한국음향학회지 제20권 제6호 pp. 51~56 (2001)

강인한 음성인식을 위한 이중모드 센서의 결합방식에 관한 연구

A Study on Combining Bimodal Sensors for Robust Speech Recognition

이 철 우*, 계 영 철*, 고 인 선*
(Chul Woo Lee*, Young Chul Kay*, Inseon Koh*)

*홍익대학교 전자공학과

(접수일자: 2001년 6월 18일; 채택일자: 2001년 7월 23일)

최근 잡음이 심한 환경에서 음성인식을 신뢰성있게 하기 위하여 입모양의 움직임과 음성을 같이 사용하는 방법이 활발히 연구되고 있다. 본 논문에서도 이러한 목적으로 영상언어인식기와 음성인식기의 결과에 각각 가중치를 주어 결합하는 방법을 제안한다. 특히 가중치를 입력음성의 잡음의 정도에 따라 자동적으로 결정하는 방법을 제안한다. 가중치의 결정을 위하여 입력샘플간의 상관도와 LPC 분석의 잔여 오차를 이용한다. 모의실험 결과, 이런 방식으로 결합된 인식기는 잡음이 심한 환경에서도 약 83%의 인식성능을 보이고 있다.

핵심용어: 음성인식, 영상언어인식, 이중모드 센서, 가중치 결합, 입모양 특징 파라메터
투고분야: 음성처리 분야 (2,5)

Recent researches have been focusing on jointly using lip motions and speech for reliable speech recognitions in noisy environments. To this end, this paper proposes the method of combining the visual speech recognizer and the conventional speech recognizer with each output properly weighted. In particular, we propose the method of autonomously determining the weights, depending on the amounts of noise in the speech. The correlations between adjacent speech samples and the residual errors of the LPC analysis are used for this determination. Simulation results show that the speech recognizer combined in this way provides the recognition performance of 83 % even in severely noisy environments.

Keywords: *Speech recognition, Visual speech recognition, Bimodal sensor, Weighted combination, Articulatory parameter*
ASK subject classification: *Speech signal processing (2,5)*

I. 서 론

최근 들어 사회가 점차 멀티미디어화함에 따라 인간과 기계의 인터페이스 (man-machine interface)를 좀 더 간편하고 정확하게 실현하기 위하여 얼굴 표정이나 방향, 응시 추적, 손동작 그리고 음성등을 이용한 멀티모달 (Multimodal) 형태의 인식연구나 이의 상용화가 점차 활발하게 되었다[1]. 음성인식의 경우에 있어서도 이의 상용화가 제대로 이루어지기 위해서는 인식기의 정확도와 주변환경의 영향을 극복할 수 있는 강인성 (robustness)이 무엇보다도 절실히 요구되고 있으나, 보상 (compensation)을 이용하는 기존의 인식기들은 성능면에 있어서 미흡하거나 이의 향상을 위하여 상당히 많은 계산량이 요구되고 있는 실정이다. 보상을 통하여 얻을 수 있는 성능향상의 한계성 때문에 새로운 방법들이 시도되고 있으며, 이 새로운 방법들 중의 하나는 사람이 잡음이 심한 환경에서도 음성을 인식할 수 있는 방법을 응용한 것이다. 즉, 사람이 서로 얼굴을 대하고 대화를 할 경우, 음성 그 자체뿐만 아니라 상대방의 얼굴표정 그리고 입술 모양 등이

책임저자: 계영철 (yckay@wow.hongik.ac.kr)
121-791 서울시 마포구 상수동 72-1
홍익대학교 전자공학과
(전화: 02-320-1604; 팩스: 02-320-1119)

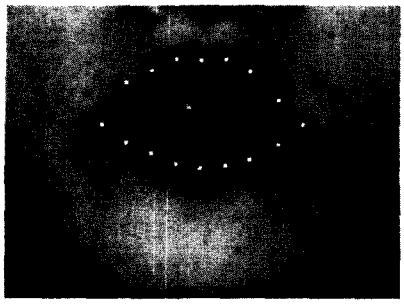

그림 1. 16개의 입술 경계점 추출
Fig. 1. Extraction of 16 lip boundary points.

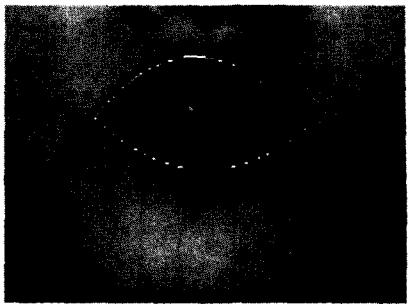

그림 2. 입술 경계를 포물선으로 모델링
Fig. 2. Modeling lip boundaries with parabolas.

종합적으로 음성의 인식에 영향을 미치고 있다. 이와같은 사실을 기반으로 하여 외국에서는 음성인식의 성능향상을 위하여 음성정보와 영상언어 (visual speech) 정보를 동시 결합하여 사용하는 새로운 인식 방법이 활발하게 소개가 되고 있다. 하지만 성능의 향상만이 발표되었을 뿐 체계적인 분석이나 최적의 결합방법들에 관하여서는 아직 연구가 되어있지 않은 상태이다.

따라서 본 논문에서는 영상언어인식 결과와 음성인식의 결과를 효과적으로 결합하는 방법을 제안한다. 음성정보에 잡음이 많이 들어갈수록 샘플간 상관성이 적어진다는 점과, 음성인식에 널리 사용되는 LPC분석이 잡음에 매우 민감하여 잡음이 많을수록 예측값과 실제값과의 오차가 커진다는 점에 착안하여, 잡음의 크기를 예측한다. 그리고 이를 이용하여 음성과 영상언어의 인식결과의 결합에 필요한 가중치를 자동적으로 조절한다.

II. 본 론

2.1. 영상 특징 추출 알고리듬

2.1.1. 입술영역 판별

입술영역 판별부분은 들어온 입력영상에서 입술부분만을 구별해 내는 단계이다. 입력영상의 각 픽셀에 들어있는 Red 와 Green 성분을 이용하여 주어진 임계값 안에 들어오는지를 조사한다.

$$L_{down} < \frac{R}{G} < L_{up} \quad (1)$$

일반적으로, 얼굴부분은 L_{down}=1.2, L_{up}=1.45 정도의 값을 가지고 있고, 입술부분은 L_{down}=1.7, L_{up}=2.0 정도의 R/G의 비율을 가지고 있다[2].

2.1.2. 경계점 추출

앞의 과정을 거쳐 나온 영상정보를 이용해 입술의 양끝을 나타내는 경계점을 찾고 이것을 기준점으로 사용하여 수평 등간격으로 나누어가면서 윗입술과 아랫입술에서 각각 7개씩의 경계점을 추출하여 특징좌표를 얻게 된다[3][그림 1]. 이러한 16개의 특징좌표들이 입모양 특징 파라메터를 추출하는 방법에 적용되었다.

2.1.3. 입모양 특징 파라메터의 추출

입술의 바깥쪽 경계선을 포물선으로 모델링하는 방법으로[그림 2], 앞에서 추출한 경계점 좌표 (x, y)를 이용하여 윗입술과 아랫입술에 대한 각각의 포물선 함수식 ax^2+bx+c를 찾는다. 식 (2)를 이용하여 Mean Spuare Error, $E[y-(ax^2+bx+c)]^2$ 가 최소가 되도록 하는 a, b, c를 찾는다[4].

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} E[x^4] & E[x^3] & E[x^2] \\ E[x^3] & E[x^2] & E[x] \\ E[x^2] & E[x] & 1 \end{bmatrix}^{-1} \begin{bmatrix} E[x^2y] \\ E[xy] \\ E[y] \end{bmatrix} \quad (2)$$

이러한 방법으로 구한 a, b, c를 이용하여 포물선으로 모델링한 입술의 폭, 높이, 면적을 구하여 입력 영상의 특징 벡터로 사용한다.

2.2. 음성 특징 추출 알고리듬

2.2.1. LPC (Linear Predictive Coding) 켑스트럼 (Cepstrum)

일반적으로 음성 신호는 짧은 구간에서는 선형 예측이 가능하므로 식 (3)과 같이 표현할 수 있다.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) \quad (3)$$

$a_1, a_2, \ldots, a_p$: LPC 계수

p : 과거의 음성 샘플의 개수

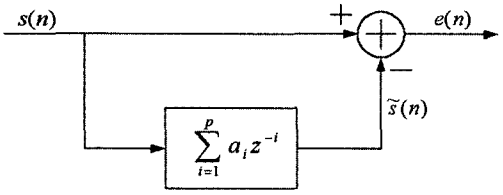

그림 3. LPC분석에 의한 예측 오차
Fig. 3. Prediction error by LPC analysis.

과거 음성 샘플의 선형 조합에 의해 추정한 값을 $\tilde{s}(n)$이라 하고 식 (4)와 같이 정의하면, 실제값과 예측값의 오차 $e(n)$은 식 (5)와 같이 표현될 수 있으며[그림 3], 한 프레임 동안의 오차 $e(n)$의 MSE (Mean Square Error)가 최소가 되도록 하는 LPC 계수를 구한다[5,6].

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{4}$$

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{5}$$

이렇게 얻은 LPC 계수를 직접 인식기에 입력할 수도

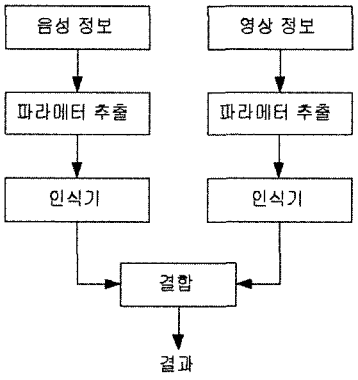

그림 4. 독립적인 인식 결과의 결합
Fig. 4. Combination of the respective recognition scores.

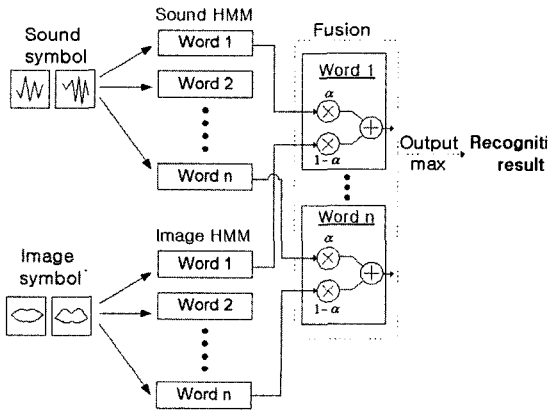

그림 5. 전체적인 인식 시스템 구성도
Fig. 5. Schmetic diagram of the entire recognition system.

있지만, 더욱더 나은 인식률을 얻기 위해서, 음성인식 시스템의 입력 파라메터로 가장 널리 쓰이고 있는 LPC-켑스트럼 계수를 음성신호의 특징 파라메터로 사용한다.

2.3. 음성과 영상언어의 인식값 결합

2.3.1 결합방법

음성정보와 영상정보의 결합방법으로서는 그림 4와 같이 음성과 영상언어를 각각 독립적으로 인식한 후, 그 결과에 가중치를 주어 결합하는 방법을 사용한다[7].

2.3.2. 전체적인 인식 시스템

그림 5는 인식 시스템의 전체적인 구성도를 나타내고 있다. 화자로부터 영상정보와 음성정보를 입력받아 각각의 특징 파라메터를 이용하여 HMM (Hidden Markov Model)을 사용해 각각의 인식 score를 구한 후, 영상인식값과 음성인식값에 가중치 (weight)를 주어 최종 score를 구하여 결과를 얻게 된다[8].

이 때 주어지는 가중치는 사용자가 발음하는 음성이 입력될 당시의 잡음의 정도에 따라 변화되어진다. 식 (6)은 가중치를 부과하여 최종 score를 구하는 방법을 나타낸다.

$$S = \alpha S_a + (1-\alpha) S_v \tag{6}$$

S_a : 음성정보에 의한 인식값
S_v : 영상정보에 의한 인식값
α : 음성 가중치
$1-\alpha$: 영상 가중치

음성 가중치 α는 0 에서 1 사이의 값을 가지며, 환경 잡음이 적을수록 1에 가까워지고 잡음이 심해질수록 0에 가까운 값을 가지게 된다.

2.3.3. 제안한 인식값 결합 방법

(1) 예측값과 실제값과의 오차 이용

실제신호를 $s(n)$이라 하고, LPC로 예측한 신호를 $\tilde{s}(n)$이라고 하면, 실제신호와 예측된 신호와의 오차 e(n)의 MSE (Mean Square Error)는 식 (7)과 같이 표현할 수 있다.

$$E_{MSE} = \frac{1}{M} \sum_{m=1}^{M} \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2 \tag{7}$$

선형 예측방법이 잡음에 민감하므로 음성신호에 잡음이 많이 들어갈수록 예측값이 부정확하여 E_{MSE}가 커지게 된다.

이러한 사실을 이용하여 테스트 음성이 들어오면 일정 구간동안의 E_{MSE}를 측정하여 잡음정도를 예측할 수 있다. E_{MSE}가 클수록 잡음이 많은 것이므로 인식값 결합시 음성 가중치 α의 값을 작게 하고 E_{MSE}가 작을수록 α의 값을 크게 한다.

(2) 샘플간의 상관성 이용

잡음이 적은 환경에서 발음된 음성신호의 경우 인접샘플간의 상관성이 크지만 잡음이 많은 환경일수록 그 상관성이 적어져 인접샘플간에도 값의 차이가 크게 나타난다.

$$r_n(m) = s(m) - s(m-n) \tag{8}$$

$$R_n = \frac{1}{M}\sum_{m=1}^{M} r_n^2(m) \tag{9}$$

$r_n(m)$을 n 만큼 떨어진 샘플값간의 차이라고 하면, R_n은 한 샘플 떨어진 샘플값간의 차이를 M 샘플만큼 구하여 제곱의 평균을 취한 것이다.

음성신호에 잡음이 많을수록 R_n이 커진다는 사실을 이용하여 테스트 음성이 들어오면 일정구간동안의 R_n을 측정하여 잡음정도를 예측할 수 있다. R_n이 클수록 잡음이 많은것이므로 인식값 결합시 음성 가중치 α의 값을 작게 하고 R_n이 작을수록 α의 값을 크게 한다.

2.3.4. 가중치 결정 방법

(1) 예측값과 실제값의 오차 이용

(i) 수동으로 가중치를 조절하면서 인식률 실험을 하여 각각의 SNR마다 최적의 인식률을 나타내는 가중치 α_i (i=0, 5, 10,...,clean 은 각각의 SNR을 나타냄.) 를 찾아낸다(그림 8 참조).

(ii) SNR이 i 인 각각의 경우에 대하여, 모든 음성 데이터를 사용하여 일정시간 동안의 E_{MSE}를 측정하고, E_{MSE}의 평균값 $\overline{E_{MSE}}(i)$를 구한다. 그후, SNR i 와 $\overline{E_{MSE}}(i)$를 대응시키는 참조 데이터 표를 만들어 놓는다.

(iii) 가중치를 자동적으로 결정하는 실험 단계에서는 테스트 음성이 들어오면 E_{MSE}를 측정한 다음 (ii)에서 구한 참조 데이터 표를 검색하여 가장 가까운 $\overline{E_{MSE}}(x)$를 찾아낸 뒤, 그에 해당하는 α_x를 인식결과 결합시 사용할 가중치로 결정한다[그림 6].

(2) 상관성 이용

E_{MSE} 와 $\overline{E_{MSE}}(i)$ 를 R_1과 $\overline{R_1}(i)$ 로 대치하여 위의 방법과 동일한 과정을 수행한다.

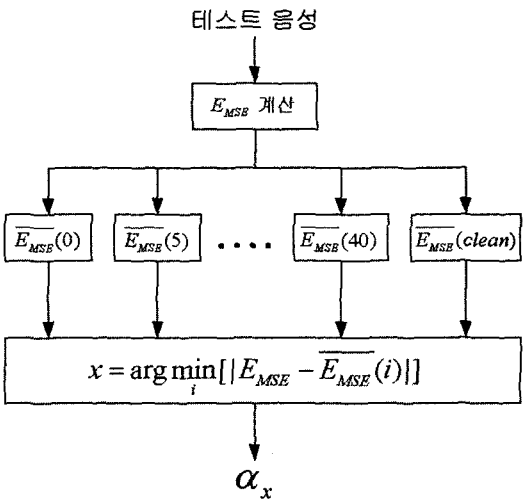

그림 6. 가중치 결정 방법
Fig. 6. Weight decision method.

III. 실험 및 결과

3.1. 모의 실험

실험 데이터로 10명의 화자에 대하여 4자리 숫자 10가지를 발음하게 하였으며, 각각 9번씩 발음하여 5개를 training에 사용하였고 나머지 4개를 이용하여 테스트하였다.

3.1.1. 영상언어 인식

입력 영상은 디지털 캠코더로 촬영한 320×240 pixels, 30 frames/sec, 24-bit RGB color 이미지이며, 앞서 언급한 입술부분의 임계값 L_{down} 은 1.9를 적용하였다. 16개의 입술경계점을 추출하여 입술을 포물선으로 모델링하였으며, 포물선 함수의 계수 a, b, c 를 이용하여 구한 입술의 폭, 높이, 면적을 영상 특징 벡터로 사용하였다. 인식 알고리듬은 코드북 사이즈 128, state 수 5인 HMM을 이용하였다.

그리고, 발음을 할때마다 입술의 크기와 위치가 달라질 수 있으므로, 발음구간 첫 프레임의 입술의 폭을 기준으로 정규화를 하였다.

3.1.2. 음성 인식

사용된 음성은 실험실 환경에서 16bit Quantization, 16kHz sampling rate로 녹음되었으며, 잡음섞인 음성의

경우는 백색 가우시안 랜덤 잡음을 발생시켜 순수 음성과 혼합하여 사용하였다. 12차 LPC-켑스트럼 계수를 추출하여 음성 특징 벡터로 사용하였으며, 인식 알고리듬은 코드북 사이즈 256, state수 8인 HMM을 이용하였다.

3.1.3. 인식결과의 결합

(1) 예측값과 실제값과의 오차 이용

모든 음성 데이터를 이용하여 10ms동안 (160 samples)의 E_{MSE}[식(7)]가 각각의 신호 대 잡음비 (SNR)에 따라 어느정도의 값을 가지는가를 훈련단계에서 미리 계산할 수 있다.

이 정보를 이용하여 테스트 음성이 들어오면 10ms동안 (160 samples)의 E_{MSE}값을 측정하여 잡음정도를 예측한 후 이를 이용하여 인식결과 결합시 가중치를 조절하도록 하였다.

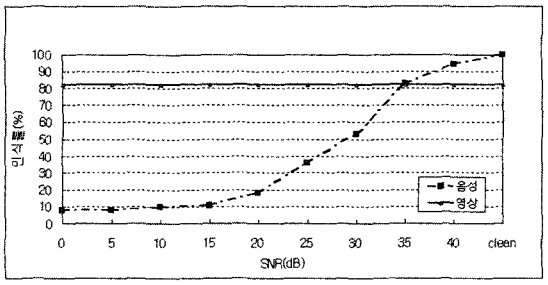

그림 7. 음성과 영상의 독립적인 인식률
Fig. 7. Respective recognition rates of acoustic and visual speeches.

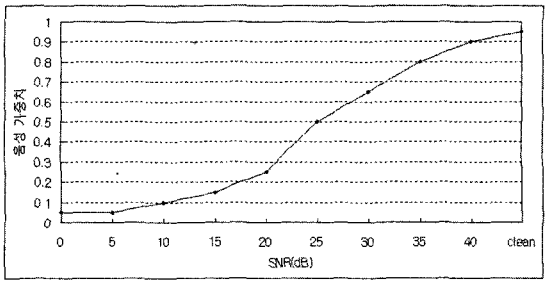

그림 8. SNR에 따른 최적 가중치
Fig. 8. Optimal weights depending on SNR.

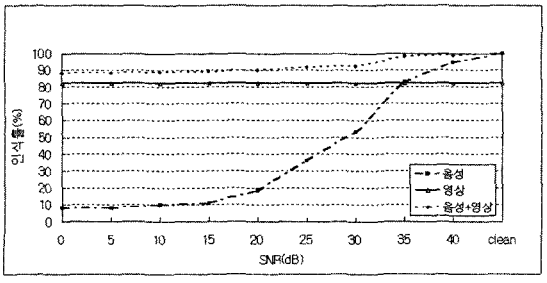

그림 9. 최적 가중치 적용시 인식률
Fig. 9. Recognition rates with optimal weights.

(2) 상관성 이용

위의 방법과 마찬가지로, 모든 음성 데이터를 이용하여 10ms동안 (160 samples)의 R_t[식(9)]이 각각의 신호 대 잡음비 (SNR)에 따라 어느정도의 값을 가지는가를 미리 계산할 수 있다.

이 정보를 이용하여 테스트 음성이 들어오면 10ms동안 (160 samples)의 R_t값을 측정하여 잡음정도를 예측한 후 이를 이용하여 인식결과 결합시 가중치를 조절하도록 하였다.

3.2. 실험 결과

3.2.1. 음성과 영상의 독립적인 인식률

영상정보만을 이용하여 인식하였을 경우에는 잡음에 영향을 받지 않으므로 음성신호의 SNR에 관계없이 82%의 인식률을 나타내었다. 그러나 음성정보만을 이용한 경우에는 잡음에 매우 민감하여 그림 7에서 나타나듯이 30dB이하에서는 인식 성능이 현저하게 저하되는 것을 볼 수 있다.

3.2.2. 최적의 가중치 적용시 인식률

테스트 음성의 SNR정보를 미리 알고있는 상태에서 각각의 SNR 경우마다 전면탐색 (exhaustive search)을 적

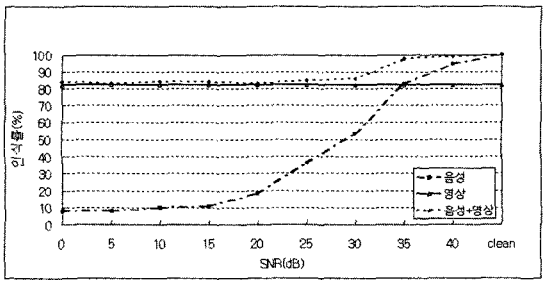

그림 10. 예측오차 방법 이용 결합시 인식률
Fig. 10. Recognition rates using the prediction error method.

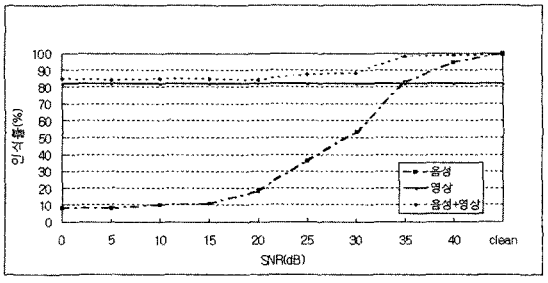

그림 11. 상관성 방법 이용 결합시 인식률
Fig. 11. Recognition rates using the correlation mothod.

용하면 결합인식률이 최대가 되는 최적의 가중치 값을 구할 수 있다.

이렇게 하여 얻은 음성 가중치 α값을 그림 8에 나타내었다. 그리고, 그림 8에 나타난 가중치를 적용하여 음성정보와 영상정보에 의한 인식결과를 결합하였을때의 인식률은 그림 9에서 알 수 있듯이 89%이상의 높은 인식률을 나타내었다.

3.2.3. 제안된 방법에 의한 인식률

제안된 방법에 의한 인식률을 그림 10과 그림 11에 각각 나타내었다. 그림 10은 음성신호의 예측오차를 이용하여 가중치를 조절하는 방법에 의한 인식률을 나타낸 것으로 약 83%이상의 인식률을 나타내었다.

그림 11은 음성신호의 샘플간 상관성을 이용하여 가중치를 조절하는 방법에 의한 인식률을 나타낸 것으로 약 84%이상의 인식률을 나타내었다.

V. 결 론

본 논문에서는 이중모드 (Bimodal) 음성인식에서 음성정보를 이용한 인식결과와 영상정보를 이용한 인식결과의 결합방법에 대한 새로운 알고리듬을 제안하였다. 기존의 방법이 이미 알고있는 음성 SNR에 따라 수동적으로 최적화된 가중치를 부여하는 방법이었는데 비하여, 본 논문에서는 테스트음성의 SNR정보를 모르는 상태에서 음성신호의 예측오차와 샘플간의 상관성을 이용하여 음성에 섞여있는 잡음 정도를 예측한 후 이를 이용하여 자동적으로 가중치를 조절하도록 하였다.

모의 실험 결과 전체적으로 83% 이상의 인식률을 나타내었으며 특히 잡음정도가 심해질수록 음성정보만을 이용했을때의 인식률에 비해 상당한 인식 성능 향상을 볼 수 있었다.

감사의 글

이 연구는 정보통신부에서 지원하는 대학기초연구 지원사업으로 수행되었음 (2001-036-2).

참 고 문 헌

1. Rajeev Sharma, Vladimir I. Pavlovic, Thomas S.Huang, "Toward Multimodal Human-Computer Interface", *Proceedings of the IEEE*, Vol. **86**., No 5., May 1998.
2. T.Wark and S. Sridharan, "A syntactic approach to automatic lip feature extraction for speaker identifi cation", *Proc. of the IEEE.*, 1998.
3. Juergen Luettin, Neil A. Thacker and S.W.Beet, "Locating and Tracking Facial Speech Features", *Proceedings of ICPR'96*, May 1996.
4. Ram R. Rao and Russell M. Mersereau "Lip modeling for Visual speech recognition", Asilomar Conference, Vol. 1 signal, systems and computer, 1994.
5. L. R. Rabiner, and R. W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, 1978.
6. Lawrence Rabiner, and Biing-Hwang Juang, "Fundamentals of speech recognition", *Prentice-Hall.*, 1993.
7. Rudolf Kover, Uirich Harz, Jutta Schiffers, "Fusion of visual and acoustic signals for command-word recongnition", *ICASSP*, 1997.
8. Akio Ogihara, Shinobu Asao "An Isolated Word Speech Recognition Based on Fusion of Visual and Auditory Information Using 30-frame/s and 24-bit Color Image", *IEICE TRANS. FUNDAMENTALS*, Vol. **E80-A**, NO. 8, AUGUST 1997.

저자 약력

● **이 철 우 (Chul Woo Lee)**

1972년 5월 10일생
1998년 2월: 홍익대학교 전자공학과 학사
2000년 3월~현재: 홍익대학교 전자공학과 석사과정 재학 중
※ 주관심분야: 음성인식, 디지탈 신호처리

● **계 영 철 (Young Chul Kay)**

1957년 12월 29일생
1980년 2월: 서울대학교 전자공학과 학사
1982년 2월: 한국과학기술원 전기 및 전자공학과 석사
1991년 5월: Univ. of Southern California, Electrical Eng. Ph.D.
1991년 9월~현재: 홍익대학교 전자전기공학부 부교수
※ 주관심분야: 디지탈 신호처리, 음성 및 영상인식, 로봇 비전

● **고 인 선 (Inseon Koh)**

1955년 7월 31일생
1979년 2월: 서울대학교 전자공학과 졸업 (B.S.)
1987년 5월: Marquette University 졸업 (M.S.)
1991년 5월: Rensselaer Polytechnic Institute, Dept. of ECSE, Ph.D.
1981년~1985년: 대우전자 근무
1991년~1992년: 대우전자 근무
1992년~현재: 홍익대학교 전자전기공학부 부교수
※ 주관심분야: 이산사건 시스템 제어, Computer-Integrated Manufacturing (CIM), Computer Network Analysis, Multimedia, Petri Nets