잡음환경에서의 숫자음 인식을 위한 특징파라메타

이재기* · 고시영** · 이광석*** · 허강인*

*동아대학교 **경일대학교 ***진주산업대학교

Features for Figure Speech Recognition in Noise Environment

Jae-ki Lee* · Si-young Koh** · Kwang-suk Lee*** · Kang-in Hur*

*Donga University **Kyungil University ***Jinju National University

E-mail : ljk3046@korea.com

요 약

본 논문은 잡음에 강한 다양한 특징 파라메타를 제안한다. 기존의 음성인식에서 사용되는 특징 파라메타 MFCC(Mel Frequency Cepstral Coeeficient)는 좋은 성능을 보인다. 그러나 잡음에 보다 강인한 성능을 위해 기존에 사용되는 파라메타 MFCC의 특징공간을 변형시키는 알고리즘인 PCA(Principal Component Analysis)와 ICA(Independent Component Analysis)를 사용하여 특징 공간을 변형시킨 파라메타와 기존의 파라메타 MFCC의 성능을 비교하였다. 그 결과 ICA에 의해 변형된 특징 파라메타가 PCA로 변형된 파라메타와 MFCC보다 우수한 성능을 보였다.

ABSTRACT

This paper is proposed a robust various feature parameters in noise. Feature parameter MFCC(Mel Frequency Cepstral Coefficient) used in conventional speech recognition shows good performance. But, parameter transformed feature space that uses PCA(Principal Component Analysis)and ICA(Independent Component Analysis) that is algorithm transformed parameter MFCC's feature space that use in old for more robust performance in noise is compared with the conventional parameter MFCC's performance. The result shows more superior performance than parameter and MFCC that feature parameter transformed by the result ICA is transformed by PCA.

키워드

MFCC, PCA, ICA, HMM

Ⅰ. 서 론

잡음환경에서의 음성인식성능에 있어서 고려되어야할 사항은 특징 파라메타와 음성 인식기의 선택이다. 현재까지 좋은 음성인식 성능을 위해서 일반적으로 인식기로는 HMM(Hidden Markov Model)과 파라메타로는 MFCC(Mel-Frequency cepstrum coefficient)가 사용되어 비교적 좋은 음성인식 성능을 보여주고 있다. 본 논문에서는 좀 더 나은 인식성능을 위해서 인식기로는 HMM을 사용하고 특징 파라메타를 다양화하여 그 성능을 비교하였다. 기본적인 특징 파라메타 MFCC의 특징 공간을 변형시키는 알고리즘인 PCA와 ICA를 사용하여 변형된 파라메타를 HMM을 통해 인식실험을 하여 그 성능을 기존의 특징 파라메타 MFCC와 비교하여 그 성능이 얼마나 개선되는지를 분석해보았다.

2장에서 MFCC에 대해 설명하고 3장과 4장에서는 PCA와 ICA 알고리즘에 대해 설명하고 5장에서 실험결과를 기술하였고 6장에서 결론을 맺는다.

Ⅱ. 특징 파라메타 MFCC

MFCC를 구하는 방법은 Fig. 1과 같다. 16KHz Sampling을 위해 8KHz의 LPF(Lowpass Filter)에

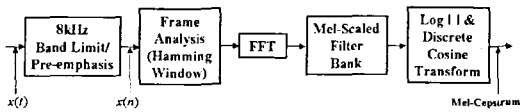

Fig. 1 Production of MFCC

의해 대역 제한된 음성신호는 A/D변환을 거쳐 디지털 신호 $x(n)$으로 변환되고 식(1)과 같은 고역강조(Pre-emphasis)를 통해 입술의 방사에 의해 20dB/decade로 감쇄되는 것은 보상하게 되어 음성의 성도 특성만을 취하게 된다.

$$\tilde{x}(n) = x(n) - 0.95 * x(n-1) \quad (1)$$

식(1)의 신호 $\tilde{x}(n)$은 음성 신호의 특성이 고정되어 있다고 가정하는 20~30ms의 길이를 갖는 윈도우함수(Hamming, Hanning, Bartlett, Blackman Window 등)를 씌워서 블록 단위의 프레임으로 나눈 다음 프레임별로 FFT(Fast Fourier Transform)를 이용해 주파수 영역으로 변환한다. 변환된 주파수 대역을 인간의 청각 특성을 반영한 여러 개의 Mel-Scaled Filter Bank(Fig. 2)로 나누고 각각의 뱅크에서 에너지를 계산한 후 계산된 에너지에 로그를 취한 다음 DCT(Discrete Cosine Transform)를 하게 되면 최종적인 MFCC가 프레임 별로 얻어진다.

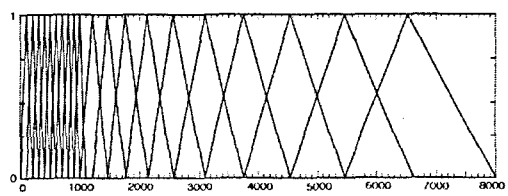

Fig. 2 Mel-Scaled Filter Bank for 16KHz Sampling

III. Principal Component Analysis

음성인식에서 가장 중요한 문제는 각각의 음성의 특성을 가장 잘 나타내는 특징을 추축하는 것이다. 이러한 특징추출(feature extraction)은 데이터 분류(data classification) 또는 패턴 인식(pattern recognition)의 중요한 문제이다. PCA(Principal Component Analysis)는 입력의 선형성과 특성 식별을 이용하여 다차원 데이터에 대해 Fig. 3의 변동량이 큰 주축(Principal Axis)을 찾아 차원을 축소하며 특징을 추출하는 방법 중 널리 이용되어지는 것 중 하나로, 패턴인식이나 영상처리에서 Karhunen- Loeve 변환으로 잘 알려져 있다.

Fig. 3 2차원 데이터의 주축

zero-mean 특성이 있는 n차원 신호 $\boldsymbol{x}$에 대해 차원 축소는 식(2)과 같다.

$$\boldsymbol{a} = [a_1, a_2, \cdots, a_l]^T$$
$$= [\boldsymbol{x}^T\boldsymbol{q}_1, \boldsymbol{x}^T\boldsymbol{q}_2, \cdots, \boldsymbol{x}^T\boldsymbol{q}_l] = \boldsymbol{Q}^T\boldsymbol{x} \quad (2)$$

여기서 $n > l$ 이며 $\boldsymbol{a}$는 $\boldsymbol{x}$에 대해 l차원으로 축소된 신호이다. 그리고 축소된 신호에서 원 신호로의 복원은 식(3)과 같다.

$$\hat{\boldsymbol{x}} = \boldsymbol{Q}\boldsymbol{a} = \sum_{j=1}^{l} a_j \boldsymbol{q}_j \quad (3)$$

식(2)(3)에서 $\boldsymbol{a}$와 $\boldsymbol{Q}$는 $\boldsymbol{x}$의 공분산 행렬(covariance matrix) $\boldsymbol{R}$의 고유치와 고유벡터이다. 이와 같은 차원 축소 신호 복원은 Fig. 4와 같다.

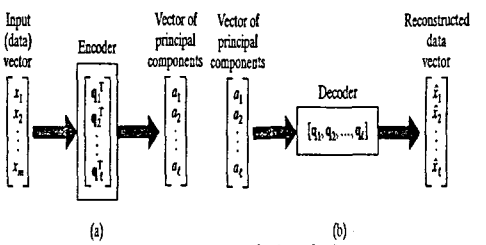

Fig. 4 주성분 해석
(a) 차원 축소 (b) 신호 복원

차원 축소된 신호 $\boldsymbol{a} = [a_1, a_2, \cdots, a_l]$는 중요도 순으로 나열된 n개의 고유벡터 $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n]$중에 l개만 사용하여 얻을 수 있다

본 논문에서는 PCA를 신경망으로 구현한 GHA (Generalized Hebbian Learning)을 이용하였다. GHA는 순방향 단층구조의 신경망으로 자율학습특성이 있으며 학습된 가중벡터 $\boldsymbol{W}$는 중요도 순에 의해 내림차순으로 정렬되는 특성을 가지고 있다. GHA의 학습은 식(4)과 같다.

$$\Delta w_{ji}(n) = \eta[y_j(n)x_i(n) - y_j(n)v(n)] \quad (4)$$
$$y_j(n) = \sum_{i=0}^{l-1} w_{ji}(n)x_i(n)$$

$$v(n) = \sum_{k=0}^{j} w_{kj}(n) y_k(n)$$

학습된 가중벡터는 입력벡터 $\boldsymbol{x}$의 고유벡터 특성을 가지고 있으며 식(5)에 의해서 얻어진 계수 $\boldsymbol{c}$는 고유치 특성을 가진다. 본 논문에서는 고유치 계수 $\boldsymbol{c}$를 PCA를 이용한 특징 파라메터로 사용하였다.

$$c_j = \sum_{i=0}^{l-1} w_{ji}(n)\, x_i(n), \quad \boldsymbol{c} = \{c_0, c_1, \cdots, c_{l-1}\} \tag{5}$$

IV. Independent Component Analysis

여러 개의 신호원이 여러 경로를 통해서 전파되고 있고 동시에 여러 개의 수신기에서 이 신호원을 수신한다고 가정하면 수신단에서 관측되는 신호 $\boldsymbol{x}(t)$는 식(6)과 같이 신호원 $\boldsymbol{s}(t)$의 선형결합 형태로 나타난다.

$$x_1(t) = a_{11} s_1(t) + a_{12} s_2(t)$$
$$x_2(t) = a_{21} s_1(t) + a_{22} s_2(t) \tag{6}$$

여기서 선형혼합행렬 $\boldsymbol{A}$의 가중치 a_{ji}와 신호원 $s_i(t)$는 'unknown'이고, 신호 $\boldsymbol{s}(t)$는 선형혼합행렬의 역행렬인 $\boldsymbol{W}$에 의해 식(7)와 같이 표현될 수 있다.

$$s_1(t) = w_{11} x_1(t) + w_{12} x_2(t)$$
$$s_2(t) = w_{21} x_1(t) + w_{22} x_2(t) \tag{7}$$

따라서 BSS(Blind Source Separation)는 식(8)의 $\boldsymbol{W}$를 찾는 문제이다. 이 문제를 해결하기 위해서 각각의 신호원 $s_i(t)$를 독립성분(independent component)라고 가정하는데 이것이 독립성분분석(ICA)의 이론적 배경이 된다. 독립성분분석은 위의 가정 하에 여러 가지 형태로 해결할 수 있는데 대표적인 것으로 상호정보량의 최소화(Minimization of Mutual Information), 비정규성의 최대화(Maximization of Nongaussianity)등이 있으며 본 논문에서는 비정규성의 최대화법을 사용하였다.

식(8)에서 관측신호 $\boldsymbol{x}(t)$를 몇 개의 확률적으로 독립인 신호 $\boldsymbol{s}(t)$들에 가중벡터가 곱해진 다음 혼합되어 생성된 것이라 가정하고 독립신호원간의 통계적인 의존성을 정의한 다음 의존성이 최소가 되는 가중치 $\boldsymbol{W}$를 추정하여 식(9)를 통해 관측신호에 대한 독립신호를 구할 수 있다.

$$\boldsymbol{x}(t) = \boldsymbol{A}\, \boldsymbol{s}(t) \tag{8}$$

$$\boldsymbol{s}(t) = \boldsymbol{W}\boldsymbol{x}(t) \quad , \boldsymbol{W} = \boldsymbol{A}^{-1} \tag{9}$$

독립신호 $\boldsymbol{s}(t)$의 합으로 나타나는 관측신호 $\boldsymbol{x}(t)$가 정규(gaussian)분포를 가진다면 $\boldsymbol{x}(t)$를 구성하는 독립신호는 비정규(nongaussian)분포를 가지게 된다. 따라서, 독립성분 $\boldsymbol{s}(t)$의 비정규성(nongaussianity)을 최대화하도록 가중벡터 $\boldsymbol{W}$를 학습함으로써 얻을 수 있다. 비정규성은 고차 통계이론에서의 첨도(kurtosis)와 정보이론의 negentropy에 의해 측정될 수 있다. $\boldsymbol{y} = \boldsymbol{s}(t)$라 할 때 첨도는 식(10)와 같다.

$$kurt(\boldsymbol{y}) = E[\boldsymbol{y}^4] - 3 \tag{10}$$

첨도는 데이터의 outlier에 민감한 단점을 가지고 있기 때문에 보통 비정규성 측정에서는 negentropy를 주로 사용한다.

정보이론에서 엔트로피는 식(11)과 같이 정의된다.

$$H(\boldsymbol{y}) = -\int p_y(\eta) \log p_y(\eta)\, d\eta \tag{11}$$

엔트로피는 정규분포를 가지는 변수에 대해 큰 값을 가지며 여기서 $p_y(\eta)$는 $\boldsymbol{y}$의 확률밀도 함수이다.

negentropy는 엔트로피를 정규화한 것으로 식(12)과 같이 정의된다.

$$J(\boldsymbol{y}) = H(\boldsymbol{y}_{gauss}) - H(\boldsymbol{y}) \tag{12}$$

여기서 $\boldsymbol{y}_{gauss}$는 상관과 공분산이 같은 정규랜덤변수이다. negentropy는 계산의 복잡성 때문에 고차 통계특성과 nonquadratic 함수를 이용하여 식(13)과 같이 근사화 할 수 있다.

$$J(\boldsymbol{y}) \approx \{G(\boldsymbol{y}) - G(\nu)\}^2 \tag{13}$$

여기서 ν는 표준 정규분포를 따르는 랜덤변수이다. G는 Hartley의 엔트로피 함수의 세 가지 조건(base, mono- tonicity, additivity)을 만족하며, 이런 함수를 대비함수(Contrast Function)라 하고 다음과 같은 함수들이 많이 사용된다.

$$G_1(\boldsymbol{y}) = 1/a_1 \log\cosh a_1 \boldsymbol{y} \tag{14}$$

$$G_2(\boldsymbol{y}) = -\exp(-\boldsymbol{y}^2/2) \quad , 1 \le a_1 \le 2 \tag{15}$$

학습은 Gradient 알고리즘을 사용하여 다음과 같이 수행된다.

$$\Delta \boldsymbol{w} \propto \Upsilon \boldsymbol{z} g(\boldsymbol{w}^T \boldsymbol{z})$$
$$\boldsymbol{w} \leftarrow \boldsymbol{w}/\|\boldsymbol{w}\|$$
$$\Upsilon = E[G(\boldsymbol{w}^T \boldsymbol{z})] - E[G(\nu)] \tag{16}$$

여기서 $g = G'$이고 $\boldsymbol{z}$는 관측신호 $\boldsymbol{x}$에 대하여 zero-mean, unit-variance로 whiten한 신호이다.

V. 실험결과

실험에 사용된 음성 데이터는 ETRI의 음성인식

용 한국어 중가 마이크 숫자음 데이터를 사용하였으며, 잡음 데이터는 NOISEX-92 잡음 데이터 중 babble noise를 사용하였다. 이 잡음은 19.98KHz, 16bit의 anti-aliasing 필터링 된 데이터로 본 실험에서는 16KHz, 16bit로 변환하여 잡음 환경을 위해 5dB, 10dB, 15dB, 20dB의 레벨 단위로 음성과 잡음 데이터를 혼합하여 학습 데이터와 테스트 데이터를 만들었다.

실험에 사용된 음성신호의 분석조건은 Table. 1과 같다.

Talbe. 1 Analysis Conditions

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

음성신호에서 추출해내는 각각의 특징 파라메타 추출 과정은 Fig. 5와 같다.

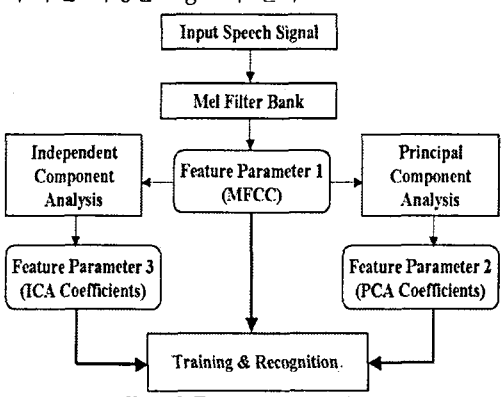

Fig. 5 Feature Extraction

Fig. 5과 같이 추출된 특징 파라메타들은 인식 실험에 널리 사용되는 HMM을 통해 학습 및 인식 실험을 하였고 Table. 2와 같이 그 결과를 잡음 단위별로 비교하였다. 실험결과에서 보여지듯이 ICA에 의해 특징 공간이 변형된 특징 파라메타가 잡음환경에서의 숫자음 인식에 가장 우수한 성능을 보임을 알 수 있다. 이는 ICA가 잡음환경에서의 숫자음의 특징 공간에서 선형분별성이 가장 뚜렷하기 때문이다.

Table. 2 Recognition Rate for Each Parameter

Feature Parameter	Recognition Rate(%)			
	5dB	10dB	15dB	20dB
MFCC	75	83	90	92
PCA	74	81	87	91
ICA	80	86	93	96

Ⅵ. 결 론

본 연구의 실험결과에서 보듯이 잡음환경에서 음성인식성능은 ICA에 의해 특징 공간이 변형된 파라메타가 가장 좋은 성능을 보임을 알 수 있었다. 기존의 파라메타 MFCC보다 약 1%정도의 성능개선을 볼 수 있었으며 PCA에 의해 변형된 파라메타보다는 약 5%정도의 성능개선 효과를 볼 수 있었다. 이는 ICA에 의해 변형된 파라메타가 특성이 다른 음소로 이루어진 숫자음에 대해 기존의 다른 특징 파라메타보다 특징 공간에서 좀 더 분류하기 쉬운 형태로 존재하기 때문이다. 따라서 ICA에 의해 변형된 파라메타가 잡음에 강인한 성능을 보인다.

이처럼 잡음에 강한 면모를 보이는 ICA는 보다 다양한 분야에 적용할 수 있을 것이다. 하지만 아직까지도 잡음이 많은 환경에서의 인식성능이 만족할 만한 수준은 아니다. 차후 인식기의 선택을 다양화하여 보다 좋은 성능을 보이는지 검증해 보아야 할 것이다. 이런 연구를 통해 추후 잡음에 보다 강한 특징 파라메타와 인식기의 조합을 찾는 연구를 해보아야 할 것이다.

참고문헌

[1] A.Hyvärinen, J.Karhunen, E.Oja, "Independent Component Analysis", John Wiley & Sons, 2001

[2] T.W.Lee. "Independent Component Analysis-Theory and Applocations", Kluwer Acadmic Publishers, 1998

[3] S.Choi, A.Cichocki, S.Amari, "Flexible independent component analysis", IEICE Trans. Fundamentals, Vol.E83-A, No.12, pp. 2715-27-22, 2000

[4] 박경훈, 표창수, 김창근, 허강인 "PCA 기반 파라메타를 이용한 숫자음 인식", 신호처리 시스템 학회 논문지. 제1권 2호, pp.181-184, 2000

[5] C.K.Kim, S.B.Kim, S.H.Kim, K.I.Hur, "Performance Improvement of Speech Recognition Based on Independent Component Analysis", ICSP2001, Vol.2 of 2, pp.663-666, 2001