 SNR(5dB, 15dB, 25dB)로 첨가하여 Table 1의 특징파라미터를 추출하고 이를 HMM (Hidden Markov Model)을 이용하여 인식실험을 수행하였다.

Table 2. Result Ⅰ

Feature Parameter	Recognition accuracy(%)		
	5dB	15dB	25dB
10th MFCC	78	90	97
DCTCs	75	88	96
delta-Cepstrum	89	91	98
delta-delta-Cepstrum	88.5	92	97

Ⅳ-3. LDA를 이용한 특징파라미터의 인식

Ⅳ-2의 실험에서 사용된 파라미터를 LDA를 이용하여 차원 축소하고 이를 HMM을 이용하여 인식실험을 수행하였다.

Table 3. Result Ⅱ(Using LDA)

Feature Parameter	Recognition accuracy(%)		
	5dB	15dB	25dB
10th MFCC	89	94	97.5
DCTCs	86.5	95	98
delta-Cepstrum	92	96	99.5
delta-delta-Cepstrum	90	95.5	98.5

Ⅴ. 결론 및 향후과제

인식실험결과 기본적인 특징 파라미터인 MFCC와 DCTCs는 잡음레벨이 강할수록 인식률의 저하가 두드러지는 반면 변이성분을 포함하는 파라미터인 delta-Cepstrum과 delta-delta-Cepstrum은 강한 잡음레벨을 가진 데이터에도 인식률의 저하가 크지 않음을 알 수 있다. 또한 LDA를 이용한 특징파라미터의 차원 축소실험 결과 전체적으로 인식성능이 향상되었음을 알 수 있고 특히 잡음레벨이 강할 경우에 상당한 인식률의 향상을 볼 수 있었다.

특징파라미터와 잡음레벨에 상당히 많은 영향을 받는 ASR 시스템이 있어서 음성의 변이성분을 포함할 수 있는 특징파라미터와 LDA의 사용은 전체적인 시스템의 성능향상을 기대할 수 있음을 알 수 있다.

ASR 시스템의 성능향상을 위해 특징파라미터와 다양한 알고리즘의 적용이 필요하다. 더불어 음성신호에 대한 정확한 시작점-끝점검출 방법이 높은 성능의 ASR 시스템을 구현하는데 필요한 요소라 하겠다.

참고문헌

[1] B. Milner, "Inclusion of temporal information into features for speech recognition," in Proc. Int. Conf. Speech Language Processing '96, 1996, pp.256-259

[2] J. Picone, "Signal modeling techniques in speech secognition," Proc. IEEE, vol. 81. NO. 9, Sept. 1993

[3] Montri Karnjanadecha, S. A. Zahorian, "Signal Modeling for High-Performance Robust Isolated Word Recognition," IEEE Trans. Speech and Audio Processing, vol. 9, NO. 6, Sept. 2001

[4] 한학용, "패턴인식 개론," 한빛미디어, 2005

[5] 박정원, 김창근, 한학용, 허강인, "잡음환경에 강인한 음성인식기의 자동단어분할기법," 한국신호처리시스템학회학술논문집, 3권 2호, pp217-220, 2002