論文 91-28B-7-6

DHMM과 어휘 해석을 이용한 Voice Dialing 시스템

(The Voice Dialing System Using Dynamic Hidden Markov Models and Lexical Analysis)

崔 成 好* 李 康 誠** 金 淳 協**

(Seong Ho Choi, Gang Sung Lee, and Soon Hyob Kim)

要 約

본 논문은 DHMM (Dynamic Hidden Markov Model)과 어휘 해석을 이용한 한국어 연속 숫자음 인식에 관한 것으로 Voice Dialing 시스템 개발을 위한 기초 연구이다. 본 시스템은 세그멘테이션부, 표준음성작성부, 인식부, 어휘해석부로 구분된다.

세그멘테이션부는 0차 LPC cepstrum의 시간 변화 파라미터, 유성음 검출 파라미터 A_i, ZCR등을 이용해 세그멘테이션하고, 표준음성 작성부는 19개의 음소 또는 음절을 DHMM으로 학습시켜 표준음성을 작성하고, 인식부에서는 Viterbi algorithm으로 인식하여 최종적으로 어휘해석부에서 연속 숫자음을 인식하였다.

본 실험은 잡음이 있는 연구실에서 20대 남성 화자 10명이 21종의 7연속 숫자음을 7회 발성한 자료를 사용하여 85.1%의 인식율을 보였다.

Abstract

In this paper, Korean spoken continuous digits are recognized using DHMM(Dynamic Hidden Markov Model) and lexical analysis to provide the base of developing voice dialing system.

After segmentation by phoneme unit, it is recognized. This system can be divided into the segmentation section, the design of standard speech section, the recognition section, and the lexical analysis section.

In the segmentation section, it is segmented using the ZCR, 0 order LPC cepstrum, and Ai, parameter of voice speech dectaction, which is changed according to time. In the standard speech design section, 19 phonemes or syllables are trained by DHMM and designed as a standard speech. In the recognition section, phomeme stream are recognized by the Viterbi algorithm. In the lexical decoder section, finally recognized continuous digits are outputed.

This experiment showed the recognition rate of 85.1% using data spoken 7 times of 21 classes of 7 continuous digits which are combinated all of the occurence, spoken by 10 man.

*準會員, 光云大學校 電子計算機工學科
(Dept. of Computer Eng., Kwangwoon Univ.)

(※ 본 연구는 과학재단 목적기초 연구 지원으로 수행된 연구의 일부분임.)

**正會員, 光云大學校 電子計算機工學科
(Dept. of Computer Eng., Kwangwoon Univ.)

接受日字 : 1991年 4月 2日

Ⅰ. 서 론

인간과 기계가 통신할 수 있는 방법은 여러가지가 있는데 그 중에서 음성처럼 가치가 높고 자연스러운 방법은 없다. 이로 인해 선진각국에서는 자동 음성 인식에 지대한 관심을 가지고 많은 연구를 하고 있다. 또한, 우리 나라에서도 한국어는 한국인만이 그 특성을 알고 해결 해 나갈 수 밖에 없기 때문에 국내에서도 음성 자동 인식에 많은 연구를 하고있다. 이에 대한 일환으로, 본 논문은 종합 통신망 서어비스 중 하나인 음성 다이얼링 시스템 구축을 위한 한국어 연속 숫자음 인식을 목적으로 한다.

음성 인식은 인식되는 음성의 종류에 따라 격리단어 인식, 연결단어 인식 및 연속음성 인식으로 나눈다. 이에, 연속음성 인식은 연속된 음성을 음소, 음절 등 음성의 기본 단위들로 부터 인식하는 것을 말한다. 따라서, 연속 음성 인식을 위해서는 발음된 음성을 기본단위 별로 나누고(segmentation) 이들을 식별한 뒤, 최종적으로 언어학적인 분석(linguistic analysis)를 통해 의미를 찾아 내는 절차를 통해야 한다.

연속 숫자음 인식에서는 인식해야할 어휘는 대어휘이지만 숫자음에 대한 음성의 수는 적기 때문에 단어 단위의 인식보다는 부단어(sub-word unit)를 인식의 기본 단위로 사용한다.

음성 인식 방법은 인식기법에 따라 정합(pattern maching)에 의한 방법과, 확률적인 방법으로 크게 구분한다.

확률적인 기법은, 최근에 많은 시도가 있었고, 많은 진전이 있었던 HMM을 이용하는 방법이다. 이 기법은 각 기준 음성에 대한 HMM을 만든 후, 이들 모델들로부터 입력 음성의 관측 확률을 구하여 가장 높은 관측 확률을 가지는 모델을 인식 음성으로 하는 것이다.

본 실험에서는 음소 단위의 세그멘테이션 후 인식하기 때문에 음소를 표준음성으로 하였고, 또한 세그멘테이션의 오류를 고려하여 몇개의 음절도 표준 음성에 포함 시켰다.

전체의 시스템 구성을 살펴보면, 세그멘테이션부, 표준 음성작성부, 인식부, 어휘해석부로 구분된다.

세그멘테이션부에서는 0차 LPC cepstrum의 시간 변화 파라미터, 유성음 검출 파라미터 A_t, ZCR등을 이용하여 세그멘테이션을 하고, 표준음성 작성부는 19개의 음소 혹은 음절을 Baum-Welch의 재추정 알고리듬으로 훈련시켜 DHMM 표준 음성을 모델링하였다.

인식부에서는 세그멘테이션된 음소와 DHMM의 Viterbi scoring으로 인식하였으며, 어휘 해석부에서 음소 병합과 음성 구간 검출및 음절 수를 조절하여 최종 인식하였다.

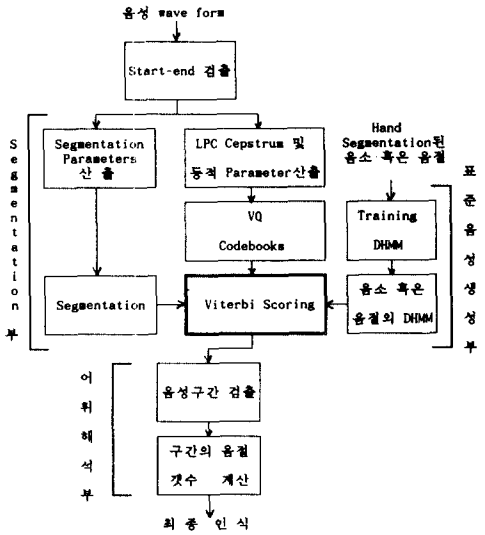


그림 1. 전체 시스템 구성도
Fig. 1. Diagram of the system.

Ⅱ. 특징 벡터 추출

본 연구에서는 시작점 끝점을 검출하는 파라미터로 peak-to-peak를 사용하였다. 이것은 log energy보다 계산이 간단하고 그 분석 능력도 떨어지지 않으므로, 이 파라미터를 선정하여, Rabiner 및 Samber의 끝점 검출 방법을 이용하였다. 시작점 및 끝점은 음성 정보의 손실없이 음성 구간의 앞뒤 묵음 구간을 모두 포함하여록 여유있게 설정하였다.

특징 파라미터는 다음과 같은 절차로 구해진다.

1) 3.5KHz LPF
2) 8KHz Sampling, 12bits
3) Pre-emphasis $H(z) = 1 - 0.95z^{-1}$
4) Hamming Window (16ms frame, 8ms displacement)
5) LPC analysis of order 10
6) LPC cepstrum analysis
7) Dynamic parameter

음성의 특징 파라미터를 분석할 때 스펙트럼의 변화를 측정함으로써 능동적인 변화를 측정하여 음성 특징에 포함시켰다. 이러한 특징을 동적특징이라하는데, 이러한 동적 특징 기울기를 측정하는 회기계수는 다음과 같다[1].

$$Di(t)=\frac{\sum_{n=-k}^{K} n[Ci(t+n)]}{\sum_{n=-K}^{K} n^2} \quad (1)$$

여기서, Ci(t)는 발성음의 t번째 프레임의 i번째 계수이고, D는 회기계수이다. 기울기는 -k부터 k까지 2k+1프레임을 측정한다.

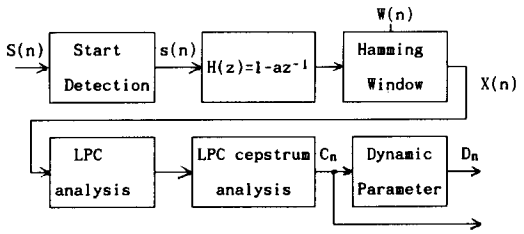


그림 2. 특징 벡터 추출 과정
Fig. 2. Procedure of feature extraction.

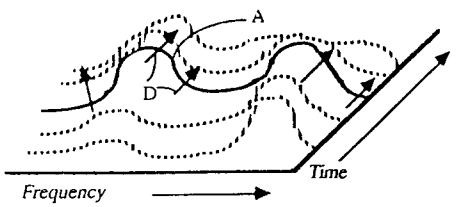


그림 3. 정적 스펙트럼과 동적 스펙트럼
Fig. 3. Instantaneous spectrum and dynamic spectrum.

Ⅲ. VQ와 HMM

1. VQ

VQ(Vector Quantization)은 연속 또는 많은 수의 불연속적인 양을 갖는 원소로 이루어진 vector를 어떤 정해진 distortion measure를 사용하여 적은 수의 불연속적인 양을 갖는 원소로 이루어진 vector로 근사시키는 것을 말한다. 이러한 VQ을 본 논문에서는 이산 HMM의 유한개 관측 심벌의 집합으로 구성된 입력을 제공하기 위해 전처리로써 사용한다.

코드북은 훈련 데이타의 특성이 잘 나타나도록 만들어야 하는데, 본 논문에서는 K-means 알고리듬을 이용하여 LPC cepstrum과 dynamic 계수를 바탕으로 하여 각각의 코드북을 만들었다.

K-means알고리듬은 cluster center수(K), 초기 cluster값 샘플들이 정해지는 순서, 데이타의 기하학적 성질등에 의해 영향을 받아 실행될 때 마다 동일한 결과가 얻어지지 않는다는 단점이 있으나, 계산이 간편하며 다른 방법에 비해 성능이 떨어지지 않는다[2].

벡터 양자화 과정에서, 코드북 작성을 하기위해서는 또는, 코드북의 codeword와 입력 벡터의 유사도를 측정하기 위해서 거리 계산이 필요한데, 본 연구에서는 식(2)의 거리식을 사용하였다.

$$d(x,y)=w(c_{x0}-c_{y0})^2+\sum_{i=1}^{P}(c_{xi}-c_{yi})^2 \quad (2)$$

여기서, x와 y는 두개의 벡터를 나타내고 w는 무게 상수이며, c_i는 LPC cepstrum계수를 나타내고, p는 차수이다.

2. HMM

음성 신호는 짧은 구간(10−30ms)내에서 안정된 상태를 유지하고, 시간에 따라 점차 변화하는 quasi-stationary process라고 할 수 있다. 이러한 음성 특성을 내포하도록 모델링할 수 있는 적합한 방법이 HMM인데, 이것은 확률적 구조에서 안정된 구간, 특이한 구간, 그리고 구간에 따라 연속적으로 변화하는 성질을 묘사하는데 매우 효율적이다. 이러한 HMM은 천이에 의해 연결된 상태들의 모임으로써 두 확률 즉, 천이확률과 출력 밀도함수를 가진다[2].

$\{S\}$ 상태의 집합, S_I 초기 상태, S_F 최종상태

$A=\{a_{ij}\}$ 천이의 집합, 여기서 a_{ij}은 i상태에서 j상태로 천이할 확률

$B=\{b_{ij}(k)\}$ 출력확률 메트릭스, i 상태에서 j 상태로 천이시 k심볼이 나올 확률

$$\lambda=(A,B,\pi) \quad (3)$$

a, b는 다음과 같다.

$$a_{ij}=P(X_{t+1}=j \mid X_t=i) \quad (4)$$

$$b_{ij}(k)=P(Y_t=k \mid X_t=i, X_{t+1}=j) \quad (5)$$

여기서, $X_t=i$은 Markov Chain이 시간 t에서 상태 i에 있고, $Y_t=k$,는 시간 t에서의 출력 심볼이 k라는 것을 의미한다.

3. HMM의 기본적인 문제

HMM을 정의하는데 세가지 기본적인 문제점을 해결하여야 한다. 그 첫째는 평가 문제인데 이는 forward 알고리듬을 이용한다. 두번째는 관측열이 주어

졌을때 최적인 상태열을 선책하는 decoding 문제인데 이는 Viterbi 알고리듬을 이용한다. 세번째 문제는 학습에 관한 문제로써 최적의 모델링을 하기위해 각 파라미터들을 조정하는 것인데 이는, Baum-Welch의 재평가 알고리듬으로 해결한다[2].

또한, HMM을 실제 모델링하는데 또 다른 몇가지 문제점이 야기된다. 첫째로 HMM의 파라미터를 어떻게 초기화할 것인가라는 문제이고,

둘째는 T가 증가 함에 따라 이들의 값이 지수함수적으로 감소하여 곧 underflow를 발생시키는 점인데, scaling값들을 변수의 값들이 컴퓨터의 dynamic range에 있도록 한다.

세째로 훈련시키는 데이타의 양이 충분히 많지 않을 경우 훈련과정에서 어떤 심볼이 모델의 어느 상태에는 나타나지 않는 것으로 추정되었더라도 시험과정중에서 이 심볼이 나타나는 경우이다. 이에 대한, 평활화법은 확률값들을 평활화함으로써 훈련되지 않은 심볼이라도 나타날 가능성을 고려하는 것이다.

5. DHMM

DHMM (Dynamic Hidden Markov Models)은 정적 스펙트럼의 특징 파라미터와 동적 스펙트럼의 특징 파라미터를 함께 모델링한 것이다. 정규화된 log-power (P), 그것의 동적 특징(PD), 정적 특징(A), 동적 특징(D)사이의 상관관계가 절대치를 보여주며, A와 D 사이의 상관관계가 A와 A나 D와 D의 상관관계 보다도 매우 작다는 사실을 이용하여 동적 스펙트럼의 특징을 바탕으로 DHMM을 커다란 계산이나 메모리의 증가없이 모델링 할 수 있다.

여기서 상태 S와 관찰 심볼 O_t가 다음과 같다고 할 때

$$S=(M_I, M_D) \tag{6}$$

$$O_t=(O_{It}, O_{Dt}) \tag{7}$$

상태 S에서 관찰 심볼 O_t가 나올 확률은 다음과 같다.

$$P(O_t \mid S)=P(O_{It}, O_{It} \mid S) \tag{8}$$

정적 특징(O_{It})과 동적 특징(O_{Dt})는 서로 상관관계가 거의 없으므로 다음식과 같이 쓸 수 있다.

$$\begin{aligned} P(O_{It}, O_{It} \mid S) &\cong P(O_{It}, \mid S)P(O_{Dt} \mid S) \\ &\cong P(O_{It} \mid M_I, M_D)P(O_{Dt} \mid M_I, M_D) \end{aligned} \tag{9}$$

또한, $S=(M_I, M_D)$이고 이것은 서로 독립이므로

$$\cong P(O_I \mid M_I)P(O_D \mid M_D) \tag{10}$$

로 표현할 수 있다. 따라서,

$$\begin{aligned} P(O_t \mid S) &\cong P(O_{It} \mid M_I)P(O_{Dt} \mid M_D) \\ &= b_s{}^I(O_{It})\, b_s{}^D(O_{Dt}) \end{aligned} \tag{11}$$

로 된다.

여기서, $b_s{}^I(O_{It})$는 상태 S에서 정적 벡터 O_{It}가 나올 확률이며, $b_s{}^D(D_{Dt})$는 상태 S에서 동적벡터 O_{Dt}가 나올 확률이다[3][4].

Ⅳ. 음소 단위의 세그멘테이션

1. 세그멘테이션 파라미터

유성음 검출 파라미터는 유성음을 검출하여 음성신호를 유성음 구간과 무성음 구간으로 구분하기 위해 유성음 검출 파라미터를 이용한다. 일반적으로 유성음은 저주파 부분에 에너지가 밀집되고 무성음은 고주파 영역에 에너지를 많이 포함하고 있으므로, 저주파 부분의 에너지를 추출하면 유성음과 무성음의 구분이 가능하다.

영차 LPC cepstrum 계수는 유성음 구간에서 모음과 모음, 모음과 비음 또는 모음과 유음이 연이어 발음될 때 유성음 분리 작업해야 할것이다. 따라서, 저주파 영역에 대부분의 언어 정보가 많이 포함되어 있는 점을 고려해 스펙트럼의 저역 부분에 가중치를 둔, 영차 LPC cepstrum 계수를 이용한다.

i번째 프레임의 영차 LPC cepstrum 계수 Ci는

$$Ci=xi[0]=\frac{1}{2\pi}\int_{-\pi}^{\pi} Xi(\Omega)\ W(\Omega)\ d\Omega \tag{12}$$

영차 LPC cepstrum 시간 변화 파라미터는 유성음 구간에서의 음소경계의 검출을 위해 영차 LPC cepstrum 시간 변화 파라미터를 사용한다. i번째 분석 프레임을 중심으로 하는 영차 LPC cepstrum계수의 시계열 $C_{i+n}(\mid n \mid < M)$에 대해서 직선식을 적용해 회귀 계수 Ai에 의해서 영차 cepstrum

$$x=A_in+B \tag{13}$$

계수의 시간 변화의 양을 표현하는데, 직선 적용 식에서 시계열구간 양단의 절단 영향을 작게 하기 위해 평가 함수인 가중된 2승 평균 오차를 이용한다.

$$\varepsilon=\frac{1}{2M+1}\sum_{n=M}^{M} W_n(A_in+B-C_{i+n})^2 \tag{14}$$

여기서, Wn은 $\mid n \mid > M$에서 0이 되는 우함수의 window 함수이다. 직선 적용의 계수 Ai는 가중된 2승 평균 오차 ε를 최소로 하는 조건에서 A_i에 대해 1차편미분하면 아래와 같이된다.

$$A_i = K_M \sum_{n=-M}^{M} W_n \, n \, C_{i+n}, \quad [K_M = (\sum_{n=-M}^{M} W_n \, n^2)^{-1}] \quad (15)$$

ZCR 파라미터는 스펙트럼에서 에너지가 집중되는 주파수를 찾는데 유용한 특징 파라미터로 사용되는 영교차율을, 무성음 구간에서의 유성자음과 무성자음 그리고 묵음을 검출하기 위해 사용한다.

2. 세그멘테이션 알고리듬 구현

유성음 검출 파라미터의 대소에 의해 음성신호를 크게 유성음구간과 무성음 구간으로 구분한 다음, 2분된 각각의 구간에 대해 더욱 자세한 음소 단위의 경계들로 분리한뒤, 영차 LPC cepstrum 계수의 시간변화 파라미터가 극대 및 극소로 표현되어 유성음 구간에서 음소 단위 경계 후보를 구하고, 영교차율 값에 의해 무성음 구간에서 음소 단위 경계 후보를 구한다. 이에 구분된 음소단위 경계후보는 음소 합병을 해서 최종적으로 음소구분화 작업을 마친다.

유성음과 무성음의 판별은 유성음 검출 파라미터에서 두개의 임계값($T_{VL}=57$, $T_{VH}=62$)을 가지고 Samber와 Rabiner의 끝점 검출 알고리듬 방식을 적용하여 유성음과 무성음을 구분하였다.

유성음 구간의 세그멘테이션은 유성음 구간에 있어서의 음소의 경계 후보의 대부분은 영차 cepstrum 시간 변화 파라미터 a_i가 극대치 또는 극소치를 취하는 분석 프레임에 의해서 주어진다.

우선 영차 cepstrum 시간 변화 파라미터 a_i의 극치를 표시하는 함수 A_i는 다음과 같이 표시된다.

$$\begin{aligned} &\text{if } ((a_i > 0 \text{ and } a_i \geqq a_{i+1}) \text{ or} \quad (16)\\ &\quad (a_i < 0 \text{ and } a_{i-1} > a_i \leqq a_{i+1}))\\ &\quad \text{then } A_i = a_i, \text{ else } A_i = 0 \end{aligned}$$

무성음 구간의 세그멘테이션은 무성음을 유음 구간과 무음 구간으로 구분하고, 유음 구간에서 다시 평가하여 작은 유성 자음 구간과 무성 자음 구간으로 분류한다.

앞에서 언급한 파라미터를 가지고 세그멘테이션한 결과는 그림4와 같다.

V. 표준음성 작성과 음소인식 및 어휘해석

1. 표준음성 작성

본 연구에서 연속 숫자음 인식의 경우에 인식 해야할 대상 어휘수는 대어휘인 경우에 반하여 숫자음이라는 것에 제한하기 때문에 음성의 수는 대폭 줄어든다[7]. 이에 단독 숫자음에서의 음성 종류를 살펴보면 다음과 같다.

단독 숫자음에서 음성을 IPA (International Phoneme Alphabet)로 표현하면 아래와 같이 나타난다.

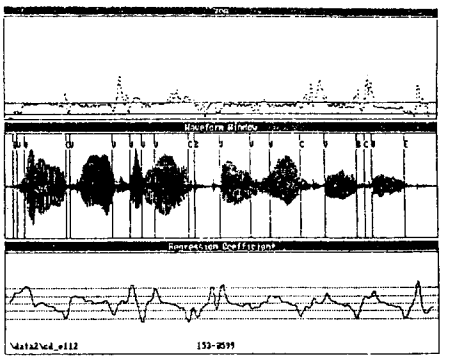


그림 4. /153-0599/연속숫자음의 segmentation 결과
Fig. 4. Result of segmentation of continuous digits /153-0599/.

공[k o n],	일[i l],	이[i:],
삼[s a m],	사[s a],	오[o:] ,
육[yu k],	칠[c ^ h i l],	팔[p h a l],
구[k u],	에[e]	

연속 숫자음에서 소리값(음성)은 나타나는 자리 즉, 환경에 따라 몇 가지 조금씩 다른 음성으로 나타난다.

/1/의 후속모음이 단모음일 경우 /r/ (예, 칠오)
/s/이 앞 음소가 /1/또는 /k/일 경우에는 /s'/ (예, 팔삼, 육삼)
/k/가 유성음 사이에 올때는 /g/ (예, 오공)
/k/의 후속모음이 중모음일 경우 /n/ (예, 육육)

이들 음성 중에서 /s'/는 /s/와 음성적 특징이 거의 유사하다는 가정하에 통합하였으며, /k/는 종성 내파음으로써 음성적 특징이 나타나지 않으므로 제외 시켰고, /r/, /g/는 추출한 정보의 양이 적고 충분한 통계적 자료가 없어 이들 또한 제외시켰다[7].

본 연구에서는, 음소 단위로 세그멘테이션을 한 다음, 인식을 하기 때문에 음소를 표준 음성으로 모델링 하였는데, 세그멘테이션시 오류가 포함된 경우를 고려해 아래와 같이 몇개의 음절도 표준음성에 포함시켰다.

a, i, u, e, o, yu, l, m, n , k,
c^h, p^h, s, r, i:, o:, il, am, al

각 음소 또는 음절당 15-20개 정도의 음소를 hand 세그멘테이션으로 추출하여 Baum-Welch의 재추정 알고리듬으로 학습시켜 DHMM을 작성하였다.

2. 음소 인식

설정된 표준 음성인 각 음소 혹은 음절에 대해 15-20개을 hand-segmentation 방법으로 D/A을 통해 확인을 거쳐 추출한 것을 10차 LPC ceptsrum으로 변환시킨 다음, 동적 파라미터를 구해 Baum-Welch의 재추정 알고리듬을 이용하여 동적HMM 학습시킨다.

이러한 DHMM의 state수는 3개로 하면 그림과 같이 표현된다.

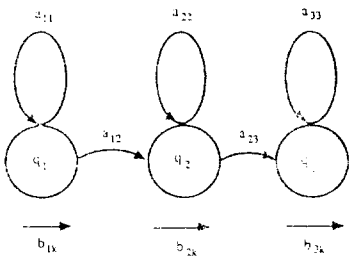


그림 5. 동적 모델 예
Fig. 5. Excemple of dynamic model.

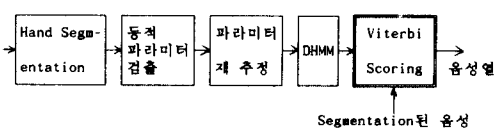


그림 6. 음소 인식 절차
Fig. 6. Procedure phoneme recognition.

Segmentation되어진 음소와 훈련된 DHMM과의 Viterbi 알고리듬으로 인식된 음소열을 출력시킨다.

3. 어휘 해석

인식부에서 출력된 음소열이 입력되면 언어어 동일한 음성으로 연결되어 졌을 경우 이 음성을 병합하여 한 개의 음성단위로 바꾸어 주고, 음성구간 검출을 하여 각구간에 음절의 갯수를 계산하여 전체 음절의 갯수를 조절하여 최종적으로 단어열을 출력해 낸다.

음소 병합은 음소열 상에서 앞뒤로 같은 음성인 경우 이들을 한개의 음성으로 합한다.

음성구간 검출은 음성 구간 검출은 다음과 같은 2가지 형태에 의해 분류 한다.

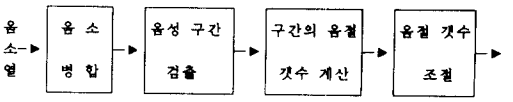


그림 7. 어휘 해석 절차
Fig. 7. Procedure of lexical decoding.

i]유성음+유성자음+무성음
→(유성음+유성자음)+무성음
ii] 유성음+유성자음+유성음
→유성음+(유성자음+유성음)

첫번째 형태 유성음+유성자음+유성음의 순서가 될 경우 유성자음은 앞 유성음의 작은 소리로 간주하여 앞 유성음으로 연결시킨다. 두번째 형태는 유성음+유성자음+유성자음으로 나열 될 경우 중간의 유성자음은 뒤 유성음의 초성으로 간주하여 뒷유성음으로 음성구간을 구분하여 검출한다.

구간의 음절 갯수 계산은 앞단계에서 각 음성 구간을 설정하였는데, 이들 음성구간 내의 음절 갯수를 산출해 낸다. 설정된 각음성구간의 음성 길이에 평균 음절 길이로 나누어 몫을 최소 가능 음절의 수로 설정하고, 소숫점을첫자리가 *.4-*.6 사이의 수일경우 최소 가능 음절의 수에 1을 더해 최대 가능 음절수의 갯수를 계산한다.

음절 갯수 조정은 인식해야 할 단어열은 [e]를 포함하여 총 8개가 된다. 그러나 앞단계에서 계산된 음절의 수가 8개 미만이거나 8개를 넘을 경우 음절의 갯수를 조정하게 되는데 조절하는 알고리듬은 아래 같다.

```
if(I< 8){                                          (17)
 if(I_maxsum =8)      maxsum
   {I_min : = I_max;
    exit;}
 else if (I_maxsum > 8)
    {diff : I-8)
     최적 차이 갯수 선택
     I_max조정}
 else {diff : =8-I;
     최적 차이 갯수 선택
     I_min조정}
 }
 if(I>8) {
      diff : =I-8;
      최적 차이 갯수 선택
      I조정}
```

여기서, I은 현재의 총 음절 수

I_{maxsum}은 최대 음절의 갯수

I_{min}은 각 구분된 음성 구간의 최소 가능음절수

I_{max}은 각 구분된 음성 구간의 최대 가능음절수

단, 최적 차이 갯수 선택은 유성 자음, 무성 자음, 묵음을 제외한 모음의 길이로만 비교 선정한다.

단어열 출력은 최종적으로 구분된 음성 구간의 음절 갯수가 정해지면, 숫자음 열을 출력해 내는데 그 절차는 아래와 같다.

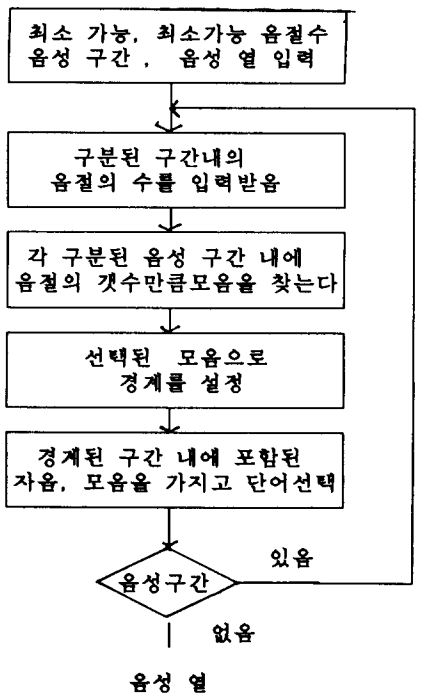


그림 8. 단어열 생성 절차
Fig. 8. Procedure of word string production.

Ⅵ. 실험 결과 및 고찰

1. Simulation 환경

실험에서 사용한 데이타는 연속해서 일어날 수 있는 모든 경우의 수를 조합하여 구성한 21개의 데이타 목록을 표 1.과 같이 만들어 약 4.2음절/초의 발성속도로 컴퓨터 잡음이 있는 연구실 환경에서 마이크폰을 통하여 입력 받아 3.5kHz 저역 여파기(LPF)을 통과하여 8kHz로 샘플링하여 12bits로 A/D 변환

표 1. 선정된 연속 숫자음 데이타 목록
Table 1. Data list of chose continuous digits.

번호	국번	전화번호
1	5 1 2	0 2 5 7
2	6 3 0	1 3 4 9
3	7 4 5	6 7 8 0
4	8 2 6	9 3 1 8
5	9 0 4	0 3 7 1
6	9 1 0	2 3 8 8
7	8 4 3	4 6 1 6
8	7 2 9	5 5 2 2
9	6 0 7	7 6 4 1
10	3 5 8	8 7 3 6
11	1 5 3	0 5 9 9

번호	국번	전화번호
12	2 7 0	9 4 8 3
13	3 9 6	0 0 1 1
14	4 0 8	6 2 8 1
15	6 8 9	6 5 4 2
16	2 0 9	1 9 2 1
17	1 4 7	3 3 2 4
18	9 8 6	5 0 6 6
19	5 6 9	1 7 7 5
20	7 9 5	9 7 8 5
21	4 4 8	1 2 3 4

시킨 디지탈 데이타로 구성하였다. 이 데이타를 가지고 16ms를 한 프레임으로 8ms씩 이동하면서 10차로 LPC분석 하였는데, 여기서 각 프레임 마다 Hamming window를 씌웠다. 그 후 10차의 LPC계수를 LPC cepstrum으로 변환하고, 이를 이용하여 동적 파라미터를 추출한 후, LPC cepstrum과 동적 파라미터에 대해 k-means 알고리듬을 이용해 각각에 대한 코드북을 작성하였다.

2. 결과 및 고찰

세그멘테이션없이 단어열을 가지고 인식 할려면 많은 양의 학습데이타, 학습 시간, 그리고 기억용량을 필요로 하지만, 음소단위로 세그멘테이션을 한뒤 음소 혹은 음절로 인식하였으므로 기억용량이 적게 들고 시간이 적게 걸리게 된다.

잡음이 있는 연구실 환경에서 자료을 입력 받았으므로 잡음이 포함된 음성을 세그멘테이션하기 때문에 세그멘테이션시 오류가 다소 발생 한것 같고, 이렇게 잘못 구분화가 이루어 졌을 경우에 음소인식에도 오류가 발생하였다. 또한 잘못 인식된 음소를 가지고 어휘 해석을 할 경우, 단어열 해석에도 영향을 끼쳐 오인율이 높았다. 또한 일정하지 못한 발성 속도로 인식률에 영향을 미쳤다.

Viterbi algorithm을 이용한 음소 인식률은 73.4% 이였고, 어휘 해석부의 음성구간 검출율은 97.6% 이며, 168개(21종의 숫자열 *8음절) 중 25개의 음절이 잘못 인식되어 85.1%의 인식률을 보였다.

표2. 결과에서와 같이, 앞부분의 오인식이 높았는데 이는 앞부분 발성시 발성 속도가 일정하지 못한것에 기인한것 같고, 뒷부분이 비교적 잘 인식되었는데 이는 일정한 간격으로 발성한것 때문으로 사료된다.

표 2. 오인식 결과
Table 2. Result of error.

오인식결과	수정사항	오인식결과	수정사항
630에 1345	5- >9	780에 5522	8- >2, 0- >9
7755 6780	5- >에	100에 9483	1- >2, 0-7
881에 9318	8- >2, 1- >6	6899 6542	9- >에
903에 0371	4- >3	107에 3324	9- >에
800에 0716	0- >4, 0- >3	506에 1775	0-6
	0- >4, 7- >6	440에 1234	0- >8
912에 0667	9- >5, 6- >2	630에 1345	5- >9
	6- >5	919에 2388	9- >0
190에 0599	9- >5, 0- >3	391에 0011	1- >6

Ⅶ. 결 론

본 논문은 모든 경우의 수를 조합한 21개의 한국어 7연속 숫자음을 DHMM과 어휘 해석을 통하여 인식하였다.

음소 단위로 세그멘테이션해서 인식을 하는데 0차 LPC cepstrum의 시간 변화 파라미터, 유성음 검출 파라미터 A_1을 이용해 세그멘테이션하고, 19개의음소 또는 음절을 Baum-Welch의 재추정 알고리듬으로 학습시켜 표준음성을 작성하였으며, 음소 인식시에는 세그먼트된 음소와 DHMM과의 Viterbi scoring통해 음소를 인식하여 73.4% 인식률을 보였고, 이렇게 인식한 음소열을 가지고 어휘 해석하여 연속 숫자임을 인식한 결과 85.1%의 인식률을 얻었다.

본 실험은 음성 다이얼링 시스템 구축을 목적으로 하는 것이나, 신원조회, Audio Response 시스템, 장애자 시스템등에서도 응용하여 널리 사용할 수 있을 것으로 생각된다.

앞으로도, 숫자음 뿐 아니라 구문론, 의미론적인 해석을 고려한 대어휘 연속음 인식에도 관심을 가지고 연구를 해야 할것이며, 인식률 향상에도 많은 개선이 있어야 할 것으로 사료되어진다.

參 考 文 獻

[1] Soong, F.K., Rosenberg, A.E. , "On the Use of Instaneous and Transitional Spectral Information in Speaker Recognition," ICASSP-86, vol. 2, pp. 17.5.1-17.5.4. Tokyo, 1986.

[2] Masafumi Nishimura, "HMM-Based Speech Recognition Using Dynamic Spectral Feature," ICASSP-89, vol. 1, pp. 298-301, 1989.

[3] Kai-Fu Lee, Automatic Speech Recognition the Development of the SPHINX System, Kluwer Academic Publishers, 1989.

[4] F. Jelink, "Continuous speech recognition by statistical methods," *Proc. IEEE,* vol. 64, pp. 532-536, Apr. 1976.

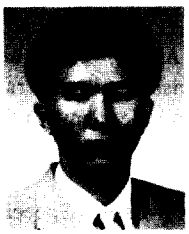
[5] Furui, S. "Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans., ASSP-34,* Feb., 1986.

[6] 이의천, "음성 신호의 음소단위 구분화에 관한 연구,"석사학위논문 광운대학교, 1990.

[7] 이강성, "음성만을 이용한 한국어 연속 숫자음 인식에 관한 연구"석사학위 논문, 광운대학교, 1988.

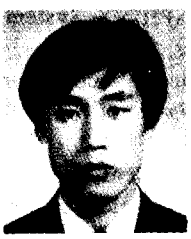
[8] S.J. Cox, J.S. Bridle, "Simultaneous Speaker Normalisation and Utterance Labilling Using Baysion/Neural Net Techniques," ICASSP-90 vol. 1, pp. 161-164, 1990.

著 者 紹 介



崔 成 好(正會員)

1965年 2月 27日生. 1989年 2月 광운대학교 전자계산기공학과 졸업. 1991年 2月 광운대학교 대학원 전자계산기공학과 공학석사 학위취득. 현재, 리버티 시스템(주) 연구소 연구원 주관심분야는 음성인식, 음성통신 등임.



李 康 誠 (正會員)

1964年 1月 15日生. 1986年 2月 광운대학교 전자계산기공학과 졸업. 1988年 8月 광운대학교 대학원 전자계산기공학과 공학석사 학위취득. 현재 광운대학교 대학원 전자계산기공학과 박사과정, 광운대학교 전산원 전자계산기공학과 교수. 주관심분야는 음성인식, 패턴인식등임.

金 淳 協(正會員)

1947年 12月 28日生. 1974年 2月 울산공과대학 전기공학과 졸업. 1976年 2月 연세대학교 대학원 전자공학과 공학석사 학위 취득. 1983年 2月 연세대학교 대학원 전자공학과 공학박사 학위 취득. 1986年 8月～1987年 7月 The University of Taxas at Austin 전기 및 전자계산기공학과 객원 교수. 현재 광운대학교 공과대학 전자계산기공학과 교수. 주관심분야는 디지탈 신호처리, 음성인식, 인공 지능등임.