목 차

1. 서 론

1.1 연구 배경

정보 전달을 위한 인간과 기계와의 인터페이스를 구현하는 것을 HMI(human machine interface) 또는 HCI(human computer interface)라 하며, 정보 전달의 수단 중에서 음성 신호는 인간에게 있어 가장 보편적이고 편리하다. 따라서 음성신호를 이용하여 인간과 기계와의 인터페이스를 가능하게 하고자 하였는데 이것이 음성신호를 이용한 인식(voice recognition)기술이며, 1950년대부터 본격적으로 진보해 오고 있다. 1952년 Bell Labs의 Davis 등은 단독화자에 대한 고립 숫자인식 시스템을 만들었고, 1956년 RCA Labs의 Olson과 Belar는 단독화자의 10개 음절을 인식하는 실험을 하였다. 1959년 MIT Lincoln Lab에서는 화자독립 10개 모음 인식기를 개발하였고, 영국의 College대학에서 Fry와 Denes가 4개의 모음과 9개의 자음을 인식하는 음소 인식기를 개발하였다. 1960년대 RCA Labs의 Martin에 의해 음성끝점 검출기가 개발되었고, 소련의 Vintsyuk는 두 음소간의 시간정렬을 위한 동적 프로그래밍(dynamic programming) 사용을 제안하였다. 그의 업적은 서구에서 크게 알려지지 않았었지만, 이후 1970년대에는 다른 이들에 의해 동적 시간 정합으로 널리 알려졌고 음성인식에서의 패턴 매칭 방법에 있어서 필수적인 방법이 되었다. 또한 카네기 멜론대학에서 음소의 동적인 탐색 기법을 이용하여 연속어 음성인식에 대한 가능성을 확인시켜 주었다. 1970년대에는 고립단어 인식기술이 실제로 사용 가능한 수준이 되었으며, IBM에서는 대어휘 음성인식을 위한 시도가 이루어졌다. 또한 AT&T Bell Labs에서는 화자독립 음성인식 시스템을 위한 노력이 본격적으로 시작되었다. 1980년대의 중요한 사건은 DARPA(Defense Advanced Research

Projects Agency)에서 수행한 대어휘 연속어 음성인식 프로젝트이다. 이 프로젝트는 많은 연구기관(카네기 멜론 대학, BBN, Lincoln Labs, SRI, MIT, AT&T Bell Labs 등)에서 수행되었다. 은닉마코프모델(hidden markov model)은 동적시간정합(dynamic time warping) 이후로 음성인식에 사용되는 가장 중요한 알고리즘이다. 그런데 1980년대 중반까지는 은닉마코프모델에 대해서 특정 기관(IBM, Institute for Defense Analyses, Dragon Systems)을 제외하고는 아무런 관심을 끌지 못하였다. 1980년대 말에는 신경회로망(neural networks)이 음성인식에 본격적으로 응용이 되었다. 신경회로망(neural networks)은 1950년대에 등장하였지만 여러 가지 구현상의 문제를 해결하지 못하여 주목을 받지 못하였으나 1980년대에 들어서서 기술의 많은 한계를 극복하고 주목을 받게 되었다 [1].

음성신호를 이용한 인식기술에는 음성인식(speech recognition)과 화자인식(speaker recognition)이 있다. 음성인식은 화자종속형과 화자독립형 음성인식으로, 화자인식은 화자식별과 화자검증으로 세분화될 수 있는데, 음성인식의 분류에서는 한 화자의 음성만을 인식 가능하도록 한 것이 화자종속형 음성인식, 임의 화자의 음성을 인식 가능하도록 한 것이 화자독립형 음성인식이다. 화자인식의 분류에서는 모집단 내에서 그 화자를 찾아내는 것이 화자식별(speaker identification), 입력 음성이 화자가 본인(genuine)인지 사칭자(impostor)인지를 판별하는 것이 화자검증(speaker verification)이다. 또한, 음성인식 시스템이나 화자인식 시스템은 입력 음성신호가 미리 지정된 문장 또는 단어 등으로 제한되는 문맥종속형(text-dependent)과 내용에 제한이 없는 문맥독립형(text-independent) 등으로 구분된다. 문맥종속형은 미리 정의한 문장이 필요하고 그 문장을 암기하고 일관성 있게 발음을 하여야 하는 등 사용자의 협조가 필요하다.

현재는 음성인식과 화자인식 기술 중 음성인식 기술의 연구가 보다 중점적으로 이루어지고 있으며, 이는 국내의 학술지에 발표되는 논문을 보아도 알 수 있다. 그러나 국제 학술지에는 화자인식 기술과 관련된 논문의 수가 매년 꾸준히 증가하고 있다. 이것은 앞으로 기계와 사람 사이의 가교 역할을 하는 기술로써 음성인식 기술뿐만이 아니라 화자인식 기술도 필수적이라는 점을 반영하는 것이다. 국내에서도 음성관련 기술에 종사하는 전문가들에 의해 그 중요성이 대두되고 있는데, 화자인식 기술은 그 기술의 응용분야 및 타 산업으로의 파급 효과가 크기 때문에 늦게 시작할수록 화자인식에 꼭 필요한 기반기술 및 특허 그리고 응용 분야의 개발시기를 더 많이 놓치게 되기 때문에 중요하다고 할 수 있다.

1.2 연구 내용

본 논문에서는 1.1절에서와 같은 음성을 이용한 인식 기술 중에서 화자식별을 연구 대상으로 하고 있으며, 음성인식에 쓰이는 기법들을 화자식별 시스템 구현에 이용하였고, 화자의 음성을 입력으로 취하여 화자별로 상이한 음성신호의 특징 파라미터를 추출하여 이를 비교 분석하게 된다. 음성인식 기술에 있어서는 화자의 특성이 변화하고, 주변환경의 변화에 민감하다는 것이 어려운 점이다. 이를 해결하기 위해 최근에는 좀 더 잡음이나 주변환경 변화에 강인한 방법을 찾으려는 연구가 활발히 진행되고 있는데, 본 논문에서는 화자의 특성이 어느 정도 일정하고, 주변환경 변화가 적다는, 즉, 학습환경과 테스트환경은 어느 정도 유사하다는 가정을 전제로 신경회로망을 이용하여 화자식별시스템을 구현하고 있다. 신경회로망 모델을 이용하여 화자를 식별하는데 있어 특징파라미터를 추출하여 신경회로망 모델의 입력으로 사용하고 있다. 일반적으로 화자식별에 사용되는 음성 특징으로는 동일 화자

에 대한 일관성(consistency)을 높임과 동시에 다른 화자와의 변별력을 높일 수 있는 특징을 추출하여 사용하게 된다. 하지만 아직은 화자식별에 이상적인 특징추출방법은 알려져 있지 않으므로 일반적으로 음성발생기관이나 인간의 청각기관의 모델에 근거를 둔 방법들을 이용하고 있다. 음성발생기관에 기반을 둔 방법에는 대표적으로 선형예측계수(linear predictive coefficient)를 이용한 방법[2]이 있으며, 인간의 청각기관의 모델에 기반을 둔 방법에는 대표적으로 Mel-주파수 켑스트럼 계수(Mel-frequency cepstral coefficient)를 이용한 방법[2]이 있다.

본 논문에서는 특징파라미터로 선형예측계수, Mel-주파수 켑스트럼 계수를 사용하고 있으며, Mel-주파수 켑스트럼 계수에 사용된 필터뱅크 대신 웨이블릿변환을 이용하여 상관을 줄이는 방법을 최근의 연구동향에 따른 특징파라미터 추출방법으로 사용한다. 그리고 각각의 특징파라미터들과 그들을 혼합한 형태를 신경회로망에 입력하여 화자를 식별하며, 실험결과 산출된 화자인식율을 비교하여 특징파라미터와 신경회로망 모델에 따른 성능을 비교하고 이에 대한 평가를 할 것이다.

현재까지 연구되어온 화자인식 기법은 화자의 특징을 나타내는 음성신호의 특징벡터로 구성되는 기준패턴을 미리 작성한 다음, 시험 패턴과 기준패턴 사이의 유사도를 측정하여 시험패턴의 신원을 확인하는 패턴정합에 의한 방법과 각 화자에서 추출한 음성 특징 파라미터들을 시간변화에 대해 관측한 후 파라미터들의 통계량을 구해 화자의 신원을 확인하는 통계적 성질을 이용한 분류방법 등이 있다.

본 연구에서 사용된 화자식별 기법은 화자의 특징파라미터를 신경회로망 알고리듬으로 학습하여 새로운 입력에 대해 인식하는 패턴인식기법의 한 부류로 볼 수 있으며, 지정된 단어에 대한 문맥종속형 화자 식별 시스템을 구

현하고 특징파라미터에 따른 시스템의 성능을 분석하였다 [3]. 기본적인 패턴 인식 시스템은 데이터 입력으로부터 특징을 추출하고 분류하여 최종적으로 인식하는 구조이고, 다음의 블럭 다이어그램과 같다 [4].

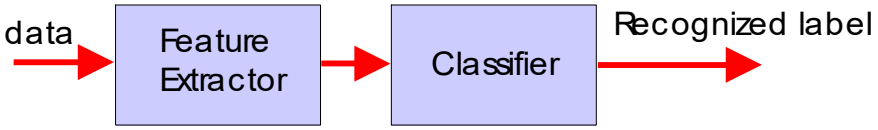


그림 1. 기본적인 패턴인식시스템의 구

Fig. 1. Fundamental structure of pattern recogn

2. 화자식별 시스템의 구성

화자인식 시스템은 입력 음성으로부터 특징파라미터를 추출해 내어 미리 저장된 기준패턴과 입력 패턴을 비교하는 패턴인식 등의 인식과정을 거쳐 인식대상을 결정하는 규칙에 따라 인식결과를 출력하게 되는 것이 기본적인 화자인식 시스템의 구성이다. 하지만 화자인식 시스템에 있어 화자검증 시스템과 화자식별 시스템은 화자모델의 수와 사용목적에 약간의 차이가 있다. 화자검증 시스템은 등록된 화자와 주장하는 화자의 일치여부를 가려내는 것이 목적이며 단일 화자모델이 필요하다. 물론 다수의 개별적인 화자검증 시스템을 결합하여 구성하게 된다면 다수의 화자모델이 필요하겠지만 이 방법은 화자모델이 시스템 블럭에서 독립적으로 사용되어 다른 화자모델과의 연관성이 없으므로 화자검증 시스템은 단일의 화자모델이 필요하다고 말할 수 있다. 이에 반하여 화자식별 시스템은 다수의 등록된 화자로부터 화자모델 그룹을 구성하게 되고 화자의 입력음성에 대해 누구의 음성인지를 가려내어 화자를 식별해 내는 방법이므로 다수의 화자모델이 필요하며 시스템 내부에서 각각의 화자모델은 독립적이지 않다. 화자식별과 화자검증을 모두 포함하는 화자인식기술의 분류는 다음 표 1과 같다.

표 1. 화자인식기술의 분류

Table 1. Classification of speaker recognition techniques

종류	문맥 종속형 / 문맥 독립형	
	화자식별	화자검증
용도	출퇴근 관리 등	음성자물쇠 등
유사기술	고립단어 인식	핵심어 인식
개념	가장 유사한 화자는?	화자의 승인 / 거절?
사용알고리듬	벡터 양자화, 동적 시간정합, 은닉 마코프모델, 신경회로망	

또한 화자식별 시스템은 화자식별 대상 범위가 닫힌 집합(closed set)이냐 열린 집합(open set)이냐에 따라 다시 분류될 수 있는데, 대상범위가 닫

힌 집합인 경우는 등록되지 않은 화자모델에 대해서도 유사한 패턴을 찾아내는 한 마디로 등록된 화자모델 이외의 집합을 고려하지 않은 시스템이라 등록되지 않은 화자를 거부할 수 없고, 열린 집합인 경우는 등록되지 않은 화자모델에 대해서는 비등록자로 간주하여 인식대상에서 제외시키는 방법이므로 등록된 화자모델 이외의 집합에 대해서도 고려한 시스템이라 할 수 있다.

닫힌 집합과 열린 집합을 다이어그램으로 나타내면 다음과 같다.

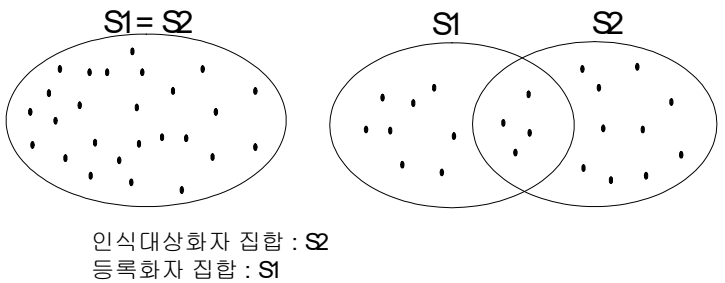


그림 2. (a) 닫힌 집합 (b) 열린 집합
Fig. 2. (a) Closed set (b) Open set

화자인식 시스템은 다양한 방법으로 구현될 수 있고, 알고리듬 또한 다양하다. 현재 이 알고리듬들은 음성 인식에서도 동일하게 사용되며, 그 구현 절차에 있어서도 화자 인식과 음성 인식의 차이점은 별로 없다.

쓰이는 동적시간정합(DTW), 은닉 마코프모델(HMM) 등과 같은 잘 알려진 방법이 사용되기도 한다 [2, 15, 21].

다음은 일반적인 패턴매칭 방식에 의한 시스템을 간략하게 나타낸 구성도와 본 논문에서 사용된 신경회로망을 이용한 시스템 구성도이다.

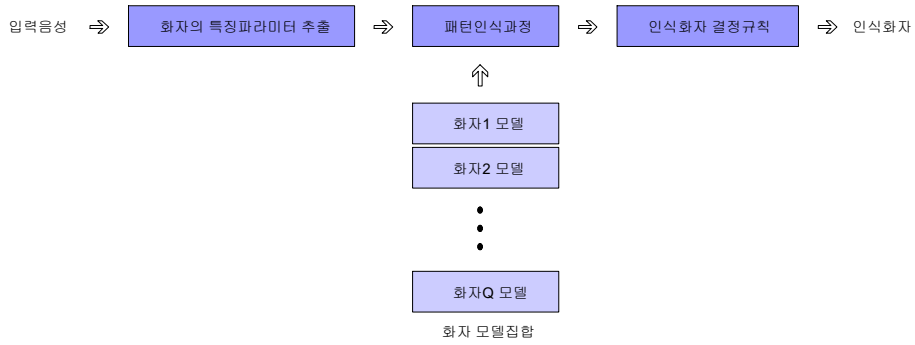


그림 3. 패턴매칭 방식의 화자식별 시스템 구성

Fig. 3. Speaker identification system configuration of pattern matching method

본 논문에서는 화자의 3가지의 특징파라미터를 추출하였고, 이를 신경회로망 입력으로 사용하여 화자를 식별하고자 하였다. 그림 5는 3가지 특징파라미터에 대한 화자식별 시스템의 세부적인 블럭 다이어그램이다.

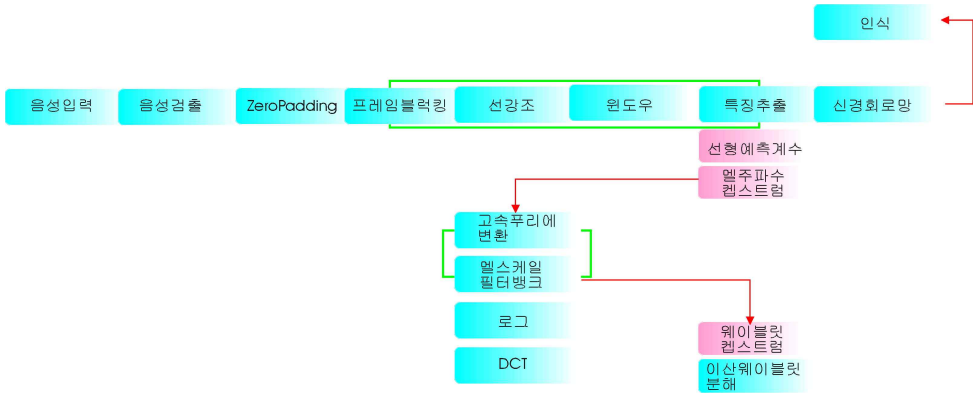


그림 5. 화자식별 시스템의 블럭 다이어그램

Fig. 5. Block diagram of speaker identification system

3. 전처리 과정

앞의 그림에서 음성신호의 입력에서부터 윈도우를 취하는 것까지를 전처리 과정이라 생각할 수 있다. 음성신호는 8kHz, 8bit/sample로 샘플링하고, 끝점검출을 통해 음성구간과 비음성구간을 분리하여 음성구간만을 취한다. 영삽입(zero padding)단계는 데이터 길이를 맞추기 위한 것이고, 다음의 프레임 블럭화 단계부터는 각 프레임 단위로 반복되어 특징을 추출하게 된다. 선강조 단계에서는 고주파 성분을 증가시키기 위해 1차 필터링을 하게 되고, 다음 단계로는 깁스(Gibbs)현상을 방지하기 위한 윈도우를 취하게 되어 다음의 특징추출 단계로 넘어가게 된다.

3.1 음성구간 검출

일정구간에서의 평균 에너지를 구하여 에너지 임계치를 넘는 구간 밖에 음성의 시작점과 끝점이 있다고 가정하여 구한다. 에너지를 이용한 방법은 유성음을 검출하기 적합한 방법이므로 이것만을 이용한 방법으로는 정확한 음성의 시작점과 끝점을 찾기 어려우므로 무성음 검출에 적합한 영교차율(zero crossing rate)을 동시에 이용한다. 단구간 에너지(E_n)를 식(3-1)과 같이 정의할 수 있고, 그림 6의 과정을 거쳐 생성된다.

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \qquad (3\text{-}1)$$

또 는

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m), \quad h(n) = w^2(n) \qquad (3\text{-}2)$$

$x(m)$: 입력신호 $x^2(m)$: 입력신호의 에너지

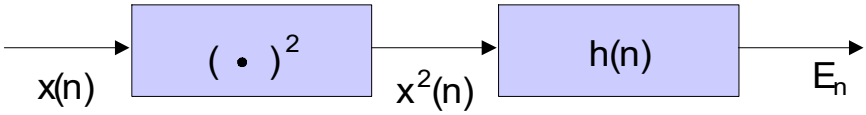
$w(n)$: 윈도우함수 $h(n)$: 임펄스 응답

시간종속 에너지 표현에 있어 윈도우 효과는 두 가지 표현 방식의 윈도우의 속성을 거론함으로써 입증될 수 있는데, 사각 윈도우와

$$h(n)=\begin{cases} 1 & 0\le n\le N-1 \\ 0 & ohter\,wise \end{cases} \tag{3-3}$$

해밍 윈도우이다.

$$h(n)=\begin{cases} 0.54-0.46\cos(2\pi n/(N-1)), & 0\le n\le N-1 \\ 0, & other\,wise \end{cases} \tag{3-4}$$



3.2 프레임 블럭화

검출된 음성신호는 프레임 단위로 선강조 및 해밍 윈도우를 취한다. 이후부터의 처리는 모두 프레임 단위로 이루어진다 [15]. 본 논문에서는 프레임의 크기는 32ms, 프레임 이동은 24ms, 오버랩은 8ms를 사용한다.

프레임을 블럭화할 때 푸리에 변환을 하기 위해서 프레임 구간의 길이를 2의 지수승 단위로 맞추게 되는데, 에너지와 영교차율을 이용하여 음성구간을 검출하고 난 후에는 각 샘플들의 데이터 개수가 상이하게 되므로 프레임 단위로 데이터를 처리할 때 마지막 프레임에서의 데이터 개수가 프레임 구간의 길이와 맞지 않기 때문에 영삽입을 하게된다.

샘플링 주파수 f[Hz](=$1/T$)로 표본화한 디지털 신호 $X(n)$의 샘플수가 N이라 하면 해석시간은

$$T_0 = T \times N \ [\sec] \tag{3-6}$$

이산푸리에변환 X(k)에서 k의 변화에 상당하는 주파수 간격은

$$\Delta f = \frac{1}{T \times N}[Hz] \tag{3-7}$$

본 논문에서는 해석시간

$$T_0 = \frac{256}{8000} = 0.032[\sec]$$

$$\Delta f = \frac{1}{T_0} = \frac{1}{0.032} = 31.25[Hz]$$

Mel-주파수 켑스트럼 계수를 구하는 경우, 필터뱅크 분석에서 Δf 간격으로 주파수 스펙트럼이 나타나고 이를 근거로 Mel-스케일 필터뱅크를 적용해서 계수를 구하였다.

3.3 선강조

디지털 음성신호는 고대역 통과 특성을 갖는 디지털 선강조 필터를 거친다. 이 필터를 사용하는 이유는 신호의 스펙트럼을 평탄하게 하기 위해서인데, 관련정보를 가진 고주파 포만트는 저주파 포만트에 비해 진폭이 더 작기 때문에 선강조가 필요하다 [21]. 주로 1차 FIR 필터로 선강조하며, 이를 z-도메인과 시간 도메인에서의 수식으로 표현하면 다음과 같다.

z-도메인에서 필터함수는

$$H(z) = 1 - a \cdot z^{-1} \quad 0 \le a \le 1 \tag{3-8}$$

시간 도메인에서 출력신호는

$$x'(n) = x(n) - a \cdot x(n-1) \tag{3-9}$$

필터계수 a는 0.95~0.98 범위의 값을 사용하는데, 주로 쓰이는 a의 값은 0.95이다.

4. 특징파라미터 추출방법

특징은 인식에 유용한 성분을 음성신호로부터 뽑아내는 과정이다. 그리고 특징추출은 일반적으로 정보의 압축, 차원 감소 과정과 관련된다. 하지만 특징추출에서는 이상적인 정답이 없기 때문에 음성인식을 위한 특징의 좋고 나쁨은 음성인식률로 판단된다. 특징추출의 주요 연구 분야는 인간의 청각특성을 반영하는 특징 표현, 다양한 잡음환경/화자/채널 변이에 강인한 특징, 시간적인 변화를 잘 표현하는 특징의 추출이다. 흔히 사용되는 특징추출 과정에서 청각특성을 반영한 것으로는 달팽이관의 주파수 응답을 응용한 필터뱅크 분석, mel 또는 Bark 척도 단위의 중심주파수 배치, 주파수에 따른 대역폭의 증가, 선강조 필터 등이 사용된다. 강인성을 향상시키기 위한 방법으로 가장 널리 사용되는 것은 채널의 영향을 줄이기 위한 켑스트럼평균차방법이다. 음성신호의 동적 특성을 반영하기 위하여 켑스트럼의 1차, 2차 미분값을 사용한다. 켑스트럼평균차방법 및 미분은 시간축 방향의 필터링으로 생각할 수 있으며 시간축 방향으로의 상관도가 적은 특징벡터를 얻는 과정이다. 필터뱅크 계수로부터 켑스트럼을 얻는 과정은 필터뱅크 계수를 상관도가 적게 바꾸기 위한 직교변환(orthogonal transform)으로 생각할 수 있다. 선형예측계수(LPC)를 이용한 켑스트럼을 사용한 초기의 음성인식에서는 선형예측계수 켑스트럼 계수에 대하여 가중치를 적용하는 리프터링(liftering)을 사용하기도 하였다.

최근의 특징 추출 분야의 연구 동향은 동적이고 유연한 필터 및 변환방법을 사용하며 통계적인 방법을 이용하여 학습하거나 최적화하는 것이다. 상관을 줄이기 위하여 PCA(principal component analysis)를 통하여 prewhitening하거나, LDA(linear discriminant analysis)를 통하여 특징을 변환하거나, 필터뱅크를 더 나은 성능을 갖는 필터(웨이블릿, 독립성분분석(independent component analysis))로 대치하는 등의 연구가 활발하다. 또한

가산잡음의 영향을 줄이기 위한 연구도 다양한 방면에서 이루어지고 있다.

본 논문에서는 특징 파라미터로 선형예측계수, Mel-주파수 켑스트럼 계수와 웨이블릿을 이용한 켑스트럼 계수를 취하였다. Mel-주파수 켑스트럼 계수의 경우 필터뱅크에서의 중심주파수 배치와 임계대역폭을 수정하여 기존의 필터뱅크와 다르게 구성하였으며, 웨이블릿을 이용한 켑스트럼 계수의 경우 Mel-주파수 켑스트럼 계수를 구하는 과정에서 필터뱅크를 웨이블릿 분해과정으로 대치하였다. 선형예측계수는 나머지 두 가지 특징파라미터와의 비교를 위해 기존의 방법을 그대로 사용하였다.

4.1 선형예측계수

4.1.1 선형예측계수의 사용배경

음성정보의 저장 혹은 전송을 실현하기 위해서는 데이터량 압축과 이미 함유된 정보 손실방지의 두 가지 상호 배치되는 점을 염두에 둘 필요가 있다. 즉, 필요정보의 손실을 막는 한도 내에서 전달에 쓰이는 정보 표본의 수를 최소할 할 필요가 있다. 저장된 음성 데이터량의 축약을 위해서는 음성 부호화 방식을 사용하는데 파형 부호화(waveform coding) 방식과 파원 부호화 (source coding)방식이 있다. 파형 부호화 방식은 적응차이진수화(ADPCM)과 같이 산술적 방법을 사용하여 낱낱의 음성정보의 표본을 본래의 값과 대응하는 작은 수로 대신 나타내어 저장에 요구되는 비트수를 감소시키는 방법이다. 즉, 음성신호의 원래 파형으로부터 산술적 방법에 의한 가공을 하는 한도 내에서 음성정보를 저장 혹은 전송하는 방식이다.

그러나 이 방식은 그 형태소적 특성상 정보의 축약력에 한계가 있기 마련이다. 이에 따라, 반복되는 음성파형을 그대로 수집하지 않고 반복파형을 발생시키는 원천을 찾아내 기록한다면 그 정보저장량의 비약적인 축약을 가져올 수 있는 것은 아닐까 생각할 수 있다 [16].

파원 부호화방식은 음성 생성 모델로부터 음성의 특징계수를 추출하여 전송하는 방식으로 성도 특성은 시변필터로, 유성음의 음원은 임의주기의 임펄스 파형으로, 그리고 무성음의 음원은 백색잡음으로 모델링한다. 음성 신호로부터 음성의 특징 계수를 추출하는 과정을 음성분석(speech analysis)이라 하며, 추출된 특징계수로부터 음을 재생하는 과정을 음성합성(speech synthesis)이라 한다. 다음은 본 논문에서 사용한 음성분석 방법의 한 가지인 선형예측계수 분석 방법을 설명하고자 한다.

4.1.2 선형예측계수 분석 방법

각각의 자기상관된 프레임을 선형예측부호화 파라미터로 변환하는 단계에서는 선형예측부호화 필터 계수를 구한다. 이런 계수들은 성도의 모양에 대한 정보를 주파수 영역을 근간으로 표현한 것이라고 생각하면 된다. 발음이 서로 다른 것은 발음할 때 성도의 모양이 다르기 때문이다. 따라서 성도의 모양에 대한 정보가 발음에 대한 정보라고 볼 수 있다. 또한 몇 차로 분석하느냐에 따라 분석의 성능이 달라지는데, 계산상의 복잡성과 메모리 한계 때문에 ITU-T의 G723.1에서와 같이 10차를 사용하고 있다 [8, 19]. 특정 프레임의 선형예측계수는 Levinson-Durbin의 방법으로 구할 수 있다. 선형예측부호화는 이전의 몇 개의 데이터로 바로 현재의 데이터를 찾는 필터이다. 즉, 샘플링된 음성 데이터를 프레임별로 a_n 이라는 계수를 찾는 것이다.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) \qquad (4\text{-}1)$$

여기서 $s(n)$, ..., $s(n-p)$는 한 프레임의 음성데이터이고, a_p는 선형예측계수이다. 과거의 p개의 데이터를 사용해서 현재의 $s(n)$를 찾을 수 있는 a_p를 찾을 수 있는 것이다.

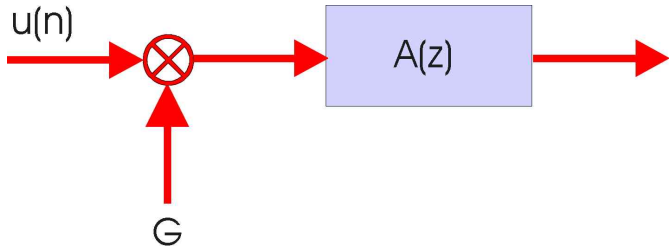


그림 7. 음성의 선형예측 모델

Fig. 7. Linear prediction model of speech

그림을 보면 s(n)과 u(n)의 정확한 관계는 다음과 같다.

$$s(n) = \sum_{i=1}^{p} a_i \, s(n-i) + Gu(n) \tag{4-2}$$

위 식을 사용하여 특징파라미터를 찾을 수 있는 근거는 다음과 같다.

음성의 발성은 허파의 횡막의 임펄스열 혹은 임의의 잡음($u(n)$)으로부터 성도(전달함수 $H(z)$)를 통해 나온다는 것을 전제로 하고 있다.

수식으로 표현하면 다음과 같다.

$$\begin{aligned} S(z) &= \sum_{i=1}^{p} a_i \, z^{-i} S(z) + GU(z) \\ H(z) &= \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \end{aligned} \tag{4-3}$$

여기서 G는 gain. 따라서 $A(z)$를 어떻게 잘 계산할 것인가의 문제로 귀결된다. 이 문제의 해를 구하는 방법은 다음과 같다.

$s(n)$의 추정치를 과거 샘플들의 선형 결합으로 생각하고 다음 수식과 같이 정의한다.

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k \, s(n-k) \tag{4-4}$$

그러면 예측 에러는 다음과 같다.

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k \, s(n-k) \tag{4-5}$$

예측 계수를 결정하기 쉽게 하기 위해서 시간 n에서의 단구간 음성(s_n)과 에러 세그먼트(e_n)를 정의한다.

$$\begin{aligned} s_n(m) &= s(n+m) \\ e_n(m) &= e(n+m) \end{aligned} \tag{4-6}$$

시간 n에서 평균자승오차(mean squared error)신호를 최소화하려면 a_k에 대해서 E_n을 미분하여 0가 되도록 식(4-7, 4-8)을 푼다.

$$E_n = \sum_m e_n^2(m) \tag{4-7}$$

$$E_n = \sum_m \left[s_n(m) - \sum_{k=1}^{p} a_k \, s_n(m-k) \right]^2 \tag{4-8}$$

$$\frac{\partial E_n}{\partial a_k} = 0, \qquad k = 1, 2, \ldots, p \tag{4-9}$$

$$\sum_m s_n(m-i) s_n(m) = \sum_{k=1}^{p} \hat{a}_k \sum_m s_n(m-i) s_n(m-k) \tag{4-10}$$

$$\phi_n(i,k) = \sum_m s_n(m-i) s_n(m-k) \tag{4-11}$$

$$\phi_n(i,0) = \sum_{k=1}^{p} \hat{a}_k \phi_n(i,k) \tag{4-12}$$

$$\begin{aligned} \hat{E}_n &= \sum_m s_n^2(m) - \sum_{k=1}^{p} \hat{a}_k \sum_m s_n(m) s_n(m-k) \\ &= \phi_n(0,0) - \sum_{k=1}^{p} \hat{a}_k \phi_n(0,k) \end{aligned} \tag{4-13}$$

따라서 평균자승오차의 최소값은 고정 항($\phi_n(0,0)$)과 예측 계수에 관련된 항으로 구성된다.

식(4-12)를 풀려면 $\phi_n(i,k)$, $1 \le i \le p$ and $0 \le k \le p$를 계산해야 한다. 그리고 나서 p 개의 연립방정식의 결과를 풀어야 한다. 실제로 식을 푸

는 방법은 m의 범위와 관련된 함수인데, 이 m은 분석구간과 평균자승오차가 계산되는 영역을 정의하는데 이용된다.

이 범위를 정의하는 두 가지 방법에는 자기상관 방법과 공분산 방법이 있는데 본 논문에서는 공분산 방법보다 계산상 더 간단한 자기상관 방법만을 다루고자 한다 [4, 6, 19].

자기상관 방법에 의해 계산을 계속해 보면 다음과 같다.

$$E_n = \sum_{m=0}^{N-1+p} e_n^2(m) \tag{4-14}$$

$$\phi_n(i,k) = \sum_{m=0}^{N-1+p} s_n(m-i)\, s_n(m-k), \quad 1 \le i \le p, \quad 0 \le k \le p \tag{4-15}$$

or

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)\, s_n(m+i-k), \quad 1 \le i \le p, \quad 0 \le k \le p \tag{4-16}$$

$$\phi_n(i,k) = r_n(i-k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)\, s_n(m+i-k) \tag{4-17}$$

자기상관함수는 대칭이므로 $r_n(-k) = r_n(k)$이고, 선형예측계수 방정식은 다음과 같이 표현될 수 있다.

$$\sum_{k=1}^{p} r_n(|i-k|)\ \hat{a}_k = r_n(i), \quad 1 \le i \le p \tag{4-18}$$

행렬 형태로는 다음과 같이 표현될 수 있다.

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \dots & r_n(p-1) \\ r_n(1) & r_n(0) & r_n(1) & \dots & r_n(p-2) \\ r_n(2) & r_n(1) & r_n(0) & \dots & r_n(p-3) \\ \vdots & \vdots & \vdots & & \vdots \\ r_n(p-1) & r_n(p-2) & r_n(p-3) & \dots & r_n(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(p) \end{bmatrix} \quad (4\text{-}19)$$

자기상관 값의 $p \times p$ 행렬은 대각 성분이 모두 같고, 대칭이므로 Toeplitz 행렬이고, 이 행렬은 여러 가지 잘 알려진 방법(가우스 소거법, 가우스-조르단 소거법, 확대행렬을 가우스 행렬로 변환하는 방법, 행렬의 역행렬을 곱해서 푸는 방법, 크래머 공식(Cramer's rule)을 이용한 방법 등)으로 풀 수 있는데, 다른 방법들보다 연산횟수가 적은 다음의 Levinson-Durbin 알고리듬으로 효과적으로 풀고 있다 [15]. 참고로 표 2에 $p \times p$ 가역 행렬에 대한 근사적인 연산 회수를 비교하여 나타내었다 [22, 23].

$$\begin{aligned} & E^{(0)} = r(0) \\ & k_i = \left\{ r(i) - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} r(|i-j|) \right\} / E^{(i-1)}, \quad 1 \le i \le p \\ & \alpha_i^{(i)} = k_i \\ & \alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \, \alpha_{i-j}^{(i-1)} \\ & E^{(i)} = (1 - k_i^2) \, E^{(i-1)} \end{aligned} \quad (4\text{-}20)$$

단, i=1에 대해서는 식 (4-20)에서의 합이 생략된다.

$$\begin{aligned} & a_m = LPC \ coefficients = \alpha_m^{(p)}, \quad 1 \le m \le p \\ & k_m = PARCOR(PARtialCORrelation) \ coefficients \\ & g_m = \log \ area \ ratio \ coefficients = \log\left(\frac{1-k_m}{1+k_m}\right) \end{aligned} \quad (4\text{-}21)$$

표 2. $p \times p$ 가역 행렬에 대한 근사적인 연산 회수

Table 2. Approximate operation counts for an invertible $p \times p$ matrix

방법	덧셈 회수	곱셈 회수
$Ax=b$를 가우스-조르단 소거법으로 푼다.	$\approx p^3/3$	$\approx p^3/3$
$Ax=b$를 가우스소거법으로 푼다.	$\approx p^3/3$	$\approx p^3/3$
[A|I]를 [I|A^{-1}]로 변형하여 A^{-1}를 구한다.	$\approx p^3$	$\approx p^3$
$Ax=b$를 $x=A^{-1}b$로 푼다.	$\approx p^3$	$\approx p^3$
행기약으로 det(A)를 구한다.	$\approx p^3/3$	$\approx p^3/3$
$Ax=b$를 크래머공식으로 푼다.	$\approx p^4/3$	$\approx p^4/3$
Levinson-Durbin 공식으로 푼다.	$\approx p^2$	$\approx p^2$

4.2. Mel-주파수 켑스트럼 계수

Mel-주파수 켑스트럼 계수를 구하는 방법을 간단히 설명한다. 프레임 단위로 전처리 단계를 거친 후의 음성신호는 고속푸리에변환(fast fourier transform)을 이용하여 주파수 영역으로 변환된다. 주파수 대역을 여러 개의 필터뱅크로 나누고 각 뱅크에서의 에너지를 구한다. 밴드 에너지에 로그를 취한 후 이산여현변환(discrete cosine transform)을 하면 최종적인 Mel-주파수 켑스트럼 계수가 얻어진다. 필터뱅크의 모양 및 중심주파수의 설정 방법은 귀의 청각적 특성(달팽이관에서의 주파수 특성)을 고려하여 결정된다. 아래 그림에서와 같이 삼각형 모양의 필터를 사용하였으며 중심주파수는 1 kHz 까지는 선형적으로 위치하고 그 이상에서는 Mel 척도로 분포하는 19개의 뱅크로 이루어져 있다. Mel-주파수 켑스트럼 계수는 $c1 \sim c10$까지의 10개를 사용하며 음식인식의 입력으로 사용되는 특징벡터는 10차 벡터가 된다. 그림 8은 이전의 블럭 다이어그램에서 Mel-주파수 켑스트럼 계수를 구하는 부분만을 나타낸 것이며, 그림의 순서대로 설명해 나갈 것이다.

그림 8. Mel-주파수 ㅊ
Fig. 8. Mel-frequency

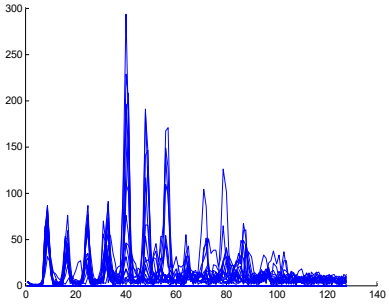
4.2.1 고속푸리에변환

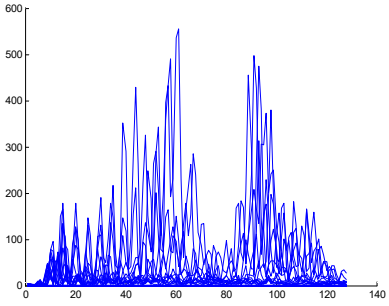
푸리에변환(FT)은 모든 주기적인 시간함
배에 해당하는 주파수로 이루어진 사인과 코
넘에서부터 출발하였고 파형을 주파수별로 분
변환과정에서 N을 분할하여 계산해서 계산회
에변환(FFT)이다. $x(n)$에 대한 고속푸리에변

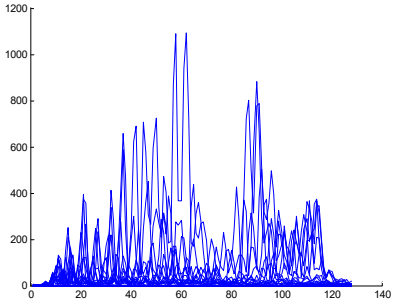
$$X(m) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi mn/N} \quad (m$$

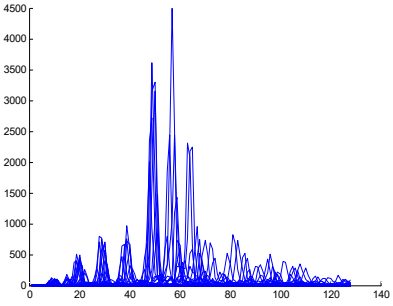
$$W = e^{-j2\pi/N}$$

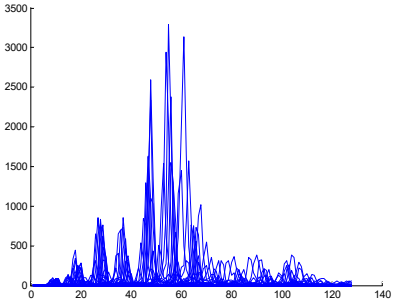
$$X(m) = \sum_{0}^{N-1} x(n)$$


300
250
200
150
100
50
0
0 20 40 60 80 100 120 140


600
500
400
300
200
100
0
0 20 40 60 80 100 120 140


1200
1000
800
600
400
200
0
0 20 40 60 80 100 120 140


4500
4000
3500
3000
2500
2000
1500
1000
500
0
0 20 40 60 80 100 120 140


3500
3000
2500
2000
1500
1000
500
0
0 20 40 60 80 100 120 140

(a) 화

4.2.2 Mel-스케일 필터뱅크

필터뱅크를 사용하면 스펙트럼에서 부적절한 부분들(피치 하모닉스나 잡음)을 로그를 취하기 전에 평탄화(smoothing)해서 감소시키는 효과를 가지고, 주파수 분해능과 스펙트럼의 각 부분에 주어지는 가중치를 제어할 수도 있다. 필터뱅크의 중심주파수는 Bark 또는 Mel 단위로 위치하고 대역폭은 임계 대역폭에 따라서 결정된다 [15]. 식(4-27, 28)은 Mel 주파수와 임계 대역폭의 특성을 수식으로 나타낸 것이고, 그림 15는 논문에서 사용된 Mel-스케일 필터뱅크를, 그림 16~21은 중심주파수 및 임계대역폭의 특성을 나타내는 그림이다. 그림16~21에서의 I은 수식에 의해 그려지는 특성이며, II는 논문에서 사용한 중심주파수 및 임계대역폭의 특성을 나타내는 것이다.

$$MelFreq = 2595 \log 10(1 + f/700) \tag{4-27}$$

$$\begin{aligned} CriticalBW &= 25 + 75[1 + 1.4(f/1000)^2]^{0.69}, \quad f > 1000 \\ CriticalBW &= 100, \quad f < 1000 \end{aligned} \tag{4-28}$$

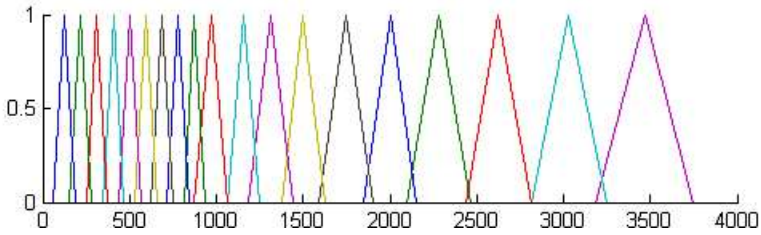
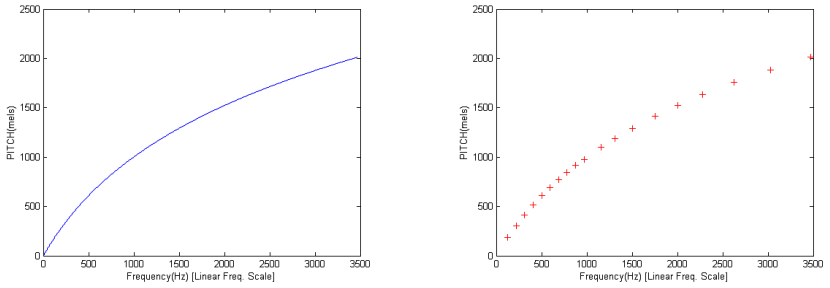


그림 10. Mel-스케일 필터뱅크

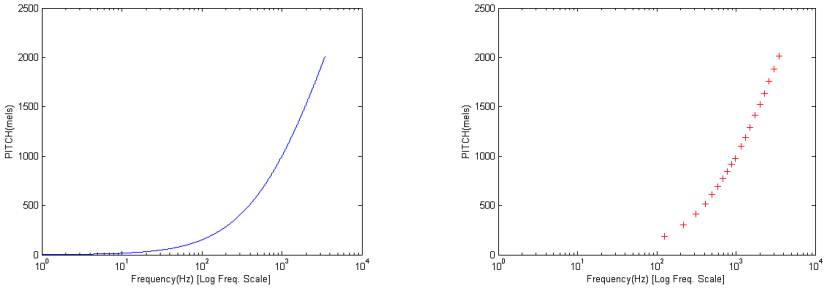
Fig. 10. Mel-scaled filter bank



(a) 특성 I (b) 특성 II

그림 11. 중심주파수(선형주파수스케일):
(a) 특성 I, (b) 특성 II

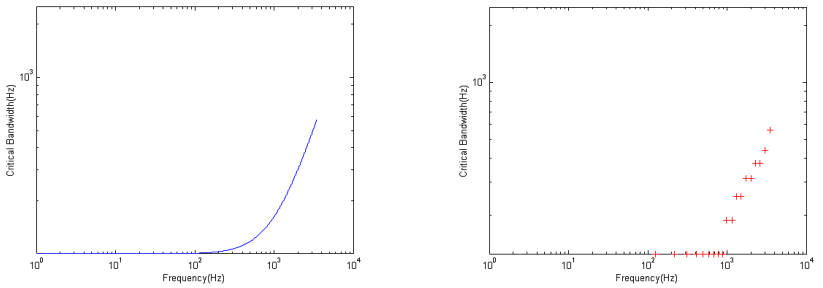
Fig. 11. Center frequency(linear freq. scale)
(a) Characteristics I, (b) Characteristics II



(a) 특성 I (b) 특성 II

그림 12. 중심주파수(로그주파수스케일):
(a) 특성 I, (b) 특성 II

Fig. 12. Center frequency(Logarithmic freq. scale)
(a) Characteristics I, (b) Characteristics II



(a) 특성 I (b) 특성 II

그림 13. 임계대역폭(로그특성):
(a) 특성 I, (b) 특성 II

Fig. 13. Critical bandwidth(logarithmic freq. scale)
(a) Characteristics I, (b) Characteristics II

필터뱅크 구현 방법은 선형 필터를 이용하여 시간 영역에서 구현되기도 하나, 일반적으로는 음성신호를 고속푸리에변환한 다음 각 밴드에 해당하는 계수의 크기에 가중치를 적용하여 합하는 방법으로 구현된다. 가중치의 모양은 삼각형, 사각형, 사다리꼴, 가우시안 형태 등을 사용할 수 있는데 흔히 삼각형 형태를 사용한다. 필터뱅크에서의 가중치를 통계적으로 최적화하는 방법도 최근에 연구되고 있다.

본 논문에서는 Δf(=31.25Hz)만큼의 이산적인 주파수 간격을 이용하여 흔히 사용하는 Mel-스케일을 62.5Hz에서 3750Hz까지의 주파수 범위에서 수정해서 사용하였고, 표 3은 흔히 사용되는 것과 논문에서 사용된 필터뱅크 분석 주파수 및 임계대역폭을 나타낸다.

표 3. Mel-스케일 필터뱅크

Table 3. Mel-scaled filter bank

Index	Mel-스케일(논문에서 사용)		Mel-스케일(일반적 사용)	
	Center freq (Hz)	BW (Hz)	Center freq (Hz)	BW (Hz)
1	125	125	100	100
2	219	125	200	100
3	313	125	300	100
4	406	125	400	100
5	500	125	500	100
6	594	125	600	100
7	688	125	700	100
8	781	125	800	100
9	875	125	900	100
10	969	188	1000	124
11	1156	188	1149	160
12	1313	250	1320	184
13	1500	250	1516	211
14	1750	313	1741	242
15	2000	313	2000	278
16	2281	375	2297	320
17	2625	375	2639	367
18	3031	438	3031	422
19	3469	563	3482	484

Mel-스케일 필터뱅크를 거친 값을 4.2.3절의 이산여현변환 과정을 통해 최종적인 Mel-주파수 켑스트럼계수를 구하게 된다.

4.2.3 이산여현변환

1974년 미 텍사스대학의 Rao 교수를 비롯한 3명의 연구진이 이산여현변환이라는 새로운 직교변환에 관한 논문을 IEEE학술지에 발표했다. 이산여현변환은 특히 영상의 압축에 탁월한 성능을 갖는 것으로 오늘날 멀티미디어 관련 국제표준인 H.261, JPEG, MPEG의 핵심요소로 자리잡고 있다. 문자, 도형, 일반 데이터 등을 무손실 압축하면 완전 복구가 가능하지만 압축률은 평균적으로 2대1정도이다. 반면 영상, 음성, 음향 등의 데이터를 인간의 눈과 귀가 거의 느끼지 못할 정도로 작은 손실을 허용하면서 압축하면 10 대1이상의 압축률을 쉽게 얻을 수 있다.

영상데이터를 효과적으로 압축하기 위한 목적으로 가장 널리 쓰이는 손실부호화 기법은 변환부호화이다. 이 방식의 기본구조는 공간적으로 높은 상관도를 가지면서 배열되어 있는 데이터를 직교변환에 의하여 저주파 성분으로부터 고주파 성분에 이르기까지 여러 주파수 성분으로 나누어 성분별로 달리 양자화하는 것이다. 이때 각 주파수 성분간에는 상관도가 거의 없어지고 신호의 에너지가 저주파 쪽에 집중된다. 단순 이진수화(PCM)에 비해 같은 비트율에서 얻는 변환부호화의 이득은 각 주파수 성분의 분산치의 산술평균과 기하평균의 비와 같다. 즉 저주파 쪽으로 에너지의 집중이 심화될수록 압축효율이 높다. 공간상의 데이터에 대한 단순 이진수화는 모든 표본을 같은 길이의 비트로(=m(bit/sample)) 표현하며, 신호대 양자화잡음비는 약 6m 정도가 된다. 반면 직교변환에 의해 주파수 영역으로 바뀐 데이터는 에너지가 많이 모이는(즉 분산치가 큰) 주파수 성분이 보다 많은 비트를 할당받아 그 주파수 성분을 보다 충실히 표현하도록 하고 있다. 분산치가 4배(즉 진폭이 2배)될 때마다 1비트씩 더 할당받는데 이렇게 되면 모든 주파수 성분에서 동일한 양자화 에러 특성을 갖게 된다. 여러가지 직교변환 가운데 이론적으로 영상신호의 에너지 집중특성이 가장 뛰어나 압축에 가장 효과적인 것은 카루넨-뢰브 변환(KLT)이다. 그러나 이것은 영상에 따라 변환함수가 새로 정의되어야 하므로 현실적으로 사용할 수 없다. 이 카루넨-뢰브 변환에

충분히 가까운 성능을 가지면서 구현 가능한 변환을 찾는 것이 Rao 교수팀의 목표였고 그 결과가 이산여현변환이다.

이산여현변환 공식은 다음과 같다.

$$c(k) = w(k) \sum_{n=1}^{N} x(n) \cos \frac{\pi(2n-1)(k-1)}{2N}, \quad k = 1, \ldots, N$$

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 \le k \le N \end{cases} \qquad (4\text{-}30)$$

$x(n)$: 입력신호

N: 입력의 크기

4.3 웨이블릿을 이용한 켑스트럼 계수

웨이블릿(wavelet)이란 “작은 파형의 조각”이라는 뜻이다. 큰 파형들도 작은 파형들의 집합으로 볼 수 있으며, 이는 곧 작은 파형들을 적당한 규칙에 의해 모으면 큰 파형의 형태를 이룰 수 있다는 것을 의미한다. 웨이블릿 변환은 이러한 개념으로부터 1990년대 중반, Grossmann과 Morlet에 의해 소개되었다. 초기에는 지진이나 음향의 신호를 함수적으로 분석하는데 주로 사용하였지만 지금은 다양한 분야로 널리 응용되고 있으며, 웨이블릿을 사용하게 된 배경은 다음과 같다. 신호 및 영상처리에서 주로 이용되던 푸리에 변환 (fourier transform)은 적당한 주파수 간격으로 얻은 푸리에 계수로부터 신호의 스펙트럼 정보를 분석하는 것이다. 하지만 푸리에 변환에 의한 방법은 주파수 영역에서만 신호를 분석할 수 있어서 신호의 시간정보와 주파수 정보를 동시에 파악할 수 없다는 단점이 있다. 이를 극복하기 위해 단구간 푸리에변환(short-time fourier transform) 이 도입되었는데, 이는 주파수 영역과는 독립적인 창함수(window fuction)를 기존의 푸리에 변환에 사용하여, 분석영역이 시간-주파수에 대해 일정하게 만든 것이다. 하지만, 단구간 푸리에변환(STFT)의 경우 그림 14에서도 볼 수 있듯이 분석영역이 단조롭다는

단점을 갖는다. 웨이블릿 변환은 이런 단점을 보완하기 위해 개발되었다고 할 수 있다. 즉, $\Psi_n(t)$를 높이변환 (scaling) 과 평행이동 (shifting) 한 결과로 생기는 함수를 사용하므로 분석영역이 유연하다.

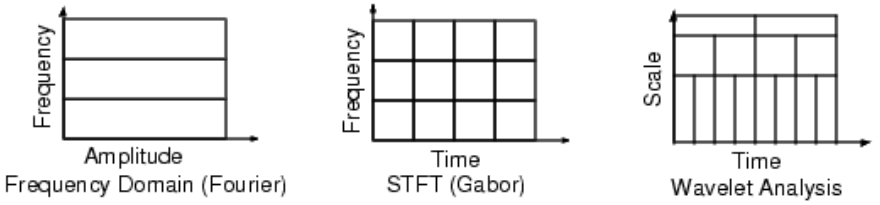


그림 14. 푸리에변환, 단구간푸리에변환, 웨이블릿의 비교
Fig. 14. Comparison of fourier transform, short-time fourier transform, and wavelet

본 논문에서의 웨이블릿은 다중해상도분석(multi-resolution analysis)의 개념을 이용하였으며, 다중해상도분석(MRA)의 개념은 *n*차원의 신호는 *n-1* 차원의 근사된 신호 X_{n-1}과 상세신호 e_{n-1}로 분리될 수 있고, 다시 *n-1* 차원의 신호는 *n-2*차원의 신호로 분리될 수 있다. 계속 반복하면 상당히 근사된 하나의 신호와 *n-1*개의 상세 신호로 표현이 될 수 있다. 근사된 신호는 저역통과필터(low-pass filter)를 통과한 신호, 상세신호는 고역통과필터(high-pass Filter)를 통과한 신호로 각각 대응된다. 즉, V_1 공간과 W_1공간이 직교이면 이들의 direct sum은 $l_2(Z)$를 생성하고 $l_2(Z) = V_1 \oplus W_1$로 표현된다. V_1은 근사신호에 해당하고, W_1은 상세신호에 해당한다. 이 단계는 저주파 브랜치에서 고주파/저주파 분할과 다운샘플링 단계를 거치면서 반복된다. 그림 15는 다중해상도 분석에서의 고주파와 저주파 통과필터를 거쳐 근사신호와 상세신호의 생성을 나타내는 트리구조이다. *x*1은 근사신호, *x*2와 *x*3는 상세신호를 나타낸다.

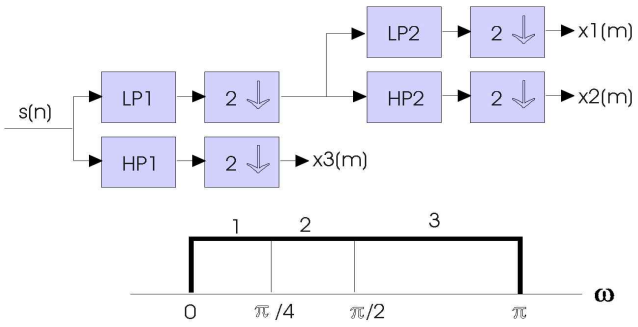


그림 15. 다중해상도분석의 트리구조
Fig. 15. Tree structure of multi resolution analysis

다중해상도분석은 고주파수에서는 시간해상도는 좋고 주파수 해상도는 좀 떨어지는, 저주파수에서는 시간해상도는 떨어지지만 주파수 해상도는 좋은 정보를 제공하기 위해 디자인되었다. 이런 접근은 특별히 주어진 신호가 짧은 구간에 대해서는 고주파수 성분을, 긴 구간에서는 저주파수 성분을 가졌을 때 의미가 있다.

본 논문에서는 이러한 다중해상도분석구조를 가지는 웨이블릿 변환과정을 Mel-주파수 켑스트럼계수를 구할 때의 필터뱅크를 대치하여 사용하였으며, coiflet을 사용한 분해과정에서 첫 번째 단계에서 생성되는 상세신호를 이용하였다. 그림 16은 사용된 coiflet의 고주파 필터를 나타내는 것이다.

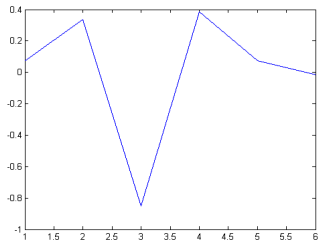


그림 16. coiflet 필터
Fig. 16. coiflet filter

5. 신경 회로망을 이용한 식별 알고리듬

인공지능 분야에서 기계학습에 관한 연구가 활발해지게 된 계기는 전문가시스템 개발시 전문가로부터 지식을 추출해내는 과정에 너무나 많은 시간과 노력이 투입되어야 함을 경험하면서부터였다. 소위 말하는 이 지식획득의 병목현상의 타개책으로 과거 데이터로부터 지식, 즉 중요한 규칙성을 자동으로 추출해 보려는 시도를 하게된 것이다.

그런데, 기계학습 중에서도 가장 실용성이 높은 귀납적 학습(inductive learning)은 주어진 데이터를 분석하여 그에 내재하는 중요한 패턴 혹은 규칙성을 찾아내는 것이다. 이러한 귀납적 학습을 위한 방법으로 인공신경 회로망을 이용할 수 있다. 인간의 뇌는 정보를 지능적으로 처리하기 위하여 엄청나게 많은 숫자의 신경세포(뉴런)들이 복잡한 체계로 상호 연결되어 있다. 각각의 뉴런들은 동시에 독립적으로 행동한다. 바로 이러한 인간의 뉴런들을 보다 단순하게 모델링하여 지능처리를 할 수 있는 소프트웨어나 컴퓨터를 만드는 데 사용하려는 것이 바로 인공신경 회로망이다. 이러한 신경회로망의 역사는 1940년대로 거슬러 올라간다. Warren McCulloch과 Walter Pitts는 1946년에 self-organizing 시스템에 관한 이론을 세우던 중에 신경세포(뉴런)에 대한 모델을 만들게 되었다 [18]. 나중에 Frank Rosenblat[18]이 신경 회로망에서 신경세포를 나타내는 처리요소에 대한 일종의 기호인 퍼셉트론을 고안해 냈다. 이 퍼셉트론은 간단한 패턴을 인식해 낼 수 있었다. 현재의 여러 가지 형태의 신경 회로망들은 대부분 이 퍼셉트론에서부터 출발하여 고안된 것들이다.

초창기의 고전적인 인공지능 연구에서는 지식을 적당하게 심벌로 표현하고, 그 심벌들을 다루는 핵심적인 규칙들을 정리하면 가능할 것이라 가정했었지만 실세계의 방대하고 다양한 지식들, 특히 수많은 상식들과 애매모호한 기준들을 수학공식이나 자료구조와 같은 방식으로 표현하는 것은 거의 불가

능하다고 보여지게 되었다. 그래서 대안으로 선택된 대표적인 것이 바로 인공 신경 회로망이며, 퍼지로직이나 혼돈이론(카오스 이론), 병렬처리, 분산처리 등도 있다. 특히 신경 회로망은 정보를 처리함에 있어서 여러 개의 뉴런들이 동시에 동작을 하며, 하나의 지식이 신경 회로망의 구조 자체와 가중치 집합에 영향을 주면서 분산되어 표현되므로 인공지능을 실현하는 데 중요한 초석으로 보여지고 있다. 본 연구에서는 신경회로망 모델 중 제어 알고리듬에 가장 일반적으로 적용되는 역전파 알고리듬을 사용하였다. 역전파 알고리듬은 1986년 RummelHart와 McClelland에 의해 제안된 학습방법[18]이며, 신호처리, 패턴분류, 동적시스템 제어 등 광범위하게 적용하고 있다.

5.1 다층 신경회로망의 구조

외부와 연결되어 있는 첫 번째 계층은 입력층(input layer)이라고 불리고, 마지막 계층은 출력층(output layer)이라고 불리며, 그 중간에 있는 층들을 중간층, 혹은 은닉층(hidden layer)이라고 불린다. 각각의 처리요소는 인간의 뇌에 있는 최소 기본단위인 뉴런(신경세포)과 같은 것이며 세포, 뉴로밈 또는 인공 신경세포라고 불린다. 하나의 뉴런 안에는 입력에 적용하여 출력을 결정하는 임계함수가 들어 있어서 뉴런의 출력을 제한하는 역할을 한다. 즉, 신경 회로망으로 들어가는 입력들과 신경 회로망 자체의 특징인 가중치(weight)가 각각 곱해져서 더해진 것이 신경 회로망이 받는 홍분이라고 할 때, 신경 회로망은 단순히 받는 것을 그대로 다음 뉴런들에게 전달하는 것은 아니다. 즉, 신경 회로망은 참을 수 있는 어떤 한계, 즉 임계치를 가지고 있어서 그 임계치(threshold)를 넘어선 입력은 다른 뉴런들에게 출력하지만 그렇지 않은 입력은 출력하지 않는다. 이러한 메카니즘은 일종의 필터와 같은 것으로 볼 수 있는데, 신경 회로망으로 들어가는 어떤 패턴이 잡음이 끼더라도 신경 회로망이 조그만 잡음 때문에 인식을 하지 못하는 일이 없는 이유가 된다. 여기서 서로 다른 계층간의 연결에 사용되는 가중치는 신경 회로망의 네트워크 특성에서 매우 중요한데 신경 회로망에서는 크게 두 가지

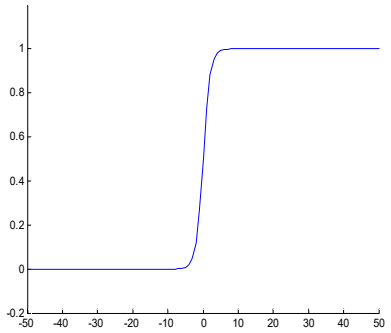
방식으로 가중치들을 사용한다. 하나는 각각의 뉴런들에 설정되어 있는 가중치들을 그대로 두고서 입력에서 출력까지 신호를 한 번 통과시키는 것이 있고, 또 하나는 한 번 통과시킨 최종결과를 정답과 비교하여 그 차이가 적어지는 방향으로 각각의 가중치들을 계속 수정해 나가는 것이 있다. 전자의 것이 바로 신경 회로망의 인식작용이며, 후자의 것은 학습작용이다. 즉, 인공신경 회로망에 있어서의 학습이라는 것은 적당한 입력들을 통하여 신경 회로망 내부의 각각의 뉴런들의 가중치들을 적절하게 조율해 나간다는 것을 의미하는 것이다.

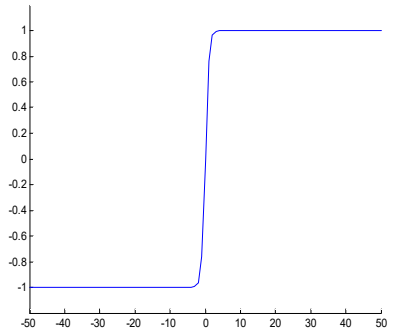
신경 회로망을 통과한 출력이 기대한 값과 다를 경우에 가중치는 수정될 필요가 있는데 이 때는 몇 가지 규칙들이 있다. 즉, 언제 가중치의 연속적인 수정과정을 멈출 것인가를 지정해 무한히 수정과정만 계속되는 것을 막게 되는데 이러한 일련의 수정과정을 학습이라고 한다. 그런데 네트워크를 학습시켜서 어떤 설정된 패턴을 인식할 수 있게 하거나 어떤 모르는 시스템의 내부를 그 시스템이 발생시키는 결과들을 보고서 추측하고 흉내내도록 하려면, 출력 뉴런의 정보를 그 이전의 어떤 계층에 있는 뉴런으로 되돌려주는 피드백(feedback)이 매우 중요하며, 피드백된 정보는 연결(connection)노드의 가중치를 조절하는데 사용된다. 피드백 되는 값은 일반적으로 출력에서의 에러이며, 중간층(은닉층)의 제일 앞단으로 되돌려져서 중간층의 가중치들을 수정한다. 다층 피드포워드 신경 회로망의 경우에 일반적으로 맨 처음에는 각각의 뉴런들의 가중치들을 완전히 임의적으로 설정하고서 시작한 다음, 학습 과정에서 역전파 알고리듬을 채택하여 출력에서부터 입력으로, 즉 역방향으로 오류를 줄이는 방향으로 가중치들을 조금씩 수정해 나가는 방식을 사용한다. 학습에 의하여 신경 회로망이 올바른 판단을 할 수 있도록 가중치 집합의 값들이 적절히 조율된다.

5.2 신경 회로망을 이용한 식별 시스템

다층 신경회로망에서 역전파 학습법은 주어진 모든 학습 패턴간의 실제 출력 값과 목표 출력 값을 비교하여 오차를 최소화하도록 수정하는 방향으로 연결 가중치 및 뉴런의 임계치를 조절하는 방법이다.

화자식별시스템에서 학습 알고리듬으로는 은닉층을 가지고 있는 다층 퍼셉트론을 역전파 알고리듬으로 학습시키는 방법을 사용하였다 [5]. 다층 퍼셉트론은 입력단으로 들어오는 입력으로, 출력 쪽으로만 단방향으로 흘러가면서 동작하기 때문에 다층 피드포워드 신경회로망이라고도 부르며, 학습할 때는 출력 쪽에서 발생한 오차가 입력측으로 피드백되면서 가중치들이 수정되기 때문에 역전파 네트워크라고도 부른다. 단층 퍼셉트론과는 다르게 XOR문제와 같이 선형적으로 분리가능하지 않은 문제도 학습에 의해 분류해 낼 수 있다는 장점이 있다. 다층 퍼셉트론의 히든노드와 출력노드의 가중치 합(weighted sum)에 부과되는 활성화 함수 중 하나인 시그모이드 함수는 두 가지가 있는데 모양은 다음과 같다.





18은 -1에서 1까지의 값을 가지는 바이폴라 시그모이드 함수이다. 바이너리 시그모이드와 바이폴라 시그모이드에 대한 식은 다음과 같다.

바이너리 시그모이드 활성화 함수

$$f = \frac{1}{1 + \exp(-w_{sum})} \tag{5-1}$$

바이폴라 시그모이드 활성화 함수

$$f = \frac{1 - \exp(-w_{sum})}{1 + \exp(-w_{sum})} \tag{5-2}$$

본 논문에서는 바이너리 시그모이드 활성화 함수를 사용하였으며, 신경 회로망의 구조에 대한 것은 다음 그림과 같다.

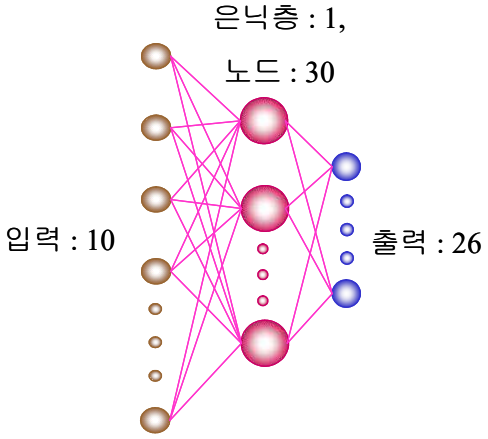


그림 19. 모델 1의
신경 회로망 구조
Fig. 19. Neural network structure of model 1

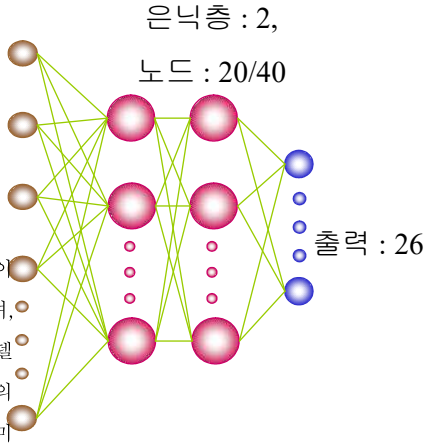


그림 20. 모델 2의
신경 회로망 구조
Fig. 20. Neural network structure of model 2

로망 모델에서 모델 1의 은닉층은 1개, 모델 2는 은닉층이 2개이
드는 2가지 모델 모두 26개이다. 모델 1은 입력노드가 10개이며,
력노드가 10개와 20개이다. 모델 1의 은닉노드는 30개이고, 모델
드는 20개와 40개이다. 입력노드가 10개인 경우는 특징파라미터의
일 경우이며 마찬가지로 입력노드가 20개인 경우는 특징파라미
여 사용하여 그 차수가 20차이다. 또한 입력층에서 중간층, 중간
층으로 노드마다 연결되어 있으므로 “fully connected 되어 있
른다. 입력노드는 단순히 들어오는 입력을 그대로 통과시키며 은
력노드는 가중치 합을 한 다음 시그모이드 함수를 통과시킨 출

6. 실험결과

본 논문에서 실험구성은 크게 음성의 특징파라미터의 추출 단계와 인식 단계로 나누어진다. 특징파라미터의 추출 단계에서는 선형예측계수와 Mel-주파수 켑스트럼계수, 웨이블릿을 이용한 켑스트럼 계수를 구하여 사용하였으며, 인식단계에서는 각각의 특징파라미터를 학습 데이터로 가지는 2가지 신경망 모델을 구성하여 실험하였다. 실험환경은 주변의 소음이 다소 적은 연구실 내에서 이루어졌고, Cirrus Logic의 오디오 칩셋인 CS1989를 사용하는 PCI방식의 사운드 카드와 저가형 마이크를 통해 음성신호를 입력받았으며, 8000Hz, 8bit/sample 포맷으로 저장하였다.

본 논문에서 음성의 특징파라미터 추출실험은 미리 정의한 단어인 “안녕”에 대해서 하였고, 실험대상자의 협조를 구해 두 가지 패턴으로 발음한 음성을 녹음하여 음성구간을 검출한 뒤 특징파라미터를 추출하였다. 음성구간의 검출은 전체 녹음구간에서 비음성구간을 버리고 음성구간만을 취함으로써 데이터량을 감소시키는 효과가 있다. 음성구간의 검출은 임계치를 정해 단구간의 에너지와 영교차율이 각각의 임계치를 넘을 때 음성구간으로 취하는 방법으로 실험하였고, 임계치가 800일 경우 정확하게 음성구간을 검출한 것으로 정의하였다.

음성의 인식 실험은 학습데이터에 따라 2가지 모델로 구성된 신경회로망을 사용하여 성능을 평가하였으며, 은닉층의 개수를 변화시키면서 성능 변화를 살펴보았다. 신경회로망의 입력정보는 3가지 음성 특징파라미터 추출 방법과 그 혼합에 의해 추출된 결과를 사용하였으며, 인식실험은 학습할 때와 거의 동일한 환경에서 시행하였다.

전체 실험의 제약 조건은 다음과 같다.

1. 블라스팅(Blasting)에 의한 음성검출 및 인식 성능 저하를 고려하여 실험 대상화자와 마이크의 거리를 약 20 cm 정도로 두어 실험을 하였다.

2. 지정된 단어에 대해서도 각 화자의 발성시간은 상이하므로 발성시간이 일정 시간을 초과하지 않는다는 가정 하에 음성추출구간을 1.024초로 고정시켰다.
3. 여러 화자의 중복음은 어떤 정형화된 일차원의 규칙성을 추출해 낼 수가 없기 때문에 한 화자가 내는 단독음만을 대상으로 화자식별 실험을 하였다[16].

6.1 음성검출 실험결과

음성검출과정은 기존에 사용되어 왔던 에너지 및 영교차율을 이용하는 방법에 의해 검출하였다. 전체 녹음구간에서 음성구간만을 검출하기 위해 단구간에서의 에너지를 이용하여 대략적인 구간을 결정하고, 단구간에서의 영교차율을 이용하여 음성의 시작점과 끝점을 검출하게 된다. 음성검출시 구간결정을 위해서 전체구간에 대해서 단구간 에너지를 이용하게 된다. 이 때 에너지가 임계치를 넘어서는 지점을 시작점으로 설정하고, 시작점으로부터 임계치에 미치지 못하게 되는 지점까지가 구간으로 결정되고 이 지점이 끝점이 된다.

본 논문에서는 전체 녹음구간의 시작으로부터 단구간 에너지가 임계치를 넘어서는 지점을 구간 시작으로 잡고, 녹음구간 종료 지점부터 역으로 검색하여 다시 임계치를 넘어서는 지점을 구간 끝으로 잡으므로, 계산량과 메모리를 절약할 수 있다. 영교차율이 구간 시작에서부터 임계치를 넘을 때 시작점으로, 구간 끝에서부터 임계치를 넘을 때 끝점으로 지정된다. 검출 결과 신호를 그래프로 그리고, 음성의 시작점과 끝점을 검출한 결과를 수직선으로 표시하였고, 에너지 임계치에 따른 음성검출의 정확도는 표 4와 같고, 그림 21～25에 음성검출결과에 대한 그림을 나타낸다.

표 4. 음성검출결과

Table 4. Result of speech dectection

	음성검출정확도	비고
에너지 임계치=300	95.19 %	영교차율임계치=3
에너지 임계치=500	97.60 %	
에너지 임계치=800	100.0 %	

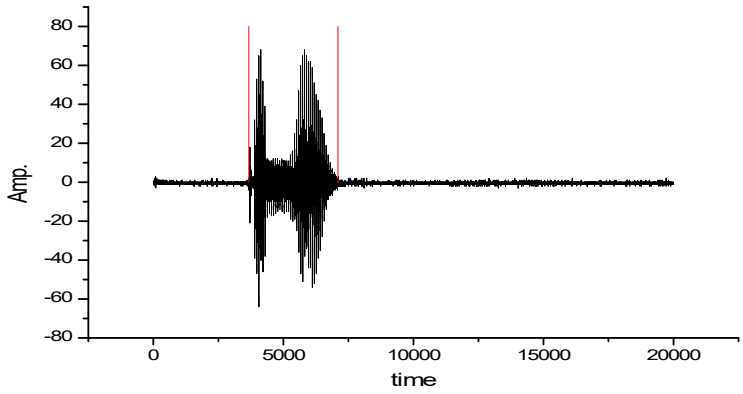


그림 21. 음성검출이 정확한 경우(화자 I)-임계치 300

Fig. 21. Case of exact speech detection(speaker I)-threshold is 300

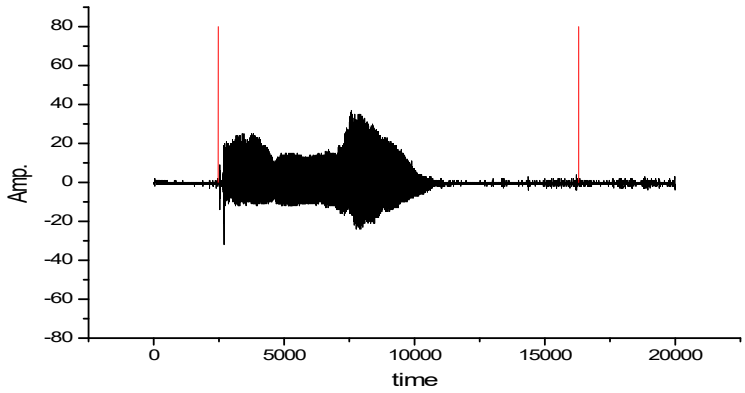


그림 22. 음성검출이 부정확한 경우(화자 V)-임계치 300

Fig. 22. Case of non-exact speech detection(speaker V)-threshold is 300

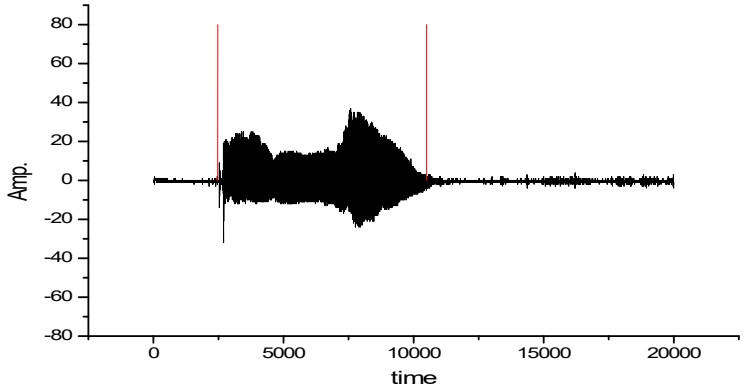


그림 23. 음성검출이 정확한 경우(화자 V)-임계치 500
Fig. 23. Case of exact speech detection(speaker V)-threshold is 500

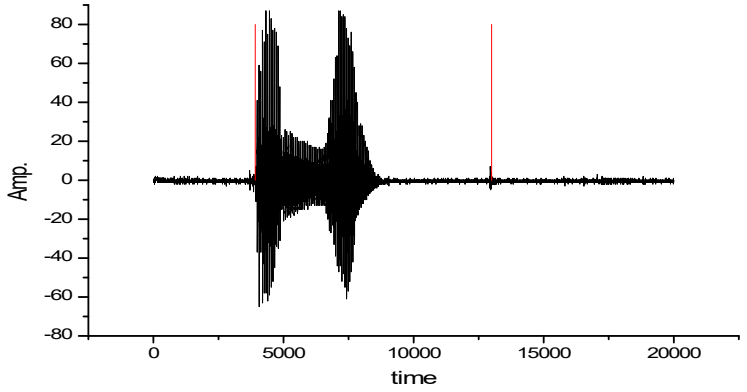


그림 24. 음성검출이 부정확한 경우(화자 X)-임계치 500
Fig. 24. Case of non-exact speech detection(speaker X)-threshold is 500

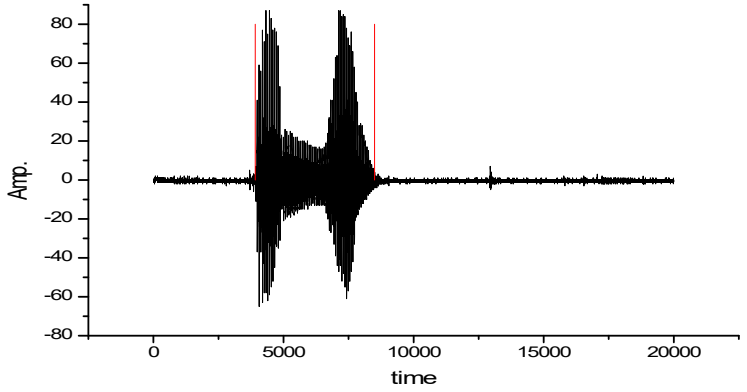


그림 25. 음성검출이 정확한 경우(화자 X)-임계치 800
Fig. 25. Case of exact speech detection(speaker X)-threshold is 800

위와 같이 음성을 검출하는 방법이 연구되어 오고 있지만, 아직까지 정확한 음성검출 결과를 얻을 수 있는 방법은 존재하지 않는다. 따라서 본 논문에서는 기존에 주로 사용되어 오던 단구간 에너지와 영교차율을 이용하여 음성을 검출하였고, 임계치를 300, 500, 그리고 800으로 했을 때 음성검출 정확도를 산출하여서 정확한 음성검출을 얻기 위해서는 임계치가 800이 되어야 함을 알 수 있었다. 그러나 음성검출 알고리듬으로 에너지와 영교차율을 이용하는데 있어서 정확도는 에너지 임계치와 영교차율 임계치에 의해 결정되는데 이는 경험적으로 설정되는 값이므로, 일반적으로 적용 가능한 방법이라고 할 수는 없다. 또한 에너지를 이용한 방법은 유성음 검출에, 영교차율을 이용한 방법은 무성음 검출에 쓰이고, 본 논문에서 지정된 단어는 유성음이 강한 음성학적 특징을 가지기 때문에 무성음의 검출에 사용되는 영교차율을 이용한 방법은 음성검출 정확도에 기여하는 효과가 적다는 것을 알 수 있다.

6.2 특징파라미터 추출 실험결과

선형예측계수의 추출은 과거 p개의 신호를 이용하여 현재의 신호를 예측하기 위한 선형예측모델에 의해 이루어지며, 이 방법은 고전적인 방법으로 본 논문에서 중점을 둔 Mel-주파수 켑스트럼 계수와 웨이블릿을 이용한 켑스트럼 계수와의 비교를 위해 이용되었다. 기존의 Mel-주파수 켑스트럼 계수는 고속푸리에변환을 거친 신호를 중심주파수를 Mel-스케일로 위치시키는 필터뱅크를 이용하였지만, 본 논문에서는 대략 400에서 1000Hz 이외의 구간에서만 중심주파수를 Mel-스케일로 위치시켜 변형된 필터뱅크를 이용하였다. 또한 이러한 Mel-스케일 필터뱅크를 웨이블릿 분해과정을 통해 생성되는 상세신호로 대치하여 켑스트럼 계수를 구하였다. 이러한 과정을 거쳐 생성된 특징파라미터는 그림 26~29와 같은 값을 가지며 대표적으로 4가지 특징 파라미터에 대해 화자 2명씩의 특징파라미터를 비교하여 나타내었다. 각각의 그림은 한 화자의 8개 음성 샘플에 대한 것이며 그림 26은 선형예측계수, 그림27은 Mel-주파수 켑스트럼 계수, 그림 28은 웨이블릿을 이용한 켑스트럼 계수 그리고 그림 29는 선형예측계수와 Mel-주파수 켑스트럼 계수의 혼합된 특징파라미터에 대한 그림이다.

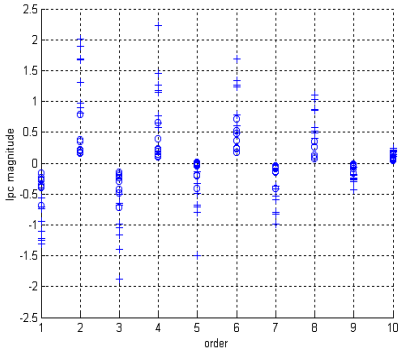


그림 26. 선형예측계수(화자C/E)
Fig. 26. LPC
(speaker C/E)

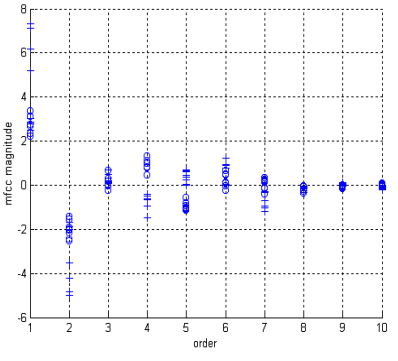


그림 27. Mel-주파수
켑스트럼계수(화자 B/D)
Fig. 27. MFCC(speaker B/D)

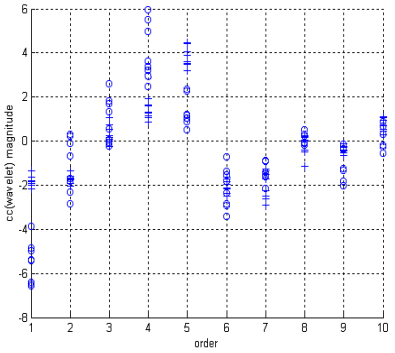


그림 28. 웨이블릿을 이용한
켑스트럼계수(화자 A/C)
Fig. 28. CC using wavelet
(speaker A/C)

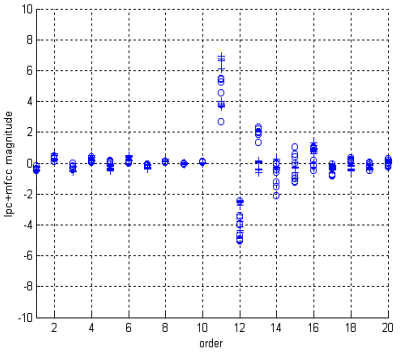


그림 29. 선형예측계수+Mel-주파수
켑스트럼계수(화자 K/L)
Fig. 29. LPC+MFCC
(speaker K/L)

6.3 화자식별 실험결과

앞에서 소개된 세 가지 특징 파라미터를 이용하여 2개의 신경회로망 모델을 구성하여 화자식별 테스트를 하였으며, 신경회로망 구성은 표 5와 같다.

표 5. 신경회로망의 구성
Table 5. Configuration of neural network

<table>
<tr><th></th><th>입력파라미터</th><th>입력
개수</th><th>은닉층
개수</th><th>은닉층
노드수</th><th>학습방법</th><th>비고</th></tr>
<tr><td>모델 1</td><td>LPC
MFCC
CC(Wavelet)</td><td rowspan="2">10</td><td>1</td><td>30</td><td rowspan="3">역전파
방법
(Back
Propagation)</td><td rowspan="3">개인당
학습
데이터:
3개</td></tr>
<tr><td>모델 2</td><td>LPC
MFCC
CC(Wavelet)</td><td rowspan="2">2</td><td rowspan="2">60</td></tr>
<tr><td></td><td>LPC+MFCC</td><td>20</td></tr>
</table>

신경회로망의 학습데이터는 각 화자(A～Z)마다 3개씩 총 78(=26×3)개를 실험대상자의 협조를 구해 음성 신호를 서로 다른 패턴으로 발음하게 하여 특징파라미터를 추출해서 사용하였고, 테스트 데이터 또한 각 화자마다 5개씩 총 130(=26×5)개를 학습 데이터와 동일한 방법으로 특징 파라미터를 추출해서 사용하였다. 실험의 제한조건은 음성검출 실험과 동일하며, 신경회로망 모델의 은닉층의 개수 및 입력 파라미터를 변화시키면서 학습한 2개의 모델을 준비하였다.

각 모델에 따른 특징파라미터에 대한 화자 인식율은 표 6, 7과 같다.

표 6. 신경회로망 모델1의 화자인식율

Table 6. Speaker identification rate of neural network model 1

	학습정확도	인식율	학습회수	비고
LPC	100%	52.31%	23,989회	에러허용율=0.01
MFCC	100%	50.77%	13,501회	
CC(Wavelet)	100%	43.85%	3,872회	

신경회로망 모델 1은 각 특징파라미터를 입력으로 은닉층의 개수가 1개일 때(은닉층의 노드수:30)의 모델이며, 실험결과는 위와 같다. 활성화 함수는 바이너리 시그모이드 함수이다.

표 7. 신경회로망 모델2의 화자인식율

Tabel 7. Speaker identification rate of neural network model 2

	학습정확도	인식율	학습회수	비고
LPC	100%	56.15%	3,785회	에러허용율=0.01
MFCC	97.60%	57.69%	4,134회	
CC(Wavelet)	100%	56.92%	1,519회	
LPC+MFCC	100%	65.38%	2,967회	

신경회로망 모델 2는 각 특징파라미터를 입력으로 은닉층의 개수가 2개일 때(은닉층1의 노드수:20, 은닉층2의 노드수:40)의 모델이며, 실험결과는 위와 같다. 활성화 함수 또한 신경회로망 모델1과 같다. 신경회로망 모델2에서는 각각의 독립적인 특징파라미터를 입력으로 하는 것은 입력노드가 10개이고 특징파라미터를 혼합하여 입력으로 취한 것의 입력노드는 20개이다. 또한 은닉층의 개수는 공통적으로 2개(은닉층1의 노드수:20, 은닉층2의 노드수:40), 출력 노드는 앞과 동일한 26개이며 활성화 함수 또한 같은 바이너리 시그모이드 함수이다.

먼저 학습 정확도를 보면 신경회로망 모델2에서 Mel-주파수 켑스트럼 계수를 특징 파라미터로 하는 경우만 제외하고 모두 100%의 학습 정확도를 가지므로 전반적으로 에러허용율 0.01에서의 학습정확도는 높음을 알 수 있

다. 각 신경회로망 모델에 대해서 학습회수는 모델2로 갈수록 적어지며, 각 특징 파라미터에 대한 학습회수는 웨이블릿을 이용한 방법이 모델에 관계없이 가장 적으며, 선형예측계수와 Mel-주파수 켑스트럼계수의 경우는 모델에 따라 달라짐을 알 수 있다.

인식결과를 살펴보면 신경회로망 모델 1에서는 선형예측계수를 이용한 방법이 Mel-주파수 켑스트럼계수와는 근소한 차이로 가장 높은 인식율을 나타냈고, 웨이블릿을 이용한 방법이 가장 낮은 인식율을 나타냈다. 신경회로망 모델2에서는 Mel-주파수 켑스트럼계수를 이용한 방법의 인식율이 높았으며, 선형예측계수를 이용한 방법의 인식율이 낮았고, 가장 높은 인식율을 보인 것은 선형예측계수와 Mel-주파수 켑스트럼계수를 혼합한 방법이었다. 모델2에서는 특징파라미터에 따른 인식율의 차이가 모델1에 비해 적었고, 모델1에 의한 방법보다 전체적으로 높은 인식율을 나타냈다.

본 논문에서의 실험결과로 각 특징벡터를 혼합하여 사용하면서 신경회로망의 은닉층의 개수를 증가시킬 때 학습회수가 줄고 인식율이 비약적으로 증가한다는 사실을 알 수 있었다.

7. 결론

본 논문에서는 인식 알고리듬으로 신경회로망을 사용한 화자식별 시스템에 있어서 기존의 특징추출방법과 현재의 연구 추세에 따른 방법을 사용하여 식별한 결과를 인식율과 학습속도를 성능기준으로 삼아 비교하였다. 화자식별시스템의 성능을 비교하기 위한 실험과정을 순서대로 나열하면 음성신호 수집, 음성검출, 특징파라미터 추출, 화자식별과정이며, 각각의 내용을 요약하면 다음과 같다.

음성신호 수집과정에서는 26명의 화자(A～Z)를 실험대상자로 선정하였고, 학습데이터와 인식데이터를 합해 각 화자에 대해 8개씩 샘플을 수집하였다. 음성신호의 수집에 있어 화자의 다양한 발음패턴을 인식결과에 반영하기 위해서 실험대상자의 협조를 구해 음성 신호를 서로 다른 패턴으로 발음하게 하여 특징 파라미터를 구하였다. 본 논문에서 사용된 특징 파라미터들의 각 화자에 대한 일관성은 발음의 억양보다 장단에 좀 더 비례하여 감소하는 경향을 보였으며, 따라서 음절의 장단에 차이를 둔 발음패턴을 사용하였다.

음성검출과정에서는 입력 음성에 대한 특징 파라미터를 산출하는 과정에서의 계산량을 감소시키기 위해 단구간 에너지와 영교차율을 사용하여 음성의 끝점검출을 하였다. 끝점검출을 통해 추출된 음성구간은 이후 단계의 고속푸리에변환과정에서 요구되는 2의 지수승의 데이터 개수를 갖지 않고 각각 다른 길이를 가지므로 데이터 길이를 2의 지수승 개로 일정하게 맞추기 위해 영삽입을 하게된다. 물론 입력되는 음성구간은 정해진 길이를 초과하지 않는다는 가정 하에서 적용되고, 실험에 의해서 비정상적으로 긴 발음을 하지 않는다는 조건하에서 이 가정은 타당한 것으로 볼 수 있다. 그래서 단구간 에너지의 임계치가 800이하일 때와 영교차율의 임계치가 3일 때 음성구간을 추출한 후, 영삽입에 의해 길이를 일정하게 맞춘 데이터를 이용하여 특징 파라미터를 추출하여 2가지 신경회로망 모델에 적용하였다. 단구간의

특징파라미터를 추출하기 위해서 프레임 블럭화를 하였고, 고주파 성분을 증가시키기 위해 1차 필터로 선강조하였으며, 프레임 끝에서 발생하는 불연속에 의한 깁스현상을 막기 위해 해밍 윈도우를 씌워주었다. 특징파라미터 중에서 선형예측계수는 자기상관함수와 Levinson-Durbin 방법을 이용해서 구해지고, Mel-주파수 켑스트럼계수는 고속푸리에변환을 통해 스펙트럼을 구한 뒤 인간의 청각적 특성을 반영한 Mel-스케일 필터뱅크를 통과한 뒤 로그값을 취해서 이산여현변환을 거쳐 구해진다. 그런데 앞에서 언급했듯이 최근에는 특징추출을 위한 연구분야에서 웨이블릿이나 독립성분분석 등의 더 나은 필터로 필터뱅크를 대치하는 등의 연구가 활발하다. 따라서 본 논문에서는 필터뱅크를 통과시키는 대신 웨이블릿 분해과정을 거친 결과 값에 절대치를 취해 웨이블릿을 이용한 켑스트럼계수를 또 하나의 특징파라미터로 추출하였다. 그리고 각 특징파라미터를 혼합하여 신경회로망 입력으로 취하는 방법으로도 실험하였다.

화자식별 과정에서는 비선형 맵핑 능력이 좋은 신경회로망을 사용하였으며, 특징파라미터와 그 혼합형태를 신경회로망 입력으로 사용하였고, 2가지 신경회로망 모델을 구성하여 사용하였으며, 각 신경회로망 모델은 은닉층의 개수, 노드 수나 입력 파라미터가 다르게 구성되어 있으며, 특징파라미터나 모델의 구성이 다른 경우에의 성능변화를 관찰할 수 있었다.

실험결과 신경회로망 모델 2에서 특징파라미터로 선형예측계수와 Mel-주파수 켑스트럼계수를 혼합한 형태를 사용했을 때 가장 우수한 성능을 나타내었으며, 은닉층의 개수가 증가했을 때 더 나은 결과를 보여주었다. 그리고 신경회로망 모델 1에서는 특징파라미터에 따른 인식율의 차이가 최대 10% 정도까지이지만 신경회로망 모델 2에서는 특징파라미터간의 인식율 차이가 1% 안팎으로 다소 적음을 알 수 있다. 따라서 인식율도 높으면서 특징파라미터에 따른 인식율의 차이도 적은 신경회로망 모델2에서 동일한 은닉층의 구조를 가지고 입력노드의 수만 다르게 해서 특징파라미터의 혼합형태를 모델의 입력으로 사용하였다. 그 결과 모델 2의 다른 방법들과 비교하여

인식율이 증가함을 알 수 있는데, 그 차이가 10%정도로 성능이 상당히 개선됨을 알 수 있었다.

본 실험에서 신경회로망 모델의 인식정확도는 사용된 모든 특징파라미터에 관계없이 은닉층의 수에 비례하여 증가하고, 학습회수 또한 그에 비례하여 감소함을 알 수 있었다. 또한 clean환경이 아닌 실험실 환경에서 실험하였고, 화자의 학습 데이터의 개수도 3개로 제한되어 있었으며, 지정된 단어이지만 실험대상자의 발음패턴이 다르다는 점이 화자식별율을 감소시킬 수 있는 요인임에도 불구하고 위의 실험결과와 같은 성능을 보인 점은 화자식별율이 높고, 인식시스템의 모델이 보다 단순한 화자식별시스템의 구현이 가능하다는 기대를 던져주었으며, 앞으로 화자식별시스템에 있어 대규모 화자집합에 적용할 수 있는 보다 확실하고 일반적인 특징파라미터를 찾아 보다 정확하게 식별할 수 있는 해결책을 찾아야 할 것이다.

참 고 문 헌

[1] R.L. Klevans and R.D. Rodman, *Voice recognition*, Artech House, 1997.

[2] S.H.S. Salleh, A. Zuri, Z. Yusoff, S. Rahman and Chieh Lim Soon, " Implementation of Speaker Identification System by Means of Personel Computer," *TENCON 2000. Proceedings*, vol. 1, pp. 43~48 , 2000.

[3] P. Sivakumaran and A.M. Ariyaeeinia, "The Use of Sub-band Cepstrum in Speaker Verification," *Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1073~1076, 2000.

[4] R.J. Mammone, Xiaoyu Zhang and R.P. Ramachandran, "Robust Speaker Recognition," *IEEE Signal Processing Magazine*, vol. 13, sept., pp. 58~71, 1996.

[5] K. Vieira, B. Wilamowski and R. Kubichek, "Speaker Identification Based on a Modified Kohonen Network," *International Conference on Neural Networks*, vol. 4, pp. 2103~2106, 1997.

[6] K. Vieira, B. Wilamowski and R. Kubichek, "Speaker Verification for Security Systems Using Artificial Neural Networks," *23rd International Conference on Industrial Electronics, Control and Instrumentation*, vol. 3, pp. 1102~1107, 1997.

[7] Jiyong Ma and Wen Gao, "An Approach to Robust Speaker Recognition Using Stochastic Matching," *5th International Conference on Signal Processing Proceedings*, vol. 2, pp. 803~806, 2000.

[8] A.I. Aarskog and H.C. Guren, "Predictive Coding of Speech Using Microphone/Speaker Adaptation and Vector Quantization," *IEEE Transactions on Speech and Audio Processing*, vol. 2, April, pp. 266~273, 1994.

[9] J. Koolwaaij, L. Boves, H. Jongebloed and E. den Os,"On Model Quality and Evaluation in Speaker Verification," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, pp. 3759~3762, 2000.

[10] Qi Li, Biing-Hwang Juang and Chin-Hui Lee, "Automatic Verbal Information Verification for User Authentication," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 585~596, Sept., 2000.

[11] V. Vuckovic, "Dynamic Time-Warping Method for Isolated Speech Sequence Recognition," *5th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service*, vol.1, pp. 257~260, 2001.

[12] S. Molau, M. Pitz, R. Schluter and H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," *Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 73~76, 2001.

[13] M. Inal, E. Butun, K. Erkan, M. Yildirum and C. Ceken, "Comparison of Linear Predictive Analysis Methods for ANN-Based Speaker Identification," *Proceedings of the 5th Seminar on Neural Network Applications in Electrical Engineering*, pp. 109~112, 2000.

[14] J.W. Picone, "Signal Modeling Techniques in Speech Recognition," *Proceedings of the IEEE*, vol. 81, Sept., pp. 1215~1247, 1993.

[15] L.R. Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recogntion*, Prentice-Hall, 1978.

[16] 박경범, *음성의 분석 및 합성과 그 응용*, 도서출판그린, 1997.

[17] L.R. Rabiner and R.W. Schafer *Digital Processing of Speech Signals,* Prentice-Hall, 1978.

[18] 이상원, *Turbo C로 학습하는 기계신경망*, Ohm사, 1998.

[19] B. Wei, J. Wang and J.D. Gibson, "Enhanced Celp Coding with Discrete Spectral Modeling," *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing,* pp. 111~113, 2001.

[20] Li Liu, Jialong He and G. Palm, "Signal Modeling for Speaker Identificaion," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 665~668, 1996.

[21] M. Bilginer Gulmezoglu, V. Dzhafarov, M. Keskin and A. Barkana, "A Novel Approach to Isolated Word Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, Nov., pp. 620~628, 1999.

[22] Howard Anton and Chris Rorres, *Elementary Linear Algebra Application Version*, John Wiley & Sons, 1994.

[23] S.M. Biyiksiz, "ARMA Modeling Based on Partial AR and MA Parameter Appriximation", 4th *Annual ASSP Workshop on Spectrum Estimation and Modeling*, pp. 397~401, 1988.

Feature Extraction and Performance Comparison for ANN-based Speaker Identification Systems

Jae-hoon Yu

Department of Electrical Engineering, Graduate School
Pusan National University

Abstract

The voice recognition technology is separated two parts that are speech and speaker recognition. At present, speech recognition is generally studied but speaker recognition is a useful technique because that can be applied in the real life. In this paper, one of speaker recognition methods, namely speaker identification, is proposed to identify certain speaker. The results of ANN-based speaker identification systems for feature parameters are compared to evaluate the system's performance. In the preprocessing step, the short-time energy and the zero crossing rate are employed to separate the voiced regions and the unvoiced regions. The calculation cost can be reduced by the preprocessing.

In the identification step, the 78 speech samples of 26 speakers are trained by the back-propagation algorithm of neural network, and one speaker is identified as a test bed. Input variables of neural network are

the feature parameters-LPC, MFCC and cepstral coefficients combined with wavelets. Neural network model 2 using mixed feature parameters for input has the best performance than others in the speaker identification system.