상태공유 HMM을 이용한 서브워드 단위 기반 립리딩*

Subword-based Lip Reading Using State-tied HMM

김 진 영** · 신 도 성***
Jinyoung Kim · Dosung Shin

ABSTRACT

In recent years research on HCI technology has been very active and speech recognition is being used as its typical method. Its recognition, however, is deteriorated with the increase of surrounding noise. To solve this problem, studies concerning the multimodal HCI are being briskly made. This paper describes automated lipreading for bimodal speech recognition on the basis of image- and speech information. It employs audio-visual DB containing 1,074 words from 70 voice and tri-viseme as a recognition unit, and state tied HMM as a recognition model. Performance of automated recognition of 22 to 1,000 words are evaluated to achieve word recognition of 60.5% in terms of 22-word recognizer.

Keywords: Multimodal, Bimodal, Lipreading, Sub-word, Speech Recognition, Viseme

1. 서 론

컴퓨터가 인간생활의 필수적인 요소로 자리매김 함에 따라, 그 활용분야가 급증하고 있다. 그러나, 그 역사가 짧고, 컴퓨터의 연구개발이 주로 컴퓨터 하드웨어의 성능 향상에 있었기 때문에, 사용자 인터페이스 측면에서 보면 사람과 친숙하지 못하다. 특히, 현재 컴퓨터에 명령을 전달하기 위한 입력도구로서 사용되는 키보드와 마우스는 기계적인 방법으로 사람이 사용하는 통신의 도구(즉, 음성, 표정, 제스처 등)에 비하여 비자연적이라고 할 수 있다. 이러한 HCI(human-computer interface)의 문제를 해결하기 위하여 인간공학적 방법들이 동원되는데, 그 대표적인 방법이 음성을 사용한 HCI라고 할 수 있다. 음성인식은 현재 잡음이 적은 환경 하에서는 상당히 높은 인식성능을 보이고 있다. 그러나 배경잡음이나 전송 채널에서의 열화가 심한 경우 급격한 인식률의 저하가 발생하여 실제 응용에 있어 큰 제약을 받고 있다. 이러한 인식률 저하를 해결하기 위하여 다각도의 인간공학적 접근이 이루어지고 있다. 예를

* 이 연구는 과학재단 '98-99 핵심전문연구결과 중의 하나입니다.
** 전남대학교 공과대학 정보통신공학부, RRC HECS
*** 전남대학교 공과대학 대학원 전자공학과

들어, 음성 외에, 사람의 대표적인 의사표현 방식인 몸동작, 얼굴표정, 눈동자의 움직임, 입술 모양 그리고 심지어는 뇌파의 움직임까지도 HCI의 방법으로 연구되고 있으며, 이와 같은 다양한 행동양식을 포함한 인식 기술을 멀티모달(multi-modal) HCI 기술이라고 한다[1~8].

입술과 음성 정보를 묶는 바이모달(bi-modal) 음성인식은 HCI의 한 가지 방법으로서, 그림 1은 바이모달 음성인식의 기본 개념도를 보여주고 있다. 본 연구에서는 그림 1에 보인 바이모달 음성인식 기술 중, 음성인식을 위한 립리딩 기술에 대하여 연구결과를 설명한다. 그리고 바이모달 시스템이 uni-modal 시스템보다 우수함을 증명한 비교 실험 결과는 참고 문헌 [11]을 참조하기 바란다. 본 연구에서는 자동 립리딩 실험을 위하여 1,074 단어에 대한 70 명분의 DB를 녹화하여 구축하였으며, 이를 바탕으로 서브워드(subword)의 립리딩 실험을 하였다. 여기서 서브워드기반의 립리딩을 구현한 이유는 서브워드를 사용할 경우 단어독립 립리딩이 가능하기 때문이다. 그러면, 다음의 장에서 본 연구에서 구현한 립리딩에 대하여 자세히 설명한다.

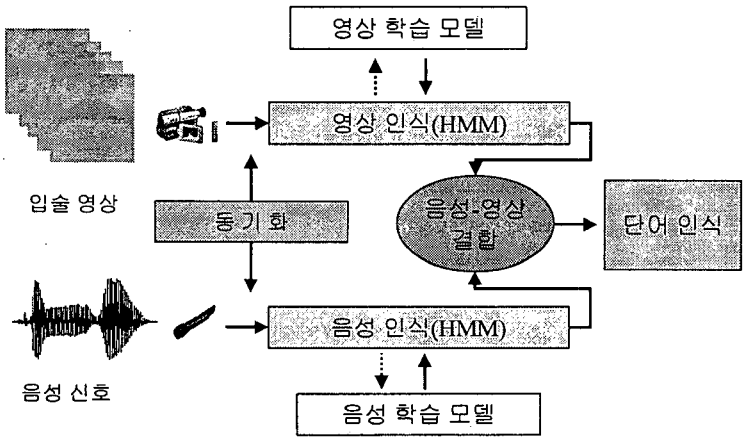

그림 1. 바이모달 음성인식 시스템의 구성도

2. 구현된 자동 립리딩의 구조

다음의 그림 2는 본 연구에서 구현된 자동립리딩 알고리듬의 블록도를 보여주고 있다. 본 립리딩 시스템에서 파라미터 추출과정은 크게 영상전처리와 영상변환의 두 과정으로 나누어진다.

추출된 파라미터는 HMM인식 알고리듬을 사용하여 인식된다. 본 연구에서는 HMM 인식을 위하여 HTK(캠브리지 대학: hidden markov model tool kit) Ver.2.2를 사용하였다.

먼저, 영상 전처리 과정은 입력된 입술 영상을 특징 파라미터로 변환시키기 위한 처리과정이다. 흑백 영상이나 컬러 영상을 입력으로 받아들여 이진 영상처리를 위해 명암 영상 형태로 변환한다. 그 후 입력된 영상에서 입술 영역만을 추출하기 위한 단계로서 2진 영상 변환을 수행하여 입술 안쪽 영역을 찾는다.

입술 안쪽 영역을 찾게 되면, 다음으로 찾아진 입술 안쪽 영역을 기준으로 입술 ROI (region of interest)를 구성하고 원 입력영상으로부터 분리해 낸다. 일단 분리된 입술 ROI에 대해 데이터 처리량을 줄이기 위한 다운샘플링(downsampling) 과정을 거친 다음 입술의 기하학적 대칭성에 의거하여 입술 영상을 절반으로 접는 과정을 거친다.

입술 ROI를 절반으로 접게 되면 이후 처리과정에 소요되는 데이터 처리량도 줄일 수 있을 뿐만 아니라 최종적으로 HMM 인식 파라미터로 사용될 특징파라미터의 개수 또한 줄어듦으로써 여러 가지 측면에서 시스템의 연산 부담을 낮출 수 있는 장점이 있다[2].

입술 대칭 접기 과정을 마지막으로 영상 전처리 과정이 끝난 후 영상 선형 변환 과정을 수행한다. 이 과정은 입력 정보를 HMM 인식 알고리즘에 적용하기 위한 특징 파라메터로 변환하는 단계이다.

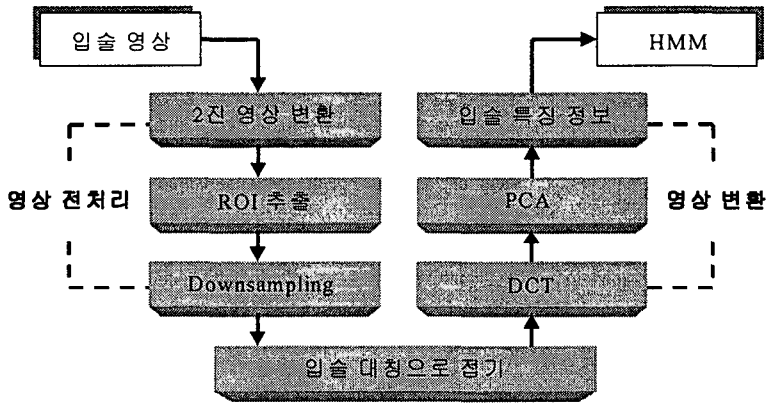

그림 2. 자동 립리딩 알고리즘 블럭도

표 1. 구축 DB의 정보

화자의 수	70 명
음성 데이터	1,074 단어(증권명)
반복 회수	1 회/1 인
발음 단어수	1,074/1 인
비디오	30 frames/sec, 8 bit gray
음 성	8 KHz 샘플링, 16 bit 양자화

먼저 입술 ROI 영상을 선형변환 시키기 위해 2D-DCT (discrete cosine transform)를 사용하여 절반으로 접어진 입술 ROI를 변환한다. 변환된 입술 ROI에 대해, 통계적 알고리즘인 PCA (principal component analysis)를 적용하여 입술 모양이 갖는 특징 정보를 가능한 많이 포함하는 주성분으로 집약시킨다. PCA를 거치게 되면 입력된 입술 영상이 갖는 입술 특징 정보를 대부분 포함하는 소수 m 차원 파라미터로 축약시킬 수 있고 여기서 추출된 주성분들이 최종적 HMM 인식 알고리즘에 사용될 입술 특징 파라미터들로 확정된다[2]. 자세한 파라미터 추출은 참고문헌 [2]를 참조하기 바란다.

3. 자동 립리딩 실험을 위한 DB

립리딩 연구에 이용된 DB는 자체 구축하였다. DB는 실내조명환경에서 수집되었으며, 총 1,074 단어에 대하여 70 명이 발성하는 음성과 영상을 녹화하였다. 영상데이터의 수집은 30 frames/sec의 성능을 갖춘 캠코더를 이용하였으며, 음성과 함께 캠코더에 녹화되었다. 자세한 DB의 스펙은 다음의 표 1과 같다.

캠코더를 사용하여 녹화한 DB는 본 연구를 위해 개발된 툴을 이용해서 캡쳐보드를 통해 영상을 A/D 변환을 하였으며 음성은 사운드 카드를 사용하여 A/D 변환을 하였다. 그리고 정확한 실험을 위해 음성과 영상의 동기를 맞추었다.

4. HTK를 이용한 서브워드 단위 립리딩 구현

본 연구에서는 HTK를 사용하여 서브워드단위의 음성인식 실험을 수행하였다. HTK 활용의 절차는 크게 두 개의 주된 과정으로 이루어진다. 하나는 학습 툴로서 학습을 위한 발화문장과 그 문장의 전사(transcription)를 이용하여 HMM 파라미터를 구하는 것이고, 다른 하나는 HTK 인식 툴로서 인식성능을 평가하는 것이다. 본 연구에서는 파라미터는 영상정보 파라미터를 자체 개발한 툴을 사용하여 HTK 파일 형태로 변환하여 사용하였다. 단, 학습을 위한 준비 단계에서는 전사를 위하여 음소가 아닌 viseme을 사용해야 한다.

음성은 발음형태상 보이지 않는 많은 음성학적 특징(유음, 유성음)을 가지고 있지만 음소들 사이에는 시각적 유사성으로 그룹 지어지는 경향이 있다. 이러한 그룹들을 visual과 phoneme의 합성어인 viseme으로 정의했다[9]. viseme을 영상음성학적 관점에서 접근할 때 이 개념은 적당하지만, 음성에서와 마찬가지로 영상에서도 연음 현상이 발생한다. 그러므로 발음시 음성과 영상은 잠시 정지되어 있는 모음에서 기대되는 변수와는 달리 왜곡이 발생하게 된다. 또한 잡음환경 하에서의 음성인식 강화를 위해 부가적으로 사용하는 영상정보는 보통 30 frame/sec의 연속된 입술 영상 프레임으로 이루어져 있다. 이는 음성의 한 음소에 해당하는 프레임 수에 비해 영상의 프레임 수가 현저히 적다는 것을 뜻한다. 이 때문에 영상과 입력된 음성신호와의 결합을 위한 동기화 문제에 있어 음성과 일치하지 않는 영상이 잘못된 정보 전단을 일으킬 수 있는 것이다.

본 연구에서는 viseme의 개념이 음성에서의 allophone과 유사하다고 가정하고 서로 다른 음소이지만 시각적으로 동일하거나 비슷하게 그룹화되는 자음과 모음의 그룹을 시각소로 정의하여 주어진 음성을 영상으로 모두 표현하는데 필요한 입술 모양을 결정하였다. 본 연구에서는 다음의 표 2의 viseme를 사용하였으며, 이 viseme의 정의는 참고문헌 [10]의 결과로서 얻어진 것이다.

그리고 HTK 학습을 위해서는 state tying을 위한 질의어 셋을 필요로 한다. HTK를 이용한 음성인식이 음소 단위를 이용한 트라이폰 구조를 가지므로, 트라이폰 시스템을 구축하기 위해 HTK에서는 결정 트리 기반 클러스터링(decision tree-based clustering)을 적용한다.

여기서 탑-다운 연속 최적화(Top-down sequential optimization process) 알고리즘을 사

용하여 트리가 형성되며 최종적 모델 클러스터링을 위해 각 트리의 노드마다 질의어를 적용하였다

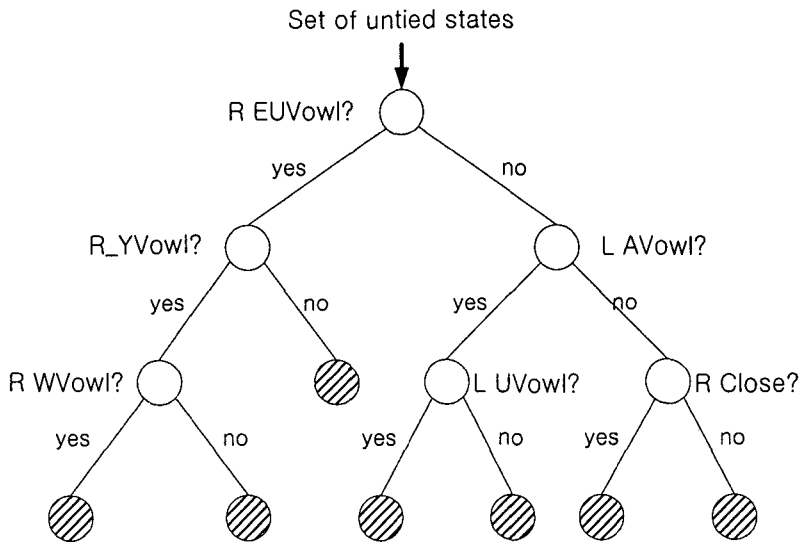

그림 3. 음성 결정 트리의 예

음성 트리의 모양은 그림 3과 같으며 루트 노드에는 특정 음소의 상태들의 풀(pool)이 입력되며 각 노드별 최적의 음성 질의들을 사용한 분할 과정을 거쳐서 마지막으로 생성된 각각의 leaf 노드는 비슷한 성질의 문맥을 가지는 상태를 포함하게 된다.

트리 생성을 위한 노드 분할 과정은 다음과 같은 방법으로 진행한다. 먼저 질의를 사용하여 생성된 yes/no 노드들과 부모 노드간에 로그 likelihood의 이득은 식 (1)을 사용하여 계산한다.

$$\Delta L = L(A) - (L(B) + L(C)) \qquad (1)$$

A는 부모 노드에 포함된 상태들의 집합이며, B와 C는 자식 노드(yes/no) 각각에 포함된 상태들을 나타낸다. 식 (1)을 사용하여 가장 큰 로그 likelihood의 이득을 생성하는 질의가 그 노드의 질의로 선택되며 그것을 사용하여 부모의 노드의 데이터를 yes와 no의 두 집합으로 분할한다. 가장 적합한 질의를 사용하였을 때의 이득이 특정 threshold보다 작아질 때까지 이러한 과정을 반복한다. 분할 과정이 끝나면 트리의 각 leaf는 특정 음성의 상태들 가운데 비슷한 성질을 가진 상태들을 포함하게 된다. 그러나 종종 서로 다른 leaf 노드들이 매우 비슷한 형태가 되는 경우가 발생하며, 이것은 음성 문맥들은 달라도 특징 벡터 내에서의 변화들은 같을 수 있다는 사실에 기인한다. 따라서 leaf들 가운데 비슷한 노드들이 있는지를 검사하여 이러한 노드들에 대한 합병(merge)을 수행함으로써 공유 상태들의 수를 줄일 수 있다.

인식시에는 트라이폰 입력에 대한 실제 상태로의 매핑 루틴이 필요하며, 우리는 다음과

같은 순서로 수행되는 매핑루틴을 구현하였다.

단계 1: 질의어 집합과 트리 파일 로딩
단계 2: 트리 위치 찾기
단계 3: 질의어 평가와 트리 횡단
단계 4: 실제 상태 찾기

단계 1에서는 질의어 집합과 이미 생성된 트리 파일을 로드한다. 단계 2에서는 현재 트라이폰의 가운데 모노폰에 대응하는 트리를 찾고 단계 3에서는 트리의 루트에서 시작하여 각 노드에 linked-list 형태로 부착된 질의어를 평가하면서 트리를 순회한다. 특정 leaf에 도달하면 트리순회는 끝나고 다음 단계 4에서 실제 상태를 찾음으로서 매핑루틴을 완료하도록 구현했다.

본 연구에서는 음성의 질의어 세트가 비음, 끊어지는 음 등의 음운학적 분류를 적용한 반면 영상정보의 모델링을 위해 표 2의 viseme을 사용하여 입술모양을 고려한 다음과 같은 질의어를 만들어 사용하였다. 표 3에 립리딩을 위한 viseme 단위의 질의어 셋을 보여주고 있다.

표2. 한국어 viseme 변환

초 성			중 성			종 성		
음소	비짐	음성	음소	비짐	음성	음소	비짐	음성
ㄱ(ㄲ), ㅋ, ㅎ	g	g(gg), kh, h, k, kk	ㅏ	a	a, aa	ㄱ	gf	g, kh
ㄴ, ㄹ	n	n, ng(ㅇ), r	ㅑ	ya	ya	ㄴ	nf	n
ㄷ(ㄸ), ㅌ	d	d(dd), th, t, tt	ㅓ	eo	eo, vv	ㄷ	df	d, s, j, ch, th, h
ㅁ, ㅂ(ㅃ), ㅍ	m	m, b(bb), ph, p, pp	ㅕ	yeo	yeo, yv	ㄹ	l	l
ㅅ(ㅆ), ㅈ(ㅉ), ㅊ	s	s(ss), ch, c, j(jj), cc	ㅗ	o	o, oo	ㅁ, ㅂ, ㅍ	mf	m, b, ph
			ㅛ	yo	yo	ㅇ	ng	ng
			ㅜ	u	u, uu			
			ㅠ	yu	yu			
			ㅡ	eu	eu, xx			
			ㅣ	i	i, ii			
			ㅐ, ㅔ	e	e, ee, ae, ai			
			ㅒ, ㅖ	ye	ye			
			ㅘ	wa	wa			
			ㅝ	weo	weo, wv			
			ㅞ, ㅚ, ㅙ	we	we			
			ㅟ	wi	wi, ui			
			ㅢ	eui	eui, xi			

표 3. 립리딩을 위한 viseme 단위 질의어 세트

질의어 (QS) 구분	질의어
"R EUVowel"	{ *+eu , *+eui }
"R YVowel"	{ *+ya ,*+yeo ,*+yo ,*+yu ,*+ye }
"R WVowel"	{ *+wa ,*+weo ,*+we ,*+wi }
"R Close"	{ *+m ,*+mf }
"R SemiOpen"	{ *+n ,*+d ,*+s ,*+g }
"L AVowel"	{ a-*, ya-*, wa-* }
"L UVowel"	{ u-*, yu-*, eu-* }
"L IVowel"	{ i-*, eui-*, wi-* }
"L EVowel"	{ e-*, ye-*, we-* }
"L OVowel"	{ o-*, yo-* }
"L EOVowel"	{ eo-*, yeo-*, weo-* }
"L SemiOpen"	{ n-*, d-*, s-*, g-* }
"R_monophone" "L monophone"	{*+g}, {*+n}, ... {g-*}, {n-*}, ...

5. 실험 결과 및 분석

5.1 단어 단위 인식결과와 비교

먼저 서브워드 단위의 인식성능을 비교하기 위하여 단어 단위의 HMM 인식기와 비교하였다. 이를 위하여 1,074 단어 중 22 단어를 선택하여 비교하였으며, 그 결과를 표 4 및 5에 나타내었다. 먼저 표 4를 살펴보면, 상태수와 mixture의 수가 증가함에 따라 인식률이 향상됨을 알 수 있으며, 최대 인식률은 약 65% 정도가 되는 것을 볼 수 있다.

표 4. 22 단어에 대한 단어 단위 인식결과

	3s	4s	5s	6s
3 m	33.31	36.95	44.22	47.92
4 m	38.44	45.78	50.33	55.00
5 m	40.58	46.82	54.55	59.94
6 m	45.84	54.42	59.42	65.00

(s: 상태수, m: 가지수(mixture), 단위: %)

표 5의 결과는 tri-viseme을 모델링 함에 있어서 상태 수 3 개, 상태당 5 개의 Gaussian을 사용한 경우의 인식률이다. 22 단어의 경우에서만 보면 단어 단위 인식률이 보다 높게 나타나고 있다. 그러나 단어 단위 인식은 인식할 단어수가 많아지면 학습모델이 증가하게 되므로

한 단어 인식을 위한 탐색 시간이 길어지게 되며 따라서 대용량 단어인식기를 구현하는데 적합하지 않다고 볼 수 있다. 여기서 최대 인식률은 60.5%로서 단어 단위모델에 비하여 약간 떨어지긴 하지만 유사한 인식률을 보임을 확인할 수 있었다. 따라서 본 연구에서 개발한 서브워드 단위 립리딩이 타당함을 보여주는 결과라고 할 수 있다.

표 5. HTK를 이용한 서브워드 단위 립리딩 인식 결과

Viterbi beam threshold	30	40	50	60
인식률	54.86	60.49	56.98	55.40

(tri-viseme: 3 state 5 mixture)

5.2 단어 수에 따른 인식률 비교

단어의 수가 증가함에 따라서 립리딩의 인식률이 저하될 것임은 충분히 예상할 수 있다. 물론, 음성인식의 경우도 동일한 결과가 발생한다. 본 연구에서는 립리딩에서 발생하는 이 같은 인식률의 저하를 평가하기 위하여 단어 수에 따른 인식 실험을 수행하였다.

다음 그림 3은 단어 수에 따른 립리딩의 성능을 보여주고 있다. 그래프를 살펴보면, 단어 수가 22 단어인 경우 약 60%의 인식률을 보이는 반면, 1,000 단어급에 이르면 인식률이 20% 정도로 급격하게 저하함을 관찰할 수 있었다. 따라서, 립리딩이 대용량 음성인식에 사용되기 위해서는 알고리즘의 개선 및 기술적 향상을 위한 연구가 상당부분 필요함을 그림 3의 분석을 통해서 알 수 있다.

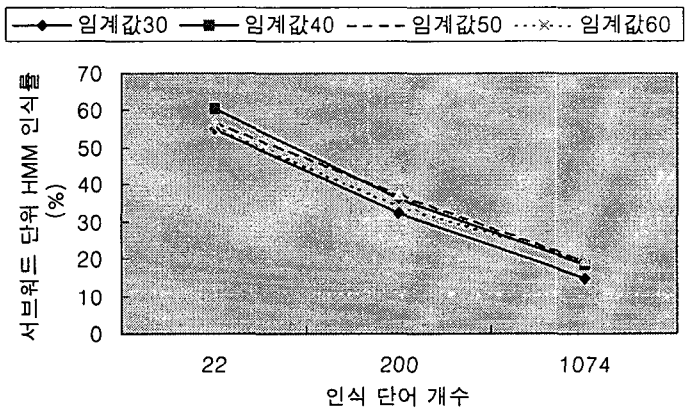

그림 3. 단어 개수에 따른 HMM 인식률 비교

6. 결 론

본 논문에서는 서브워드 단위의 립리딩에 대하여 실험 결과를 설명하였다. 서브워드 단위

의 립리딩을 위하여 본 연구에서는 한국어 viseme을 정의하였으며, HTK를 이용하기 위하여 질의어 셋을 정의하였다. 실험결과 서브워드 단위 립리딩이 단어단위 인식에 비하여 성능저하가 크지 않음을 확인하였다. 또한 립리딩 기술이 단어수가 작은 경우에는 유효하지만, 단어수가 중규모 이상으로 증가할 경우, 실제 HCI에 이용되기 위해서는, 많은 기술적 향상이 필요함을 확인하였다.

참 고 문 헌

[1] Sharma, R., Vladimir I. Pavlovic. & Thomas S. Huang. 1998. "Toward Multimodal Human-Computer Interface." *Proceedings of the IEEE,* Vol. 86 No. 5, 853-869.

[2] 김진범, 김진영. 2000. "입술의 대칭성에 기반한 효율적인 립리딩 방법." *대한전자공학회 논문집,* 제 37권 5호.

[3] Potamianos, G., H. P. Graf. & E. Cosatto. 1998. "An image transform approach for HMM based automatic lipreading" *Proceeding of International Conference on Image Processing,* Vol. 3, 173-177.

[4] Rabiner, Lawrence & B. H. Juang. 1993. "*Fundamentals of Speech Recognition*" published by PTR Prentice-Hall, Inc. 321-389.

[5] 민덕수, 김진영. 1999. "Lipreading에 기반을 둔 HMM을 이용한 단어 인식." *신호처리 합동학술대회, 한국음향학회 발표*

[6] Liëvin, M & F. Luthon. 1998. "Lip features automatic extraction." *Proceedings of the 5th IEEE International Conference on Image Processing.* Chicago. Illinois.

[7] 박병구, 김진영, 임재열. 1999. "입술 파라미터 선정에 따른 바이모달 음성인식 성능 비교 및 검증." *한국음향학회지,* 제 18권, 제 3호, 68-72.

[8] 박병구, 김진영, 최승호. 1999. "바이모달 음성인식의 음성정보와 입술정보 결합방법 비교." *한국음향학회지* 제 18권, 제 4호, 31-37.

[9] Fisher, G. G. 1968. "Confusions among visually perceived consonants." *Journal of Speech & Hearing Research",* 472-482.

[10] 정수경. 2001. *코퍼스 기반의 음성/영상 동기화 알고리즘의 개발.* 전남대학교 전자공학과 석사학위 청구 논문, 21-26.

[11] 민덕수. 2001. *동적환경에서 립리딩 성능저하 요인 분석 및 인식성능 향상에 관한 연구.*" 전남대학교 전자공학과 석사학위 청구 논문, 21-26.

접수일자: 2001. 7. 10.
게재결정: 2001. 9. 6

▲ 김진영
광주광역시 북구 용봉동 300번지 (우: 500-757)
전남대학교 전자공학과
Tel: +82-62-530-1757, Fax: +82-62-530-0472
E-mail: kimjin@dsp.chonnam.ac.kr

▲ 신도성
광주광역시 북구 용봉동 300번지 (우: 300-757)
전남대학교 전자공학과 DSP Lab
Tel: +82-62-530-0472, Fax: +82-62-530-0472
E-mail: jesus33@dsp.chonnam.ac.kr