論文
9-1-3

에지 분석에 의한 자동 독화 실험
Automatic Lip Reading Experiment by the Analysis of Edge

이경호(Kyong Ho Lee)[1] 금종주(Jong Ju Kum)[2] 이상범(Sang Bum Rhee)[3]

ABSTRACT

In this paper, the edge parameters were drawn from speaking image around lip and effective automatic lip reading system to recognize the Korean 'a/e/i/o/u' 5 vowels were constructed using the parameter.

Speaking images around lip were divided into 5x5 pane. In each pane the number of digital edge element using Sobel operator were evaluated. The observational error between samples was corrected by using normalization method and the normalized value is used for parameter.

In the experiment to convince the strength of parameter, 50 normal persons were sampled. The images of 10 persons were analyzed and the images of another 40 persons were experimented for recognition. 500 data are gathered and analyzed. Based on this analysis, the neural net system is constructed and the recognition experiments are performed for 400 data. The neural net system gave the best recognition result of 91.1 %.

요약

본 논문에서는 입술 주위 영상만으로 독화를 위한 에지 파라미터를 추출하였고, 한국어 5모음 'ㅏ/ㅔ/ㅣ/ㅗ/ㅜ'를 인식하는데 효과적임을 보였다.

발화하는 입주위의 이미지를 5x5로 나누고, 각 영역에 소벨 연산자를 적용하여 디지털 에지 수를 구한 후, 이 값들의 관찰 오차를 정규화를 통하여 수정하고, 정규화된 값을 파라미터로 사용하였다. 파라미터의 견인성을 확인하기 위하여, 자동 독화 시스템을 구축하였다. 인식 실험에 정상인 50명이 동원되었고, 10명의 이미지로 분석하고, 다른 40명의 이미지로 인식 실험을 하였다. 500개의 데이터를 분석하고, 이 분석을 바탕으로 신경망 시스템을 구축하였으며, 400개의 데이터로 인식 실험하였다. 신경망 시스템의 최고 인식 결과는 91.1%였다.

key word : automatic lip reading

논문 접수 : 2008. 1. 25.
심사 완료 : 2008. 2. 5.

1) 정회원 : 한라대학교
2) 정회원 : 국방품질관리원
3) 종신회원 : 단국대학교 대학원 컴퓨터응용 교수

1. 서 론

최근 음성 인식 분야에서는 심한 잡음 환경에서도 높은 인식률을 갖게 하기위한 연구가 활발히 진행되고 있다. 실험실과 같이 거의 잡음이 없는 환경에서는 높은 인식률을 보이나 소음이 많은 환경에서는 인식 성능이 많이 저하되고 있기 때문이다.

자동독화(Lip-reading)는 음성인식 분야 중 잡음 환경에서 현저하게 떨어지는 인식률을 높이기 위한 보상 방법의 하나로, 시각적 관점에서 화상을 이용하여 조음 현상을 반영하는 입술 주위로부터 얻은 정보를 추가하여 인식률을 높이고자 하는 연구이다. [1-6]

발화라는 것이 무한의 자유도를 갖지만 한 개의 언어에서 사용되는 모음의 종류가 유한이고, 이산적인 언어 정보를 전달하고 있음은 모음 생성에 관여하는 제어도 비교적 소수의 이산적이고 단계적이라는 것이다. [12]에서 한국어 모음이 조음 차원에서 제어의 이산성을 가짐을 명확히 하였고, 발화를 위한 턱의 열림각의 제어가 3단계, 혀의 제어가 3단계, 입술의 오므림의 유무를 2단계로 분석하였다. 즉 조음 영상은 언어정보의 이산성을 반영한다는 것을 확실히 하였다. [17]에서 역시 시스템을 구축하여 정규화한 데이터로 발화 제어에 이산성이 있음을 보였다.

우리말의 시각적 관점에서의 인식 연구는 미미한 편이나, 입술 영역 검출 및 입술 형태 표현은 많은 연구가 되고 있는 편이다. 자동 독화에 관한 연구로는 [14-16], [18] 등이 있으며, [14], [15]는 입술 특징 파라미터와 음성 파라미터를 구별 식별 방법으로 통합하여 84%의 인식률을 얻은 결과를 보이고 있다. [16]은 입술 영상 한 프레임을 주성분 분석을 통한 특징계수를 구하여 독화만으로 40%의 인식률을 얻고 있다. [17]은 입주위에 설정한 관찰점간의 거리 변화 정규화 데이터를 이용하여 87.44%의 인식률을 얻고 있다. [12]는 납 조각을 관심영역에 붙여 X선을 이용한 장치로 조음 제어를 조사한 것이므로, 자동 독화 장치로는 적합하지 않으며, [14], [15]에서는 1명을 대상으로 인식 실험하였고, [16]너무 낮은 인식 결과를 보이고 있다.

본 연구에서는 발화를 위한 입술의 벌림과 오무림, 혀의 제어, 이의 보임 유무를 정보로 이용하고자 하였다. 컴퓨에 연결한 카메라를 이용하여 주로 얼굴 잡히는 영상을 이용하여, [17]을 응용한 방법으로 입주위의 영상을 획득하고, 획득한 영상에서 조음 시 제어의 의해 시각적으로 관찰되는 입술의 오무림과 벌림, 혀와 이의 보임 유무를 정보로 이용하고자 하였다. 취득한 영상에서 추출한 입 주위의 영상을 5x5로 나누고, 영역별 측정되는 디지털 외각선의 길이를 추출하여 이를 계수화 하고, 분포를 분석하고, 이 분석을 바탕으로 신경망 시스템을 구축하고 인식하였다.

실험의 결과는 트레이닝한 환경과 인식 대상이 같을 경우 96.9% 매우 높은 인식률을 보였으며, 이중에서는 '아'와 '에'의 인식률이 많이 저하되었다. 인식 대상이 다르며 환경도 다른 경우 인식 실험에서는 좀 저하되기는 하였으나 91.1%의 인식률을 보였다.

또한 음향을 이용한 음성 인식에서는 '오/우'에서 오류가 많음[13]에 비하여 화상을 이용한 본 인식 실험에서는 '오/우'에서 거의 오류가 없음을 확인하여, 음향만을 이용한 음성 인식기와 상호 보완성이 강함을 확인하였고, 두 시스템을 통합하면 우리말인식기에서도 매우 좋은 결과를 얻을 수 있다는 개연성을 확인하였다.

2. 시스템 구축을 위한 자료 수집 및 분석

음성 인식을 위한 음향 관점에서 특징 파라메터를 잡는 방법은 모음과 유성 자음에만 존재하여 성도의 공진 특성을 반영하는 포먼트와

에너지, 시간 영역에서의 피치, LPC(Linear prediction coefficient), PARCOR(Partial auto corelation), 켑스트럼 등이 있으며, 시각적 관점에서 입술 정보를 이용하는 방법은 입술 형태를 모델링 하는 방법과 입술 주위의 관심 영역을 설정하여 이들의 변화를 계수화하여 이용하는 방법이 있다.[4-9], [11], [14], [15]

본 논문에서는 조음 시 제어의 의해 시각적으로 관찰되는 입술의 오무림과 벌림, 이의 보임 유무를 정보로 이용하고자, 영상에서 외각선의 길이를 추출 이용하였으며, 입 주위의 영상을 5x5로 나누어 영역별 추출되는 외각선의 길이를 계수화 하고 정규화 하여 분포를 분석하고 이 분석을 바탕으로 신경망 시스템을 구축하고 인식하는 실험을 하였다.

[17]을 기반으로 관심 영역의 인식은 YCbCr의 기반의 영상에서 얼굴 후보 영역을 검출하고, 후보 영역에서 색차와 휘도 영역에서의 특징 추출을 통하여 eye-map을 통하여 눈을 추출하고, 얼굴 후보 마스크를 이용한 필터링을 통하여 mouth-map을 통하여 입을 추출하여 이들의 존재를 바탕으로 얼굴 영역을 확정하며, 찾은 눈과 입의 좌표를 통하여 눈과 눈의 중점과 입 사이의 휘도 성분을 이용하여 입의 주위의 범위를 결정하였다.[8], [10-11], [17].

2.1 이미지 획득 환경

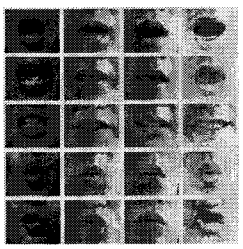
[17]을 참고하여 구성한 이미지 획득 환경에서 관심 영역 이미지의 추출은 조명과 같은 다양한 환경에서 상당한 오차를 발생한다고 보고하고 있다. 이를 바탕으로 본 연구에서 이미지 획득을 위하여 구성한 기본적 실험 환경은 발화자의 앞쪽 위 약 70도 상에 조명이 있었으며, 약 1.5미터 앞에 발화자보다 약간 높게 카메라를 장착하였고, 획득된 화상은 발화시 코끝에서 턱끝까지가 150픽셀 정도가 되도록 충분히 크게 하였다.

2.2 관심 영역의 획득

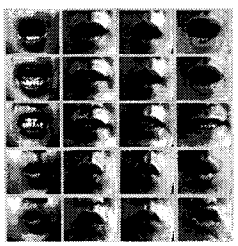
관심 영역의 획득은 YCbCr의 기반의 영상에서 얼굴 후보 영역을 검출하고, 후보 영역에서 색차와 휘도 영역에서의 특징 추출을 통하여 eye-map을 통하여 눈을 추출하고, 얼굴 후보 마스크를 이용한 필터링을 통하여 mouth-map을 통하여 입을 추출하여, 찾은 눈과 입의 좌표를 통하여 휘도 성분의 이미지에서 입의 중점 상하 80픽셀과 좌우 100픽셀의 이미지를 얻어냈다. [8], [10-11], [17]. 본 실험에서 관심 이미지 습득에 아래 (그림1)에 보이는 습득 이미지와 같이 조명과 습득 환경을 조금 변경 해 보았으며, 얼굴이 대부분인 영상에서 눈과 입을 찾기 때문에 그다지 어려움이 없었다.

2.3 관심 영역의 처리 및 분석

습득한 이미지는 빛과 배경에 의하여 발생하는 에러를 줄이기 위하여 평균 100 분산 80을 기준으로 이미지를 정규화 하였다.



(그림 32) 습득 이미지

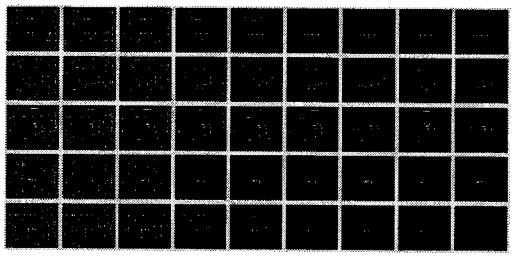


(그림 33) 정규화 이미지

본 연구의 아이디어는 발화 시 입술을 벌리고 오무리며, 혀와 이의 보임 정도를 이용하여 외곽선의 길이를 구하고, 이렇게 구한 입 주위 이미지의 외곽선 길이가 영역별로 분포한 것을 분석하면 어떤 정보를 제공해 줄 것이라는 믿

음으로 실험한 논문이다.

본 실험에서 외곽선의 길이를 구하기 위하여 사용한 방법은 소벨 연산자이며, 발화별 안정적 디지털 외각선의 길이를 확보하기 위하여 소벨 연산자의 경계치(threshold)를 0.00에서 0.16까지 0.01씩 증가시켜가며 외각선 길이에 상응하는 외각선 변환에 적용된 회수를 조사하였다.



(그림 34) 소벨 연산자 경계치 0.00~0.16 적용이미지.

이 조사에는 앞에서 언급한 기본 환경 중심으로 하였으나 추가로 기본 조명에 좌, 우, 전 조명을 추가한 환경 4가지 환경을 집단으로 하여 외각선 변환 값을 추출하였으며, 각 집단은 10명의 정상 20~40대 사람들을 무작위로 추출하여 조사하였다. 발화는 개인별 분석을 위한 것 말고는 '아', '에', '이', '오', '우'를 각각 20회씩 발화 하도록 하였고, 1000개의 데이터 중 발화별 각각 100개씩 추출하여 분석하였다. 또한 각 음절과 음절 사이 무음의 지속은 최소한 1초 이상이 되게 하였으며, 발화 시간도 1초 이상 지속되게 하여 화상에서 발음이 가장 정상적으로 진행되는 곳을 얻을 수 있도록 하였다.

본 연구에서 중요 관점은 외각선의 길이가 각 발화와 관련하여 의미를 갖는가 하는 것으로 (표 3)은 이를 보이는 한 증거이다.
표에서 보는 바와 같이 개인별 환경별 차이 없이 경계치 0.02~0.12범위 내에서 '이'가 가장 지배적인 값으로 확인되었다. 단 이 범위를 벗어나면 '이'가 지배적인 값을 갖는 것이 보장되지 않았다.

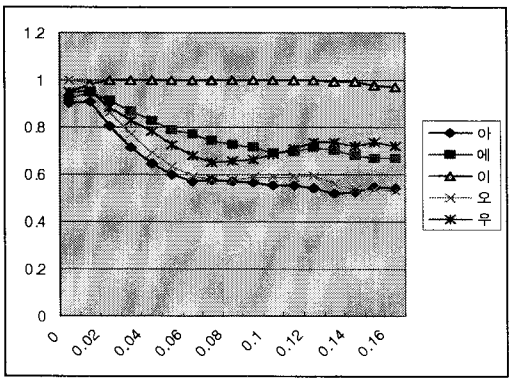
측정된 값은 동일인이더라도 환경이 다른 경우에서 경계치가 같을 때 최대 26.7%까지 길이의 오차를 보였으며, 경계치가 0.02 또는 0.12와 같이 경계선 쪽으로 갈수록 오차가 심했고 중심으로는 오차가 비교적 적었다. 또 0.02 쪽 보다는 0.12쪽이 더 심했고, 0.06~0.10 사이는 오차가 10% 이하였다.

다른 조명 환경 다른 사람일 경우 68.9%까지 길이의 오차가 있었으나 앞에서와 마찬가지로 경계선 쪽이 심했고 내부로 갈수록 오차가 줄어들었으나 최대 53.2%까지 오차를 보였다. 또한 조명과 관련하여 밝고 그림자가 적을수록 외각선의 길이가 길었다.

경계치		00	02	04	06	08	10	12	14	16
A	아	1956	1804	1317	988	578	378	267	202	160
	에	2094	1985	1640	1342	817	608	491	394	327
	이	2226	2146	1775	1546	967	705	523	387	308
	오	2266	1958	1339	1037	517	322	212	146	113
	우	2216	1970	1503	1176	584	371	250	158	106
B	아	2160	1675	905	669	376	209	136	94	69
	에	2095	1774	1119	877	515	277	182	133	86
	이	2069	1780	1208	990	636	391	253	175	128
	오	2236	1580	780	570	339	209	134	97	66
	우	2067	1653	922	683	411	274	201	153	113
C	아	2051	1195	478	366	242	166	103	64	51
	에	2039	1363	632	476	255	161	109	79	52
	이	2108	1619	935	716	412	250	167	124	85
	오	2248	1206	553	398	260	185	146	107	79
	우	2161	1287	620	446	251	169	136	120	87
D	아	1918	1330	673	524	322	212	158	120	97
	에	2006	1669	903	673	425	253	139	96	70
	이	2088	1862	1143	902	607	400	290	232	169
	오	2181	1681	886	662	372	247	161	106	63
	우	2046	1671	971	745	454	302	219	165	126

(표 14) 소벨 연산자 경계치 0.00~0.16 범위에서 변환 횟수 분포

2.4 획득 정보 이용 방법의 모색

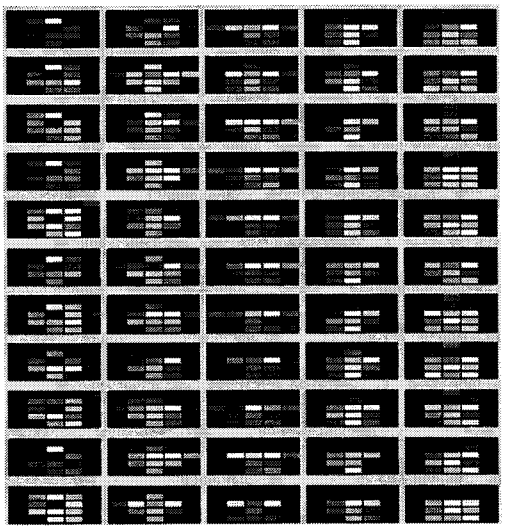


(그림 35) 외각선 길이 비율 변환표

(그림 6)은 앞에서 취득한 데이터를 이용한 5 모음 중 디지털 외각선의 최장 길이를 1로 하였을 때 상대 길이들의 평균을 이용하여 작성한 그래프이다. 그림에서 보이는 바와 같이 경계치 0.02~0.12에서는 실험에 평가된 모든 데이터에서 '이'는 항상 1위로 가장 긴 값을 보였다. 특히 경계치0.06~0.10에서는 다른 값들과 현저한 차이를 보이고 있다. '이'만을 구별할 경우 경계치0.06~0.10가 좋은 것으로 사료되나 다른 모음들을 고려한다면 경계치 0.02~0.06 범위에서 획득하는 것이 좋은 것으로 판단되었다.

그러나 동일인일 경우라도 환경이 다르면 외각선으로 변환하는 횟수가 차이가 있고 다른 사람 다른 환경일 경우를 특정 경계치 범위라도 많은 오차를 보여 바로 사용하기가 어려울 것으로 판단되었다.

따라서 본 실험에서는 정규화를 모색하였고 100x80의 크기로 취득한 발화 입주위 이미지를 5x5의 영역으로 나누고, 각 영역별로 경계치 0.05을 기준으로 외각선 변환에 참여하는 비율을 구하고, 이를 평균 0.5, 분산 0.4로 정규화하여 이 값을 데이터로 이용하였다. (그림 4)는 정규화한 발화 이미지 외각선 변환 비율 분포를 이용하여 이미지화한 것이다.



아 에 이 오 우

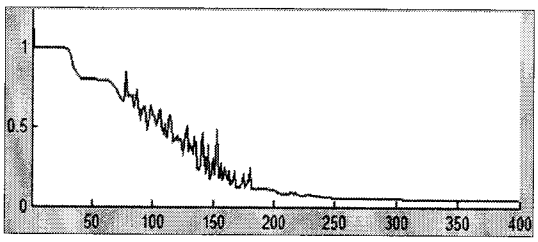
(그림 36) 외각선 분포 이미지

위에서부터 앞의 6개는 1사람이 약간의 추가 조명의 변화를 준 다른 환경에서 발음한 것을 이용하여 얻은 값을 밝기로 변환하여 이미지화한 것이고, 나머지는 각각 다른 사람으로 기본 조명에 전 조명 1개의 이미지와 좌우로 추가 조명을 하여 얻은 각각 2개의 이미지에서 추출한 값을 이용하여 얻은 이미지이다. 직관적으로도 그림에서 보는 바와 같이 시각적으로도 어느 정도 모음별 분포의 유사성이 보인다.

이는 발화시 입의 벌림과 오무림, 이의 관찰 유무가 외곽선의 길이에 영향을 주었으며, 이 영향은 시스템 구축에 이용할 수 있는 가능성을 보여준 것이다.

3. 인식을 위한 시스템 구축 및 실험

앞에서 분석한 자료를 바탕으로 신경망을 구성하여 인식하는 실험을 하였다. 구축된 Multi Layer Perceptron은 입력은 25, 히든은 10, 출력은 5 총 40유닛으로 구성하였다. 학습율은 0.75, 시그모이드 함수 곡선 기울기에 반영되어 수렴을 조절할 모멘텀은 7, 학습패턴은 분석을 위하여 준비한 5모음 데이터 중 무작위로 각각 55개를 추출하였으며, 학습은 350회 또는 에러 값 경계치 0.0001이하까지로 하였다. 학습은 최적치라고 할 수 없으며, 데이터 수를 5개씩, 학습 회수는 50회씩 늘려 나가며, 비교적 좋은 값에서 멈춘 것이다.



(그림 37) 평균 오차값의 변화

인식실험은 2가지 방법으로 수행하였다. 첫 번째는 5명의 데이터로 학습하고 동일 환경 동일 집단에 의해서 만들어진 데이터를 인식하는 실험을 하였으며, 두 번째는 조명을 약간 변화시킨 10명으로부터 얻은 데이터를 이용하여 학습하고 40명으로부터 얻은 데이터로 인식 실험을 하였다.

	인식 여부	실험 1	환경 2	비교
모음	인식	97.3	91.1	
아	인식	90.1	88.9	정상 데이터 인식
	미인식	9.9	11.1	정상데이터 미인식. 주로 '에'로 인식
	오인식	1.9	8.3	오 데이터 인식
에	인식	97.1	80.6	
	미인식	2.9	19.4	'이'와 '우'로 인식
	오인식	7.6	11.2	주로 '아'를 '에'로 인식함
이	인식	99.1	97.2	
	미인식	0.9	2.8	
	오인식	1.0	11.1	주로 '에'를 이로 인식
오	인식	100	94.4	
	미인식	0.0	5.6	
	오인식	0.0	0	
우	인식	100	94.3	
	미인식	0.0	5.7	
	오인식	0.0	2.8	주로 '에'와 '아'를 오인식

*. 환경1 : 동일환경 동일집단. 환경2 : 이질환경 이질집단

(표 15) 인식 실험 결과

(표 5)에서 보는 바와 같이 실험1의 동일 환경 동일 집단일 경우 인식률이 매우 높으며, 실험2 이질 환경 이질 집단일 경우 인식률은 떨어지나 비교적 높은 인식률을 가지고 있었다. 실험1에서는 '아'의 인식률이 많이 저하되어 있으며 특히 '아'임에도 불구하고 인식하지 못하는 현상이 심했다. 또 '에'는 '아'를 많이 오인식 하였다. 실험2에서는 '아/에'의 인식률이 많이 저하되어 있었으며, '아'는 주로 에로 인식하고, '에'는 '이/우'로 인식되었으며, '이'는 '에'를 오인식하였다.

4. 결론

본 논문에서는 입술 주위 영상만으로 독화 파라미터를 추출하여 한국어 5모음을 인식하는 효과적인 방법을 보였다.

인식 실험의 결과에서 보는 바와 같이 분할 이미지 에지 분석에 의한 파라미터의 견고성이 확인되었고, 동일 환경 동일 집단의 경우 높은 인식률을 감안한다면 사용 직전에 훈련하여 사용하는 시스템에 의미가 있을 것으로 사료된다. 또한 음성 인식에서는 '오/우'에서 오류가 많음[13]에 비하여 본 데이터에 의하여 구축된 실험에서는 '이/오/우'에서 매우 높은 인식률을 보여 음향과 화상이 상호 보완될 수 있는 개연성을 확인하였다.

참고문헌

[1] Rajeev Sharma, Vladimir I. Pavlovic, Thomas S, Huang, "Toward Mult-imodal Human-Computer Interface", Proceeding of the IEEE Vol 86. No 5. May 1998.

[2] Gerasimos Potamianosm, Hans peter Graf, Eric Cosatto, "An Image Transform Approach for HMM based Automatic LipReading", Proceeding of the Int. Conf. On Image Processing. pp. 173-177, 1998.

[3] C. Bregler and Yochai Konig, "Eigenlips' for Robust Speech Recognition", Proc. IEEE Int. Conf. On Acoustics, Speech and Signal Processing, pp. 669-672, 1994

[4] T. Chen, H. P. Graf, and K. Wang, "Lip-synchronization using speech-assisted video processing", IEEE Signal Processing Lett., Vol 2, pp. 57-59, 1995

[5] Devi Chandramohan, Peter L. Silsbee, "A Multiple Deformable Template Approach for Visual Speech Recognition", Proc. ICSLP, Vol 1, pp. 434-437, 1996

[6] Iain Matthnews, Timothy F.Cootes, J. Andrew Banghan, Stephen Cox and Richard Marvey, "Extraction of Visual Features for Lipreading," IEEE Tans. on Pattern Recognition and Machine Analysis, Vol 24, No. 2, pp. 198-213, Feb 2002.

[7] Prasad, K.V., Stork D.G., Wolff G. Preprocessing video images for neural learning of lipreading. Ricoh California Research Center, Technical Report CRC-TR-93-26. 1993

[8] Watanabe, T. & Kohda, M. "Lip-reading of Japanese vowels using neural networks." 1990

[9] 민소희, 김진영, 최승호, "입술 정보를 이용한 음성 특징 파라메터 추정 및 음성인식 성능 향상", 대한 음성 학회지 : 말소리 제44호, pp 83-92, 2002

[10] Rein-Lien Hsu, Mohamend Abdel-Mottaleb, Anil K. Jain, "Face Detection in Color Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 24, no 5, pp. 696-706, May 2002,

[11] Prasad, K.V., Stork D.G., Wolff G. Preprocessing video images for neural learning of lipreading. Ricoh California Research Center, Technical Report CRC-TR-93-26. 1993

[12] 김부일, 양룡, 이태원, "한국어 모음의 조음적 제어에 관한 연구", 한국정보과학회 논문지 , Vol.14. No.3, pp.194~202, 1987

[13] 조용덕, 김기철, 맹승렬, 조정완, "Multi-Layer Perceptron을 이용한 백색 잡음이 섞인 모음의 인식", 한국정보과학회 가을 학술발표논문집, Vol.16 No.2, pp 629~632, 1989

[14] 백성준, 김진영, "입술정보 및 SFM을 이용한 음성의 음질향상알고리듬" 음성과학 제10권 제2호, pp. 77-84 , 2003. 6.

[15] 김진영, 민소희, 최승호, "음성인식에서 입술 파라미터 열화에 따른 견인성 연구", 음성과학, 제10권2호 , pp. 27-33 , 2003

[16] 이은숙, 이호근, 이지근, 김봉완, 이상설,

이용주, 정성태, "견고한 입술 영역 추출을 이용한 립리딩 시스템 설계 및 구현", 한국멀티미디어학회, 춘계학술논문발표집, pp. 524~527, 2003
[17] 한태규, 이상범, "칼라기반 정지영상에서의 다중 얼굴 추출", 단국대학교 대학원 석사학위논문, 2006
[18] 이경호, 금종주, 이상범, "한국어 5모음의 조음적 제어 분석을 이용한 자동 독화에 관한 연구", 한국컴퓨터산업교육학회 논문지, Vol 8., No 4., pp. 281~288, 2007