잡음환경및 채널왜곡에 강인한 ARS용 전화음성인식 방식 연구

The Development of a Speech Recognition Method Robust to Channel Distortions and Noisy Environments for an Audio Response System (ARS)

안 정 모*, 임 계 종**, 계 영 철**, 구 명 완***
(Jung Mo Ahn*, Kye Jong Yim**, Young Chul Kay**, Myoung Wan Koo***)

※이 연구는 95년도 한국과학재단 연구비 지원에 의한 결과임.(951-0906-105-1)

요 약

본고는 음성인식 기능이 추가된 음성응답장치(ARS)의 음성 인식률을 향상시키는 방법을 제안한다. ARS에 입력되는 전화음성은 안내방송, 전화잡음, 그리고 채널왜곡에 의하여 영향을 받기 때문에, 양질의 음성을 대상으로 하여 개발된 인식 알고리듬을 그대로 적용하면 상당한 인식률의 저하를 가져오게 된다. 이러한 문제점을 극복하기 위하여 본고에서는 세 가지 방법을 제안한다 : 1)음성이 시작되는 순간 안내방송을 즉시 끊기 위한 음성 입력순간의 정확한 검출, 2)Teager 에너지를 이용한 잡음 섞인 전화음성의 효과적인 끝점검출, 3)SDCN 알고리듬을 이용한 채널왜곡의 보상. 위의 세 가지 방법을 모두 결합하여 화자독립인 전화음성을 대상으로 실험한 결과, 기존의 방법이 약 23%의 인식률을 보인 반면, 제안된 방식은 약 77%의 인식률로서 상당한 성능향상을 보여주었다.

ABSTRACT

This paper proposes the methods for improving the recognition rate of the ARS, especially equipped with the speech recognition capability. Telephone speech, which is the input to the ARS, is usually affected by the announcements from the system, channel noise, and channel distortion, thus directly applying the recognition algorithm developed for clean speech to the noisy telephone speech will bring the significant performance degradation. To cope with this problem, this paper proposes three methods: 1)the accurate detection of the inputting instant of the speech in order to immediately turn off the announcements from the system at that instant, 2)the effective end-point detection of the noisy telephone speech on the basis of Teager energy, and 3)the SDCN-based compensation of the channel distortion. Experiments on speaker-independent, noisy telephone speech reveal that the combination of the above three proposed methods provides great improvements on the recognition rate over the conventional method, showing about 77% in contrast to only 23%.

I. 서 론

지금까지의 음성인식 기술은 실험실내에서 양질의 음성을 대상으로한 제한적인 실험에 불과하였으며, 상용화될 수 있을 정도의 인식률을 보이는 시스템도 이러한 실험실 수준의 양질의 음성을 대상으로 하는 제한적인 응용범위에 사용되고 있는 형편이다. 그러나, 이러한 시스템을 주변환경의 영향이 크고 잡음이 심한 경우에 적용하였을 경우 상당한 인식률의 저하를 가져온다고 보고되고 있으며[1], 그 대표적인 예로서는 전화망을 통한 음성의 인식시스템을 들을 수 있다. 최근 들어 음성인식의 상용화를 위하여 환경 및 잡음을 고려한 강인한 음성인식 시스템의 연구가 활발히 진행 중에 있으며, 대부분 잡음이 섞인 음성의 향상 및 보상 방법의 개발에 중점을 두고 있다.

음성의 향상을 위한 스펙트랄 감산법(spectral subtraction)이 Boll[2]과 Berouti[3]등에 의하여 소개되었으나 이러한 방법은 신호대잡음비는 향상시키나 음성의 인지도는 향

*대우전자 영상연구소
**홍익대학교 전자공학과
***한국통신 멀티미디어 연구소
접수일자 : 1996년 11월 1일

상되지 못하였다. 최근 들어 Van Compernolle[4][5]과 Stern 그리고 Acero[6]는 이러한 스펙트럼 감산법을 음성인식에 적용하였다. Van Compernolle은 desk-top 마이크를 사용한 IBM 음성인식 시스템의 강인함을 향상 시켰으며 로그영역에서의 독별한 스펙트럼 감산법에 의한 채널등화(channel equalization)를 제안하였다[4][5]. Stern과 Acero의 연구는 비록 LPC(Linear Predictive Coding)로부터 도출된 켑스트럴 계수에 기초를 둔 것이었지만[6] Van Compernolle의 연구[4][5]와 유사하다. Porter 그리고 Boll[7]은 MMSE 방법을 이용한 DFT계수의 측정을 제안하였으며 이러한 방법은 로그영역에서의 스펙트럼 측정이 인식률의 정확성에 더욱 효과적임을 보였다.

이와같은 음성의 보상에도 불구하고 인식률의 저하를 가져오는 또 다른 하나의 중요한 요인은 유성의 특징 파라메터를 추출하기 위한 순수한 음성구간 검출의 부정확성이다. 전화망을 통과한 전화음성은 잡음 등의 영향으로 효과적인 음성구간의 검출이 어렵다. 음성구간의 검출을 위하여 현재 사용중인 끝점검출 알고리듬의 대부분은 Rabiner[8]가 제안한 음성신호의 영교차율과 프레임 에너지의 조합을 바탕으로 하고 있다. 프레임 에너지는 유성음과 무성음을 구분하는데 사용되며 영교차율은 유성음, 무성마찰음 그리고 묵음을 구분하는데 사용된다. 그러나 전화음성과 같이 잡음의 영향이 심한 경우 영교차율을 이용한 무성마찰음등의 검출은 효과적이지 못하다. 따라서 Rabiner[8]가 제안한 끝점검출 알고리듬을 사용할 경우 잡음이 심한 음성에서는 정확한 끝점검출이 불가능하므로 이를 극복할 새로운 방법이 필요하다.

본 고에서는 잡음 및 채널왜곡에 강인한 전화음성 인식 ARS(Audio Response system)의 개발에 수반되는 앞서 언급된 문제점들을 제시하고, 그의 해결방법을 제안한다. 2장에서는 ARS에 음성인식 기능을 추가할 경우의 문제점을 제시하며, 3장에서는 그의 해결책으로서 음성입력순간의 결정, 효과적인 끝점검출 알고리듬, 전화잡음 및 채널왜곡 보상 방법을 다룬다. 제안된 방법의 타당성을 4장에서 실험으로 검증하며, 5장에서 결론을 맺는다.

Ⅱ. 전화음성인식 ARS

그림 1은 음성인식 ARS의 블록도이며, far-end A는 안내방송을, near-end C는 정보를 요구하는 화자의 음성을, 그리고 far-end B는 ARS에 입력되는 near-end 화자의 음성을 나타낸다. near-end C에서 정보를 음성으로 요구하면, ARS의 음성인식부에서 far-end B를 인식하여 far-end A에서 해당정보를 안내방송하게 된다. 그러나, far-end A로 부터 안내방송 중에 near-end C에서 다른 정보를 음성으로 요구하는 경우에는, 이 음성에 far-end B로 feedback되는 안내방송이 혼합되어 ARS의 음성인식부에서 음성을 제대로 인식할 수 없게 된다. 이와 같은 문제뿐만 아니라, 전화음성 자체에도 잡음과 채널왜곡이 포함되어 있어 음성인식을 더욱 어렵게 한다. 본고에서는 이와 같은 문제점들을 극복하여 보다 강인한 전화음성인식 ARS를 개발하기 위하여 다음의 방법들을 제안한다: 1) 안내방송 중단을 위한 전화음성 입력순간 결정, 2) 전화음성 분석을 위한 정확한 끝점검출, 3) 채널왜곡의 보상.

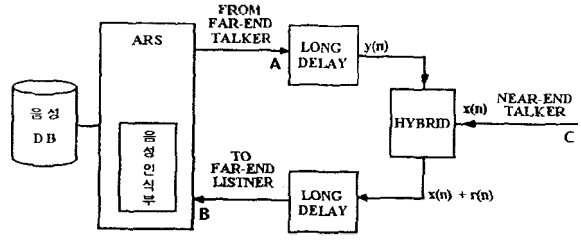


그림 1. 음성인식 ARS 블록도
Fig 1. The block diagram of a speech recognition-based ARS

Ⅲ. 전화음성 인식 알고리듬

ARS용 전화음성 인식의 경우 인식률 저하는 크게 세 가지 원인으로부터 기인한다: 첫 번째는 안내방송과의 혼합에 의한 전화음성의 입력순간의 정확한 검출의 어려움, 두 번째는 잡음의 영향에 의한 부정확한 끝점검출에 기인한 것이며, 세 번째는 전화선로의 대역폭 제한 및 채널왜곡 현상 그리고 주변 잡음에 의한 음성신호의 오염에 의한 것이다. 본 절에서는 이러한 ARS용 인식기의 인식률 저하를 개선하기 위하여 앞서 언급한 세 가지 원인을 극복하는 방법을 제시한다.

3.1 전화음성 입력순간의 결정

그림 1에서 보듯이, A로부터의 안내방송 중에 C로부터 전화음성이 입력되면 피드백된 안내방송과 전화음성이 혼합된 파형을 이용하여 far-end A에서는 즉시 안내방송을 중지 하여야하고, 그 직후 B에서는 전화음성만을 이용하여 음성인식을 하게 된다. 이와 같은 ARS가 정상적으로 동작하기 위해서는 무엇보다도 전화음성이 입력된 순간이 정확히 검출되어 그 즉시 안내방송이 중단되어야 한다. 그렇지 못하면 음성인식에 필수적인 음성구간의 끝점검출이 부정확하게 되어 오인식을 일으키게 된다. 이러한 문제점의 극복을 위하여 본 절에서는 다음의 방법을 제안한다.

그림 1과 같은 시스템에서 전화음성의 입력순간을 결정하기 위하여서는 안내방송이 나오는 A와 안내방송과 음성이 혼합되어 들어오는 B에서의 신호에너지를 비교하여 그 차이가 심하면 음성이 입력되었다고 판단하면 될 것이다. 그러나, 신호가 near-end의 hybrid를 거쳐 피드백 되는 동안 상당히 감쇄되기 때문에 이러한 단순비교로는 음성의 입력순간을 정확히 검출할 수 없다.

hybrid를 거친 피드백 신호를 분석하여 보면 약 6dB이상의 감쇄가 생김을 알 수 있다[9]. 따라서 피드백된 신호

에 6dB의 보상을 취하여 프레임 에너지를 구한 후 far-end 로부터 발생되는 안내방송의 프레임에너지와 비교를 행하게 된다.

$$y_{far\ end} \le 2(x_{near\ end} + x_{feedback}) \quad (1)$$

식 (1)의 좌변은 far-end로부터의 안내방송 신호이며 우변은 near-end의 신호와 피드백된 안내방송이 hybrid를 거친 신호이다. 우변에 2배를 한것은 6dB의 보상을 한 것이다[9]. 각 프레임은 20ms로 구성되어 있으며 10ms씩 이동하게 된다. 프레임 에너지를 이용하면 전화선로의 딜레이에 의한 영향을 무시할 수 있어 효과적이다. 만일 far-end로부터 안내방송이 행하여지는 동안에 near-end로부터 음성이 발성된다면 far-end로 들어오는 신호의 프레임 에너지는 far-end 안내방송의 프레임 에너지에 비하여 갑작스런 피이크가 발생하게 된다. 따라서 이 피이크의 시작점이 near-end로부터의 음성의 시작점에 해당하게 된다.

3.2 끝점검출의 개선

기존의 끝점검출 알고리듬들의 대부분은 음성신호의 영교차율과 에너지의 조합을 바탕으로 하였다[8]. 에너지는 유성음과 무성음 그리고 주변잡음을 구분하는데 사용되며, 영교차율은 유성음, 무성마찰음 그리고 묵음을 구분하는데 사용된다. 3KHz의 대역폭을 갖고 잡음의 영향이 큰 전화음성의 경우 이러한 영교차율을 이용한 방법은 효과적이지 못하다. 실험결과 묵음구간에서 마이크 음성은 0~30, 전화음성은 평균 250정도의 영교차율을 나타내었다. 마찰음 및 파열음의 검출에 사용되는 영교차율은 이러한 특징으로 인하여 전화음성의 끝점검출 알고리듬에 적용이 어렵다.

본 연구에서는 전화음성의 효율적인 끝점검출을 위하여 기존의 끝점검출 알고리듬에 사용되었던 에너지와 영교차율을 대신하여 새로운 방법인 Teager 에너지를 이용한다[10][12].

3.2.1 기존 에너지의 정의

기존의 끝점검출 방법에서 사용되는 에너지는 프레임별로 각 샘플의 절대값의 합으로 정의하며, 식 (2)과 같이 나타낼 수 있다.

$$E_k = \sum_{n=-\infty}^{\infty} |x(n)| w(k-n) \quad (2)$$

일반적인 에너지와는 다르게 절대값의 합으로 단기간 에너지를 정의한 이유는 곱셈을 피하여 계산량을 줄이고 음성신호의 진폭차이에 의한 시간축상의 변화를 감쇄시키기 위한 것이다. 그러나, 이러한 에너지 측정방법은 마찰음이나 파열음의 경우 주변잡음과 비슷한 크기를 갖고 있어 주변잡음과의 구별이 어렵기 때문에 정확한 끝점검출이 어렵다[8].

3.2.2 Teager 에너지의 정의

마찰음이나 파열음이 주변잡음과 구별이 어려운 이유는 신호의 크기만을 이용하는 에너지의 성의가 적절하지 못하기 때문이다. 가령 예를 들면, 같은 크기를 갖는 10 Hz 와 1000 Hz 의 음향신호의 경우 기존 에너지 정의에 의하면 모두 같은 크기의 에너지를 갖게 되나, 실제로 이들 에너지를 생성하는 시스템이 필요로 하는 에너지는 상당히 차이가 남을 알 수 있다. 기존의 에너지 정의를 이용하여 신호를 구별하면 이와 같은 문제점이 있음을 고려하여, 신호 자체의 에너지보다는 그 신호를 생성하는 데 필요한 에너지를 이용하여 신호를 구별하려는 시도가 있었다[10].

신호 $x(t) = A\cos(\omega t + \phi)$가 용수철 상수 k인 용수철에 매달린 질량 m인 물체의 단진동 운동을 나타낸다고 하면, 이 신호는 2차 미분방정식인 다음과 같은 뉴튼의 운동방정식의 해이다.

$$\frac{d^2x}{dt^2} + \frac{k}{m}x = 0 \quad (3)$$

여기서 $\omega = (k/m)^{1/2}$로 결정된다. 이와 같은 단진동 운동을 생성시키는 용수철 시스템의 총에너지는 운동 및 위치에너지의 합으로서 다음과 같이 된다.

$$E = \frac{1}{2}kx^2 + \frac{1}{2}mx^2 \quad (4)$$

식 (4)에 식 (3)의 해인 $x(t) = A\cos(\omega t + \phi)$를 대입하면 시스템의 에너지를 구할 수 있다.

$$E = \frac{1}{2}m\omega^2 A^2$$

또는,

$$E \propto A^2\omega^2 \quad (5)$$

따라서 $x(t)$의 에너지로서 식 (2)를 이용하는 대신에 그것의 원천(source) 에너지인 식 (5)를 사용하면, 에너지는 진폭의 제곱 뿐만 아니라 주파수의 제곱에도 비례하게 된다.

주어진 신호 샘플들로부터 식 (5)로 정의되는 순시 원천 (instantaneous source) 에너지를 간단히 구하기 위한 방법은 다음과 같이 유도될 수 있다[10]. $x(n)$은 진동체 (oscillatory body)의 움직임을 나타내는 신호의 샘플이라 하면, 즉

$$x(n) = A\cos(\Omega n + \phi) \quad (6)$$

여기서 $\Omega = 2\pi f/f_s$는 디지탈 주파수, f는 신호의 아날로

고 주파수, f_s는 샘플링 주파수, 그리고 ϕ는 임의의 초기 위상이다. 샘플링을 등간격으로 한다고 가정하면, 인접한 샘플들은 다음과 같이 표시된다:

$$\begin{aligned} x(n) &= A\cos(\Omega n+\phi) \\ x(n+1) &= A\cos[(n+1)\Omega+\phi] \\ x(n-1) &= A\cos[(n-1)\Omega+\phi] \end{aligned} \tag{7}$$

다음의 trigonometric identity 를 이용하면,

$$\cos(\alpha+\beta)\cos(\alpha-\beta)=\frac{1}{2}[\cos(2\alpha)+\cos(2\beta)] \tag{8}$$

다음과 같은 관계식을 유도할 수 있다.

$$x(n+1)x(n-1)=\frac{A^2}{2}[\cos(2\Omega n+2\phi)+\cos(2\Omega)] \tag{9}$$

식 (9)에 다시 다음의 항등식을 적용하면

$$\cos 2\Omega = 2\cos^2\Omega - 1 = 1-2\sin^2\Omega \tag{10}$$

식 (9)는 다음과 같이 표현될 수 있다.

$$x(n+1)x(n-1)=A^2\cos^2(\Omega n+\phi)-A^2\sin^2(\Omega) \tag{11}$$

식 (11)을 살펴보면 우변의 첫째항은 $x(n)$의 제곱이므로,

$$x(n+1)x(n-1)=x(n)^2-A^2\sin^2(\Omega) \tag{12}$$

또는,

$$A^2\sin^2(\Omega)=x(n)^2-x(n+1)x(n-1)$$

샘플링 주파수를 충분히 크게 하면, $\sin\Omega\approx\Omega$ 이므로, 식 (12)로부터 다음과 같이 근사화 할 수 있다.

$$A^2\Omega^2\approx x(n)^2-x(n+1)x(n-1)\triangleq E(n) \tag{13}$$

식 (13)은 식 (5)와 마찬가지로 신호 샘플의 크기의 제곱 뿐만 아니라 진동 주파수의 제곱의 항이 포함되어 있으므로 에너지로 정의할 수 있으며, 이를 순시(instantaneous) Teager 에너지라 부른다. 이렇게 정의된 Teager 에너지는 신호 자체의 크기 뿐만 아니라 그 신호의 에너지가 몰려있는 주파수도 함께 고려하므로, 음성신호 처리에 있어서 좀더 적합한 방법일 뿐만 아니라 또한 실제로 진폭 및 주파수의 변화에 빠른 응답이 가능하다[10, 12].

음성신호 처리에 있어서는 순시 에너지보다는 프레임 에너지를 주로 사용하므로, 순시 Teager 에너지를 한 프레임 동안 합한 것을 프레임 Teager 에너지라고 정의한다.

$$\begin{aligned} E_k &= \sum_{n=-\infty}^{\infty} E(n)w(k-n) \\ &= \sum_{n=-\infty}^{\infty} \{x(n)^2-x(n+1)x(n-1)\}\,w(k-n) \end{aligned} \tag{14}$$

3.3 잡음 및 채널왜곡의 보상

잡음환경에서도 강인한 성능을 보이는 음성인식을 위한 잡음제거 및 채널등화 알고리듬 중에서 대표적인 것이 MMSE(Minimum Mean Square Error) 알고리듬이다. 본 연구에서도 MMSE 알고리듬에 근간을 둔 SDCN (SNR Dependent Cepstral Normalization)알고리듬을 실시간으로 구현하는 방식을 연구하였다. 지금까지의 MMSE 알고리듬은 주파수 영역에서의 처리였다[11]. 따라서 시간 영역의 음성신호를 주파수 영역으로의 변환(DFT)이 필요하며, 상당한 계산량을 요구하여 실시간 구현이 힘든 문제점을 지니고있다. 이러한 문제점을 해결하기 위하여 현재 본 연구에서 특징벡터로 사용되는 켑스트랄 계수를 영역의 변환이 필요 없이 바로 켑스트럼 영역에서 잡음제거 처리 및 채널등화를 할 수 있는 방법인 SDCN 알고리듬을 선택하였다. 이 방법은 연산의 복잡성과 양을 감소시키고 실시간 처리를 가능하게 할 것으로 예상된다.

3.3.1 SDCN 알고리듬

LPC 켑스트럼은 본 연구에서 사용한 음성인식 시스템 뿐만 아니라 상용화 되어있는 시스템의 특징벡터로도 많이 사용되고 있다. 따라서 MMSEN(Minimum Mean Square Error N) 알고리듬을 이러한 시스템에 이용하기 위해서는 두번의 DFT 과정이 필요하다. 한번은 켑스트랄 영역에서 로그주파수 영역으로, 다른 한번은 그 반대 과정에 사용된다. 따라서 이러한 연산부담의 제거와 실시간 처리를 위하여 본 실험에 사용한 특징벡터와 같은 범주(category)에서의 처리가 필요하며, MMSEN 알고리듬을 켑스트럼 영역에서 직접 처리한 것이 SDCN알고리듬이다[11]. 식(15)는 SDCN 알고리듬의 기본식이다.

$$\hat{x}=z+w(SNR) \tag{15}$$

보정(correction)벡터 w는 식(16)과 같으며, 기준환경의 켑스트랄 벡터와 테스트 환경의 켑스트랄 벡터의 차에 의하여 계산 될 수 있다.

$$w[j,k]=\frac{\sum_{i=0}^{N-1}(x_i[j]-z_i[j])\,\delta[SNR_i-k\triangle_{SNR}]}{\sum_{i=0}^{N-1}\delta[SNR_i-k\triangle_{SNR}]} \tag{16}$$

x_i와 z_i는 각각 i번째 프레임의 참조 켑스트랄 벡터와 테스트 켑스트랄 벡터이며, SNR_i는 테스트 음성의 i번째 프레임 SNR이다.

MMSEN 알고리듬은 주파수대역 SNR을 사용하며, SDCN 알고리듬에서는 프레임 SNR을 사용한다. 따라서 MMSEN 알고리듬의 보상함수는 많은 주파수대역들(Acero의 실험에서는 32개 대역)을 보상하여야 한다. 그러나 SDCN 알고리듬의 경우 Acero의 실험에 의하면 단 2개의 캡스트럼 벡터의 보상만으로도 전 채널보상을 한 MMSEN 알고리듬과 같은 성능을 보인다. 그림 2은 SDCN 알고리듬 적용시 보정 벡터수에 따른 단어의 인식률을 나타낸다.

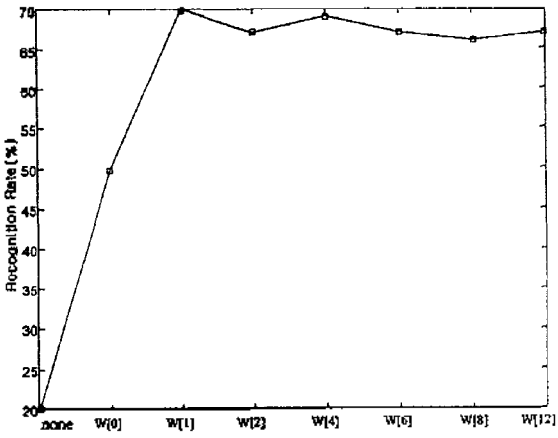


그림 2. 보정 벡터수에 의한 인식률[11]

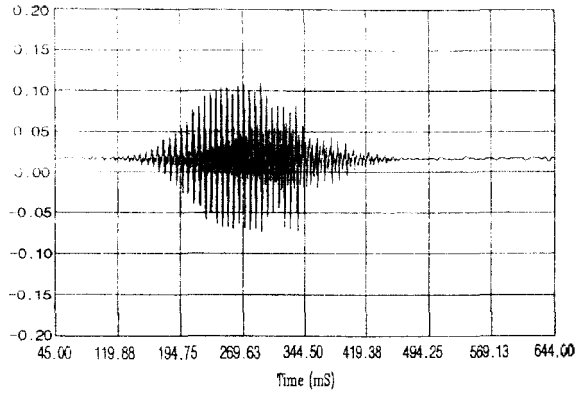
Fig 2. Recognition rate as a function of the number of correction vectors[11]

Ⅳ. 실험 및 결과고찰

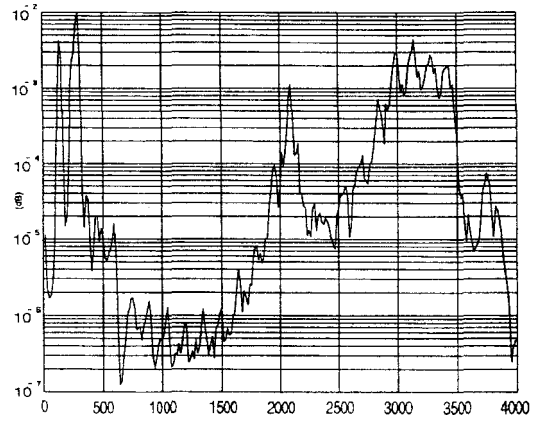
4.1 숫자음성의 분석

실험에 사용된 음성은 현재 홍익대학교 전자공학과 대학원에 재학중인 대학원생들의 음성을 오디오 테이프에 녹음하여 사용하였다. 녹음된 음성은 20대 중반의 남성 40명이 발음한 한국어 숫자음 /영/에서 /구/까지 이다. 녹음된 음성은 1)마이크로 직접 녹음된 음성과 2)전화망을 통해 녹음된 음성의 두 종류로 분류할 수 있으며, 이들은 각각 스테레오의 왼쪽채널과 오른쪽채널에 동시 녹음되었다. 마이크를 통한 음성은 주변 잡음이 존재하는 일반 사무실환경에서 desktop 마이크를 통하여 녹음되었다.

녹음된 실험용 음성은 A/D 변환기에 의하여 16KHz로 샘플링된 후 16bit 양자화되어 DSP 보드를 거쳐 IBM-PC의 하드디스크에 이진화일 형태로 저장되었다. 그림 3과 그림 4는 각각 마이크 음성과 (안내음성이 제거된) 전화음성의 파형 및 주파수 스펙트럼을 나타낸다. 그림에서 볼 수 있듯이 전화음성 및 마이크 음성의 파형은 묵음구간의 잡음을 제외하고는 매우 비슷함을 알 수 있다. 그러나 주파수 영역에서 두 가지 음성의 스펙트럼을 비교하여 보면 형태가 매우 다름을 알 수 있다. 이것으로부터 주변 잡음 및 채널의 왜곡이 전화음성의 인식에 얼마나 심각한 영향을 미칠 수 있는지를 예측할 수 있다.



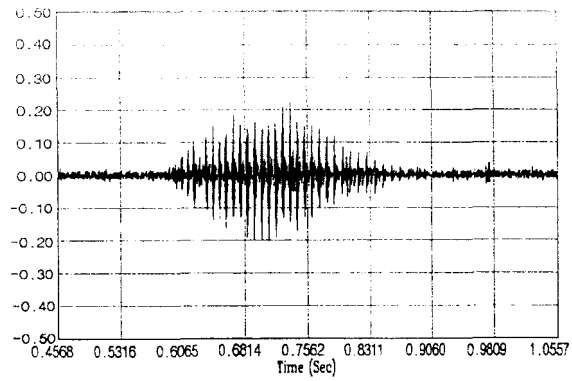
(a) 마이크음성 /이/의 파형



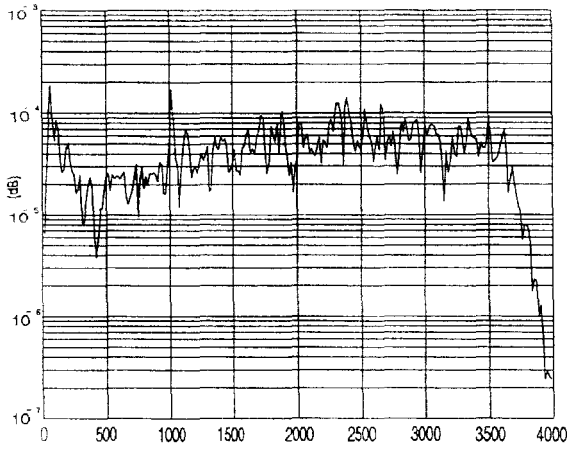
(b) 마이크음성 /이/의 스펙트럼

그림 3. 마이크음성의 파형 및 스펙트럼

Fig 3. Waveform and spectrum of the microphone speech(a) 전화음성 /이/의 파형



(a) 전화음성 /이/의 파형



(b) 전화음성 /이/의 스펙트럼

그림 4. 전화음성의 파형 및 스펙트럼
Fig 4. Waveform and spectrum of the telephone speech

4.2 BASELINE 시스템

본 연구에서 사용한 끝점검출 알고리듬 및 보상 알고리듬의 평가를 위하여 이산 HMM (Discrete Hidden Markov Model)에 근간을 둔 음성인식 시스템을 구현하였다. 그림 5는 본 연구에서 사용한 음성인식 시스템의 블록도이다. 그림에서와 같이 전처리 과정을 거친 음성샘플은 저장되어 있는 각 HMM모델에 대한 Viterbi 알고리듬을 거쳐 이 중 가장 높은 확률을 갖는 모델이 선택된다.

입력된 음성은 A/D 변환기를 통하여 디지탈 신호로 변환된 후 TMS320C30 보드에 의하여 전처리과정과 인식과정이 수행된 후 IBM-PC에서 선택과정이 수행된다. 참조패턴은 ASCII화일 형태로 IBM-PC의 HDD에 저장하였다. 또한 이 시스템을 이용하여 전화음성 인식용 HMM모델과 마이크음성 인식용 HMM모델을 구성하였다.

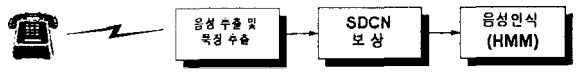


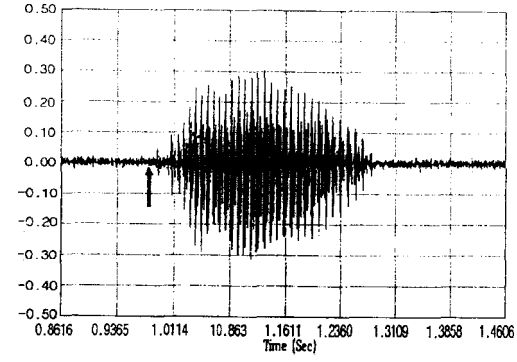
그림 5. ARS에 포함된 음성인식 시스템의 블럭도
Fig 5. The block diagram of the speech recognition system included in an ARS

음성인식 시스템의 입력파라메터로는 LPC 켑스트럼 계수를 인간의 청각특성인 mel-scale로 주파수warping하여 사용하였다. 실험에는 10개의 state를 갖는 left-right 모델을 이용한 DHMM (Discrete Hidden Markov Model) 음성인식 시스템을 사용하였다. 음성인식 시스템의 훈련에는 15명의 화자의 음성이 사용되었으며, 인식실험에는 25명분의 음성이 사용되었다.

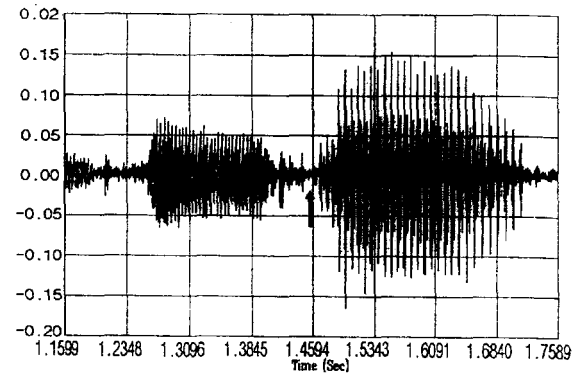
전화음성 인식의 실험에 앞서 마이크음성을 사용하여 baseline 시스템의 기본 성능평가를 하였다. Teager 에너지를 이용하여 끝점검출한 마이크 음성을 대상으로 한 화자 독립 숫자음성 인식률은 92.5%였다. 단일화자의 참조패턴에 대한 트레이닝을 수행할 경우 이보다 좋은 성능을 보일 수 있다. 이렇게 평가된 baseline 시스템은 전화음성을 대상으로 하였을 경우 인식률의 평가 기준을 제시한다.

4.3 전화음성 입력순간 검출 알고리듬의 실험

그림 6은 전화음성의 파형과, 전화음성과 안내방송이 혼합된 파형을 각각 나타내고 있다. 전화음성은 안내방송의 피드백에 의하여 음성의 시작부분(그림에서 화살표로 표시) 전에는 상당히 영향을 받았으나, 시작점 부근에서는 '전화음성 입력순간 검출 알고리듬'에 의하여 그 순간 안내방송이 중단되어 시작점 이후의 파형은 사용자 음성의 형태가 그대로 남아 있음을 볼 수 있다.



(a)



(b)

그림 6. (a) 숫자음 /오/의 파형과 (b) 안내방송과 혼합된 숫자음 /오/의 파형
Fig 6. The waveforms of (a) the telephone speech /ɔ/ and (b) that mixed with the announcement

4.4 제안된 끝점검출 알고리듬의 실험

전화음성의 끝점검출은 서론에서 언급한 것과 같이 많은 어려움이 따른다. 본 절에서는 3.2절에서 제안한 효율적인 끝점검출 알고리듬의 실험 방법 및 결과에 대하여 살펴본다. 마이크로 녹음된 음성은 Rabiner가 제안한 방법을 사용하였으며[8], 전화음성은 잡음특성을 갖고 있기

때문에 기존의 프레임에너지를 Teager가 제안한 에너지 알고리듬으로 대체하여 끝점검출을 수행하였다. (식(14) 참고)

본 연구에서는 이산 HMM 인식기를 이용하여 기존의 알고리듬과 새로운 에너지 알고리듬에 대한 인식률을 비교하였다. 기존의 에너지 알고리듬 대신에 새로운 에너지 알고리듬을 적용하였을 때 파열음 및 마찰음 구간에서 좋은 효율을 보였다. 표 3에서 보듯이 일반적인 프레임에너지를 적용하였을 때 보다 새로운 알고리듬을 적용하였을 때 인식률이 향상되었다. 그림 7은 숫자음성 /칠/에 대한 기존의 알고리듬과 Teager에너지를 적용한 알고리듬의 끝점 검출 결과를 보여주고 있다. 기존의 알고리듬을 사용한 경우는 숫자음성 /칠/의 시작구간인 마찰음 구간에서 음성의 시작점이 잘못 검출되어 음성정보가 손실 되었으며, Teager에너지를 적용한 알고리듬의 경우 검출된 구간에 모든 음성정보를 포함하고 있음을 알 수 있다.

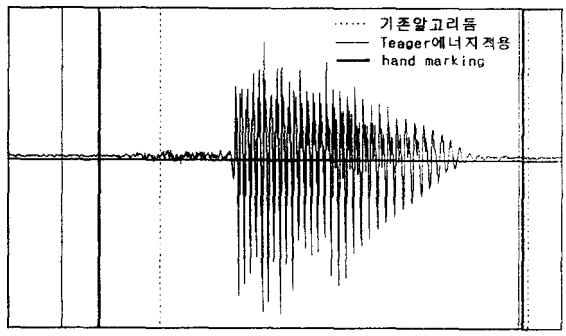


그림 7. 숫자음성 /칠/에 대한 끝점검출 결과
Fig 7. The result of the endpoint detection for speech /chil/

4.5 음성보상 알고리듬의 실험

음성보상 알고리듬의 실험은 본고에서 제안한 끝점검출 알고리듬을 적용하여 얻어진 음성에 전화음성의 보상방법을 이용하여 수행하였다:입력된 음성은 제안된 끝점검출 알고리듬을 통하여 음성의 끝점검출을 수행하여 baseline 시스템에 필요한 특징벡터를 구하는 과정을 거친다. 이러한 과정을 거친 결과로서 주파수 warping된 켑스트럼 벡터가 생성되며 이는 ASCII 형태로 IBM-PC의 ADD에 저장한다. 이렇게 저장된 켑스트럼 벡터는 SDCN 알고리듬으로 보상된 후 인식프로그램의 입력으로 사용된다. 인식프로그램의 결과로서 각 HMM 모델과 입력 특징벡터열의 발생확률이 가장 큰 HMM 모델이 선택되어진다.

전화음성의 경우 보상전과 보상후의 결과를 각각 비교하였다:표 1은 오인식된 전화음성의 예로 전화음성 /이/가 /일/로 오인식된 경우이다. 표 2는 오인식된 전화음성 /이/를 SDCN 알고리듬에 기초한 방법을 이용하여 보상을 취한 결과이다. 그 결과, 표 1과 비교하여 보면 /이/의 후보순위가 가장 높아졌음을 알 수 있다. 표 3에서 보듯이 같이 마이크 음성을 통한 baseline 시스템의 화자독립 인식실험에서는 92.5%의 인식률을 보였으나, 전화음성을 이용하였을 경우 인식률이 30%로 상당한 인식률의 저하를 보였다. 이러한 전화음성의 인식률의 저하를 향상시키기 위하여 SDCN 알고리듬을 적용하여 보상을 취한 결과, 인식률이 76.9%로 향상되었다.

표 1. 전화음성의 오인식 결과
Table 1. Misrecognized results of telephone speech

전화음성 /이/의 오인식결과	
* Recognition result : 1	(SDCN002T.FVR)
1st cadidate is '1' =>	FVL1.HMM:-230.2
2nd cadidate is '2' =>	FVL2.HMM:-284.5
3rd cadidate is '0' =>	FVL0.HMM:-335.3
4th cadidate is '6' =>	FVL6.HMM:-625.9
5th cadidate is '4' =>	FVL4.HMM:-639.4
6th cadidate is '3' =>	FVL3.HMM:-641.9
7th cadidate is '7' =>	FVL7.HMM:-644.2
8th cadidate is '5' =>	FVL5.HMM:-645.4
9th cadidate is '8' =>	FVL8.HMM:-650.5
10th cadidate is '9' =>	FVL9.HMM:-657.4

표 2. 보상에 의한 인식결과
Table 2. Recognition results of the compensated speech

전화음성 /이/의 보상결과	
* Recognition result : 2	(SDCN002T.FVR)
1st cadidate is '2' =>	FVL2.HMM:-234.8
2nd cadidate is '1' =>	FVL1.HMM:-283.1
3rd cadidate is '0' =>	FVL0.HMM:-382.5
4th cadidate is '6' =>	FVL6.HMM:-536.8
5th cadidate is '8' =>	FVL8.HMM:-618.0
6th cadidate is '4' =>	FVL4.HMM:-624.1
7th cadidate is '5' =>	FVL5.HMM:-627.6
8th cadidate is '9' =>	FVL9.HMM:-630.2
9th cadidate is '3' =>	FVL3.HMM:-631.2
10th cadidate is '7' =>	FVL7.HMM:-631.6

표 3. 인식률의 비교
Table 3. The comparison of the recognition rate.

입력음성 / 끝점검출	마이크 음성	전화음성	보상된 전화음성
프레임에너지 + 영교차율	80.0	22.5	65.0
Teager 에너지	92.5	30.0	76.9

V. 결 론

본고에서는 기존의 ARS에 음성인식 기능을 추가할 경우에 발생하는 인식률 저하의 원인을 제시하고 그의 향상방법을 제안하였다. 음성이 안내방송과 혼합되어 음성의 오인식이 발생하는 것을 피하기 위하여 음성의 입력순간을 정확히 검출하고 그 순간 안내방송이 중단되도록 하였다. 또한 전화음성에 존재하는 잡음이 음성분석을 위한 끝점검출을 부정확하게 만드는 것을 Teager에너지

을 도입하여 극복하였다. 마지막으로, 전화채널의 특성에 의한 음성신호의 왜곡을 보상하기 위하여 SDCN 알고리듬을 사용하였다. 이와 같은 세 가지 방법을 모두 결합하여 음성인식률을 시험한 결과 상당한 성능의 향상을 이루었다. (표 3 참고).

SDCN보다 진보된 방법으로 전화음성을 보상하기 위하여서는 인식을 위해 입력된 음성으로부터 환경에 대한 변수들을 예측하여 보상벡터를 구해주는 알고리듬을 이용할 수 있다. 그러나, 이러한 알고리듬은 환경에 대한 변수들을 예측하기 위하여 확률적인 방법을 사용하므로, DTW를 이용한 인식기에는 적용이 어려우며 계산량이 많아 좀더 뛰어난 성능의 하드웨어를 필요로 한다. 일반 음성인식 시스템에 사용된 데이타 베이스는 잡음이 적은 환경에서 마이크 음성을 이용하여 구현하므로 전화음성 인식 시스템에 이를 적용하면 상당한 성능 저하를 보인다. 따라서 전화음성인식 시스템은 음성 데이타 베이스의 구축을 위하여 일반적으로 전화음성을 사용한다. 그러나, 전화음성을 통한 데이타 베이스의 구축은 많은 시간과 노력이 필요하게 된다.

따라서 향후 연구 과제는 전화음성인식 시스템의 인식률의 향상을 위한 방법과, 음성 데이타 베이스 구축을 기존의 음성인식 시스템과 같은 방법으로 구축할 수 있는 방법과, 간단한 보상을 통한 기존의 데이타 베이스의 활용방법 등에 중점을 두어야 할 것이다.

참 고 문 헌

1. 도산수, 은종관, "전화음성의 격리단어인식 개선에 관한 연구," 한국 음향 학회지, 9권 4호, pp. 66-76, 1990. 7.
2. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. ASSP-27(2)*, pp. 113-120, April, 1979.
3. M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech Corrupted by Acoustic Noise," *In J. S. Lim(editor), Signal Processing. Vol. 1 : speech Enhancement*, Prentice-Hall, Englewood Cliffs, NJ, pp. 69-73, 1983.
4. D. Van Compernolle, "Spectral Estimation Using a Log-Distance Error Criterion Applied to Speech Recognition," *In Proc. IEEE Int. conf. ASSP*, Glasgow, UK, pp. 258-261. May, 1989.
5. D. Van Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System," *Computer Speech and Language*, 1989.
6. R. Stern and A. Acero, "Acoustical Pre-processing for Robust Speech Recognition," *In Proc. Speech and natural Language Workshop*, Cape Cod, MA, Morgan Kaufmann, pp. 311-318. Oct, 1989.
7. J. E. Porter and S. F. Boll, "Optimum Estimators for Spectral Restoration of Noisy Speech," *In Proc. IEEE Int. Conf. ASSP*, sandiego, CA, pp. 18A. 2. 1, May, 1984.
8. L. R. Rabiner, J. G. Wilpon, "An Improved Endoint Detector for Isolated word Recognition," *IEEE Trans. ASSP-29, NO. 4*, pp. 777-785, Aug. 1981.
9. D. Messerschmitt, D. Hedberg, *Digital Voice Echo Canceller with a TMS32020*, Prentice Hall and Texas Instruments, 1987.
10. J. F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal," *In Proc. IEEE Int. conf. ASSP*, pp 381-384, 1990.
11. A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic, 1993.
12. G. S. Ying et al., "Endpoint detection of isolated utterances based on a modified Teager energy measurement," *In Proc. IEEE Int. conf. ASSP*, pp. 732-735, 1993.

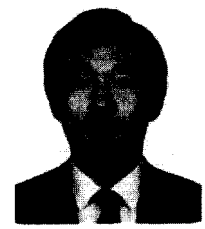
▲안 정 모 1970년 7월 5일생

1993년 2월 : 홍익대학교 전자공학과 학사
1995년 2월 : 홍익대학교 전자공학과 석사
1994년 12월~현재 : 대우전자 영상연구소 연구원
※주관심분야 : 음성신호 및 영상신호처리

▲임 계 종 1970년 8월 21일생

1995년 2월 : 홍익대학교 전자공학과 학사
1997년 2월 : 홍익대학교 전자공학과 석사
1997년 1월~현재 : 현대전자 근무 중
※주관심분야 : 음성인식, 디지탈 신호처리, 컴퓨터 네트워크

▲계 영 철 1957년 12월 29일생



1980년 2월 : 서울대학교 전자공학과 학사
1982년 2월 : 한국과학기술원 전기 및 전자공학과 석사
1991년 5월 : Univ. of Southern California, Electrical Eng. Ph.D.
1991년 9월~현재 : 홍익대학교 전자공학과
※주관심분야 : 디지탈 신호처리, 음성 및 영상인식, 로봇 비젼

▲구 명 완
현재 : 한국통신 멀티미디어 연구소 음성언어 연구팀장
(1996년 제15권 4호 참조)