論文 87-24-5-24

Mellin 변환을 이용한 격리 단어 인식

(An Isolated Word Recognition Using the Mellin Transform)

金 鎭 晩*, 李 商 郁**, 高 世 文***

(Jin Man Kim, Sang Uk Lee and Se Moon Goh)

要 約

본 논문에서는 화자 종속격리 숫자음 인식을 위하여 Mellin 변환을 이용한 인식 알고리듬을 제안하였다. Mellin 변환은 스케일 성분을 위상 성분으로 바꾸어 주는 성질이 있으므로 음성인식 알고리듬의 시간축 교정 문제를 이런 성질을 이용하여 해결하도록 시도하였다. 영교차율, 대수에너지, 자기상관 계수, 1차예측계수, 예측오차의 에너지 등의 음성특징에 Mellin 변환을 취하면 좋은 결과를 얻을 수 있음을 알았다. 한편 두 패턴 사이의 근사도를 측정하기 위하여 출력신호의 차이를 이용하였다. 실제 본 논문에서 제안한 알고리듬을 한국 숫자음에 적용한 결과 83.3%의 인식율을 얻을 수 있었다. 인식율은 LPC거리 척도 등의 방법보다 떨어지나, Mellin 변환이 음성인식의 시간축 교정에 효과적으로 이용될 수 있음을 알았다.

Abstract

This paper presents a speaker dependent isolated digit recognition algorithm using the Mellin transform. Since the Mellin transform converts a scale information into a phase information, attempts have been made to utilize this scale invariance property of the Mellin transform in order to alleviate a time-normalization procedure required for a speech recognition. It has been found that good results can be obtained by taking the Mellin transform to the features such as a ZCR, log energy, normalized autocorrelation coefficients, first predictor coefficient and normalized prediction error. We employed a difference function for evaluating a similarity between two patterns.

When the proposed algorithm was tested on Korean digit words, a recognition rate of 83.3% was obtained. The recognition accuracy is not compatible with the other technique such as LPC distance however, it is believed that the Mellin transform can effectively perform the time-normalization processing for the speech recognition.

I. 서 론

음성인식은 제5세대 컴퓨터 및 로보트, 인공지능 등 인간과 기계와의 의사전달을 필요로 하는 분야에 필수적인 기술로서, 음성처리 기술의 발전과 반도체 기술의 향상에 힘입어 이에 대한 연구가 활발히 진행되고 있다.[1,2]

음성인식은 처리하고자 하는 음성의 종류에 따라 격리 단어인식,[3,4] 연결단어인식,[5,6] 연속음성인식[7] 등으로 나누어진다. 현재 격리 단어인식 시스템은 부분적으로 실용화가 되어 있으나, 음성인식의 궁극적인 목표라고 할 수 있는 연속되는 음성에 대한 인식은 아직도 어려운 과제로 남아 있다.

*準會員, **正會員, 서울大學校 制御計測工學科
(Dept. of Control and Instru. Eng.)
***正會員, 國防科學研究所
(Agency For Deffence Development.)

接受日字 : 1987年 1月 30日

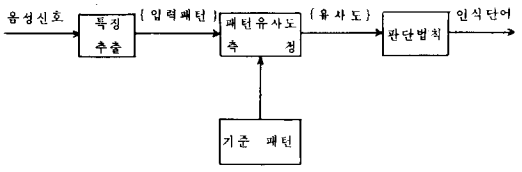


그림 1. 음성인식 시스템
Fig. 1. Speech Recognition System.

격리 단어 인식시스템은 그림 1에서 보는 바와 같이, 한 명 또는 그 이상의 화자에 의해 발음된 인식대상 발음들로부터 기준패턴을 구성한 후에, 입력음성과 기준패턴들 중에서 가장 유사한 패턴을 찾아내어 인식한다. 이때 각각의 패턴의 길이가 다른 현상을 교정하기 위한 time alignment 과정이 필요한데, 이 과정이 음성인식 알고리듬에서 가장 중요한 부분을 차지한다.

이 문제를 해결하기 위한 방법으로는 dynamic time warping (DTW),[8,9] vector quantization, hidden Markov model에 의한 방법들이 있다. 이 중에서 현재 DTW가 가장 광범위하게 사용되고 있으나, 이 방법은 계산량과 메모리가 상당히 많아지는 단점이 있다. 따라서 DTW의 계산량과 메모리를 줄이기 위한 알고리듬이 많이 연구되었는데, DTW의 특성상, 실시간 구현은 아직도 어려운 문제이다.[10,11,12]

본 논문에서는 DTW 방법을 대체할 수 있는 새로운 알고리듬을 개발하기 위하여 Mellin 변환을 도입하였다. Mellin 변환은 광학 영상처리와 음향 신호처리 분야에서 상당한 관심을 끌어 왔는데[13,14], Mellin 변환의 크기가 그 독립 변수의 스케일에 불변하다는 성질 때문이다. 본 논문에서는 이러한 Mellin 변환의 스케일 불변 성질을 time alignment 과정에 이용하여 효과적인 음성인식 알고리듬을 개발하고자 하였다.

그러나 음성신호에 Mellin 변환을 실제로 적용할 때에는 여러 가지 문제가 있다. 즉 짧은 단어의 경우에도 음성 데이타의 수는 수천개에 달하기 때문에 그 데이타에 직접 Mellin 변환을 할 수는 없다. 따라서 음성 신호의 데이타 수를 축소시킬 필요가 있는데, 본 논문에서는 가장 효과적인 축소방법을 찾아내어 데이타를 축소시킨 후 Mellin 변환을 적용하였다.

이와같이 Mellin 변환을 적용하여 스케일 성분을 제거시킨 후에는 그 출력을 이용한 패턴 비교과정이 필요하게 된다. 본 논문에서는 Mellin 변환 뿐만 아니라 그 중간 과정에서 얻을 수 있는 homomorphic 신호를 사용하여 패턴을 비교하도록 하였다. 단 Homomorphic 신호에서는 스케일 성분이 translation 성분으로 바뀐 상태이므로 패턴 비교과정에서 그 성분을 고려해야 한다.

한편 본 논문은 Mellin 변환의 스케일 불변성질 이 음성인식 알고리듬에 응용될 수 있는 가능성을 실험을 통하여 알아보고자 함이 목적이므로 인식대상 단어를 한국어 숫자음으로 제한하여 화자 종속격리 단어인식 시스템에 대하여 고찰하였다.

본 논문의 구성은 다음과 같다.

서론에 이어서 제 2 장에서는 Mellin 변환을 소개하고 제 3 장에서는 Mellin 변환을 이용한 음성인식 알고리듬 개발과정을 설명한다. 제 4 장에서 실험결과를 보였으며 제 5 장에서 결론을 맺는다.

Ⅱ. Mellin 변환

1. Mellin 변환의 이론적 고찰

우선 연속 Mellin 변환을 살펴보고 그 스케일 불변성질을 고찰해 보면 다음과 같다. $t \geq 0$에서 함수 $g(t)$가 주어졌을 때 1 차원에서의 연속 Mellin 변환은 다음과 같이 정의된다.

$$G(s) = \int_0^\infty g(t)\ t^{s-1}\ dt \qquad (1)$$

독립변수에 $t = Te^x$라는 지수적 왜곡(exponential distortion)을 도입하면 Mellin 변환은 다음과 같이 Fourier 변환을 이용하여 구현할 수 있다.

$$G(s) = T^s \int_{-\infty}^{\infty} g(Te^x)\ e^{xs}\ dx \qquad (2)$$

식(2)에 $s = -jw$를 대입하면, T^{-jw}의 크기가 1 이므로 $G(-jw)$의 크기는 지수적으로 왜곡된 함수의 Fourier 변환의 크기와 같게 된다.

이러한 지수적 왜곡을 Fourier 변환의 크기의 shift 불변 성질과 결합시키면 Mellin 변환의 크기가 스케일 불변 성질을 갖는다는 것을 알 수 있다. 예를 들어 $h(t) = g(kt)$로 놓고 식(2)를 적용시키면,

$$H(s) = T^s \int_{-\infty}^{\infty} g(kTe^x)\ e^{xs}\ dx \qquad (3)$$

$$= T^s \int_{-\infty}^{\infty} g(Te^{x+\ln k})\ e^{xs}\ dx \qquad (4)$$

여기에서 $y = x + \ln k$로 변수를 치환하면

$$H(s) = k^{-s} G(s)$$

$$|H(s)| = |G(s)| \qquad (5)$$

인 결과를 얻게된다.

즉 스케일 성분은 지수적 왜곡에 의해서 translation 성분으로 바뀌는 것을 알 수 있다. 이 translation 성분은 Fourier 변환을 거치면 순수한 phase 항으로 변환된다. 따라서 Mellin 변환의 크기는 스케일 성분에 대하여 불변인 성질을 갖는다.

그런데 식(1)에서 정의된 Mellin 변환을 표준 Mel-

lin 변환이라 부르는데, 이 변환은 다음과 같은 두 가지 결점이 있다. 즉,

(1) 시간영역 원점에 가까운 신호성분은 원점에서 멀리 떨어진 신호에 비해 Mellin 변환에 더 큰 영향을 미친다.

(2) 경계효과(boundary effect)가 있다.

따라서 이런 단점들을 제거한 일반 Mellin 변환이 [15]에서 제안되었는데 다음과 같다.

$$F(u) = \int_0^\infty f(t)\, t^{-j2\pi u-\alpha}\, dt$$

$$f(t) = \int_{-\infty}^{\infty} F(u)\, t^{j2\pi u+\alpha-1}\, du \qquad (6)$$

$\alpha=1$일 때 위 식은 표준 Mellin 변환이 되며, $\alpha=1/2$일 때 orthonormal Mellin 변환이 된다. 이 orthonormal Mellin 변환에서는 앞에서 지적한 Mellin 변환의 단점이 해결되므로 본 논문에서는 orthonormal Mellin 변환을 채택하였다.

한편 식(2)를 살펴보면 주어진 함수를 지수적으로 왜곡시킨 후 Fourier 변환을 함으로써 Mellin 변환을 구현할 수 있다는 것을 알 수 있다. 이때 지수적으로 왜곡된 함수 즉 $g(Te^x)$를 homomorphic 신호라 부르며, 이 신호를 Fourier 변환하였을 때의 결과가 바로 Mellin 변환이 된다.

그림 2에 테스트 신호에 대한 Mellin 변환 결과를 나타내었다. (a)는 128 샘플의 입력신호, (b)는 2배로 압축한 신호 즉 64샘플의 입력 신호이며, (c)와 (d)는 각각에 대한 256 points homomorphic 신호, 그리고 (e)와 (f)는 Mellin 변환 결과이다.

이 그림을 보면 입력신호의 스케일 성분이 지수적 왜곡에 의하여 homomorphic 신호에서는 translation 성분으로 바뀌었으며, Mellin 변환의 크기는 스케일 성분에 대하여 완전히 불변인 것을 알 수 있다.

2. 음성 인식에의 이용 가능성 고찰

격리 단어인식 시스템에서는 각각의 단어들의 길이가 다른 현상을 교정하기 위한 time alignment 과정이 가장 중요한 부분을 차지한다. 그림 3에 입력패턴 T(n)과 기준패턴 R(n) 사이의 시간축 교정 함수를 도시하였다. 현재 time alignment 과정에 가장 많이 쓰이고 있는 DTW 알고리즘의 경우에는 다음과 같은 최적화 문제의 해를 구함으로써 warping 함수 w(n)을 결정하는 방법을 사용한다.

$$D=\min_{w(n)}[\sum d(T(n),\ R(w(n))] \qquad (7)$$

여기에서 d(T(n), R(w(n))은 입력패턴의 프레임 n과 기존패턴의 프레임w(n) 사이의 거리를 나타낸다. DTW 방법은 상당히 정확하지만, 위의 최적화 문제를

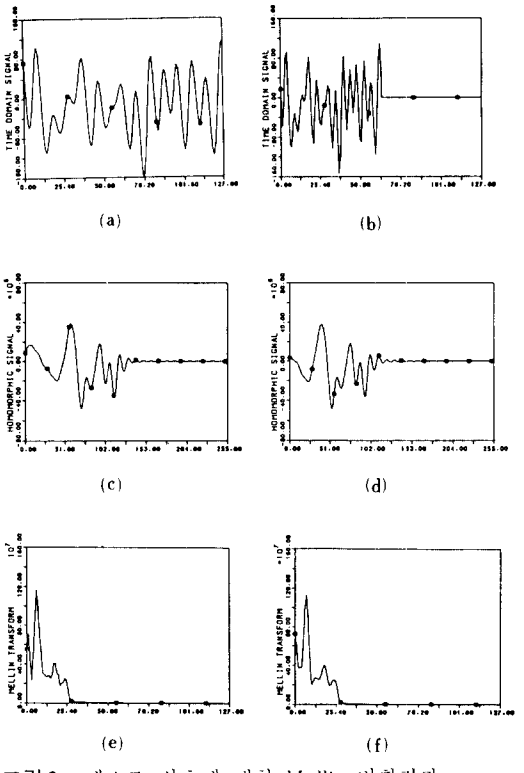


그림 2. 테스트 신호에 대한 Mellin 변환결과
(x축 : 샘플수, y축 : 크기)
(a) 원래의 신호
(b) 압축된 신호
(c) (a)의 homomorphic신호
(d) (b)의 homomorphic 신호
(e) (a)의 Mellin 변환
(f) (b)의 Mellin 변환

Fig. 2. Results of the Mellin Transform on Test Signals. (x axis : sample index, y axis : magnitude)
(a) Original Signal.
(b) Scaled Signal.
(c) Homomorphic Signal of (a).
(d) Homomorphic signal of (b).
(e) Mellin Transform of (a).
(f) Mellin Transform of (b).

푸는 데에는 많은 계산량과 메모리가 필요하게 된다. 현재 DTW의 실시간 구현을 위해 systolic array 등 병렬처리 연산을 이용하는 연구가 많이 행하여지고 있다.

한편 Mellin 변환의 스케일 불변 성질을 이용하여 time alignment 문제를 해결하면, 계산과 메모리가 많이 필요한 DTW 방법을 대체할 수 있을 것으로 생각된다. 즉 기준패턴과 길이가 다른 음성신호가 들어 왔을 때, 그 신호를 Mellin 변환하면 스케일된 성분이 없어지면서 기준패턴과 같은 길이의 출력신호로 바뀌므

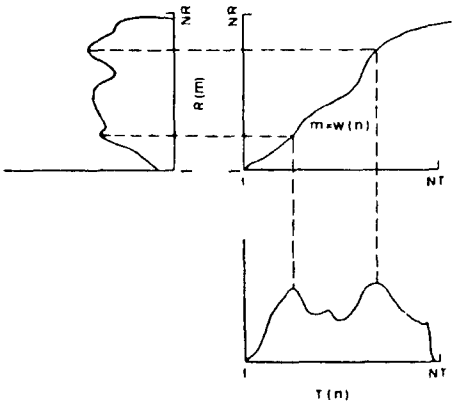


그림 3. 시간축 교정함수
Fig. 3. Time Alignment Function.

로 이 출력신호를 이용하여 유사도를 계산하면 음성인식이 가능할 것이다. 이와 같은 관점에서 Mellin 변환을 이용한 음성인식 알고리듬을 다음에 설명한다.

Ⅲ. 음성인식 알고리듬

제 2 장에서 설명한 스케일 불변성질을 실제로 음성신호에 적용하기 위해서는 몇가지 해결해야 될 문제점들이 있다. 그 중 가장 큰 문제가 서론에서 언급했던 데이타 축소문제이다. 본 논문에서는 이 문제를 해결하기 위해서 다음과 같은 여러 가지 방법을 시도하였다.

(1) 시간영역에서 프레임 별로 변환
(2) 주파수 영역에서 프레임 별로 변환
(3) 프레임의 합을 구한 후 변환
(4) 프레임 별로 feature 추출 후 변환

본 논문에서는 64개의 샘플 데이타를 1 프레임으로 하였다. 여기에서 (1)과 (2)의 방법은 음성 데이타의 한 프레임에 대하여 직접 Mellin 변환을 한 후 그 결과를 이용하고자 하였다. 그러나 시뮬레이션 결과, 단어의 길고 짧은 현상은 그 음성신호 전체에 나타나며, 어느 몇개의 프레임에서는 나타나지 않았기 때문에 이 방법은 타당하지 못하였다. 한편 (3)의 방법은 이러한 문제점을 고려하여 음성신호 전체에 적용하였지만 단지 프레임의 합으로서는 음성의 특징을 정확하게 표시하지 못함을 확인할 수 있었다. 따라서 feature 추출 후 Mellin 변환을 하는 방법을 사용하게 되었는데, 위의 네가지 방법의 시뮬레이션 결과로부터 이 방법이 가장 타당한 것으로 밝혀졌으므로, 다음절에 이 방법을 설명한다. 자세한 내용은 참고문헌[19]를 참조하기 바란다.

1. Feature추출

Feature 추출에 의한 데이타 축소 방법이 효과적이기 위해서는 우선 그 feature들을 음성신호로부터 추출하기가 쉬워야 할 뿐만 아니라 그 feature들이 음성의 특징을 잘 나타내고 있어야 한다. 본 논문에서는 다음과 같은 5 개의 측정값을 feature로 사용하였다.[16]

(1) 신호의 영교차율
프레임 내에서 영점을 지나는 횟수이다.

(2) 신호의 대수에너지
이 값은 다음과 같이 정의된다.

$$E_S = 10 \times \log_{10}(\varepsilon + \frac{1}{N}\sum_{n=1}^{N} s^2(n)) \qquad (8)$$

(3) 한 샘플 지연된 자기상관계수
인접한 음성 신호간의 상관성을 나타내며, 다음 정의에 따라 -1에서 $+1$ 사이의 값을 갖는다.

$$C_1 = \frac{\sum_{n=1}^{N} s(n)\, s(n-1)}{\sqrt{(\sum_{n=1}^{N} s^2(n))\ (\sum_{n=0}^{N-1} s^2(n))}} \qquad (9)$$

(4) 1 차예측 계수
이 값은 선형예측 분석(LPC)으로부터 얻을 수 있다.

(5) 예측오차의 에너지
이 값은 데시벨(dB)로 표시되는데 다음과 같이 정의된다.

$$E_P = E_S - 10 \times \log_{10}(\varepsilon + |\sum_{k=1}^{P} \alpha_k \phi(0,k) + \phi(0,0)|) \qquad (10)$$

여기에서 E_S는 위에서 정의된 바와 같으며

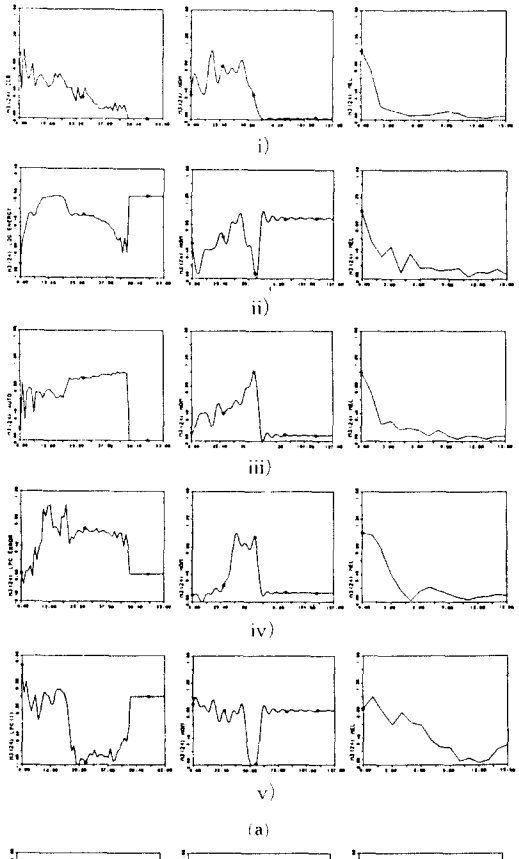
$$\phi(i,k) = \frac{1}{N}\sum_{n=1}^{N} S(n-i)\ S(n-k) \qquad (11)$$

는 음성신호의 covariance 행렬의 (i, k)항에 해당하고, α_k는 예측 계수를 나타낸다.

이런 feature들을 음성 데이타의 프레임 별로 추출한 후 그 값들을 Mellin 변환하게 된다. 이때 Mellin 변환은 해상도를 고려하여 128 points Mellin 변환을 하였다.

그림 4 에 3 가지 음성신호에 대한 feature 추출 결과 및 Mellin 변환결과를 도시하였다. 그림(a)는 길이가 48프레임인 "삼," 그림(b)는 길이가 42프레임인 "삼," 그림(c)는 48프레임인 "오"에 대한 결과이다. 5 가지의 feature추출 및 그에 대한 homomorphic 신호, Mellin 변환이 나타나 있다.

(a)와 (b)를 비교해 보면, 같은 "삼"을 나타내는 단어지만 길이는 각각 48프레임과 42프레임으로서 6 프레임의 길이 차이를 보이고 있다. 이에대해 추출한 각 5 개의 feature를 살펴보면, 전체적으로 비슷한 파형을 나타내면서 스케일 되어 있는 것을 알 수 있다. 그러나 그에 대해 Mellin 변환을 행하였을 때에는, homomorphic 신호의 경우 스케일된 현상이 없어지면서 대



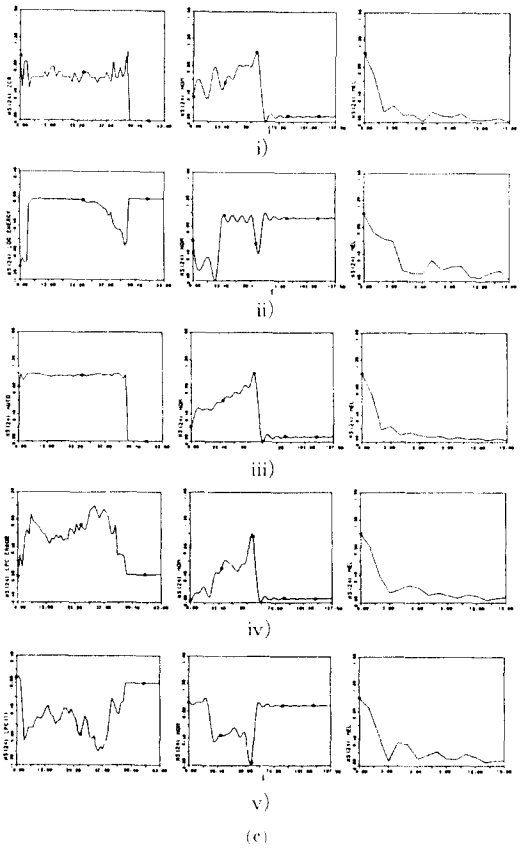


그림 4. 추출한 Feature에 대한 Mellin변환결과
(x축 : 샘플수, y축 : 크기)
(i) 영교차율 ii) 에너지 iii) 자기상관계수 iv) 예측오차 v) 예측계수의 추출결과 및 homomorphic 신호, Mellin 변환이 도시되어 있다.)
(a) "삼" (48 프레임)
(b) "삼" (42 프레임)
(c) "오" (48 프레임)

Fig. 4. Results of the Mellin Transform on the Extracted Features.
(x axis : sample index, y axis : magnitude)
(i) ZCR ii) energy iii) autocorrelation coeffi. iv) prediction residual v) predictor coeffi.)
(a) "sam" (48 frames).
(b) "sam" (42 frames).
(c) "oh" (48 frames).

신 약간의 translation이 생겼으며, Mellin변환 출력에서는 그 translation 성분조차 없어지면서 스케일 현상이 완전히 없어진다.

한편 (c)의 경우에 음성신호의 길이는 (a)와 같지만 음성의 내용이 따르기 때문에 feature를 추출하고

Mellin 변환을 행한 결과는 (a)나 (b)의 결과와 확실히 구분된다. 이상의 결과로부터 feature 추출에 대한 Mellin 변환 방법을 음성인식 알고리듬에 이용할 수 있다는 사실을 알 수 있다.

2. 패턴비교

위와 같이 입력된 음성신호들의 feature를 Mellin 변환하여 스케일 성분을 제거시킨 다음에는 기준패턴과의 유사도를 구하는 패턴 비교 과정이 필요하게 된다. 패턴비교에는 homomorphic 신호 및 Mellin 변환 출력신호를 모두 이용하였는데, homomorphic 신호에는 translation 성분이 존재하지만 Mellin변환 출력신호보다 음성의 특징을 뚜렷이 보여주고 있다. Mellin 변환 출력 신호는 스케일 성분이 완전히 사라지지만 출력신호 자체가 너무 단순해지는 단점이 있다(그림4 참조). 음성인식에서 유사도의 측정은 인식률을 좌우하는 대단히 중요한 문제로서 신중한 고려가 필요하다. 본 논문에서는 다음과 같은 패턴 비교 방법을 사용하였다.

(1) 체인코드를 이용한 패턴비교[17]

(2) 스트립 트리를 이용한 패턴비교[18]

(3) 상호 상관도를 이용한 패턴비교

(4) 출력신호의 차를 이용한 패턴비교

위와 같은 방법들은 영상인식에서 물체의 경계부분만 가지고 두 물체를 비교 인식할 때 흔히 사용되고 있는 방법들이다. 체인코드를 사용하였을 때에는 체인코드의 특성상 복잡한 곡선 파형을 충분히 나타내지 못하여 체인 상호상관도 값이 부정확 하였으며 (2)의 방법도 체인코드와 비슷한 결과를 나타내었다. (3)과 (4)의 방법은 실제곡선을 가지고 유사도를 측정하므로 두 방법 모두 비슷한 결과를 보였으나 (4)의 방법이 (3)의 방법보다 계산이 간편하므로 출력신호의 차를 이용하는 방법을 본 논문에서는 사용하였다.

이 방법은 다음과 같이 두 방법이 가능하다.

(1) Homomorphic 신호를 이용하는 경우

$$d_{ij}(k)=\sum_{n=1}^{128}|h_i(n)-h_j(n+k)| \qquad (12)$$

단, $n+k>128$일 때 $h_j(n+k)=h_j(128)$

여기에서 $d_{ij}(k)$는 k만큼 수평이동 하였을 때 패턴 i, j와의 거리를 나타내며, $h_i(n)$은 패턴i의 homomorphic 신호 중 n번째 성분을 나타낸다. 이제

$$d_{ij}=\min_k d_{ij}(k) \qquad (13)$$

인 d_{ij}를 구하여 그 값이 최소가 되는 두개의 신호패턴을 찾아 인식하게 된다.

(2) Mellin 변환 출력을 이용하는 경우

$$d_{ij}=\sum_{n=1}^{16}|m_i(n)-m_j(n)| \qquad (14)$$

여기에서 $m_i(n)$은 패턴i의 Mellin변환 출력 중n번째 성분의 값을 나타낸다. 이 경우에는 스케일 성분이 완전히 해결되었으므로 수평 이동현상을 고려해줄 필요없이 d_{ij}만을 계산하여 최소가 되는 두 패턴을 찾으면 된다.

Ⅳ. 실험결과 및 고찰

1. 음성데이타

이 실험에서 사용된 음성데이타는 0에서 9까지는 숫자음으로서 우리말로 "영," …, "구"로 발음되었다. 이 음성데이타는 모두 3가지인데 각각 M, N, S로 이름지어졌다. M, N은 같은 날 한시간 간격으로 얻은 데이타로서 평상시의 음성 그대로 발음되었다. 한편 S는 한달 뒤에 녹음한 것으로서 녹음환경은 앞의 경우와 같지만, 발음할 때 의도적으로 짧게 발음하였다. 모든음성은 200-3400Hz의 대역필터를 거친 후 샘플링 주파수 8000Hz, 13비트로 표본화하여 얻어졌다.

이 음성데이타의 길이를 표1에 제시하였다. 세 종류의 음성데이타의 길이를 비교해 보면, 같은 단어라 하더라도 상황에 따라 길게 혹은 짧게 발음된 것을 알 수 있다.

표 1. 음성데이타의 길이
(단위 : 프레임=64개의 샘플 데이타)

Table 1. Length of Speech Data.
(unit : frame=64 speech sample datas)

	M	N	S
0	48	48	36
1	54	48	44
2	44	42	34
3	48	42	34
4	48	40	38
5	54	48	38
6	38	36	28
7	48	42	40
8	50	44	42
9	42	44	28

2. 알고리듬

본 논문의 음성인식 알고리듬은 근본적으로 그림1과 유사한 것으로 제3장의 설명을 바탕으로 도시하면 그림5와 같다.

입력되는 음성신호에 대하여 5개의 feature를 추출한 후, 그 homomorphic 신호나 Mellin 변환 출력을 이용하여 출력신호의 차를 계산하면 5개의 거리(distance)가 얻어진다. 따라서 최종 판단과정에서는 각

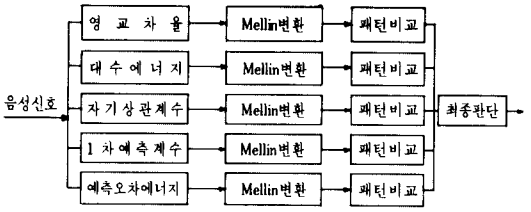


그림 5. Mellin 변환을 이용한 음성인식 모델
Fig. 5. Speech Recognition Model Using the Mellin Transform.

feature들의 거리에 같은 weighting을 주어, 이 5 개의 거리를 모두 합한 후 그 값이 가장 적은 단어를 선택하도록 하였다.

3. 실험결과

음성 데이타가 모두 세 종류이므로 다음과 같은 세 가지의 실험을 행하였다.

(a) 음성 데이타 N을 기준으로 하고, M이 입력으로 들어올 때

표 2. Homomorphic 신호를 이용한 음성인식결과
Table 2. Results of the Speech Recognition Using the Homomorphic Signal.

	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	MCH	S/U
M0	7.45	17.69	19.29	16.84	14.26	11.45	15.71	26.12	12.99	13.32	0	S
M1	20.21	13.57	17.80	17.45	16.38	16.03	19.61	20.82	16.39	18.87	1	S
M2	16.02	18.88	10.69	17.16	14.30	15.52	11.19	24.80	14.78	15.92	2	S
M3	10.41	15.71	20.54	8.24	10.39	13.81	14.38	22.42	9.29	16.58	3	S
M4	13.42	16.22	17.88	12.12	7.27	11.57	12.48	24.81	12.13	17.75	4	S
M5	12.07	20.69	15.80	18.66	14.63	8.53	16.48	26.53	15.51	12.87	5	S
M6	11.16	21.58	16.75	16.90	12.71	12.37	10.22	28.06	14.28	13.55	6	S
M7	22.67	15.70	19.75	19.13	18.55	21.58	20.26	12.82	19.15	20.12	7	S
M8	18.61	12.28	15.00	15.25	15.57	19.64	17.25	18.03	11.56	14.97	8	S
M9	14.68	19.71	18.76	16.07	11.58	10.15	14.77	26.08	13.32	9.14	9	S

Recognition Rate : 10/10

(a) Input data : M0 – M9 Reference data : N0 – N9

	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	MCH	S/U
M0	7.62	14.03	14.91	16.48	12.61	11.24	13.76	21.26	12.26	12.68	0	S
M1	19.19	16.65	16.73	18.91	16.75	17.88	20.45	16.86	18.05	18.79	1	S
M2	15.59	15.16	10.06	18.77	16.32	13.33	13.73	19.41	19.06	16.40	2	S
M3	13.59	14.28	15.32	9.38	11.85	14.25	13.05	15.98	11.37	15.61	3	S
M4	16.26	13.09	14.09	13.15	10.79	14.28	12.74	18.87	13.45	17.80	4	S
M5	12.55	10.29	13.27	18.56	12.87	7.51	14.94	23.66	14.12	13.79	5	S
M6	10.46	13.80	13.21	15.75	12.51	9.81	11.12	22.33	14.37	15.52	6	U
M7	22.88	21.44	19.88	20.10	20.58	20.86	20.86	11.84	20.15	22.15	7	S
M8	17.36	18.34	14.04	16.07	17.48	16.64	16.73	11.58	16.78	16.68	8	U
M9	11.45	13.68	15.75	14.65	11.01	10.16	13.53	20.43	12.21	9.27	9	S

Recognition Rate : 8/10

(b) Input Data : M0 – M9 Reference Data : S0 – S9

	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	MCH	S/U
S0	8.88	16.94	18.87	15.37	12.74	10.97	13.98	25.57	13.25	11.06	0	S
S1	14.34	18.88	16.14	16.20	12.87	10.01	14.04	23.95	14.14	15.27	5	U
S2	14.73	16.13	12.37	15.38	13.82	14.48	14.37	23.10	13.27	15.04	2	S
S3	13.07	17.59	23.16	8.36	9.65	13.78	13.18	24.92	11.33	14.28	3	S
S4	11.41	17.16	19.27	12.84	7.06	7.80	12.28	24.63	12.34	12.46	4	S
S5	10.39	19.26	15.58	15.28	12.24	8.39	13.27	24.61	13.01	10.10	5	S
S6	11.73	21.38	18.97	11.44	9.58	13.76	8.45	26.15	13.60	13.20	6	S
S7	19.85	12.21	19.05	16.38	18.02	20.10	20.04	12.25	17.80	18.47	1	U
S8	10.18	17.92	22.64	12.37	11.94	10.88	16.25	24.25	12.40	13.09	0	U
S9	16.36	21.42	20.45	15.20	13.79	12.83	15.22	25.17	14.52	11.83	9	S

Recognition Rate : 7/10

(c) Input Data : S0 – S9 Reference Data : N0 – N9

(b) S가 기준, M이 입력

(c) N이 기준, S가 입력

패턴비교 과정에서 homomorphic 신호를 이용했을 때의 실험결과를 표 2에 나타내었다. 표에서 보는 바와 같이 (a)의 경우에는 100%, (b)는 80%, (c) 는 70%의 인식을 보였다. 한편 Mellin 변환 출력을 이용했을 때에는 (a)가 100%, (b)와 (c)는 각각 40%의 인식률을 나타내었다.

이 결과를 살펴보면 제 3 장에서 예측한 바와 같이 homomorphic 신호를 이용할 때가 소요시간은 약간 늘어나지만 정확한 결과를 나타낸다는 것을 알 수 있다. 또한 데이타S가 관계되는 실험에서는 인식결과가 좋지 않다는 사실도 알 수 있다.

표 2에서 보듯이, 음성데이타 M과 N을 비교하였을 때에는 인식률이 100%에 이르렀으나, 짧은 음성데이타S와 비교하였을 때에는 인식률이 상당히 낮아졌다. 이런 현상을 설명하기 위하여 음성데이타를 도시해 본 결과, 음성데이타S의 경우에는 그 길이가 짧아진 현상이외에도, 일부러 짧게 발음하려고 한데서 생긴이상 현상과 녹음일시의 차이로 인한 변화를 볼 수 있었다. 이러한 견실성(robustness)의 부족현상의 가장 큰 원인은, 본 논문에서 사용한 feature 추출 및 거리계산 방식이 완벽하지 못하기 때문인것 같다.

V. 결 론

본 논문에서는 Mellin 변환을 이용한 음성인식 알고리듬을 제안하였다. Mellin 변환의 스케일 불변성질을 이용하여 time alignment 문제를 해결하였으며, 이전의 표준 Mellin 변환대신 orthonormal Mellin 변환을 채택하였다.

음성신호로부터 feature를 추출한 후, 그 feature에 대하여 Mellin 변환을 적용하였다. 여기에 사용한 feature로는 영교차율, 대수에너지, 자기상관계수, 1차 예측계수, 예측오차의 에너지 등이 있다. 패턴 비교과정에서는 출력신호의 차이를 계산하는 방법이 가장 좋은 결과를 보였다. 이 과정에서 homomorphic 신호를 이용하였을 때에는 계산량이 많은 대신 정확한 결과를 보였고, Mellin 변환출력을 이용하였을 때에는 계산량은 상대적으로 줄어들지만 인식률은 낮았다. "영"에서 "구"까지의 숫자음에 대한 실험결과, 각각 83.3%, 60.0%의 인식률을 보였다.

결론적으로 본 논문에서는 Mellin 변환의 스케일 불변성질을 음성인식에서 절대적으로 필요한 시간축 교정 문제를 해결하는데 이용할 수 있음을 실험적으로 보였다. 그러나 feature 선택 및 근사도 계산방법을 보완하면 보다 높은 인식율을 얻을 수 있을 것으로 기대된다. 또한 Mellin 변환의 스케일 불변성질은 음성인식 뿐만 아니라 스케일 현상이 발생하는 모든 신호 처리 문제에 광범위하게 이용될 수 있을 것이다.

參 考 文 獻

[1] T.B. Martin, "Practical application of voice input to machines," *Proc. IEEE,* vol.64, 487-571, 1976.

[2] L.R. Rabiner & S.E. Levinson, "Isolated and connected word recognition--- theory and selected applications", *IEEE Trans.* COM-29, pp. 621-659, 1981.

[3] B.A. Dautrich, L.R. Rabinson & T.B. Martin, "On the effects of varing filter bank parameters on isolated word recognition", *IEEE Trans. ASSP-31,* pp.793-806, 1983.

[4] L.R. Rabiner & J.G. Wilpon, "A two-pass Pattern-recognition approach to isolated word recognition", *The BSTJ* vol.60, pp. 739-766, 1981.

[5] L.R. Rabiner & M.R. Sambur, "Some preliminary experiments in the recognition of connected digits", *IEEE Trans. ASSP-24,* pp.170-182, 1976.

[6] H. Ney, "The use of one-stage dynamic programming algorithm for connected word recognition", *IEEE Trans. ASSP-32,* pp.263-271, 1984.

[7] S.E. Levinson, "Some experiments with a linguistic processor for continuous speech recognition", *IEEE Trans. ASSP-31, pp.* 1549-1555, 1983.

[8] F. Itakura, "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. ASSP-23,* pp.67-72, 1975.

[9] H. Sakoe & S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. ASSP-26,* pp.43-49, 1978.

[10] C.C. Tappert & S.K. Das, "Memory and time improvements in a dynamic programming algorithm for matching speech patterns", *IEEE Trans. ASSP-26,* 583-586, 1978.

[11] C. Myers, L.R. Rabiner & A.E. Rosenberg "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition", *IEEE Trans. ASSP-28,* pp.623-635 1980.

[12] M.H. Kuhn & H.H. Tomaschewski, "Improvements in isolated word recognition", *IEEE Trans. ASSP-31,* pp.157-167, 1983.

[13] G.M. Robbins & T.S. Huang, "Inverse filtering for linear shift-variant imaging system", *Proc. IEEE* vol.60, pp.862-872, 1972.

[14] P.E. Zwicke & I. Kiss Jr., "A new implementation of the Mellin transform and its application to radar classification of ships", *IEEE Trans. PAMI-5,* pp.191-199, 1983.

[15] Semoon Goh, "The Mellin transformation", *Ph.D thesis, Purdue University,* May, 1985.

[16] B.S. Atal & L.R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", *IEEE Trans. ASSP-24,* pp.201-212, 1976.

[17] R.O. Duda & P.E. Hart, *Pattern classification and scence analysis,* Wiley-Interscience Publication, 1973.

[18] D.H. Ballard & C.M. Brown, *Computer vision,* Prentice-Hall Inc., 1982.

[19] J.M. Kim, "Isolated word recognition using the mellin transform", M.S. Thesis, Dept. of Control and Instrumentation Eng., Seoul National University, Feb. 1987.