멀티모달을 이용한 응용프로그램 제어에 관한 연구

최광국°, 곽상훈, 하얀돌이, 김유진, 김 철, 최승호
동신대학교 정보통신공학과

The design of application program in Multi-modal system

Kwang-Kook Choi°, Sang-Hun Kwak, Yan-Dol-I Ha, Yu-Jin Kim, Cheol Kim, Seung-Ho Choi
Dept. of Information and Communication Eng., Dongshin University
e-mail: shchoi@white.dongshinu.ac.kr

요 약

본 논문은 멀티모달 시스템에서 응용프로그램 S/W를 제어하는 연구로써 음성과 입술인식기를 결합시켜 문자 데이터를 수신하는 Comdio의 명령어들을 이 시스템이 제어하도록 설계하였다. 음성과 입술인식기는 HMM으로 구현되어 결합 시 각각의 인식기에 8:2의 가중치를 부여하였다.

1. 서론

최근 음성인식 분야에서는 인간이 갖고있는 여러 가지의 정보 즉, 입술, 제스쳐, 얼굴, 손짓, 눈동자의 움직임 등과 결합하여 정보를 정확히 확인하고 인식하고자 하는 연구가 활발히 진행되고 있다. 이러한 여러 가지 정보를 이용한 인식은 사람과 컴퓨터간의 인터페이스분야에서 중요한 부분을 차지한다[1].

또한, 멀티모달 시스템은 오디오, 비주얼을 통합하여 기존 시스템의 성능 향상을 위한 연구가 국외뿐만 아니라 국내에서도 활발히 이루어지고 있다[2]. 이러한 멀티모달 시스템은 응용프로그램과 결합하여 상용화를 위한 발돋움을 하고 있다.

2. 멀티모달 시스템 설계

멀티모달 시스템은 음성과 입술인식기를 결합하여 구현되었으며, 음성과 입술의 성능차이를 고려하여 각각에 가중치를 부여하였다.

2.1 서브워드 단위의 음성인식기 설계 및 구축

음성 특징파라미터는 26차 MFCC, 인식모델은 서브워드 단위의 HMM을 적용하고 학습 DB는 HTK를 이용하였다[3][4]. 음성 전처리는 그림 1과 같이 Pre-emphasis, 해밍윈도우, DFT 등의 MFCC 추출 과정도를 거치고, 인식단위는 서브워드 단위의 트라이폰 구조이며, 이에 대해 평균과 분산, mixture 가중치로 훈련 DB를 생성한다. 각 단어의 트리는 Lexicon구조로써 인식시 각 단어의 트리구조와 입력음성의 확률을 계산하여 가장 큰 값의 확률을 갖는 단어를 인식단어로 한다.

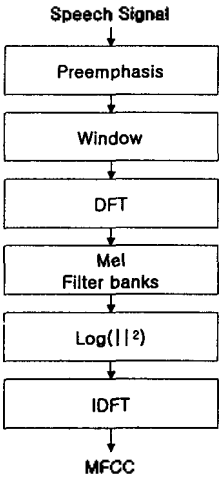

그림 1. MFCC 추출 과정도

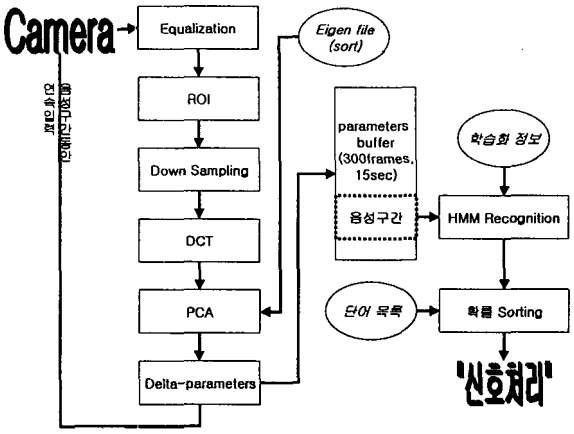

그림 2. HMM을 이용한 입술인식기의 흐름도

2.2 HMM을 이용한 입술인식기의 설계

입술인식기 설계에서는 그림 2와 같이 카메라에서 입력된 입술영상을 ROI(Region of Interest)를 이용하여 입술영역을 추출하였으며, DCT(Discrete Cosine Transform)와 PCA(Principal Component Analysis) 과정을 거쳐 입술 특징 파라미터를 추출한다[5][6].

입술 영역 추출은 이미지로부터 템플릿 방법을 사용하여 그 폭과 높이의 비를 1:1로 분리하고, 폭은 그레이 이미지로, 빛의 조사방향과 강도를 보상하기 위해 입술영상을 4영역으로 분할하여 그 폭을 구하였고 입술의 높이는 입술영역 내에서 입술의 특정 좌표로 안과 바깥 입술의 높이를 구한다.

입술파라미터는 ROI에서 16x16 픽셀 크기로 다운샘플링하고, 8x16의 이미지를 DCT와 PCA 과정을 거쳐 입술 특징파라미터를 추출한다.

실험영상의 각 단어는 입술 파라미터의 시간적인 변화를 학습시켜 DB를 생성시켰으며, HMM 인식은 발음구간 동안 영상에서 추출한 파라미터와 학습DB를 비교하여 최적의 확률을 갖는 단어를 인식한다.

2.3 음성과 입술인식기의 결합

그림 3과 같이 음성과 입술인식기의 출력된 결과에 확률을 구한 뒤 음성 80%, 입술 20%의 가중치를 부여하고 식1과 같이 결합한다.

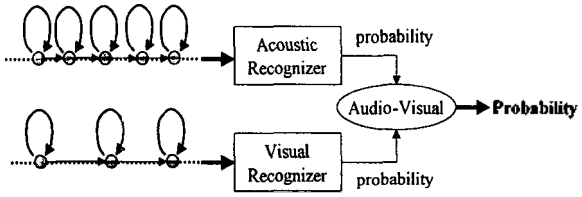

그림 3. 음성과 입술인식기의 확률 결합과정

$$S_w = k_v S_v + k_a S_a. \qquad (k_v + k_a = 1) \qquad (1)$$

S_w : 인식확률, S_v : 입술인식확률, S_a : 음성인식확률
K_v : 입술가중치, K_a : 음성가중치

그림 4는 C/S모델에서 음성과 입술인식기의 데이터교환 과정을 나타낸 것으로서 TCP/IP의 스트림 소켓방식을 사용한다.

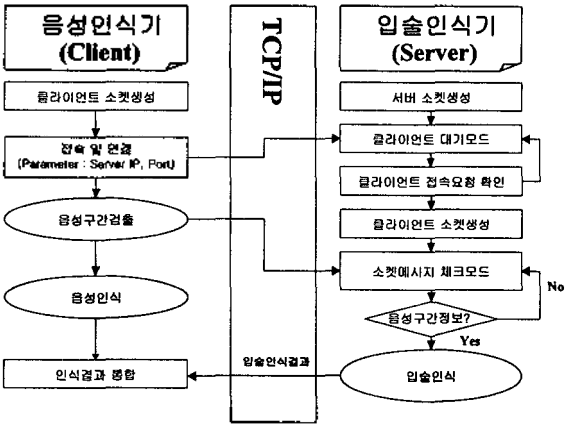

그림 4. 데이터 교환과정의 흐름도

2.4 멀티모달 시스템의 구축

멀티모달 시스템의 구성은 메인 프로세스 두 개를 사용하여, 음성과 입술에 할당함으로서 프로세스의 부하를 감소시켜 다운 현상이 일어나지 않도록 설계하고 입술인식기가 음성구간의 정보를 갖도록 한다.

하드웨어 구성은 영상입력 프레임 그리버를 이용하여 평균 초당 10프레임을 입술인식기에서, 사운드카드에서 8Khz, 16Bit의 오디오를 입력받아 음성인식기에서 사용할 수 있도록 설계한다. 다음 그림 5는 멀티모달 시스템의 구성도를 나타낸 것이다[7].

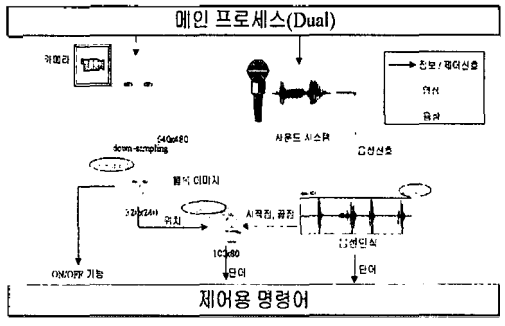

그림 5. 멀티모달 시스템의 구성도

3. 응용프로그램의 설계

컴퓨터에서 DARC(Data Radio Channel) 데이터 수신과 라디오를 청취할 수 있는 Comdio(Computer Radio)의 응용프로그램을 멀티모달 시스템이 제어 할 수 있도록 설계한다. 이에 대한 프로그램은 다음과 같다.

- 서버 소켓 초기화 : Init_Socket()

```
WSAStartup(wVer, &wsaData);
m_sAccept = socket(AF_INET,SOCK_STREAM, 0);
bind(m_sAccept, &serv_addr, sizeof(serv_addr);
listen(m_sAccept, 5);
WSAAsyncSelect();
```

- 소켓 이벤트 처리 : WindowProc()

```
WM_SOCKET:
  switch(lParam) {
    FD_READ: // 현재 소켓에 데이터 수신
      ActivationButton_of_Result();
      break;
  .....
  .....
  }
```

- 현재 메뉴상태와 비교하여 메뉴 제어 : ActivationButton_of_Result();

```
for( i = 0; i < 5; i++)
{
        // 현재 메뉴에 같은것이 있다면 loop탈출
        if( ! str.Compare(FunctionMenu[i].Title))
                break;
}

// 현재 Function Menu에서 발견됐다면
switch( i )
{
        case 0: OnProg1(); break;
        case 1: OnProg2(); break;
        case 2: OnProg3(); break;
        case 3: OnProg4(); break;
        case 4: OnProg5(); break;
}

///// 라디오 On/Off
if(str.Compare("표준FM") == 0)
        if(Comdio_Current_Status)
                OnChup();

if(str.Compare("메뉴명") == 0)
        if(!Comdio_Current_Status)
                OnChup();
```

응용프로그램의 내부는 그림 6과 같이 데이터와 라디오 수신 모듈로 나누어지고 여기에 멀티모달 시스템과의 데이터 교환을 한다. 여기에서 Comdio 명령어인 메뉴는 DARC와 라디오 상태로 나뉠 수 있으며, 이때 DARC 상태는 5가지로 분류되어 각각을 제어한다. Comdio 제어는 멀티모달 인식결과가 전송되면, 현재 메뉴상태와 비교하여 메뉴를 변경한다.

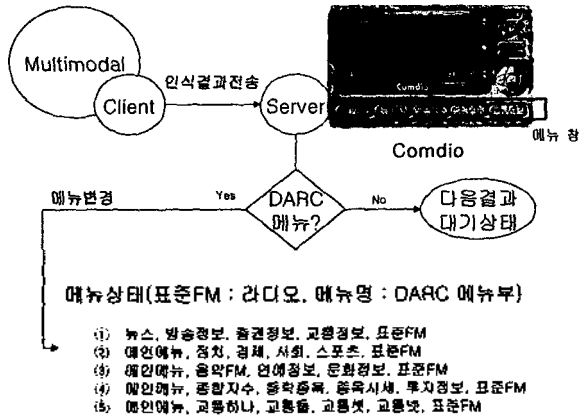

그림 6. 멀티모달 시스템과 응용프로그램의 인터페이스

4. 실험 및 결과

음성과 입술 데이터베이스는 DARC 시스템의 수신용 모듈인 Comdio에서 사용하는 22개 단어를 대상으로 70명의 20대 남성화자가 조용한 실험실에서 2회 발성한 데이터를 사용한다.

학습 데이터는 52명, 테스트 데이터는 18명 화자가 발성한 데이터(실험1)와 학습DB에 참여하지 않은 화자 5명(실험2)을 이용한다. 또한, 영상데이터는 캠코더를 이용하여 M-JPEG 형태로 30frame/sec로 저장하여 사용한다.

멀티모달 시스템의 성능평가는 실험 1과 2로 구분하여 평가한 결과 전자 96.6%, 후자 92.3%의 인식률을 나타내었다.[7]

5. 결론

본 논문에서는 멀티모달 시스템을 구현하여 응용프로그램인 Comdio의 명령어를 제어하였다. 이 결과로부터 멀티모달 시스템은 응용 프로그램 S/W와 결합하여 상용화 할 수 있는 가능성을 보였다.

6. 참고문헌

[1] Rajeev Sharma, Vladimir I. Pavlovic, Thomas S, Huang, "Toward Multimodal Human-Computer Interface," Proc. IEEE, Vol. 86, No. 5, 1998.

[2] Kyungnam Kim, JongGook Ko, SeungHo Choi, JinYoung Kim, KiJung Kim, "An Experimental Multimodal Command Control Interface for Car Navigation Systems," Proc. ITC-CSCC 2000, Vol. 1, pp. 249-252, 2000.

[3] Steve Young, *The HTK Book (for version 2.2)*, Entropic Ltd., 1999.

[4] 최광국, 김철, 최승호, 김진영, "자바를 이용한 음성인식 시스템에 관한 연구," 한국음향학회 논문집, Vol. 19, No. 6, pp. 41-46, 2000.

[5] DukSoo Min, JinYoung Kim, SeungHo Choi, KiJung Kim, "Robust Lip Extraction and Tracking of Mouth Region," Proc. ITC-CSCC 2000, Vol. 2, pp. 927-930, 2000.

[6] 최광국, 민덕수, 김유진, 김철, 최승호, 김진영, "잡음환경에서의 음성인식 성능향상을 위한 입술정보와의 결합방법에 관한 연구," 제17회 음성 통신 및 신호처리 학술대회 논문집, Vol. 17, No. 1, pp. 303-306, 2000.

[7] 최광국, 곽상훈, 하얀돌이, 김유진, 김철, 최승호, "DARC 시스템 제어기 구현을 위한 멀티모달 시스템 설계," 제13회 신호처리합동학술대회논문집, Vol. 13, No. 1, pp. 179-182, 2000.