한국음향학회지 제20권 제6호 pp. 40~45 (2001)

음성과 영상정보를 결합한 멀티모달 제어기의 구현

Implementation of a Multimodal Controller Combining Speech and Lip Information

김 철*, 최 승 호*
(Cheol Kim*, Seung-Ho Choi*)

*동신대학교 정보통신공학과
(접수일자: 2001년 3월 16일; 채택일자: 2001년 5월 31일)

본 논문에서는 음성과 영상정보를 결합한 멀티모달시스템을 구현하고 그 성능을 평가하였다. 음성정보를 이용해서 음성인식기를, 영상정보를 이용해서 입술인식기를 설계하였으며, 두 인식기는 HMM (Hidden Markov Model) 기반의 인식엔진을 사용하였다. 음성과 영상인식의 결과는 각각 8:2의 가중치를 부여하여 통합하였다. 한편, 구축된 멀티모달 인식시스템은 DARC (data radio channel)시스템과 통합되어 응용프로그램인 Comdio (computer radio)를 제어하도록 구현하였다. 멀티모달과 DARC시스템, 멀티모달시스템 내에서 두 인식기간의 정보교환은 TCP/IP 소켓 방식을 사용하였다. 통합시스템의 Comdio 제어실험의 결과는 입술인식이 음성인식기의 보조수단으로 사용될 수 있음을 보였으며, 향후 교통정보 및 자동차 항법장치에 적용되어짐으로써 그 적용분야를 넓힐 수 있을 것으로 기대된다.
핵심용어: 멀티모달시스템, DARC, Comdio
투고분야: 음성처리 분야 (2,5)

In this paper, we implemented a multimodal system combining speech and lip information, and evaluated its performance. We designed speech recognizer using speech information and lip recognizer using image information. Both recognizers were based on HMM recognition engine. As a combining method we adopted the late integration method in which weighting ratio for speech and lip is 8:2. By the way, Our constructed multi-modal recognition system was ported on DARC system. That is, our system was used to control Comdio of DARC. The interface between DARC and our system was done with TCP/IP socked. The experimental results of controlling Comdio showed that lip recognition can be used for an auxiliary means of speech recognizer by improving the rate of the recognition. Also, we expect that multi-model system will be successfully applied to o traffic information system and CNS (Car Navigation System).
Keywords: *Multimadal system, DARC (data radio channel), Comdio (computer redio)*
ASK subject classification: *Speech signal processing (2,5)*

I. 서 론

최근, 인간과 컴퓨터 사이의 상호작용에서 좀더 인간 중심적이고 고급화된 인터페이스의 개발은 단순히 음성, 입술, 웅시, 제스쳐 등의 인터페이스에서 이들을 자연스럽게 접목시킬 수 있는 멀티모달시스템이 요구되고 있다. 이것은 각 유니모달을 서로 결합한 인간과 컴퓨터 간의 대화형시스템으로써 유니모달을 사용한 것보다 인식성능을 향상시킬 수 있다. 또한, 이 시스템은 응용프로그램과 결합하여 상용화되고 있으며, 예를 들면, 자동전화안내시스템에 사용되는 음성인식과 음성합성, 보안시

책임저자: 최승호 (shchoi@white.dongshinu.ac.kr)
520-714 전남 나주시 대호동 252번지
동신대학교 정보통신공학과
(전화: 061-330-3194; 팩스: 061-330-2909)

스템에 사용되는 홍채인식과 지문인식, 그리고 보이스타이핑과 아바타 등 다양한 제품 등이 선보이고 있다[1]. 최근의 음성인식시스템은 100%의 인식성능을 이루기 위한 많은 연구가 다양한 방법으로 진행은 되고 있으나 실제적인 적용환경에서의 오 인식률의 벽은 깨지 못하고 있다. 따라서, 본 논문에서는 이러한 문제점을 해결하기 위해 음성과 영상정보를 결합하여 다양한 잡음환경에 대처할 수 있는 멀티모달시스템을 설계하였다. 또한, 자동차 환경에 HCI기술을 접목하기 위해 PC상에서 DARC시스템을 제어하도록 구현하였다.

II. 음성정보를 이용한 음성인식기의 구현

2.1. 전처리 및 특징추출

음성 전처리 부에서는 메모리의 초기화, 현재 음성 입/출력 장치의 배경잡음 조절, 음성 구간검출, 전처리의 과정으로 이루어진다. 이때 음성의 구간검출은 음성에 포함되어 있는 정보를 나타내기 위하여 우선 입력신호를 저역통과 필터를 거쳐 고주파 잡음을 제거하고 8 kHz로 샘플링한 후 16 bit 양자화 과정을 거친 뒤 실시간 끝점검출을 하였으며, 음성의 초기입력 5프레임 (125 msec)을 묵음구간으로 가정하여 에너지와 영 교차율의 통계적 특징을 구간검출과 끝점검출의 기준 값으로 사용한다. 또한 구간검출된 음성신호는 전처리를 통해 특징 파라미터를 추출하게 된다. 이 특징 파라미터는 12차 MFCC, 12차 델타 MFCC, 1차 정규화된 로그에너지, 1차 델타에너지를 사용하여 총 26차로 구하여진다[2].

2.2. 학습과정

HTK를 통한 학습은 음운 변동 과정을 거쳐 사전 (dict)를 만든 후에 훈련 DB를 구축하며, 학습과정은 전-후향 알고리즘과 비움-웰츠 (Baum-Welch) 알고리즘을 이용해서 음소를 모델링 하고, 각 음소 상태에 대한 평균과 분산 그리고 상태 천이 확률 등을 추출한다[3].

2.3. 인식과정

인식단위는 서브워드 단위로 한 어절이 트라이폰 구조를 갖으며, 각 단어에 정의된 트리구조와 입력 음성에 대한 확률을 계산하여 값이 가장 큰 것을 인식 단어로 한다. 그림 1은 HMM을 이용한 음성인식 과정을 나타낸 것이다.

III. 영상정보를 이용한 입술인식기의 구현

3.1. 전처리 및 특징 추출

입술모양은 다양한 인지정보 중에서 음성과 가장 관련이 많고 특징적인 요소가 많기 때문에 입술정보를 인식구조에 많이 반영하고 있다. 입술 특징 파라미터 추출은 입력되는 입술영상을 흑백이미지로 전환하여 수평 측과 수직 측의 명암분포 정보를 이용하여 바깥입술의 높이와 폭, 안쪽입술의 높이와 폭을 구한다.

3.2. 학습과정

입력된 매 프레임은 먼저, 입술의 정확한 위치를 찾고 입술 특징 파라미터를 추출한다. 매 프레임마다 발성된

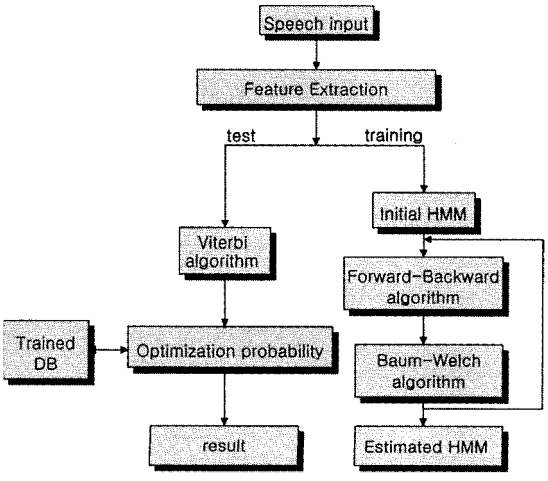

그림 1. HMM을 이용한 음성인식 과정
Fig. 1. Speech recognition process using HMM.

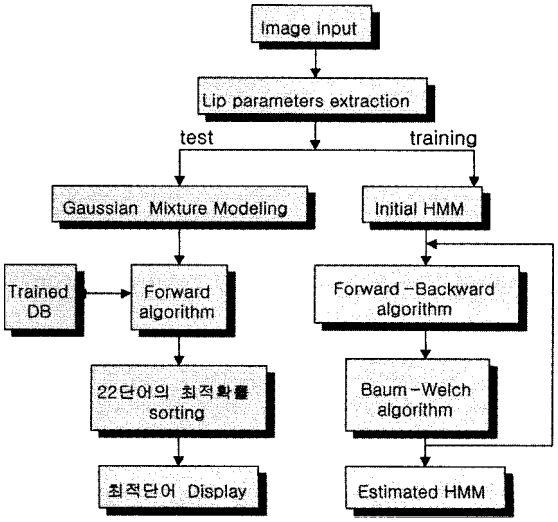

그림 2. HMM을 이용한 입술인식 과정
Fig. 2. Lip recognition process using HMM.

단어에 대한 음성구간 동안의 파라미터를 생성하고 HMM의 학습화 과정을 거쳐 평균, 분산, 가중치가 계산되어 학습 DB에 저장된다.

3.3. 인식과정

입술 인식과정에서는 전향 알고리즘을 이용하여 단어 단위로 인식한다. 이 과정에서 입력된 입술영상으로부터 추출된 파라미터인 평균, 분산, 가중치, 학습 DB와 전향 알고리즘으로부터 확률적으로 가장 근사한 값을 추출한다. 그림 2는 HMM을 이용한 입술인식 과정을 나타낸 것이다.

Ⅳ. 음성과 영상정보를 결합한 멀티모달 시스템의 구현

이 시스템을 구현하기 위해서는 음성과 영상정보의 동기화, 그리고 음성인식기와 입술인식기의 결합방법을 고려해야 한다.

4.1. 음성과 영상정보의 동기화

음성과 영상정보의 동기화를 위해 음성의 실시간 구간 검출의 시작점과 끝점정보를 영상정보인 입술에 주고, 입술이 시작점과 끝점에서 먼저 열리고 늦게 닫히기 때문에 이 부분을 보상해준다.

4.2. 음성과 입술인식기의 결합방법

HMM을 이용한 음성과 입술인식기의 구현에서는 각 인식기의 최대출력이 단어의 확률이기 때문에 수식 (1)과 같이 가중치 (α)를 부여함으로써 두 인식기를 결합한다.

$$S=\alpha S_v+(1-\alpha)S_a \qquad (1)$$

여기서, S는 멀티모달시스템의 인식확률이고, S_v는 입술인식확률이며 S_a는 음성인식확률이다.

결합방법의 타당성을 위해 영상가중치(α)를 0에서 1까지 0.1의 간격, 음성 신호 대 잡음비 (SNR: signal-to-noise ratio)는 0에서 40dB까지 5dB간격으로 구분하여 실험하여 표 1과 같은 결과를 얻었다. 여기에서 α가 0.0이면 음성인식만을, 1.0이면 영상인식만을 사용했다는 의미이다. 또한, 음영부분은 주어진 음성 신호 대 잡음비에 따른 최대인식률이다.

표 1. 음성SNR에 따른 영상가중치의 변화와 그 인식률
Table 1. Lip-weight variation and the rate of recognition for speech SNR.

음성 SNR / 입술 가중치	0 [dB]	5 [dB]	10 [dB]	15 [dB]	20 [dB]	25 [dB]	30 [dB]	35 [dB]	40 [dB]
0.00	5.30	9.09	33.08	62.63	81.57	88.38	92.68	94.44	95.20
0.10	6.82	11.11	37.88	67.17	84.60	89.65	92.93	94.70	95.20
0.20	8.08	12.12	42.93	68.18	85.86	91.16	94.70	94.96	96.21
0.30	9.85	13.38	44.44	71.46	86.11	91.67	95.45	95.96	95.71
0.40	10.86	18.43	49.49	72.98	86.36	91.92	94.19	94.95	95.45
0.50	13.13	25.00	54.04	72.73	86.11	90.15	91.67	93.43	94.19
0.60	16.92	30.81	57.07	71.21	82.58	87.12	91.16	92.17	92.93
0.70	25.00	40.91	58.08	70.45	77.53	83.84	86.87	88.89	90.15
0.80	36.87	47.22	54.55	65.66	71.97	76.77	80.81	82.07	83.59
0.90	43.43	45.96	51.01	54.55	59.34	62.12	64.14	65.91	67.17
1.00	42.93	42.93	42.93	42.93	42.93	42.93	42.93	42.93	42.93

따라서, 모든 경우에 음성 신호 대 잡음비가 20dB에서 인식률이 급격히 떨어졌으며 0dB의 경우 인식률이 0에서 40%정도이고, α가 0.9와 1일 경우 신호 대 잡음비 영역에서 40에서 60%정도의 인식률이 나타났다. 결론적으로 음성인식과 영상인식만을 인식한 경우보다 두 인식기를 결합한 경우 결합된 인식기의 인식률이 높게 나타났다. 그 중에서 α가 0.2이고 신호 대 잡음비가 40dB인 경우가 96%의 높은 인식률을 보임으로써 음성과 입술인식기는 8:2의 가중치로 결합되어 6장에서 실험하였다[4].

4.3. PC기반의 멀티모달시스템의 구현

PC기반의 멀티모달시스템을 구현한다는 것은 얼굴추적과 이를 통한 입술인식 그리고 음성인식을 통해 사용자의 입력을 해석하고 그 결과를 화면상에 디스플레이 하는 것을 말한다.

이 시스템은 H/W와 S/W로 구분되어지며, H/W에 듀얼 CPU를 사용하고 음성과 입술인식기에 각각 1개의 CPU가 할당됨으로써 시스템의 과부하를 막아주도록 하였다. S/W는 사운드 시스템에서 출력된 음성신호를 통해 음성인식기를, 얼굴추적을 통해 출력된 블록이미지로 입술인식기를 각각 구현하고 두 인식기가 결합되도록 음성구간의 시작점과 끝점정보를 입술인식기에서 사용하도록 설계하였다.

이 블럭도는 그림 3과 같으며, 두 인식기에서 나온 결과를 화면상에서 디스플레이 함으로써 이 시스템을 구현하였다[5].

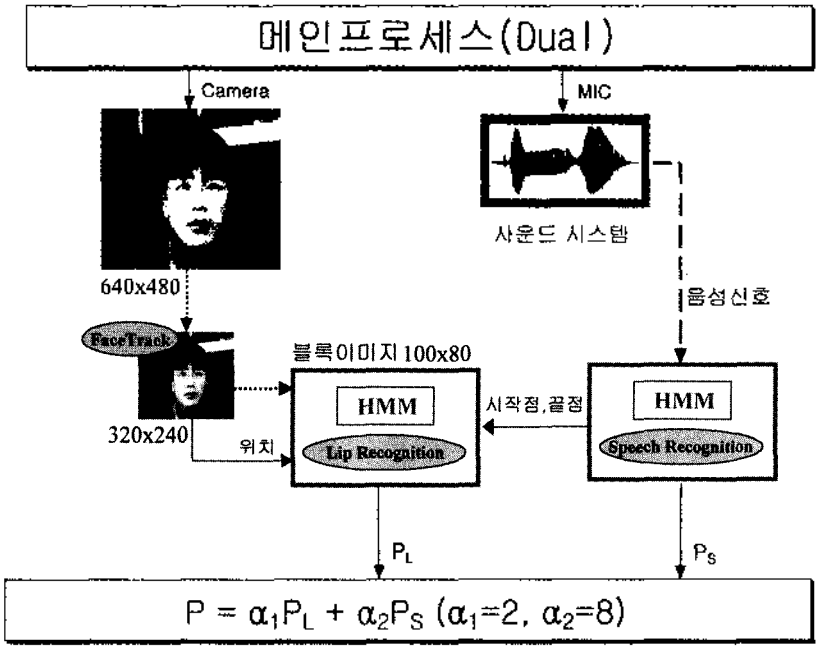

그림 3. 멀티모달시스템 블록도
Fig. 3. Multimodal system block diagram.

V. 멀티모달과 DARC시스템의 통합

5.1. DARC시스템 수신카드 및 모듈

DARC (Data Radio Channel)시스템은 FM 방송채널에 음성 외에 각종 디지털 데이터를 추가로 전송하여 이동 중에도 실시간으로 손쉽게 수신 가능한 것으로서 각종 정보 즉, 문자정보, CNS (car navigation system)용 교통정보, 증권정보 등을 문자와 그래픽으로 수신기의 액정에 표시한다. 이를 기본으로 PC에서 DARC 데이터 수신과 라디오를 청취할 수 있는 Comdio (computer radio)의

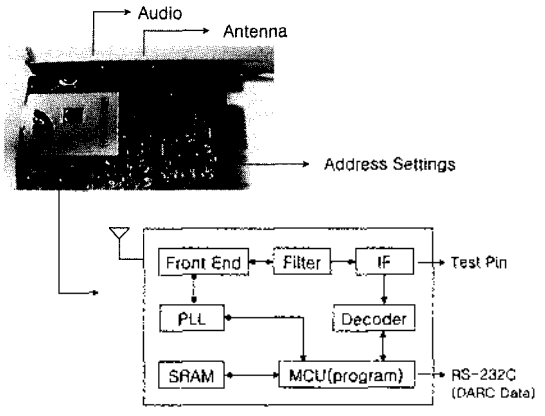

그림 4. Comdio 수신카드와 DARC모듈의 블록도
Fig. 4. Comdio receiving card and block diagram of DARC module.

응용프로그램을 멀티모달시스템이 제어하도록 설계하였다. 그림 4는 Comdio의 수신카드와 그에 장착된 DARC모듈의 구조를 나타낸 것이다. 이 수신카드는 ISA 인터페이스로 작동되며 PnP로 설계되지 않아 수동으로 카드를 설정하여야 한다. 또한 Win95/98의 운영체제와 16비트 그래픽 이상의 비디오 카드가 장착된 컴퓨터에서만 동작 가능하다.

5.2. TCP/IP을 이용한 통신모듈과 시스템 통합

시스템의 통합에서 야기되는 다운현상을 해결하고 스트림 소켓으로 데이터의 안정성을 고려하며 다른 IPC (inter processor communication)보다 이동성, 확장성, 신속성이 뛰어난 방식이 TCP/IP 소켓 방식이다. 또한, 이것은 서버/클라이언트 모델이며 서버에 음성인식기, 클라이언트에 입술인식기를 둔다.

정보교환은 음성인식기에서 음성의 시작과 끝점시간 정보를 입술인식기로 전달하고, 입술인식기는 입술인식 결과정보를 음성인식기에 전달한다. 시스템과 DARC의 수신카드인 Comdio와의 정보교환에서도 동일한 방법으로 서버에 DARC, 클라이언트에 멀티모달시스템을 장착하여 DARC에 전송해 제어할 수 있도록 하였다. 이렇게 TCP/IP 소켓방식은 그림 5에서처럼 소켓1과 2로 구분하여 소켓1에서는 DARC시스템과의 통신을, 소켓2에서는 음성과 입술인식기간의 통신방법으로 설계되었다[6].

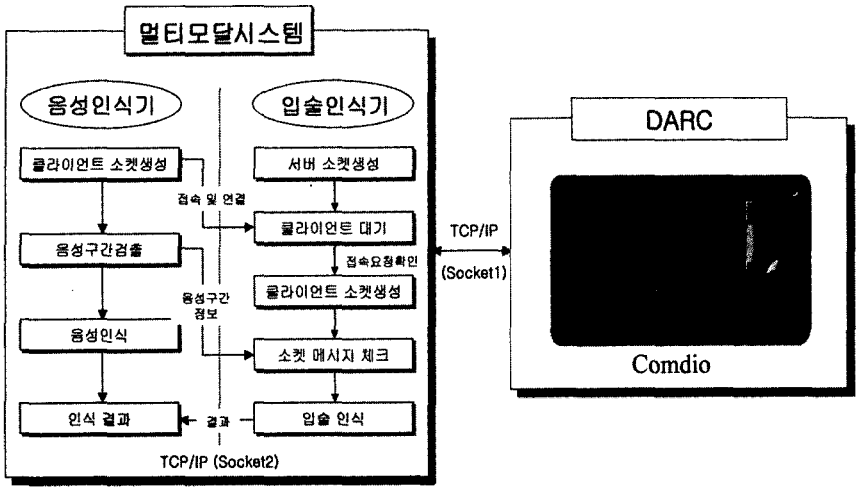

그림 5. 멀티모달과 DARC시스템의 통합
Fig. 5. Combination of multimodal with DARC system.

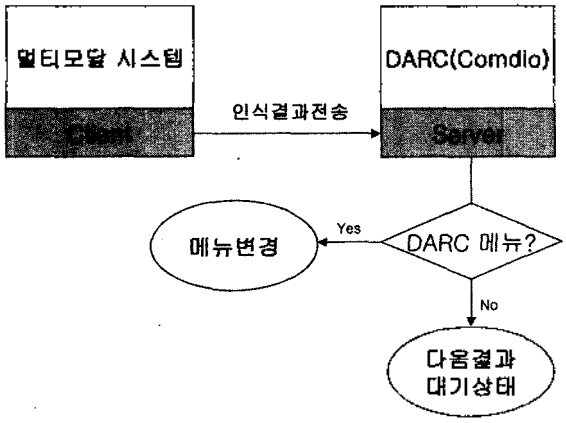

그림 6. Comdio의 제어를 위한 메뉴결정
Fig. 6. Menu decision for controlling comdio.

5.3. Comdio 프로그램 제어

이 모듈의 결과가 소켓을 통해 입력되면 Comdio는 그림 6과 같이 현재 메뉴상태와 비교하여 처리하게 된다. Comdio의 제어는 라디오 모드와 DARC 모드로 구분하였다. 라디오 모드는 Comdio를 이용하여 현재 진행중인 라디오를 청취할 수 있도록, DARC 모드는 현재 Comdio가 DARC를 통해 수신된 데이터를 출력하도록 설계하였다. 또한, Comdio를 제어하기 위한 명령어는 시스템에서 그 결과를 입력받아 처리되도록 Comdio에 모듈통신 인터페이스를 추가하였다.

VI. 실험 및 인식결과

6.1. 음성과 영상정보의 데이터베이스

DB구축은 70명의 남성화자가 조용한 연구실 환경에서 2회 발성한 22개 단어를 음성과 영상으로 동시 녹음하였으며, 이러한 음성과 영상 데이터가 각각의 인식기에 사용될 수 있도록 구간검출 알고리즘을 이용하여 시작점과 끝점을 검출함으로써 동기화하였고, 음성은 8 kHz, 16 Bit 양자화를 거쳐 PCM 파일로 저장하고, 영상은 음성정보를 이용하여 동기화할 경우 발생되는 발음시작 전과 후의 영상 파일의 단절현상 때문에 음성의 구간정보 전과 후에 최대 150 msec 만큼 시간을 허용하여 25frame/sec의 JPEG 형태로 구축하였다.

인식실험에 사용한 단어리스트는 표 2와 같다.

6.2. 인식기의 실험

음성인식기 실험은 70명의 대상자 중 52명이 발성한 데이터를 학습데이터로, 학습에 참여하지 않은 18명 (18명×2회/개×22개/명)이 발성한 396개 데이터를 SPT1으로, 실험실 환경에서 5명의 화자가 2회 발성한 220개 데이터를 SPT2로 실험한다. 입술인식기 실험은 음성과 동일하게 LRT1과 LRT2로 구분하여 잡음환경에 대해 실험하였다.

표 2. 인식실험에 사용한 단어 리스트
Table 2. Words list used in the recognition experiment.

순번	1	2	3	4	5	6	7	8	9	10	11
단어명	메뉴명	뉴스	메인메뉴	정치	경제	사회	스포츠	방송정보	표준FM	음악FM	연예정보
순번	12	13	14	15	16	17	18	19	20	21	22
단어명	문화정보	증권정보	종합지수	종목시세	등락종목	투자정보	교통정보	교통하나	교통둘	교통셋	교통넷

표 3. 실험환경에 따른 음성과 입술 인식실험
Table 3. Experiment of speech and lip recognition in Lab. environment.

실험방법	SRT1	SRT2	LRT1	LRT2	MRT1	MRT2
인식률 (%)	95.7 (379/396)	91.8 (202/220)	60.2 (238/396)	14.5 (32/220)	97.7 (215/220)	89.5 (197/220)

멀티모달시스템은 두 인식기를 8:2로 결합하여 MRT1, MRT2로 구분하였다. MRT1은 입술에 잡음이 존재하지 않은 경우, MRT2는 입술에 잡음이 존재하는 경우로 나누고 이때 테스트 데이터는 220개를 사용하였다.

6.3. 인식결과

인식된 결과는 멀티모달시스템이 음성인식기보다 2%, 입술인식기보다 37.5% 그리고 잡음이 추가된 경우 입술인식기보다 75% 향상되었다. 인식결과는 표 3와 같다.

VII. 결 론

본 논문에서는 음성과 영상정보를 결합하여 멀티모달 제어기를 구현하기 위해 PC상에서 멀티모달과 DARC시스템인 Comdio를 제어하여 이 시스템이 자동차 내에서 사용 가능함을 볼 수 있었다.

이 시스템은 음성인식기보다 실험실 환경에서 2%, 입술인식기보다 37.5%의 성능향상을 보여 입술인식기가 음성인식기의 보조수단으로 활용됨을 알 수 있었다. 또한, 음성과 입술인식기를 신호 대 잡음비에 따른 인식률의 결과를 통해 8:2의 가중치로 결합하였고, TCP/IP 소켓을 이용하여 확장성과 이동성이 뛰어난 멀티모달시스템을 구현하였다. 향후 자동차 내에서 사용가능하기 위해서는 차량잡음의 제거방법에 대한 연구와 잡음 DB 구축이 선결과제라고 사료된다.

감사의 글

본 논문은 정보통신부에서 지원하는 연구 결과물 중 일부입니다.

참 고 문 헌

1. Rajeev Sharma, Vladimir I. Pavlovic, and Thomas S.Huang, "Toward Multimodal Human-Computer Interface," *Proceedings of IEEE*, Vol. **86**, No. 5, pp. 853-869, May 1998.
2. Claudio Becchetti, Lucio Prina Ricotti, Speech Recognition Wiley, Inc., 1999.
3. L. R. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition" *Prentice-Hall International, Inc.*, 1999.
4. Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, Phil Woodland, *The HTK Book* Microsoft Corporation, 2000.
5. 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.
6. 최승호, 김진영, 최광국, 김철, "자바를 이용한 음성인식 시스템에 관한 연구" 한국 음향학회 논문집, Vol. **19**, No. 6, pp. 41-46, 2000.

저자 약력

● 김 철 (Cheol Kim)

1946년 9월 29일생
1970년 2월: 조선대학교 전기공학과 (공학사)
1982년 2월: 한양대학교 산업대학원 (공학석사)
1998년 2월~현재: 동신대학교 정보통신공학과 박사과정 재학중
※ 주관심분야: 음성인식 및 신호처리, 통신망운용/관리

● 최 승 호 (Seung-Ho Choi)

1955년 8월 24일생
1981년 2월: 전북대학교 물리학과 (이학사)
1984년 8월: 명지대학교 전자공학과 (공학석사)
1992년 2월: 명지대학교 전자공학과 (공학박사)
1992년 3월~현재: 동신대학교 정보통신공학과 교수
※ 주관심분야: 음성인식, 멀티미디어통신, 멀티모달MMI