DARC 시스템 제어기 구현을 위한 멀티모달 시스템 설계

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

The design of Multi-modal system for the realization of DARC system controller

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 DARC 시스템 제어기를 구현하기 위해 음성인식기와 입술인식기를 결합하여 멀티모달 시스템을 설계하였다. DARC 시스템에서 사용하고 있는 22개 단어를 DB로 구축하고, HMM을 적용하여 인식기를 설계하였다. 두 모달간 인식 확률 결합방법은 음성인식기가 입술인식기에 비해 높은 인식률을 가지고 있다는 가정하에 8:2 비율의 가중치로 결합하였고, 결합시점은 인식 후 확률을 결합하는 방법을 적용하였다. 시스템간 인터페이스에서는 인터넷 프로토콜인 TCP/IP의 소켓을 통신 모듈로 설계/구현하고, 인식실험은 테스트 DB를 이용한 방법과 5명의 화자가 실시간 실험을 통해 그 성능 평가를 하였다.

1. 서론

최근 멀티모달 시스템은 오디오, 비주얼을 통합하여 기존 시스템의 성능 향상을 위한 연구가 국외뿐만 아니라 국내에서도 활발히 이루어지고 있다. 멀티모달 시스템은 응용프로그램과 결합하여 상용화를 위한 발돋움을 하고 있다. 그 예로써, 자동 전화 안내 시스템에 사용되는 음성인식과 음성합성, 또는 보안에 사용되는 홍채인식, 지문인식 그리고 음성으로 문장을 입력할 수 있는 보이스타이핑 등 다양한 제품들이 선보이고 있다[1].

최근 음성인식시스템에서는 100%의 인식성능에 도전하고자 많은 연구가 진행되고 있으나 다양한 적용환경에 의한 오인식률의 벽을 깨지 못하고 있다. 본 논문에서는 이러한 문제점을 해결하고자 음성과 입술인식기를 결합하여 다양한 환경에 독립될 수 있는 멀티모달 시스템을 설계하고 DARC(Data Radio Channel) 시스템의 제어기로 활용하고자 연구를 진행하였다.

2. DARC 시스템

DARC는 FM방송을 이용하여 초당 16Kbps까지 문자 데이터를 송/수신할 수 있는 시스템이며, 국내에서는 ITS용 통신채널로 개발되어 일부지역에서 시범서비스 되고 있다. 본 논문에서는 DARC 시스템의 수신용 모듈 프로그램인 Comdio를 제어하기 위해 멀티모달 시스템을 그림 1과 같이 TCP/IP의 소켓방식을 이용하여 통합하였다[2].

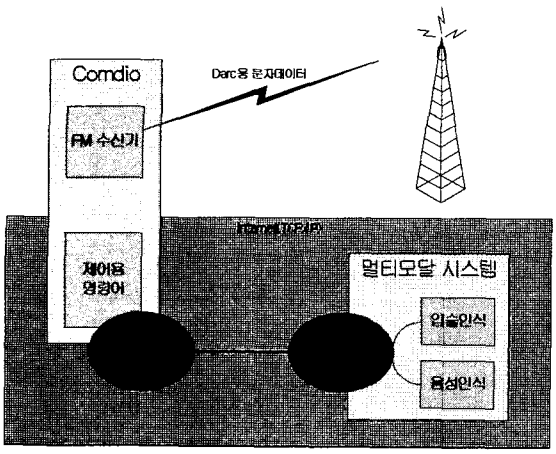

그림 1. 멀티모달 시스템과 DARC 시스템의 연결 구성도

3. 음성인식기 설계

음성인식기 설계에서는 DARC 시스템 수신기인 Comdio의 제어용 명령어인 22개 단어를 DB로 구축하고, 특징 파라미터는 26차 MFCC, 인식모델은 서버워드 단위의 HMM을 적용하였으며 학습 DB는 HTK를 이용하여 구축하였다[3][4][5].

3.1 음성 전처리 및 특징 파라미터 추출

음성 전처리는 프레임 단위를 200샘플로 하고, 이를 다음 그림 2와 같이 Pre-emphasis, 해밍윈도우, DFT 등의 MFCC 추출 과정도를 거쳐 26차 특징 파라미터를 추출하였다.

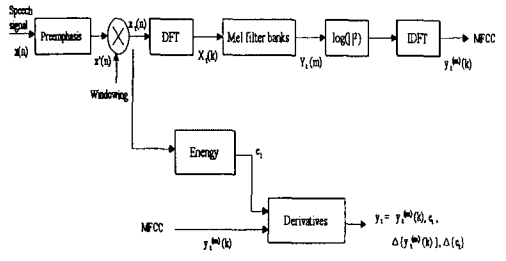

그림 2. MFCC 추출 과정도

3.2 서브워드 단위의 음성인식 시스템 구축

서브워드는 트라이폰의 구조를 갖고 있으며, 트라이폰에 대해 평균과 분산, mixture 가중치를 구하고 HTK을 이용하여 훈련DB를 생성하였다.

그림 3은 "서대문", "서울대", "서울역", "사당동", "사직로"에 대한 lexicon 구조를 이루고 있으며, 인식시 각 트리에 대한 비터비 빔 탐색을 수행하여 기준 값 이하의 값을 제거하는 방법으로 구현하였다. 따라서, 각 단어에 정의된 트리 구조와 입력 음성의 확률을 계산하여 가장 큰 값의 확률을 갖는 것을 인식 단어로 결정하였다.

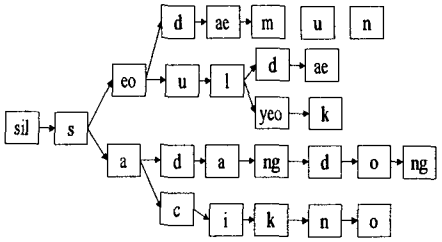

그림 3. Lexicon 구조

4. 입술인식기 설계

입술인식기 설계에서는 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 추출하였다[6].

4.1 입술 영역과 파라미터 추출

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 구하였다. 입술의 높이는 입술영역 내에서 입술의 특정 좌표를 구하여 안과 바깥입술의 높이를 구하고, 프로젝션과 부분 분산을 이용하여 입술의 변화정보를 얻는다. 따라서, 이 변화된 정보로 입술의 경계 구간에 찾을 수 있는 특정 좌표들간의 거리를 음성구간동안의 입술 특징 파라미터로 사용하였다.

4.2 HMM을 이용한 입술 인식기의 설계

카메라로부터 이미지를 입력받아 입술 파라미터를 추출하고 학습화과정을 거친 단어들과 비교하여 최적의 확률을 갖는 단어를 인식하도록 설계하였으며, 그림 4는 HMM을 이용한 입술 인식기의 흐름도를 나타낸 것이다.

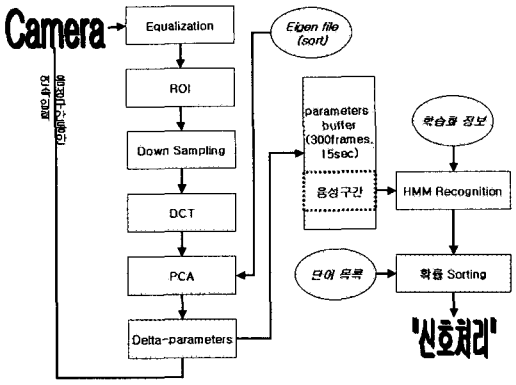

그림 4. HMM을 이용한 입술인식기의 흐름도

5. 멀티모달 시스템 구현

멀티모달 시스템은 구축된 음성인식기와 입술인식기를 결합하여 구현하였다. 각 인식기에서 출력된 인식결과를 통합하기 위해 인터넷 프로토콜인 TCP/IP의 소켓을 이용하였고, DARC 시스템에 인식명령어를 전송하기 위한 통신 모듈을 구현하였다. 인식기의 인식결합은 확률 후 결합방식을 적용하였다.

5.1 멀티모달 시스템의 설계

멀티모달 시스템의 구성은 메인 프로세스 두 개를 사용하고, 이를 음성과 입술에 할당하여 프로세스의 부하를 감소시켜 다운 현상이 일어나지 않도록 설계하였다. 또한, 두 인식기를 결합하기 위해 음성구간 동안의 정보를 입술이 이용할 수 있도록 하였다. 하드웨어 구성은 영상입력 프레임 그리버를 이용하여 평균 초당 10프레임을 입술인식기에서 사용하고, 사운드카드를 이용하여 8Khz, 16Bit의 오디오를 입력받아 음성인식에서 사용할 수 있도록 설계하였다. 다음 그림 5는 멀티모달 시스템의 구성도를 나타낸 것이다.

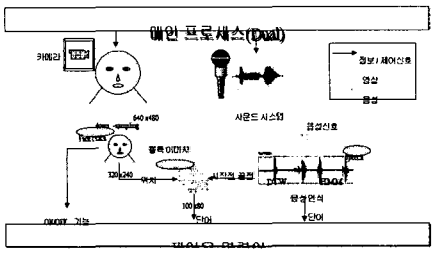

그림 5. 멀티모달 시스템의 구성도

5.2 인식결과 통합

음성과 입술인식기에서 출력된 결과를 결합하기 위해 그림 6과 같이 확률 후 결합을 이용하였으며, 음성에 80%, 입술에는 20%의 가중치를 부여하는 방식으로 다음 식 1과 같이 결합하였다.

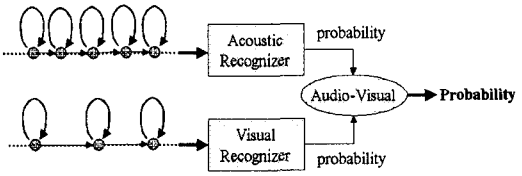

그림 6. 인식결과 통합 구성도

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

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

5.3 통신 모듈 설계 및 시스템 통합

통신 모듈은 인터넷 프로토콜인 TCP/IP의 소켓방식을 적용하여 구현하였다. DARC 시스템과 통신하는 소켓1과 음성과 입술인식기가 통신하는 소켓2로 구분하여 설계하였다. 소켓방식은 스트림 소켓방식을 이용하여 데이터의 안정성을 고려하였고, 이는 기타 다른 IPC(Inter Processor Communication)방식보다 이동성, 확장성, 신속성 면에서 뛰어난 장점을 가지고 있다. 다음 그림 7은 소켓 2의 처리 과정을 나타낸 것이며, 그림 8은 멀티모달 시스템과 DARC 시스템의 통합을 나타낸 것이다.

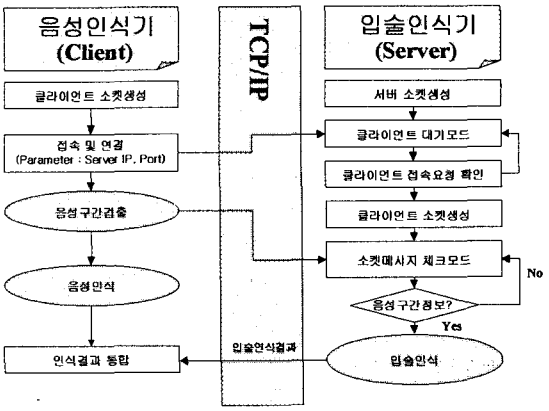

그림 7. 소켓2의 프로세스 구성도

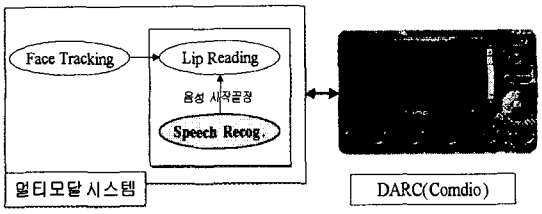

그림 8. DARC 시스템과 멀티모달 시스템의 통합

6. 실험 및 결과

6.1 데이터베이스 구축

음성과 입술 데이터베이스를 구축하기 위해 비디오 카메라와 콘덴서 마이크로폰을 사용하였으며 데이터베이스 목록은 DARC 시스템의 수신용 모듈인 Comdio에서 사용하는 22개 단어를 대상으로 70명의 20대 남성화자가 조용한 실험실에서 2회 발성한 데이터를 사용하였다.

구축된 음성데이터는 8Khz, 16Bit 샘플링 단위로 변환하였으며, 학습 데이터는 52명의 화자가 발성한 데이터, 테스트 데이터는 18명 화자의 데이터를 이용하였다. 또한, 영상데이터는 캠코더 카메라를 이용하여 수집하였으며, 수집된 데이터는 M-JPEG 형태로 30frame/sec로 저장하여 사용하였다.

6.2 멀티모달 시스템의 인식결과

멀티모달 시스템의 성능평가를 위해 구축된 DB 중 18명의 화자가 2회 발성한 테스트 데이터를 이용한 실험 1과 실험실환경에서 5명의 화자가 2회 발성한 테스트 데이터를 이용한 실험 2로 구분하였으며, 인식결과는 다음 표 1과 같다.

<표 1> 멀티모달 시스템의 인식결과

실험 / 인식방법	인식실험 1	인식실험 2
음성	94.9% (752/792)	91.8% (202/220)
입술	52.9% (419/792)	35% (77/220)
멀티모달	96.6% (765/792)	92.3% (203/220)

7. 결론

본 논문에서는 멀티모달 시스템을 설계하여 DARC 시스템 제어기를 구현하였다. 구현된 멀티모달 시스템은 음성인식을 이용한 인식결과보다 1~3%의 성능향상을 나타내었다. 이 결과에서 우리는 입술인식기가 음성인식기의 보조수단으로써 활용 가능함을 알 수 있었으며, DARC 시스템뿐만 아니라 다른 시스템과 접목할 수 있는 통신모듈을 설계하였기 때문에 확장성과 이동성이 뛰어난 멀티모달 시스템이 될 것이라고 생각된다.

8. 참고문헌

[1] 박병구, 김진영, 최승호, "잡음환경에서의 바이모달 음성인식," 한국음향학회 학술발표대회 논문집, Vol. 17, No. 1, pp. 111-114, 1998.

[2] 이상운, "FM DARC용 교통정보 수집·가공 및 전달 시스템 개발," TTA 저널 63호, 1999.

[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] 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.

[6] 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.