한국지능시스템학회 논문지 2009, Vol. 19, No. 3, pp. 415-420

CCTV 응급상황에 따른 지능형 음성인식 시스템 구현

Implementation of Intelligent Speech Recognition System according to CCTV Emergency Information

조영임 · 장성순
Young Im Cho and Sung Soon Jang

수원대학교 IT대학 컴퓨터학과
E-mail: ycho@suwon.ac.kr, veteranus@paran.com

요 약

일반적으로 어떤 순간에 발생할지 모르는 응급 상황을 CCTV의 영상 정보만으로 상황을 항상 모니터링하기에는 인력과 비용의문제점이 발생되고 있다. 본 논문에서는 응급상황을 동적으로 보여주는 CCTV환경에서 감지하기 위해 음성인식 기술을 도입하여 문제점을 해결하고자 한다. 이를 위해 본 논문에서는 HMM(Hidden Markov Model) 기반 음성인식을 이용하여, 상황판단의 선택 여부로 고려하였으며, CCTV 환경의 기본적인 잡음 환경은 Wiener 필터를 이용하여 효과적으로 제거하고자 하며, 향후 응급 상황만을 효과적으로 CCTV 관리자에게 제공을 하여 상황인지 하고자 한다.

키워드 : 응급상황, CCTV, HMM, Wiener 필터, MFCC

Abstract

For the emergency detecting in general CCTV environment of our daily life, the monitoring by only images through CCTV information occurs some problems especially in cost as well as man power. Therefore, in this paper, for detecting emergency state dynamically through CCTV as well as resolving some problems, we propose our advanced speech recognition system. For the purpose of it, we adopt HMM(Hidden Markov Model) in our system to do a feature extraction. Also, we adopt Wiener filter technique for noise elimination in many information coming from on CCTV environment. In this paper, our system send only the emergency speech information to a manager to deal with emergency state effectively.

Key Word : Eergency. CCTV. HMM, Wiener filter, MFCC

1. 서 론

유비쿼터스 시대의 도래와 함께, 각종 정보기기들의 소형화 등 홈네트워크 시장의 확대로 인하여 음성 기술은 이들 장비간의 제어를 가장 효율적으로 할 수 있는 인터페이스로서 점점 중요성이 높아지고 있다[1,2].

현재 CCTV 환경은 기본적으로 주변에 소음이 생길 수밖에 없는 잡음환경에 처해 있으며, 응급 상황이 일어 날 수 있는 동적 환경을 기반으로 운영되고 있다. 그러나 항상 관리자가 모니터링을 할 수 있다는 장점이 있기 때문에 보안의 중요한 도구로 사용되고 있다.

그러나 이러한 보안기술 측면에서 CCTV상에서의 영상분석이 많이 연구되고 있는데, 이러한 영상을 분석한 연구에서의 한계점은 다음과 같다. 첫째, CCTV 카메라의 영상인식이 가지고 있는 많은 기술적 문제들, 특히 기상 변화, 그림자 등 조명의 변화에 따른 오인식과 같은 문제점이 발생한다는 점이다[3]. 둘째, 어두운 밤이나 혹은 화면상으로 구분이 불가능한 응급 상황 발생 시에 이를 확인하기 어렵다는 점이다. 따라서 보다 효과적인 응급 상황 대처를 위해 음성인식 기술을 이용하여 보안성 강화를 고려한 연구들이 병행되어 연구되고 있다.

그러나 CCTV 상에서의 응급시 음성을 확인하는 지금까지의 연구들은 홈 네트워크라는 환경 내에서 연구가 주로 이루어지고 있으나[4], 본 연구에서는 음성인터페이스 기술을 사용하여, CCTV 환경에서 음성인식이 가능한 연구로 확장 연구하여 응급 상황 시에 대처할 수 있는 시스템을 개발하고자 한다. 즉, 홈 네트워크 환경처럼 정적이고 제한적인 환경이 아닌 동적이고 잡음의 영향이 큰 상황에서의 음성인식이 가능한 시스템구현에 관해 연구하고자 한다.

본 논문에서는 2절에서 음성인식의 주요 기술을 서술하며, 3장에서는 본 논문에서 제안하는 음성 인식의 주요 알고리즘을 소개하고, 4장에서는 알고리즘을 시뮬레이션 결과를 고찰하고, 마지막으로는 5장에서 결론을 서술하고자 한다.

접수일자 : 2009년 4월 6일
완료일자 : 2009년 6월 1일
본 사업은 2008년도 경기도의 경기도지역협력연구센터 사업의 일환으로 수행되었음(GGA0801-45700). 본 논문은 본 학회 2009년도 춘계학술대회에서 선정된 우수논문입니다.

2. 음성 인식의 주요 기술

기본적으로 음성인식 시스템은 그림 1에서와 같이 총 6단계에 걸쳐 구성된다. 1단계는 음성신호를 전기신호로 변환하여 디지털화하여 전송하는 음성입력 단계이며, 2단계는 주위 잡음을 제거하고 음성신호를 분리하여 음성이 있는 구간을 찾아내게 되는 전처리 단계이다. 3단계는 음성인지모델을 통하여 음성인식에 유용한 특징을 뽑아내는 특징추출 단계이며, 4단계는 음성 인식 훈련 과정으로 표준 패턴 DB를 생성하는 단계이다. 5단계는 미리 생성된 기준패턴과 입력되는 음성을 비교하여 가장 비슷한 것을 인식결과로 결정하는 음향모델 단계인 탐색과정이다. 마지막으로는 이러한 인식결과를 원하는 응용에 적용하여 사용자 인터페이스 기술을 이용하게 되는 단계이다.

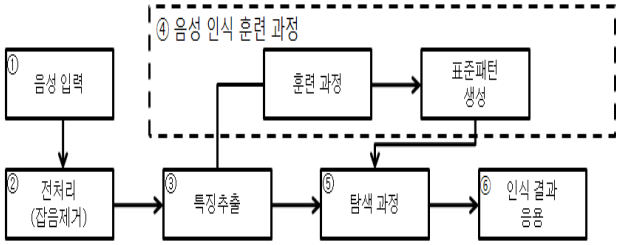

그림 1. 일반적인 음성인식 시스템 구조
Fig. 1. The general structure of speech recognition system

2.1 Wiener 필터

위의 그림 1의 ② 전처리 단계에서 사용되는 방법이다. Wiener 필터의 성능을 좌우하는 요소는 잡음구간검출과 동적 잡음으로부터의 음성 분리이다. 잡음구간 검출은 매 프레임별 정확한 잡음 모델 추정을 위한 것으로 통계기반 VAD(Voice Activity Detection) 기법을 적용하여 성능을 향상시켰다. 동적 잡음의 분리를 향상시키기 위해 음성의 보편적 특성을 나타내는 GMM(Gaussian Mixture Model)을 이용하여 MMSE(Minimum Mean Squared Error) 방법으로 잡음 보상된 음성을 추정하고, 이것과 통계기반 VAD에서 얻어진 잡음 모델을 이용해 최종 모델 기반 Wiener 필터를 설계하였다. 모델기반 Wiener 필터의 구성도는 다음 그림 2와 같다[5,6].

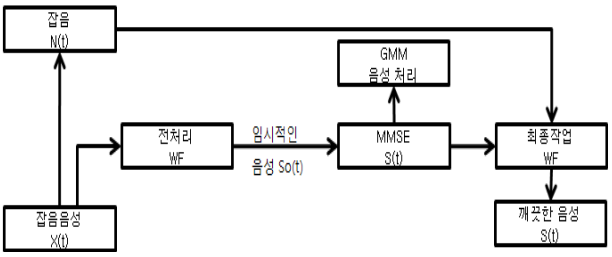

그림 2. 모델 기반 Wiener 필터의 구성도
Fig. 2. The structure of model based Wiener filter

일반적인 모델 기반 Wiener 필터의 구현 과정은 아래와 같다. 잡음음성을 X(t)라고 하고, S(t)와 N(t)를 각각 깨끗한 음성과 잡음이라 하면, 스펙트럼 영역에서 X(t) = S(t) + N(t)로 표현할 수 있으며, 필터를 설계한다는 것은 X(t)로부터 N(t)의 추정치를 구하고 이것을 이용해 S(t)의 근사치를 얻는다는 것이다. 또한 S(t)에 더 가까운 근사치를 얻기 위해 음성의 보편적인 특성을 나타내는 GMM을 이용하는데, 이것은 (1)로 표현된다.

$$P(s)=\sum_{k}^{K} p(k)N(s;\mu_{\kappa};\sum k) \qquad \text{- 식 (1)}$$

식 (1) 의 가정으로부터 모델 기반 Wiener 필터는 아래 순서로 설계된다.

① 입력된 현재의 프레임에서 통계기반 VAD를 이용해 잡음구간을 판별하고 잡음구간이면 잡음모델을 이전 값에서 갱신한다.
② Decision-directed Wiener 필터를 이용해 그림 2의 전처리-WF 블록에서 임시적인 깨끗한 음성을 추정한다.
③ 앞의 과정에서 얻어진 추정치를 이용해 가지고 있는 GMM의 각 Gaussian에 대한 사후확률을 계산하고, 이것을 이용해 MMSE 방법에 따라 최종 작업 WF 후 깨끗한 음성을 추정한다.
④ 추정된 깨끗한 음성과 ①에서 얻은 잡음 모델을 이용해 최종적인 Wiener 필터를 설계한다.
⑤ 얻어진 Wiener 필터로 현재 프레임을 처리하여 깨긋한 음성을 만들고, 다음 프레임은 단계 ①부터 위의 과정을 반복해서 처리한다.

2.2 MFCC

앞의 그림 1의 ③ 특징추출단계에 사용되는 방법이다. 사람의 귀가 주파수 변화에 반응하게 되는 양상이 선형적이지 않고 로그스케일과 비슷한 멜(Mel) 스케일을 따르는 청각적 특성을 반영한 켑스트럼 계수 추출 방법이다. 멜 스케일에 따르면 낮은 주파수에서는 작은 변화에도 민감하게 반응하지만, 높은 주파수로 갈수록 민감도가 작아지므로 특징추출시에 주파수 분석 빈도를 이와 같은 특성에 맞추는 방식이다. 처리 과정은 그림 3과 같다.

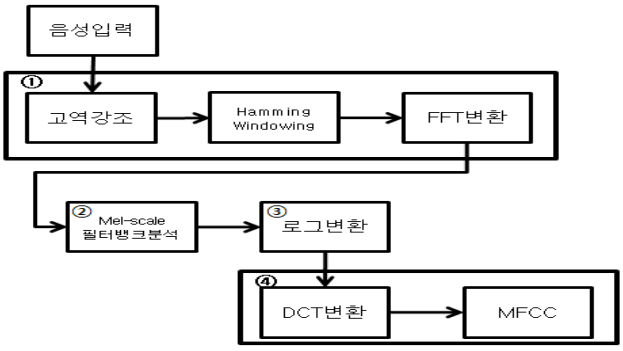

그림 3. MFCC 과정
Fig. 3. The process of MFCC

① 분석구간의 음성 신호에 푸리에(Fourier) 변환을 취하여 스펙트럼을 구한다.
② Mel 스케일에 맞춘 삼각 필터뱅크를 대응시켜 각 밴드에서의 크기의 합을 취한다.
③ 필터뱅크 출력값에 로그를 취한다.
④ 로그를 취한 필터 뱅크 값에 이산 코사인 변환(DCT, Discrete Cosine Transform)을 하여 최종 MFCC를 구한다.

2.3 HMM

HMM은 그림 1의 ④ 음성인식훈련과정과 ⑤ 탐색과정에서 사용되는 방법이다. HMM은 1960년대에 와서 Markov 모델의 파라미터 최적화 방법이 발견되어 1975년 Carnegie-Mellon 대학의 Baker와 IBM의 Jelink 등에 의하여 음성 신호 처리 분야에 처음으로 도입되었다. 현재까지는 HMM은 음성 인식에서 주도적인 역할을 해온 DTW(Dynamic Time Warpping) 알고리즘의 단점을 보완할 수 있는 방법으로 각광받아 오고 있다. HMM은 관측이 불가능한 과정을 관측이 가능한 다른 과정을 통해 추정하는 이중 확률 처리로서 음성과 같이 다변성이 발생과정을 알 수 없는 처리를 표현하는데 적당한 방법이다. HMM은 안정된 구간이나 특이한 구간의 판별, 구간에 따른 연속적 변화 특성의 묘사 및 각 구간에 대한 공통적인 단 구간 모델을 효과적으로 처리할 수 있다.

Markov 특성은 시간 영역의 사건들에 대해 과거와 현재의 사건이 주어 졌을 때 현재 사건의 조건 확률 밀도(conditional probability density)가 가장 최근의 j개 사건에 영향을 받는다는 현상을 말하며 이러한 특성을 만족시키는 과정을 Markov 과정이라고 한다.

N개의 상태들의 집합과 각각에 존재할 확률이 주어진 시스템으로 설명된다. 간단한 예로 Markov 체인에서 한 상태에 존재할 확률은 다음과 같이 표시할 수 있다.

$$P[q_t = S_i | q_{t-1} = S_i, q_{t-2} = S_k, \ldots] = P[q_t = S_i | q_{t-1} = S_i]$$

여기서 위 식의 우측 항은 시행 횟수에 대해 독립적이므로 전이 확률 a_{ij}는

$$a_{ij} = P[q_t = S_i \mid q_{t-1} = S_j], 1 \le i, j \le N$$

으로 표시할 수 있고

$$a_v \ge 0, \sum_j^N a_{ij} = 1$$

의 특성을 갖고 있기 때문에 일반적인 확률적 제한에 따른다. 이와 같은 확률 처리 출력은 각 시행에서의 상태집합이므로 관측 가능한 Markov 모델이라고 한다.

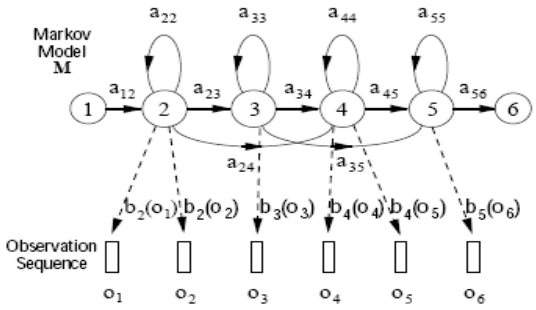

그림 4. HMM 알고리즘 인식과정
Fig. 4. The recognition process of HMM algorithm

HMM은 유한개의 상태를 갖는 Markov 모델과 출력 분포의 집합으로 특성 지을 수 있다. Markov chain에서 출력 분포의 파라미터가 음성 신호의 변이성을 모델링하는 반면, 전이 파라미터는 시간의 변화를 모델링 한다. 이러한 2가지 타입의 변화성으로 음성의 성질을 특성 지을 수 있다[7]. 학습 과정에서 forward-backward 알고리즘은 대량의 데이터로부터 자동적이고 효율적으로 전이 파라미터와 출력 파라미터 값을 추정한다. 이 알고리즘은 E-M(estimate-maximize) 알고리즘의 한 예로서 명확한 수학적인 기반을 가지고 있고, 복잡도(complexity)가 입력 음성의 길이에 선형적이라는 장점을 가지고 있다[8]. HMM은 관측 불가능한 프로세스를 관측 가능한 출력을 내는 프로세스로 추정하는 방식으로, 초기상태확률(initial state probability), 상태전이확률(state transition probability), 그리고 관측확률(observation probability)로 나타내어진다. 일반적으로 HMM은 흔히 위의 3개 파라미터로 표시된다. 모델화를 위한 HMM의 요소들을 정의하면 다음과 같다[7,8].

HMM의 개념을 이해하기 위하여 그림5 에서 서로 다른 다섯 개의 색 구슬(L=5)이 들어 있는 세 개의 상자(N=3)에서 임의의 과정에 의해 초기의 상자를 선택되고 색 구슬이 꺼내진다. 구슬의 종류를 기록하고 다시 넣는다. 이러한 작업을 T회 반복한다. 이때 t회째의 작업에서 상자q_j를 선택할 확률 a_{ij}와 그 상자 q_j중에서 구슬 v_k를 선택할 확률 $b_j(k)$로부터 T회 반복하여 얻은 구슬 열 $O_1, O_2, \ldots, O_t$가 나올 확률을 얻을 수 있다. HMM은 확률 $A = a_{ij}, A = b_j(k)$와 모델 파라미터로 하여 이 모델을 구성한다. HMM의 구조는 A, B, π에 의해 결정되므로, HMM 모델 λ = { A, B, π }와 같이 표기한다[9].

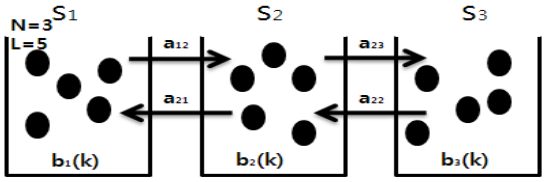

그림 5. HMM 예시
Fig. 5. HMM example

< HMM 요소정의 >

T : 관측열의 길이 ; 시행횟수
$O_1, O_2, \ldots, O_t$: 관측열 ; 구슬 열
N : 모델의 상태 개수; 상자 수
L : 관측 심볼(observation symbol)의 개수;구슬 종류 수
$S = s_1, s_2, s_3, \ldots, s_N$: 상태 집합 ; 상자 이름
$V = v_1, v_2, v_3, \ldots, v_N$: 생성 가능한 관측 심볼의 집합
$A = a_{ij}, a_{ij} = P(q_{t+1} = s_i | q_t = s_j)$:
상태 천이 확률 분포(state transition probability distribution)
$B = b_j(k), b_j(k) = P(o_t = v_k | q_t = s_j)$:
상태 j에서의 관측 심볼 확률 분포
(observation symbol probability distribution
$\pi = \pi_i, \pi_i = P(q_1 = s_i)$:
초기 상태 확률 분포(initial state probability distribution)

2.3 ECHOS

ECHOS(Easy Compact Hangul Object-orienter Speech recognizer)는 개발된 객체지향 음성인식기인 ezCSR을 바탕으로 오류 수정, 기능 추가를 거쳐서 개발된 한국어 음성 인식 플랫폼이며, 잡음제거, ESTI 특징 추출,

N-Best 탐색, Word graph 생성이 가능한 음석 디코더이다. 한글 처리가 가능하며 객체 기반의 구조를 가지고 있으며, 음성인식 DB를 구축한 HTK와의 호환이 가능하다[10].

3. 제안하는 알고리즘

본 논문의 목적은, 기존 음성인식 시스템이 위급상황에 대한 대처보다는 일반적인 음성인식 분야에 많이 사용되어 왔기 때문에, 응급상황에 대한 음성인식 시스템을 구현하려는 것이 본 논문의 주요 목적이다.

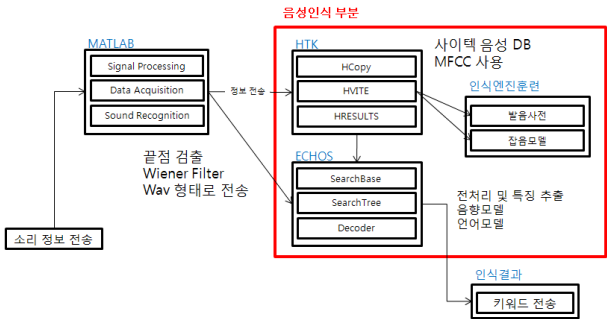

그림 6. 음성인식 시스템 전체구조
Fig. 6. The overview of proposed speech recognition system

제안하는 음성인식 시스템의 전체구조는 그림 6과 같이 구성되며 이러한 음성인식 시스템을 이용하여 응급상황 발생시에 효과적인 대처가 가능하도록 CCTV 환경을 구축할 것이다. 이 시스템의 각 과정은 다음과 같은 과정을 수행하게 된다.

[Step 1] 음성인식 DB 구축

HTK(Hidden markov Model Toolkit)[11]을 사용하여 음성인식 DB를 구축하며, 사용되는 음성 DB는 500여명의 발성, 16 kHz, 16 Bit의 데이터를 가지고 구축하였으며, 기본 음성인식 DB와 함께, 응급상황에 따른 음성인식 DB을 구축하였다.

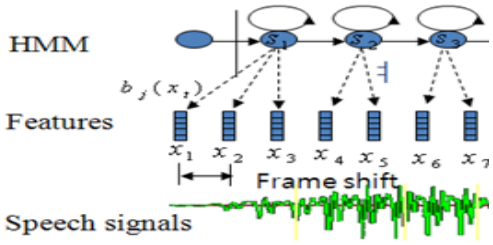

그림 7. HMM와 음성신호의 대응관계
Fig. 7. The relation between HMM and speech signal

본 논문에서 사용한 HMM의 음성신호와의 관계는 그림 7의 상태와 같이 한 대응관계를 나타낼 수 있다. 처리과정은 다음과 같다.

① S_1상태에서 x_1,x_2,x_3 가 발생되며, S_2상태에서는 x_4,x_5, S_3 상태에서 x_6,x_7이 대응된다.
② Frame shift 시간 간격에 MFCC를 이용하여 하나의 특징벡터가 음성신호로부터 추출한다..
③ 천이확률은 상태 S_i에서 S_j로 천이한다.
④ 관측확률밀도함수는 $b_i(x_t)$는 상태 S_j에서 특징벡터 x_t를 관측할 확률이다.

이처럼 좌에서 우로의 천이확률만을 갖는 HMM에서 천이확률은 각 상태에서 머무는 프레임수와 다음 상태로 이동하는 프레임의 수의 비로부터 구할 수 있다.

[Step 2] CCTV에서 음성정보 수신

CCTV 의 수신 상태를 항상 유지를 하며, 보안 감시 상황에서 수신 되는 정보를 바탕으로 일차적으로 저장되는 음성정보들을 수신 전송하는 과정을 거치게 된다. 이 때, 시간에 대하서 연속적인 크기를 갖는 아날로그 신호(analog signal)들을 디지털 신호(digital signal)으로 바꾸는 작업, 즉 양자화(quantization)을 거치게 된다.

[Step 3] 수신정보의 전처리 과정

CCTV에서 음성이 수신이 되면 MATLAB[12]에서 전처리 과정을 거치게 된다. Data acquisition 모듈과 Signal processing 모듈을 이용하여, 다음과 같은 과정을 수행한다.

① 음성의 발화구간을 추출한다.
② 추출된 음성 구간에서 특징벡터를 추출한다.
③ 추출된 특징벡터를 Wiener 필터 과정을 거친다.

이러한 과정을 거치면서 CCTV 상에서 전송된 음성정보에서 추출된 특징 벡터들은 불필요 정보(인식에 사용되지 않은 정보, 전송 시에 추가되는 잡음)들을 제거된 인식에 필요한 최적의 음성 데이터만을 기반으로 다음 과정을 수행하게 된다. 즉, 음성신호의 입력이 정확할수록 각 프레임에서 얻을 수 있는 특징벡터가 정교해지므로, Wiener 필터를 이용하여 불필요한 부분의 정보를 제거하는 것이 더욱 필요하다. 본 논문의 적용 환경은 CCTV 환경으로 응급한 상황 혹은 특정한 사건이 발생을 탐지하는 것을 목표로 하고 있기 때문에, CCTV 환경의 특징상 잡음의 빈도가 크므로 보다 빠르고 정확한 Wiener 필터의 적용 후 다음 과정을 거치도록 하였다.

[Step 4] 전처리 과정을 거친 음성의 탐색과정

구축된 음성인식 DB를 바탕으로 탐색과정은 그림 8과 같으며, 응급상황 인식 및 탐지에 사용되게 된다. 이 때 기본적으로 음소를 기반으로 단어를 인식하고자 DB를 구축하는 것을 기본으로 구축되었다.

음향모델을 중심으로 하여 단어(keyword)을 탐지하는 것을 우선으로 선정하였으며, 이는 다른 연구[13]를 통해서 확인되었으며, 본 논문에서도 이를 위하여 일반적인 음성인식 DB를 구축하였으며, 보다 빠른 탐지와 인식률을 높이기 위하여 CCTV 설치 환경에서 일어날 수 있는 응급상황을 고려하여 응급상황 DB를 새로이 구축하여 이를 우선적으로 탐색하며, 일반적인 음성인식 DB 또한 동시에 검색하여, 그 탐지확률을 높였으며, HMM 기반의 디코더 중, 한국어 인식에 맞게 개발된 ECHOS를 통하여, 보다 나은 인식결과를 도출하게 되었다. Two-pass 탐색을 통하여, 단어 간의 인식률을 높인 후, 응급상황에 검출에 맞는 음향 모델과, 언어모델을 이용하여, 단어 인식률을 높이도록 하였다.

[Step 5] 인식 결과의 전송

각 단계를 거쳐 최종적으로 검색 후 나온 결과는 텍스트 형태의 정보로 사용자 인터페이스에 전송된다.

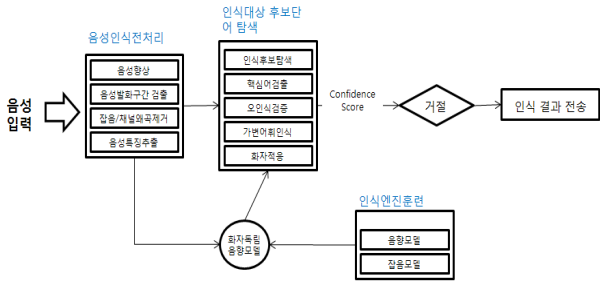

그림 8. 음성인식 부분 구조
Fig. 8. The proposed speech recognition part of system

4. 시뮬레이션 및 결과 분석

CCTV에서 전송되는 모든 소리를 음성인식에 사용되지 않으며, 기본적으로 그림 9에 나타난바와 같이 기본적으로 음성에 필요한 에너지를 가지고 있는 소리 정보에 대해서 감지를 하여, 이를 인식에 사용되는 디지털 신호로 저장하는 과정을 거치게 된다.

Wiener 필터 적용 전에 이러한 과정을 거쳐야만 응답속도에서 결과를 얻을 수 있기 때문에, 이런 음성 감지를 위한 에너지량의 체크는 필수적으로 필요하다.

이렇게 저장된 음성정보들만을 이용하여, Wiener 필터를, MATLAB에서 구현하여 그림 9에서와 같이 음성 정보에 포함된 잡음을 제거하는 과정을 통하여, 정확한 인식효과를 가져 올 수 있게 하였다.

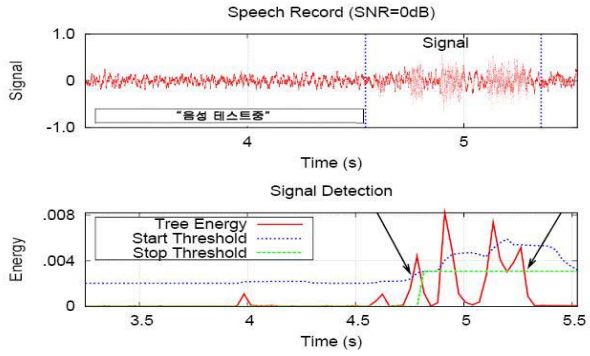

그림 9. 음성 감지 테스트
Fig. 9. Signal detection test

이와 같이 필터링 전 음성 정보들의 불필요한 부분, 즉 인식 결과에 큰 영향을 끼치지 못하는 음성의 부분이나 소리가 발생 시에 일어나는 각종 잡음, 혹은 CCTV에서 정보를 전송하면서 생기는 각종 잡음들을 효과적으로 필터링을 하여, 인식률의 상승효과를 가져 올 수 있는 결과를 확인할 수 있다.

그림 11, 그림 12와 같이 구축된 음성인식 DB를 이용하여 ECHOS를 이용하여 two-pass 탐색 시에 정방향 탐색(bigram)시와 정방향과 역방향(trigram) 시와의 결과를 얻을 수 있었다. 단어 모델의 사용 여부가 인식률과 인식 시간 모두에 큰 영향을 미치고 있음을 알 수 있었다. 즉, 탐색하는 단어들의 연관관계를 고려하는 단어 모델을 사용하는 경우 그렇지 않는 경우보다 보다 나은 인식률을 가져오는 것을 알 수 있었다.

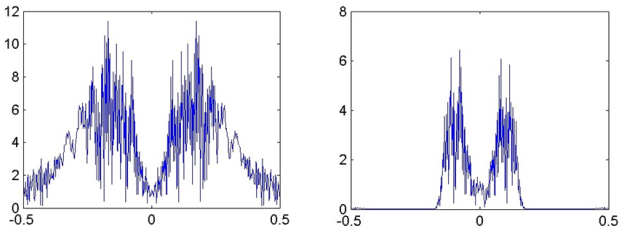

그림 10. Wiener 필터링 적용 전후 결과 비교
Fig. 10. The before and after results of Wiener filter adoption

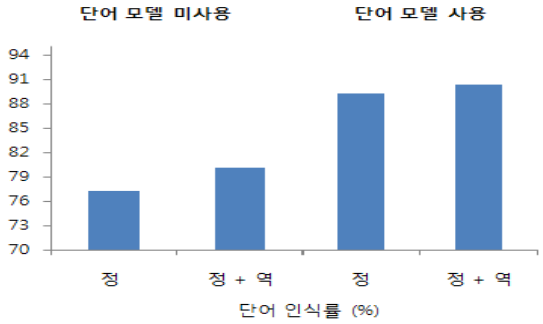

그림 11. 제안하는 시스템의 단어 인식률 결과
Fig. 11. The word recognition result of proposed system

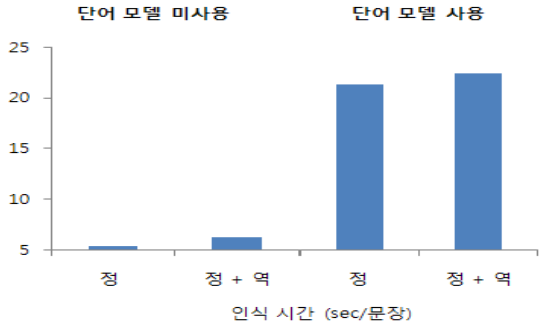

그림 12. 제안하는 시스템의 인식 시간 결과
Fig. 12. The recognition time of proposed system

또한, 이와 같은 결과를 바탕으로 HMM의 기본적인 탐색방법(좌에서 우로)을 따르는 정방향 탐색 시보다는 음소들 간의 관계를 고려하는 역방향 탐색이 단어 인식률이 8% 이상 상승으로 단어의 인식 시에 음소들 간의 관계를 고려하여 시스템을 구축하여 탐색하는 것이 정확한 인식률을 얻을 수 있음을 알 수 있었다. 그러나 음성인식 DB에서 가장 일치하는 단어열을 찾는 확률이 높아지는 것이 결과적으로는 탐색 시간의 증가도 가져오고 있음을 알 수 있었다.

따라서 음석인식율과 인식시간의 상관관계로부터 보다 효과적인 음성인식 시스템을 구축하려면 효과적인 음향모델과 언어모델을 도입하여 인식률과 인식시간의 감소를 가져오는 방향으로의 연구가 선행되어야 할 것이다.

다음 표 2는 본 논문에서 적용하고 있는 ECHOS의 성능과 벤치마킹을 위해 기존 인식기로 잘 알려진 Julius[14]와의 인식률과 인식시간을 비교한 표이다.

표 2. Julius와 ECHOS의 비교
Table 2. The Comparison between Julius and ECHOS

시스템	언어모델 적용방법	탐색방향	단어 인식률(%)	인식 시간 (sec/문장)
Julius	LM weight : 8.0 Insertion penalty : -2.0	정방향	87.9	-
		정방향+역방향	93.7	-
ECHOS	LM weight : 5.0 Insertion penalty : 10.0	정방향	94.6	30.5
		정방향+역방향	95.2	31.2

이 표로 부터 ECHOS이 인식율은 Julius와 비교해 볼 때, 훨씬 정확함을 알 수 있었다. 결론적으로 인식률관점에서 볼때, Julius보다는 ECHOS가 높고, ECHOS보다는 단어 모델을 사용하는 제안하는 시스템이 높음을 알 수 있었다.

5. 결 론

CCTV 환경처럼 많은 정보가 수집되는 환경에서는 단일 음성DB를 사용하여 음성인식 및 감지를 하는 경우, 정보를 처리하는 속도의 저하와 함께, 인식률의 저하를 가져온다는 문제점에서 착안하여, 응급상황용 음성인식 DB를 새로이 구축하여 음성 DB를 이중으로 구성하여, 각 검색을 병렬로 처리하여, 긴급을 요하는 처리를 빠르게 감지하여 보다 뛰어난 결과를 가져올 수 있었다. 물론, 응급상황용 음성인식 DB의 구축에 따라 그 결과가 다르게 나올 수 있는 문제점을 고려하여, 추후, 연계되는 연구를 통하여, 기 구축된 음성인식 DB를 자율 신경망 알고리즘을 이용하여, 자체적으로 정보를 추가하여 보다 많은 변수가 일어나는 환경에서 그 감지률을 높이고자하는 연구가 계속 되어야 할 것이다.

그리고 본 논문의 설정한 환경에서처럼 CCTV 기술의 발달로 주변 환경의 소리를 감지하여, 그 정보를 이용하여 음성인식 시스템을 효율적으로 구축된다면, 영상정보만으로 사용되어지고 있는 CCTV 환경보다 효율적이고, 현재처럼 모니터링에 필요한 인력에 대한 비용까지도 줄일 수 있을 것이다.

결론적으로 이러한 분야의 연구가 충분히 이루어진다면, 현재 CCTV 환경보다 더 나은 보안성을 갖춘 CCTV 환경을 구축하고, 그 결과를 실생활에 적용할 수 있는 좋은 연구 목표가 될 것이다.

참 고 문 헌

[1] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, Toward conversational human-computer interaction, *AI Magazine*, vol. 22, no. 4, pp 27-37, 2001.

[2] H. Kruegle, CCTV Surveillance: *Analog and Digital Video Practices and Technology*, Elsevier, pp. 227-239, 2007.

[3] 유장희, 문기영, 조현숙, 지능형 영상보안 기술현황 및 동향, *전자통신동향분석*, vol. 23, no. 4, pp 80-89, 2008

[4] M. Vacher, Jean-François S. Stéphane Chaillol, Dan Istrate, V. Popescu, “Speech and Sound Use in a Remote Monitoring System for Health Care”, *LNAI 4188*, vol. 4188, pp. 711-718, 2006.

[5] 강점자, 강병옥, 정호영, 정훈, 이윤근, 신성장동력 산엉용 대어휘 음성인식 기술 및 응용. *전자통신동향분석*, vol. 23, no. 1, pp 70-76, 2008.

[6] Doclo, S., Rong Dong, Klasen, T.J., Wouters, J., Haykin, S., Moonen, M., Extension of the multi-channel Wiener filter with ITD cues for noise reduction in binaural hearing aids, *Applications of Signal Processing to Audio and Acoustics,* vol. 16, no. 16, pp 70-73, 2005.

[7] 박재홍, 이광호, 안동순, HMM에 기반 음성인식을 위한 Toolkit의 구성요소, *한국정보과학회 논문지*, vol. 26, no. 1, pp. 472-473, 1999.

[8] 이윤근, 박준, 김상훈, 음성인터페이스 기술, *전자통신동향분석*, 제20권 제5호, pp. 1-15, 2005.

[9] 김일환, 배건성, HMM 기반의 한국어 음성합성에서 지속시간 모델 파라미터 제어, *한국음성과학회*, vol. 15, no. 4, pp. 97-105, 2004.

[10] 한국어 음성 인식 공통 플랫폼(ECHOS), http://www.sitec.or.kr/kongji_show.asp?num=111.

[11] Steve Young, Gunnar Evermann, Mark Gales, The HTK Book (for HTK Version 3.4), 2009

[12] The MathWork, Getting started guide, 2009.

[13] M. Vacher, J-F. Serignat, S. Chaillol, Dan Istrate, l. Popescu, Speech and Sound Use in a Remote Monitoring System for Health Care. *Lecture Notes in Computer Science,* pp. 711-718, 2006.

[14] Tatsuya Kawahara, Akinobu Lee, Free software toolkit for Japanese large vocabulary continuous speech recognition, *Spoken Language Processing*, vol. 4, pp. 476-479, 2000.

저 자 소 개

조영임(Cho Young Im)
1988 : 고려대학교 컴퓨터학과 학사
1990 : 고려대학교 컴퓨터학과 석사
1994 : 고려대학교 컴퓨터학과 박사
1995 : 삼성전자 선임연구원
2000 : Univ. of Massachusetts, post-doc.
현재 : 수원대학교 컴퓨터학과 교수

관심분야 : 인공지능, 뉴로퍼지시스템, 에이전트시스템, 음성인식, 유비쿼터스 시스템

장성순(Jang Sung Soon)
2008 : 수원대학교 컴퓨터학과 학사
현재 : 수원대학교 컴퓨터학과 석사과정

관심분야 : 인공지능, 정보검색, 음성인식, 유비쿼터스 시스템