전화음성의 격리단어인식 개선에 관한 연구

A Study on the Improvement of Isolated Word Recognition for Telephone Speech

도 삼 주* 은 종 관**

(Sam Joo Do, Chong Kwan UN)

요 약

본 논문에서는 잡음과 전화선로의 왜곡이 음성인식에 미치는 영향을 알아보고, 전처리 과정을 추가하여 이를 개선하는 방법을 제안하였다. 컴퓨터 모의실험은 음소적으로 고르게 분포되어있는 한국어 격리단어 100단어를 각각 10회 발음한 1000개 데이타를 대상으로하고, 화자종속으로 수행하였다. 먼저 잡음에 대한 개선방법으로 spectral subtraction을 제안하였는데, 이것은 매우 간단하면서도 좋은 성능을 보였다. 다음으로 대역폭제한과 전송로왜곡의 영향을 실험하였는데, 대역폭의 제한과 진폭왜곡은 인식율을 크게 떨어뜨렸으나 위상왜곡은 별로 영향이 없었다. 또, 전송로의 영향을 개선하기 위하여 training data를 사용하여 기준패턴을 변화시키는 방법을 제안하였다. 잡음과 전송로의 왜곡이 동시에 있는 경우에 인식율이 7.7~26.4% 밖에 되지 않았는데, 위에서 제안한 방법을 이용하여 76.2~92.3%로 개선되었다.

ABSTRACT

In this work, the effect of noise and distortion of a telephone channel on the speech recognition is studied, and methods to improve the recognition rate are proposed. Computer simulation is done using the 1000-word test data which were made by pronouncing ten times 100-phonetically balanced Korean isolated words in a speaker dependent mode. First, a spectral subtraction method is suggested to improve the noisy speech recognition. Then, the effect of bandwidth limiting and channel distortion is studied. It has been found that bandwidth limiting and amplitude distortion lower the recognition rate significantly, but phase distortion affects little. To reduce the channel effect, we modify the reference pattern according to some training data. When both channel noise and distortion exist, the recognition rate without the proposed method is merely 7.7~26.4%, but the recognition rate with the proposed method is drastically increased to 76.2-92.3% .

* 정회원, 서울대학교 전자계산기공학과 음성정보연구실
** 정회원, 서울대학교 전자공학과 통신연구실

I. 서 론

음성은 인간이 가지고 있는 가장 오래된 의사전달 수단의 하나로서, 시간이나 장소, 그리고 신체적 결함등으로 인한 제약이 적으며 사이스템에 많은 양의 정보를 전달할 수 있다. 또한 음성은 특별한 교육이나 훈련이 따로 필요하지 않다는 장점을 가지고 있어서, 음성을 기계와 인간사이의 정보전달의 수단으로 이용하고자 하는 연구가 꾸준히 진행되어 왔다.

음성인식 시스템의 응용분야에는 여러가지가 있겠지만, 전화를 통한 음성인식이 가능하다면 그 응용분야는 훨씬 넓어질 것이다. 값비싼 컴퓨터나 단말기를 개인이 보유하고 있지 않더라도, 손쉽게 이용할 수 있는 전화를 통하여 필요한 조작을 하고, 원하는 정보를 얻을 수 있다면 매우 편리할 것이다. 현재 이와 비슷한 기능을 하는 시스템으로 audiotex라는 것이 있으나 이것은 버튼으로 명령을 하기 때문에 버튼식 전화기에서만 사용할 수 있고, 버튼을 누르는 번거로움과 사용할 수 있는 명령의 제한성 등을 가지고 있어 사용에 불편한 점이 많다. 또한 많은 사람들이 바라고 있는 전화 자동통역 시스템을 구현하기 위해서는 전화를 통한 음성인식이 반드시 필요하다.

그러나, 전화를 통한 음성인식에는 몇가지 문제점이 있다. 먼저 대역폭의 제한을 들 수 있다. 현재 전화선은 약 3kHz 정도의 대역폭 밖에 가지고 있지 않으므로 그보다 높은 주파수 대역에 있는 음성정보를 잃어 버리게 된다. 그리고, 전송로에 의한 왜곡과 전송로 특성의 시간적 변화 등에 영향을 받게 되며, 이외에도 crosstalk, impulse noise, quantization noise 등의 각종 잡음에 의해서도 인식률이 [illegible]

[illegible] 음성인식을 더욱 어렵게 만든다. 그러므로, 이를 극복하기 위한 연구가 많이 되어 왔는데, 크게 잡음에 대해서는 [illegible] 이를 위한 방법으로 여러가지로, 대역폭의 방법

들은 주로 음성인식의 여러가지 요소들을 잡음에 좀더 강하도록 만드는 것이다. 예를 들면 잡음에 강한 음성특성과 distortion measure를 이용하거나, 잡음 부분의 통계적 특성이나 음성대상 단어의 구문과 의미적 특성등을 이용하여 끝점검출을 더 잘하도록 만드는 것등이 있다.[1][2][3] 다른 방법으로는 음성인식 전에 전처리 단계를 추가하여 나빠진 음질을 개선시킨 후 음성인식에 사용하도록 하는 방법이 있다. 잡음에 의하여 나빠진 음성의 개선에 관한 연구는 speech enhancement라고 하여 1970년대 중반부터 활발히 진행되었는데, 이것은 주로 사람이 그 음성을 더 잘알아 듣도록 만드는데 목표를 둔 것이었다.[4] 여기에서는 음성과 잡음의 특징 차이를 이용하여 잡음을 제거시키거나, 사람의 음성인식에 중요한 부분을 강조시키는 방법 등을 사용하였다. 컴퓨터의 음성인식의 경우에는 다시 거기에 적합한 방법을 찾아야 겠지만 기본적으로는 위의 방법을 그대로 이용하여도 큰 효과를 볼 수 있다. 그러나, 전화선로를 통한 음성인식에 관한 연구는 미국의 Bell 연구소에서 실제상황에서의 실험을 하였고 간단한 시스템은 상업용으로도 나와 있지만, 그 성능개선을 위한 연구는 아직 미진한 형편이다.[5]

본 논문에서는 잡음, 대역폭의 제한, 그리고 전송로의 왜곡등이 음성인식에 미치는 영향을 알아보고, 이에 대한 개선 방법을 제안한다. 잡음의 영향에 대해서는 speech enhancement의 한 방법인 spectral subtraction을 음성인식 시스템에 적용하여서 인식율을 향상시켰고, 대역폭의 제한과 전송로의 왜곡에 대해서는 몇개의 training data를 이용하여 기준패턴(reference pattern)을 적절히 변화시켜 줌으로써 인식율을 향상시켰다. 음성인식 실험은 화자종속에서 1000개 (100단어×10회 발음)의 한국어 격리단어를 대상으로, filter bank output을 인식특징으로하고 DTW를 이용하여 수행하였다.

II. 전화선로의 특성

전화선로에서 발생하는 문제점들을 파악하고, 그에 대한 대책을 마련하기 위해서는 그 전화선로의 특성을 잘 알아야 한다. 그러나, 이러한 특성

표 1. 전화선로에서 고려되는 최악의 경우의 여러가지 손상

impairment	level
attenuation of a 1004 Hz tone	27 dB
signal to C-notch noise ratio	20 dB
signal to second harmonic distortion ratio	34 dB
signal to third harmonic distortion ratio	33 dB
frequency offset	3 Hz
peak to peak phase jitter(2~300 Hz)	20 degrees
peak to peak phase jitter(20~300 Hz)	13 degrees
impulse nois (-4 dB threshold)	4 per minute
phase hits(20 degree threshold)	1 per minute
round trip delay(no satellites)	50 ms

의 조사는 그 규모의 방대함 때문에 쉽게 이루어지기가 어렵다. 더구나, 현재의 공중전화망에는 특성이 다른 여러가지 교환기와 선로들이 섞여져 있고, 이들이 어떻게 연결되는가에 따라 그 특성이 다르게 나타나기 때문에 그 특성을 한마디로 말하기는 어렵다. 국내의 경우 아직도 전화선로의 특성에 대한 조사가 미진한 실정이다. 표 1에 전화선로에서 고려되는 여러가지 손상들을 나타내었다. 이것은 미국의 벨 연구소에서 측정한 것으로, 이것을 우리나라의 경우에 직접적으로 사용하기는 어렵겠지만 개략적인 값을 파악하는데는 좋은 참고가 될 것이다.[6]

이와 같이 전화선로에는 각종 잡음과 왜곡등이 존재하는데, 여기에 나타낸 것외에도 진폭왜곡과 위상왜곡이 매우 중요하다. 이것은 전송로의 주파수 특성이 그림 1과 같이 주파수에 따라서 일정하지 않은 것을 뜻하는 것으로서, 이들을 합하여 선형왜곡이라고 한다. 만약 전송로의 주파수 응답 $C(f)$가

$$C(f) = |C(f)|e^{j\theta(f)} \tag{1}$$

라고 하면, 진폭왜곡은 $|C(f)|$가 일정하지 않은 것을 말하고, 위상왜곡은 다음 식과 같이 정의되는 군지연 $\tau(f)$가 일정하지 않은 것을 말한다.

$$\tau(f) = -\frac{1}{2\pi}\frac{d\theta(f)}{df} \tag{2}$$

이와같은 여러가지 왜곡들은 사람사이의 통신에는 큰 문제가 없다하나 컴퓨터의 음성인식에는 큰 문제가 될 수 있다. 본 논문에서는 위의 왜곡들 중에서 진폭왜곡과 위상왜곡의 영향에 대하여 실험을 하였다.

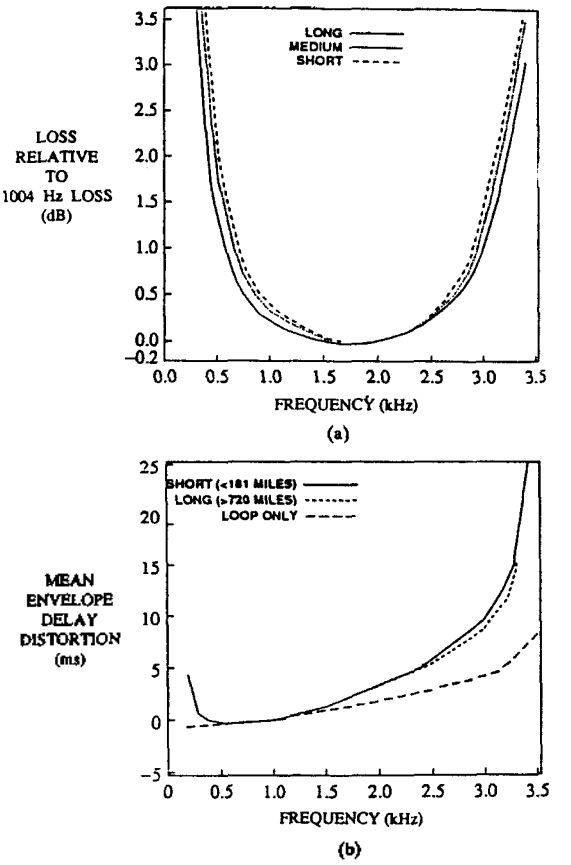


그림 1. 전화선로의 선형 왜곡 특성
(a) 진폭왜곡 (b) 위상왜곡

Ⅲ. 잡음과 전송로의 영향의 대한 개선 방법

Ⅲ.1 잡음의 영향에 대한 개선 방법

잡음을 제거시키는 방법들은 기본적으로 음성과 잡음의 여러가지 정보를 이용하는 것이다. 우리는 음성신호를 표현하기 위하여 적당한 모델을 사용하는데, 일반적으로 음성신호를 더 세세히 모델링하면 할수록 잡음으로부터 음성을 더 잘 분리해 낼 수 있다. 그러나, 이것은 음성신호 모델과 실제 음성신호사이의 차이점에 시스템을 더 민감하게 만드는 단점을 가지고 있기 때문에 적당한 모델이 필요하며

된다. 잡음 정보의 이용에서도 마찬가지여서 어떤 종류의 잡음을 고려하는가에 따라 시스템을 적절히 구성하여야 한다. 예를 들어서 고려되는 잡음이 다른 사람의 말소리인 경우와 white noise인 경우가 있을 때, 각각에 알맞은 시스템의 구성은 서로 다르게 될 것이다.

음성에 섞여진 잡음을 제거시키는 방법을 크게 나누면 단시간 스펙트럼 진폭 추정(short-time spectral amplitude estimation)에 의한 방법, 유성음의 주기성을 이용한 방법, 그리고 음성모델을 이용한 방법들이 있다. 첫째 방법에는 spectral subtraction, Wiener filtering 등이 있고, 둘째 방법에는 comb filter, 적응 잡음 제거 등이 있으며, 세째 방법에는 all-pole model이나 ploe-zero model 등을 이용하는 것이 있다. 여기서 첫째와 세째 방법은 기본적으로 음성을 stochastic process로 생각한 것으로 무성음에 더 적합하다고 할 수 있다. 사람이 음성을 들을 때는 phase의 정보는 별로 중요하지 않고, 단시간 스펙트럼 진폭이 가장 중요하다. 위에서 첫째 방법은 이러한 것을 이용한 것으로, 간단하면서도 좋은 성능을 보인다. 컴퓨터의 경우에도 이와 비슷하다고 생각되어 본 논문에서는 이 방법을 사용하였다.

다음과 같이 잡음이 더하여진 신호를 생각하자.

$$y(n)=s(n)+d(n) \tag{3}$$

이 식에서 $s(n)$은 음성이고, $d(n)$은 잡음으로서 window 가 씌어진 신호들이다. 이에 대응되는 Fourier transform을 각각 $Y(\omega)$, $S(\omega)$, 그리고 $D(\omega)$라고 하면,

$$|Y(\omega)|^2=|S(\omega)|^2+|D(\omega)|^2+S(\omega)D^*(\omega)+S^*(\omega)D(\omega) \tag{4}$$

이 된다.

여기에서 우리의 목적은 $|S(\omega)|^2$의 값을 추정하는 것이다. 이를 위해서는 각각의 값을 알아야 하는데, $|Y(\omega)|$의 값은 입력 신호로 부터 직접 계산할 수 있으나, 나머지의 값은 알 수가 없다. 따라서 추정에 의한 근사치인 $E[|D(\omega)|^2]$, $E[S(\omega)D^*(\omega)]$, 그리고, $E[S^*(\omega)D(\omega)]$들을 사용하게 된다. 만약 잡음 $d(n)$이 $s(n)$과 uncorrelate되어 있고 그 평균값이 0이라면, $E[S(\omega)D^*(\omega)]$와 $E[S^*(\omega)D(\omega)]$는 0이 된다. 그러므로, 이것을 (4)에 이용하면, $|S(\omega)|^2$의 추정값은 다음과 같이 된다.

$$|\hat{S}(\omega)|^2=|Y(\omega)|^2-E[|D(\omega)|^2] \tag{5}$$

여기서 $E[|D(\omega)|^2]$의 값은 $d(n)$의 특성을 알고 있다고 가정하거나, 음성이 없는 구간으로 부터 잡음을 측정하여 구한다. 여기에서는 배경 잡음이 이 구간 내에서 stationary하다고 가정한 것이다. 그런데, (5)는 오른쪽 항이 음수가 될 수도 있는 문제점을 가지고 있다. 이를 해결하기 위하여 음수를 부호만 바꾼 양수로 만들거나 음수인 것은 그값을 모두 0으로 치환하는데, 보통은 뒤의 방법이 더 많이 사용된다.

$s(n)$의 값을 추정하기 위해서는 위에서 구한 amplitude 외에도 위상의 값을알아야 하는데, 정확한 값을 계산하는 것은 매우 어렵다. 그런데, 우리는 위상의 정보를 중요하게 생각하지 않으므로 $y(n)$의 위상을 그대로 사용한다. 그러면,

$$\hat{S}(\omega)=|\hat{S}(\omega)|exp[jY(\omega)] \tag{6}$$

이고,

$$\hat{s}(n)=F^{-1}[\hat{S}(\omega)] \tag{7}$$

가 된다. 이렇게 하여 음성신호를 개선시키는 방법의 블록도를 그림 2에 나타내었다.[7][8]

좀더 일반적인 방법으로 다음 식과 같이 a와 k를 적당한 값으로 정하여 이용하는 수도 있다.

$$|\hat{S}(\omega)|^a=|Y(\omega)|^a-k\hat{E}[|D(\omega)|^a] \tag{8}$$

본 논문에서는 $|\hat{S}(\omega)|^2$의 값을 구할 때, (5)의 성능에 큰 차이가 없어서 다음과 같이 간단한 식을 사용하였다.

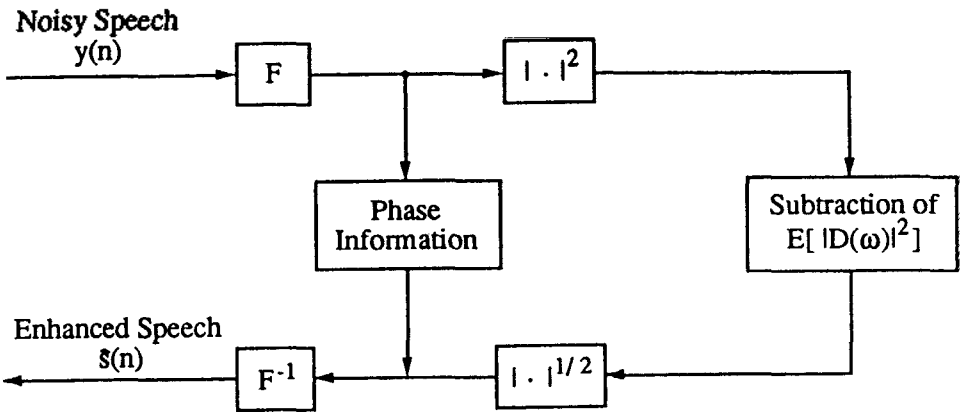


그림 2. Spectral Subtraction에 의한 음성개선 시스템의 블록도

$$|\hat{S}(\omega)| = |Y(\omega)| - E[|D(\omega)|] \quad (9)$$

Ⅲ.2 전송로의 영향에 대한 개선 방법

데이타 통신의 경우에는 전송로의 왜곡에 대하여 equalizer를 사용하여 보완을 해주지만, 음성의 경우에는 이러한 방법이 어렵다. 데이타 통신에서는 보내는 신호의 형태를 미리 알고 있기 때문에 이것이 가능한 것이지만, 음성의 경우에는 약속된 말을 먼저 하더라도 그 파형은 항상 변하는 것이기 때문에 이러한 방법이 곤란한 것이다. 그러나, 음성인식 시스템은 대상어휘에 대한 기준패턴을 가지고 있으므로 이것을 이용하는 방법을 생각할 수 있다. 즉, 약속된 단어를 training data로 먼저 받아서 이것의 특징패턴을 추출한다. 다음으로, 이 특징패턴과 기준패턴을 비교하여 평균 차이값을 구한 후, 기준패턴을 이 값만큼 변화시켜서 그 차이를 보상하여 주면 상당한 인식율의 개선을 얻을 수 있다.

본 논문에서 음성인식은 Sakoe와 Chiba의 제약조건에 따른 DTW 방법을 사용하였으므로 평균 차이값을 구하는 것도 이와 같은 방법으로 한다. 즉, DTW에 의한 두 패턴의 distance가 최소가 되도록 기준패턴을 변화시키는 것이므로, 변화시킬 양을 구하는 것도 distance를 구할 때와 같은 경로에서 같은 방법으로 하여야 한다. 최적경로를 찾기 위해 누적거리를 구하는 식은

$$D_a(1,1)=2d(1,1)$$
$$D_a(n,m)=\infty,\ n<0 \text{ or, } m<0$$
$$D_a(n,m)=\min[D_a(n-1,m-2)+2d(n,m-1)+d(n,m),\ D_a(n-1,m-1)+2d(n,m),\ D_a(n-2,m-1)+2d(n-1,m),+d(n,m)],\ 2<n<N,\ 2<m<M \quad (10)$$

이므로, 점 (n,m)에서 k번째 filter bank output의 차이값을 $diff_k(n,m)$라 하면, 점$(1,1)$에서 점(n,m)까지 k번째 filter bank output의 누적 차이값 $Diff_{k\cdot a}(n,m)$는 다음 식과 같이 된다.

$$Diff_{k\cdot a}(1,1)=2diff_k(1,1)$$
$$Diff_{k\cdot a}(n,m)=0,\ n<0 \text{ or, } m<0$$
$$Diff_{k\cdot a}(n,m)=\min[Diff_{k\cdot a}(n-1,m-2)+2diff_k(n,m-1)+diff_k(n,m),\ Diff_{k\cdot a}(n-1,m-1)+2diff_k(n,m),\ Diff_{k\cdot a}(n-2,m-1)+2diff_k(n-1,m),+diff_k(n,m)],\ 2<n<N,\ 2<m<M \quad (11)$$

여기에서 $\min_d$는 distance를 구하는 식에서 minimum으로 결정된 항에 대응되는 항을 선택한다는 뜻이다. 그러면 k번째 filter bank output의 차이 $Diff_k$는

$$Diff_k = \frac{Diff_k a(N,M)}{N_0} \quad (12)$$

된다.

그런데, 본 논문에서 사용된 DTW 방법은 각 단어에 대하여 5개의 reference를 가지고 있고, 각 단어에서의 distance 계산은 이들 5개중에서 distance가 작은 3개만의 평균값으로 하였다. 평균 차이값을 구할 때도 이와 마찬가지로 distance가 작은 3개에 대응되는 차이값만을 사용하였다. 이렇게하여 한 training data에 대한 평균 차이값을 구한 후에, 또 다른 training data에 대해서도 위의 과정을 반복하여서 구한 전체값들을 각 filter bank output 별로 평균하여서 최종적인 값을 얻는다. 다음으로는 기준패턴에서 이 값을 빼어서 기준패턴을 변화시킨다.

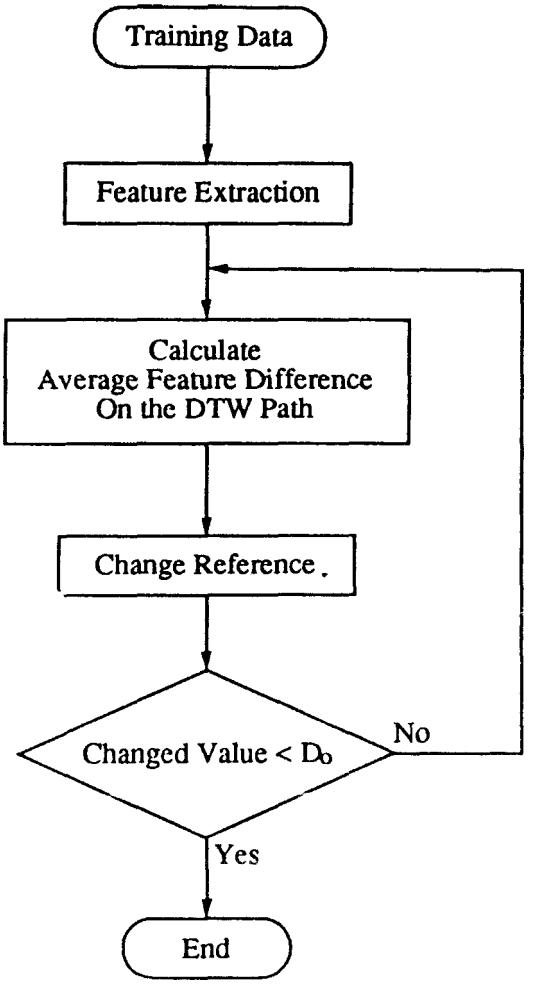


그림 3. 기준패턴을 변화시키는 방법

이렇게 기준패턴을 변화시키면 최적경로도 변할 수 있으므로, 변화된 기준패턴을 이용하여 위의 과정을 반복하여 최적값을 찾는다. 이와 같이 기준패턴을 변화시키는 것의 전체적인 블록도가 그림 3에 있고, 이에 의해 기준패턴이 변화된 예가 그림 4에 있다. 이렇게 변화된 기준패턴을 왜곡되어져 들어오는 입 음성과 비슷한 특성패턴을 가지므로, 이것을 이용하여 음성인식을 하면 좋은 인식율을 얻을 수 있다.

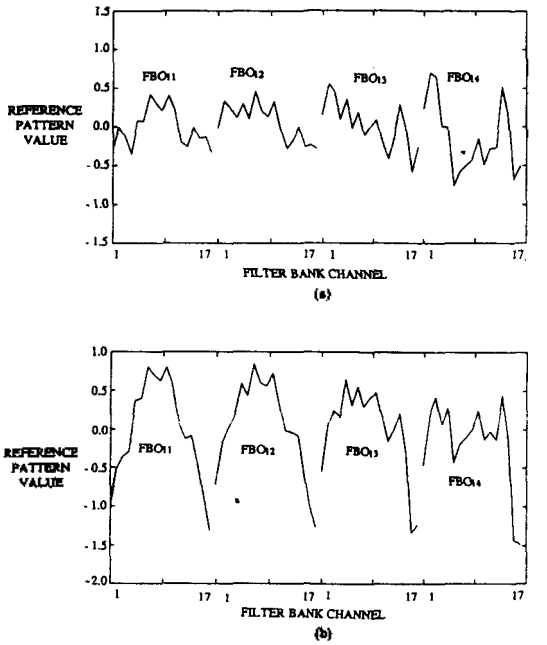


그림 4. 기준패턴이 변경된 예 (a) 변경 전 (b) 변경 후

Ⅳ. 컴퓨터 모의 실험

Ⅳ.1 실험 데이타의 교성과 인식 방법

본 논문에서는 표 2에 나타낸 것과 같은 한국어 고립단어 100개를 인식대상어휘로 사용하였다. 이 단어들은 100개 전체에서 발음에 나타나는 각 음소들의 개수가 비슷하도록 선정된 것이다. 이 단어들의 각 음소별 출현 빈도가 일상적으로 우리가 사용하는 말의 그것과 같다고는 할 수 없겠지만, 실험 데이타로서는 충분하다고 할 수 있을 것이다. 기준패턴은 한 사람이 각 단어를 5회씩 발음

한 것을 이용하여 구성하였고, 실험 데이타는 같은 사람이 각 단어를 10회씩 발음한 1000개의 데이타를 사용하였다. 즉, 화자종속 인식 실험을 하였다. 먼저 릴 테이풉에 단어를 녹음한 후에 4.5kHz의 cut-off 주파수를 갖는 low-pass filter를 사용하여 고주파 성분을 제거시켰다. 이때 사용된 filter는 8-pole Butterworth filter이다. 다음으로 이것을 SUN 컴퓨터의 A/D 변환기를 이용하여 10kHz 주파수로 sampling하여 12 bit의 디지탈 데이타로 만들어서 실험에 사용하였다.

음성특징은 17-channel filter bank output를 사용하였고, distortion measure로는 Euclidean distance를 사용하였다. 인식방법은 DTW를 사용하였는데, 계산시간의 절약을 위하여 각 입력단어에 대하여 먼저 간단한 방법으로 10개의 후보기준패턴을 선정하고 이 후보기준패턴들 하고만 DTW 방법을 적용하였다. 후보기준패턴은 각 데이타의 frame 길이가 모두 똑같이 9가 되도록 linear normalization한 후 Euclidean distance를 이용하여 선정하였다. 10개의 후보기준패턴과의 DTW에서, 한 단어에 대해서 5개씩의 기준패턴을 가지고 있으므로 입력패턴과 한 후보기준패턴 사이에는 5개의 distance가 계산된다. 본 논문에서는 이중에서 최대 최소를 제외한 중간의 3개 값을 평균하여서 distance로 사용하였다.

기준패턴은 clean speech를 이용하여 만든 것을 계속 사용하였다. 입 음성이 왜곡됐을 경우, 입 음성과 같이 왜곡된 음성으로 기준패턴을 만들어서 인식에 사용하면 그냥 인식하는데는 더 좋은 인식율을 얻을 수 있다. 그러나, 본 실험에서는 왜곡된 음성의 특성을 미리 알지 못한다고 가정한 것이므로 어려한 것은 적절하지 못하다. 물론 하나의 비교의 대상으로는 의미를 가질 수도 있을 것이나 본 논문에서는 이에 대해서는 실험을 하지 않았다.

표 2. 음소 고르게 분포되어 있는 인식대상 어휘

개	길	곧	얘기	연구
너	왜놈	넷	오늘	능력
더	달	닭	제도	현대
됴	물	먹	범위	전망
벼	불	북	교분	양보
소외	셈	색	효성	중심
이	왕	옆	농업	예의
유의	의원	우리	위약	화원
자유	정표	집	진리	문제
추위	총의	책	잔치	실천
코	칼	유쾌	싫고	색깔
터	탈	탑	넣다	형태
과괴	풀	팔	예포	살펴
혀	형	흙	유희	문화
께	끈	끝	야권	함께
때	땔	뚯	뒤뜰	갈등
뻐	뿔	오빠	안방	예법
씨	씸	싹	새싹	원칙
짜	짝	짓짜	발전	특징
억자	체조	체석	케법	화대

Ⅳ.2 잡음의 영향과 이에 대한 개선 실험

먼저 잡음이 섞이지 않은 원래의 음성에 대하여 인식 실험을 하였는데, 표 3에 그 결과를 나타내었다. 이 음성은 4.5kHz의 대역폭을 갖고, SNR이 31.3 dB이었는데, 96.8%의 인식율을 보였다.

잡음으로는 컴퓨터 언어인 C의 random 함수를 이용하여 만든 white noise를 사용하였다. 이것을 음성 데이타에 더하여서 잡음이 섞인 음성을 만들었으며, 이때 잡음에 적당한 값을 곱하여서 신호대 잡음비(SNR)를 변화시켰다. 예를 들어서 '안방'이라는 단어에서 '안'부분의 원래 파형과 white nosie가 섞인 파형, 그리고 개선되어진 파형을 보면 그림 5과 같다. 여기에서 SNR은 실험 데이타 전체에 대한 평균값을 의미하는 것으로, 같은 SNR에는 같은 크기의 잡음을 더하였다. 그러므로, 신호 크기가 작은 단어의 SNR은 상대적으로 더 낮게된다. 앞에서 설명한 spectral subtraction에 의한 개선결과를 Wiener filter를 이용한 경우와 비교하여서 표 4에 나타내었다.

SNR이 작을 때에는 인식율이 상당히 크게 개선되지만, SNR이 클 때는 인식율이 오히려 약간 낮아지는 경우도 있다. 예를 들어 표 4에서 spectral subtraction 에 의한 실험결과를 보면, SNR이 15dB일 때에는 65.0%의 개선율을 보였는데, SNR이 25 dB일 때에는 -3.9%의 개선율을 보였다. 음질 개선을 위하여 스펙트럼을 변화시키는 것에 의해서도 [illegible]

의 왜곡이 발생하는데, 이것도 한 가지 원인이 된다고 생각된다. 두 가지 개선 방법을 서로 비교하면, spectral subtraction에 의한 것이 계산이 간단하면서도 인식율이 좋게 나타났다. 특히 SNR이 작을 때에는 Wiener filter보다 훨씬 좋은 성능을 보였다. 또한, white noise 대신 에어콘의 fan 소리를 잡음으로 이용한 실험도 하였는데 대체로 비슷한 결과를 보였다.[7]

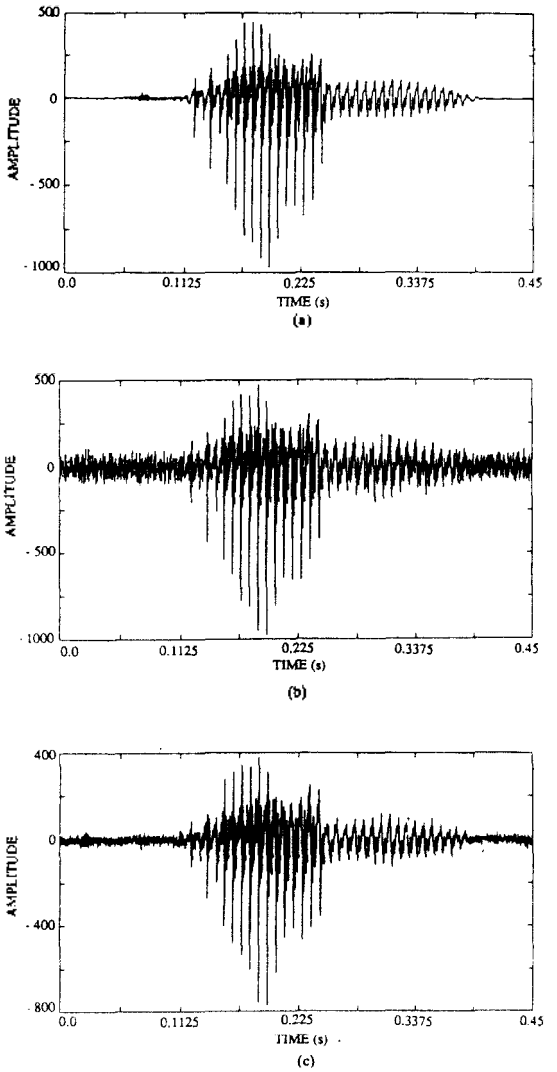


그림 5. (a) 깨끗한 음성의 파형
(b) (a)에 백색잡음이 섞인 파형(SNR=15dB)
(c) Spectral Subtraction에 의하여 개선된 파형

표 3. 왜곡이 없는 음성의 인식 결과

SUR(dB)	31.3
인식율(%)	96.8

표 4. White Noise 섞인 음성의 인식율
(괄호안은 개선율) (단위:%)

분류 \ SNR	15 dB	20 dB	25 dB
개선전	13.9	26.9	96.2
Spectral Subtraction	78.9 (65.0)	95.9 (69.0)	92.3 (-3.9)
Wiener Filter	20.5 (6.6)	51.8 (24.9)	91.9 (-4.3)

격리단어 인식에서는 끝점검출이 인식율에 매우 큰 영향을 미친다. 특히, 잡음 등에 의하여 왜곡되어진 음성의 경우에는 끝점검출의 잘못에 의하여 인식율이 크게 떨어지게 된다. 그러므로, 이런 경우에는 좀더 개선된 끝점검출 방법이 절실히 필요하게 된다. 그러나, 본 논문에서는 음성인식 알고리즘은 변화시키지 않고, 적절한 전처리 과정만을 추가함으로써 인식율을 향상시키는 방법을 연구한 것이기 때문에 이것에 관해서는 실험을 하지 않았는데, 우선 이것에 의한 영향의 정도를 알기 위하여 두가지로 나누어서 실험을 하였다. 즉, 하나는 기존의 프로그램을 이용하여 자동으로 끝점검출을 한 것이고, 또 하나는 사람이 직접 파형을 보고 끝점검출을 한 것이다. 본 논문의 표에 나타낸 실험 결과들은 자동으로 끝점검출을 한것이다. 끝점검출이 더 정확히 되는 수동의 방법이 더 좋은 인식율을 보이며 SNR이 작아질 수록 수동과 자동의 인식율 차이가 더 커지는데, 그만큼 끝점검출 부분이 중요하다는 것을 알 수 있다.[9] 그런데, 끝점검출에 관하여 고려하여야 할 것이 또 있다. 그것은 음질을 개선시키는 전처리 과정에서도 잡음의 특성을 알기 위하여 끝점검출이 필요하기 때문이다. 좋은 개선을 위해서는 잡음의 특성을 정확히 알아야 하므로 여기에서도 끝점검출이 정확해야 한다. 그런데, 잡음의 특성을 구할 때 끝점검출이 정확하지 않아서 음성 구간까지도 잡음 구간으로 [illegible] 잡음 구간의 일부분을 제외시키는 것은 별 문제가 없다. 그래서, 본 논문에서는 이때의 끝점검출을 자동으로 하되 약간의 여유를 주어서 끝점검출이 잡음 구간의 안쪽에서 되도록 하였다.

Ⅳ.3 전송로의 영향과 이에 대한 개선 실험

대역폭의 제한은 Signal Technology Inc.의 신호처리 프로그램인 ILS를 이용하여 98개의 tap을 갖는 finite impulse response(FIR) low-pass filter를 만들어서 실험하였는데, 그 결과는 표 5와 같다. 대역폭이 제한되면서 인식율이 상당히 저하되는데, 4.0 kHz에서의 인식율은 3.5kHz 때보다 오히려 낮은 것을 볼 수 있다. 그러나, 수동의 경우의 인식율은 3.5kHz 때 95.3 %, 4.0kHz때 97.8%이어서 이 범위 내에서는 별로 저하되지 않았고, 4.0kHz에서의 인식율이 더 좋은 것을 보면 이런 경우에도 끝점검출 부분의 영향이 큼을 알 수 있다.

전화선로의 영향을 알아보기 위하여 사용된 전화선로의 특성은 그림 6과 같다. 여기에서 진폭 왜곡은 G1과 G2인데, G2의 왜곡이 더 크다. 위상 왜곡은 D1과 D2인데, D1은 위상이 0으로서 왜곡이 없는 것을 뜻한다. 이러한 G와 D의 조합에 의하여 하나의 전송로가 구성되는데, 이의 실험 결과는 표 6과 같다. 여기에서 진폭 왜곡인 G1과 G2의 차이에 의하여 인식율이 크게 차이가 나고, 위상 왜곡이 있고 없음에는 큰 영향을 받지 않음을 알 수 있다. 표 6의 (b)와 (c)는 이를 개선을 시킨 결과를 보인 것이다. G1D2인 경우에는 21.1~21.8%의 개선율을 보였고, G2D2인 경우에는 52.9~54.1%의 개선율을 보였다. 각각의 경우에 사용된 training 데이타는 표 7과 같다. 여기에서 사용된 training 데이타의 갯수가 5개에서 20개까지 변하여도 인식율에는 별 차이가 없다. 즉, training 데이타가 몇개만 있으면 인식율을 크게 높힐 수 있음을 알 수 있다.

Ⅳ.4 잡음과 전송로의 영향이 모두 있는 경우의 개선 실험

앞에서 살펴본 잡음과 전송로의 왜곡이 모두 있는 경우에 대하여 실험을 하였는데, 여기에서는 앞에서 상대적으로 더 큰 영향을 주는 G2D2 전송로의 경우에 대해서만 실험을 하였다. 실험은 먼저 음성에 잡음을 더한 후, 이것을 전송로에 통과시키는 방식으로 하였다. 그 결과는 표 8에서 보는 것처럼 7.7~26.4%의 인식율을 보여서 인식을 거의 못하였다고 할 수 있다. 이에 대한 개선을 위하여 먼저 spectral subtraction을 사용하여 잡음을 제거시키고, 전송로의 왜곡을 보상하기 위한 training에는 표 7에서 (3)번의 것을 사용하여서 65.9~69.2%의 좋은 개선율을 얻었다.

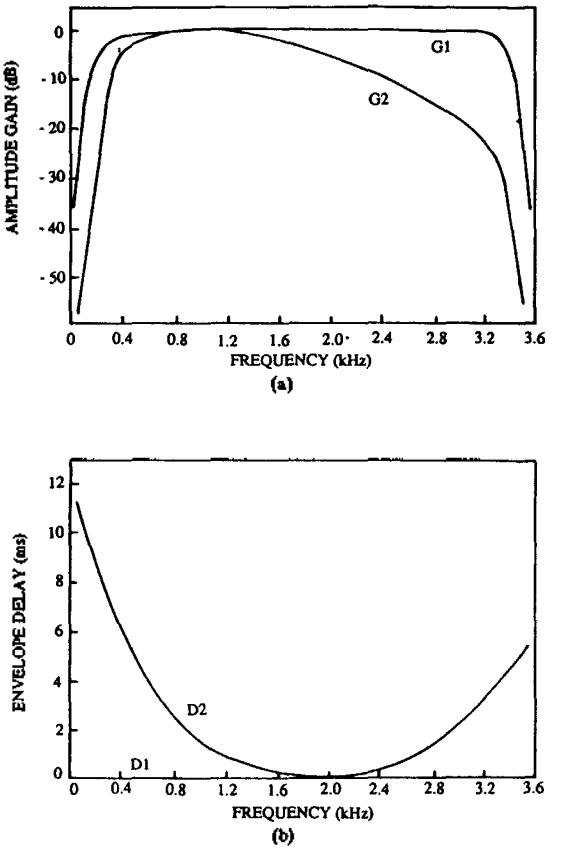


그림 6. 실험에서 사용된 전화선로의 특성
(a) 진폭 왜곡 (b) 위상 왜곡

표 5. 대역폭이 제한된 음성의 인식율

대역폭 (kHz)	2.5	3.0	3.5	4.0
인식율 (%)	50.9	62.5	79.8	75.9

표 6. 전송로에 의하여 왜곡된 음성의 인식율
(괄호안은 개선율) (단위:%)
(a) 개선 전의 인식 결과
(b) 개선 후의 인식 결과

전송로 종류	G1D1	G1D1	D2D1	G2D2
인식율 (%)	70.3	67.3	31.5	29.5

전송로 Train	(1)	(2)	(3)	(4)	(5)
G1D2	89.1 (21.8)	88.4 (21.1)	88.9 (21.6)	88.4 (21.1)	88.7 (21.4)
G2D2	82.7 (53.2)	82.4 (52.9)	82.4 (52.9)	82.5 (53.0)	83.6 (54.1)

표 7. Training 데이타로 사용된 단어

Train	사용된 단어				
(1)	책	갈등	형태	실천	얘기
(2)	확대	잔치	문화	색깔	예법
(3)	책 뿔	갈등 이	형태 잔치	실천 너	얘기 때
(4)	안방 문화	북 넣다	쾌적 색깔	확대 야권	정표 예법
(5)	책 뿔 안방 문화	갈등 이 북 넣다	형태 잔치 쾌적 색깔	실천 너 확대 야권	얘기 때 정표 예법

표 8. 잡음과 전송로의 영향이 동시에 있는 경우의 인식율 (괄호안은 개선율) (단위:%)

분류 SNR	15 dB	20 dB	25 dB
개선 전	7.7	13.5	26.4
개선 후	76.2 (68.5)	82.7 (69.2)	92.3 (65.9)

V. 결 론

본 논문에서는 전화선로에서의 음성인식을 위하여, 잡음과 전송로가 음성인식에 미치는 영향에 대하여 알아보고, 전처리 단계의 추가에 의한 이의 개선 방법을 연구하였다. 컴퓨터 모의 실험은 음소적으로 고르게 분포되어 있는 한국어 고립단어 100개를 인식대상 어휘로 하고 이를 각각 10회씩 발음한 1000개의 데이타에 대해서 화자종속 시스템으로 수행하였다.

잡음이 첨가되면 인식율이 급격히 저하되는데, 이의 영향을 개선하기 위해 spectral subtraction 방법을 사용하여서, white noise에 의해 SNR이 15dB일 때 [illegible] 수의 제한과 선로 왜곡는 인식율을 크게 저하시키지만, 위상의 왜곡은 인식율에 큰 영향을 미치지 않았다. 전송로의 영향을 개선하기 위하여 별개의 training 데이타를 받아서 이에 알맞게 기준패턴을 변경시켰다. 이 방법에 의하여 21.1~54.1%의 인식율의 향상을 얻었는데, 이때 training 데이타의 갯수에는 개선율의 차이가 별로 없었다. 또, 잡음과 전송로의 왜곡이 동시에 있을 경우에는 인식율이 7.7~26.4% 정도로 나바졌는데, 앞에서 제안한 방법을 사용하여 76.2~92.3%의 인식율을 얻었다.

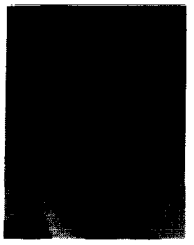
본 논문에서는 인식 algorithm은 변경시키지 않고 적절한 선처리 과정을 추가하여 인식율을 향상시키는 방법을 사용하였는데, 좀더 좋은 인식율을 얻기 위해서는 인식 algorithm에 대한 연구도 추가되어야 한다. 특히 끝점검출 부분의 개선이 중요하다. 또, 본논문에서 고려하지 않은 여러가지 왜곡의 영향에 대해서도 계속 연구가 필요하며, 실제 전화에서의 실험도 하여야 할 것이다.

참 고 문 헌

1. F.K. Soong and M.M. Sondhi, "A frequency-weighted itakura spectral distortion measure and its application to speech recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol.36, pp.41-48, January 1988.
2. B.A. Hanson and H.Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processisng*, Vol. ASSP-35, pp.968-973, July 1987.
3. J.G. Wilpon, L.R. Rabiner, and T.Martin,"An improved word-detectrion algorithm for telephone-quality speech incorporating both syntactic and semanctic constraints," *AT&T Bell Lab. Tech. J.*, Vol.63, pp.479-498, March 1984.
4. J.S. Lim ed., *Speech Enhancement*, Englewood Cliffs, NJ:Prentice-Hall, 1983.
5. J.G. Wilpon and L.R. Rabiner, "On the recognition of isolated digits from a large telephone customer population," *Bell Syst. Tech. J.*, Vol. 62, pp. 1977-2000, September 1983.
6. M.B. Carey et al., "1982/83 end office connection study : Analog voice and voiceband data transmission performance characterization of the public switched network," *AT&T Bell Lab. Tech. J.*, Vol. 63, pp. 2059-21119, November 1984.
7. S.F. Boll, "Supression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. ASSP-27, pp. 113-1

20, April 1979.
8. R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, pp. 137-145, April 1980.
9. 도심주, "전화 음성의 격리 단어 인식에 관한 연구," CRL-T-9015, 한국과학기술원, February 1990.

▲都 三 周(정회원) 1965년11월23일생



1988년 2 월 : 서울대 전자공학과 졸업(BS)
1990년 2 월 : 한국과학 기술원 전기및 전자공학과 졸업(MS)
1990년 3 월 – 현재 : 한국 전기통신공사 연구개발단 음성처리연구실 전임연구원

▲殷 鍾 官(정회원) 1940년 8 월25일생

1964년 : 미국University of Delaware전자공학 학사
1966년 : 동대학원 전자공학석사
1969년 : 동대학원 전자공학박사
1969년 : 미국 University of Maine조교수
1973년~1977년 : Stanford연구소(SRI) 책임연구원
1987년~1989년 : 한국음향학회 회장
1977년~현재 : 한국과학기술원 전기 및 전자과교수