한국음향학회지 제26권 제1호 pp. 25~31 (2007)

입술움직임 영상신호를 고려한 음성존재 검출

Speech Activity Decision with Lip Movement Image Signals

이 수 종*, 박 준*, 이 영 직**, 김 응 규***
(Soo-jong Lee*, Jun Park*, Youngjik Lee**, Eung-Kyeu Kim***)

한국전자통신연구원 음성/언어정보연구센터**, 음성인터페이스연구팀*
한발대학교 공과대학 정보통신 · 컴퓨터공학부***

(접수일자: 2006년 10월 18일; 수정일자: 2006년 11월 23일 채택일자: 2006년 12월 7일)

본 논문은 음성인식을 위한 음성구간 검출과정에서, 음향에너지 이외에도 화자의 입술움직임 영상신호까지 확인하도록 함으로써, 외부의 음향잡음이 음성인식 대상으로 오인식되는 것을 방지하기 위하여 시도한 것이다. 먼저, PC용 화상카메라를 통하여 영상을 획득하고, 입술움직임 여부가 식별된다. 그리고 입술움직임 영상신호 데이터는 공유메모리에 저장되어 음성인식 프로세스와 공유한다. 한편, 음성인식의 전처리 단계인 음성구간 검출과정에서는 공유메모리에 저장되어 있는 데이터를 확인함으로써 사람의 발성에 의한 음향에너지인지의 여부를 확인하게 된다. 음성인식기와 영상처리기를 연동시켜 실험한 결과, 화상카메라에 대면해서 발성하면 음성인식 결과의 출력까지 정상적으로 진행됨을 확인하였고, 화상카메라에 대면하지 않고 발성하면 음성인식 결과를 출력하지 않는 것을 확인하였다. 이는 음향에너지가 입력되더라도 입술움직임 영상이 확인되지 않으면 음향잡음으로 간주하도록 한 것에 따른 것이다.

핵심용어: 음성구간 검출, 입술움직임 영상신호, 공유메모리

투고분야: 음성처리 분야 (2)

This paper describes an attempt to prevent the external acoustic noise from being misrecognized as the speech recognition target. For this, in the speech activity detection process for the speech recognition, it confirmed besides the acoustic energy to the lip movement image signal of a speaker. First of all, the successive images are obtained through the image camera for PC. The lip movement whether or not is discriminated. And the lip movement image signal data is stored in the shared memory and shares with the recognition process. In the meantime, in the speech activity detection process which is the preprocess phase of the speech recognition, by confirming data stored in the shared memory the acoustic energy whether or not by the speech of a speaker is verified. The speech recognition processor and the image processor were connected and was experimented successfully. Then, it confirmed to be normal progression to the output of the speech recognition result if faced the image camera and spoke. On the other hand, it confirmed not to output of the speech recognition result if did not face the image camera and spoke. That is, if the lip movement image is not identified although the acoustic energy is inputted, it regards as the acoustic noise.

Key words: Speech activity detection, Lip movement image signals, Shared memory

ASK subject classification: Speech Signal Processing (2)

I. 서 론

음성인식 기능은 기본적으로 음향에너지를 분석의 대상으로 한다. 그런데, 음성인식의 실제 서비스 환경에서는 다양한 음향잡음이 존재하고, 특히 동적인 음향잡음이 예고 없이 유입될 수 있어서 실제 서비스를 위해서는 이들을 효과적으로 처리하는 것이 필수불가결한 요건이 되고 있다. 특히, 음성인식 실행시점을 인위적으로 동작시킬 수 없는 Non-PTT (Push to Talk) 형태의 연속음성인식을 위해서는 완벽한 음향잡음 처리 대책을 강구하여야 한다.

한편, 사람은 말할 때 입술을 움직이게 된다. 입술을 움직이지 않고 말할 수 있는 방법은 거의 없다. 또한, 영

책임저자: 이 수 종 (sjleetri@etri.re.kr)
305-350 대전광역시 유성구 가정동 161 한국전자통신연구원
(전화: 042-860-5584; 팩스: 042-860-4889)

상은 음향잡음과는 무관하게 획득되고 처리되므로, 음성인식을 위한 음향에너지 분석과정에서 입술움직임 영상신호의 도입은 음향잡음 처리를 위한 효과적인 대책이 될 수 있다.

이와 같은 영상의 장점을 음성인식에 활용하기 위한 Lipreading에 관한 연구가 꾸준히 이루어지고 있다. 주로 극한 소음환경 하에서 입술모양만으로 음성인식을 시도하는 경우와 입 모양의 구분이 명확한 모음과 일부 단어를 인식하는 수준에 머물고 있는 실정이다 [1][2]. 이는 영상과 음성신호 자체의 속성의 차이와 영상신호 처리량의 증가에 주로 기인하고 있다.

본 논문은 외부의 음향잡음이 음성인식 대상으로 오인식되는 것을 방지하기 위한 방안의 일환으로 이루어졌다. 영상처리와 음성인식이 독립적으로 수행되도록 함으로써 기존의 음성인식 과정에 추가될 수 있는 부하를 최소화하였다. 음성인식의 전처리 단계인 음성구간 검출 과정에서 화자의 입술움직임 영상신호를 확인하는 구조로서, Visual C++로 구현되었다. 음성구간 검출 과정에서 입술움직임 영상신호의 유무에 따라 음성인식 절차로의 진행여부가 결정되도록 한 것이다.

서론에 이어, 제 2 장에서는 시스템 구성에 대해 개략적으로 살펴본다. 제 3 장에서는 영상획득에서부터 영상신호 추출까지의 과정을 세부적으로 살펴본다. 제 4 장에서는 영상 및 음성시스템 연동에 대하여 서술한다. 제 5 장에서는 입술움직임 영상 검출실험 결과와 영상/음성 연동상태에서의 실험결과를 살펴본다. 마지막 6 장에서 결론 및 향후 연구방향을 서술한다.

II. 시스템 개요

(그림 1)은 음성인식 과정에서 영상정보를 활용하는 것을 보여주는 개략도이다. 입술움직임 영상신호의 추출 기능과 음성인식의 모든 과정은 하나의 PC 내에서 이루어지며, 윈도우를 분리하여 인터페이스가 이루어진다. 음성인식을 위해 화자가 발성함에 따라 마이크를 통해서 음향에너지가 수신되어 분석되는데, 동시에 화상카메라를 통해서 얼굴영상 프레임이 연속적으로 획득되고 입술움직임 영상을 검출한다. 입술움직임 영상분석 데이터를 음성인식 과정에서 공유하도록 하기 위하여 공유메모리가 준비된다. 음성인식의 음성구간 검출과정에서는 음향에너지의 분석과 함께 공유메모리로부터 입술움직임 영상신호를 확인하여 음성구간 검출을 수행하고 음성인식 절차로의 진행여부를 결정하게 된다.

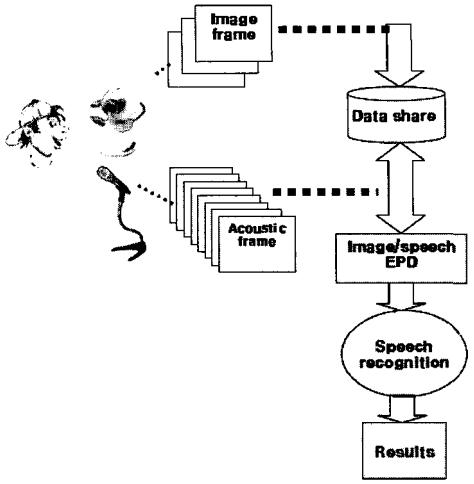

그림 1. 영상과 음성연동 하의 음성인식 개략도
Fig. 1. Speech recognition illustration under combining.

입술움직임 영상신호의 추출을 위해서는 실제 더 많은 처리절차를 거치게 된다. 즉, 영상 프레임간 비교, 잡음영상 제거, 영상분리, 입술움직임 영상 특징추출, 검출, 영상신호로의 변환, 공유메모리에의 저장 등이다. 이들에 대해서는 다음 장에서 상술한다.

한편, 음성인식의 음성구간 검출기능에는 입술움직임 영상신호를 확인하는 절차를 추가하게 된다. 영상신호를 확인하는 기능을 추가함에 있어서는 기능부가로 인하여 추가될 수 있는 부하를 최소화하는 것이 필요하다. 이에 대하여는 입술움직임 영상데이터 확인절차를 포함하여 구동되는 음성/영상 시스템연동 절차를 서술하는 과정에서 간략히 언급하고자 한다.

음성인식 결과의 출력여부를 통하여 시스템의 정상적인 동작여부를 다음과 같이 확인하였다. 즉, 카메라로 하여금 입술움직임을 감지하는 환경에서 발성하는 경우에는 음성인식을 수행한다. 반면에, 카메라로 하여금 입술움직임 영상을 획득하지 못하도록 한 상태에서는 발성한다 해도 인식결과가 출력되지 않아야 한다. 이는 사람이 입술을 움직여 발성된 경우가 아닌, 외부의 음향잡음이 음성으로 오인식되지 않도록 한 것에 따른 것이다.

III. 입술움직임 영상신호 추출절차

입술움직임 영상신호를 음성인식의 음성구간 검출에 활용하기 위해서는, 카메라를 통하여 연속적으로 입력되는 영상 프레임을 대상으로 영상배경, 얼굴 전체의 움직

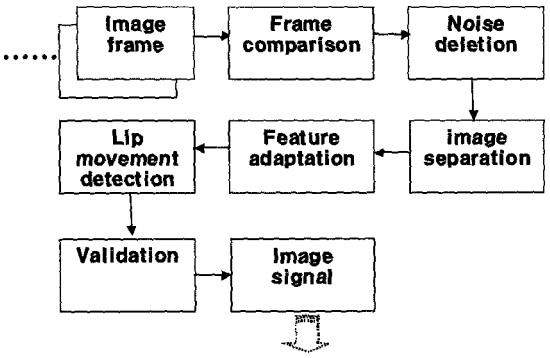

그림 2. 입술움직임 영상신호 추출단계
Fig. 2. Lip movement image signal extraction phase.

임 그리고 얼굴요소의 움직임 영상들로부터 입술움직임 영상만을 분리해내야 한다. 이를 위해 연속되는 인접 영상 프레임을 픽셀단위로 비교하는 것으로부터 시작하여 여러 단계를 거치게 된다. (그림 2)는 영상 프레임으로부터 영상신호 추출까지의 단계를 보여준다.

3.1. 연속 영상 프레임 획득

PC용 화상카메라는 초당 30 프레임의 영상을 획득한다. 그러나 실제 프로그램 실행과정에서의 부하에 의하여 그 속도가 떨어짐을 보인다. 그럼에도 최소한 15 프레임의 속도는 유지하므로 연구에 큰 지장은 없다. 화면의 크기는 320x240 (가로 픽셀수 x 세로 픽셀수)이다. 칼라 영상 프레임의 전체 화면처리를 위해서는 1회에 230,400 (가로 320 x 세로 240 x 3 RGB) 회의 처리가 필요하다. 따라서 실행속도의 향상을 위해서 컬러와 흑백 영상으로의 적절한 변환, 부분적인 처리 등의 방법을 강구하였다.

3.2. 움직임 영상 검출

움직임이 있으면 두 영상 프레임 간에 변화가 있는 부분이 있게 된다. 영상 프레임의 동일한 위치의 픽셀값 차이를 분석하여 움직임 영역을 찾아낼 수 있다. RGB 컬러모델로부터 HSI 컬러모델의 명암값 $(I=\frac{(R+G+B)}{3})$으로 변환한 다음에 각 픽셀의 명암값 변화 정도에 따른 움직임 영상 검출 관계를 다음 식에 의해 나타낼 수 있다.

$$d_{ij}(x,y)=\begin{cases}255 & |f(x,y,t_i)-f(x,y,t_j)|>T \\ 0 & \text{이 외의 경우}\end{cases} \quad (1)$$

시간 t_i와 시간 t_j에서 획득한 두 영상 프레임 $f(x,y,t_i)$와 $f(x,y,t_j)$ 사이의 비교결과인 $d_{ij}(x,y)$는 공간좌표 (x,y)에서 두 영역간의 픽셀값 차이가 문턱치 T 보다 큰 경우에만 명암값 255를 갖도록 한다 [3]. 조명의 미세한 변동과 같이, 실제 움직임과는 무관하게 나타날 수 있는 영상잡음을 제거하기 위해서도 문턱치의 설정은 필요하며, 여기서 255는 명암값이 8bits로 표시되기 때문이다.

구현에서는 문턱치를 10으로 하되, 명암값 차이의 정도를 입술움직임 영상의 특징값으로 활용하기 위하여 255 대신 $|f(x,y,t_i)-f(x,y,t_j)|$의 값을 그대로 산출하였다. 문턱치 이상의 움직임 영상의 경우에도 다양한 요인에 의하여 미세한 크기의 잡음영상이 많이 포함되는데, 입술움직임 영상의 크기를 감안하여 5x7 (세로x가로) 크기 이하의 영상은 제거하였다 [4].

3.3. 움직임 영상 분리 및 입술움직임 영상 특징 적용

움직임은 얼굴요소의 여러 부분에서 일어날 수 있다. 몸의 움직임, 얼굴 움직임, 눈의 깜박임, 턱의 움직임 등이다. 이들 각각을 식별하기 위해서는 분리하여 개별화하여야 한다. 따라서 움직임 영상 각각을 라벨링하여 영역별로 분리하였다. 영상분리는 grassfire 기법을 적용하였다 [5]. 분리된 각각의 움직임 영역을 대상으로 입술움직임 영상 특징과 비교하였다. 입술움직임 영상 특징과의 유사도가 높은 순서에 따라 일부 움직임 영상만을 대상으로 다음 단계의 세밀한 적합도 산출 단계를 거쳐 입술움직임 영상 여부를 판별하였다.

입술움직임 영상특징 요소를 선정함에 있어서는, 눈의 움직임 영상과 대비될 수 있는 파라미터를 주로 발굴하였다. 눈의 깜박임은 인위적으로 조절할 수 없는 요소일 뿐만 아니라 입술움직임 영상의 특징과 유사성이 많기 때문이다. 입술움직임 특징요소별 데이터는 카메라로부터 50cm 거리에서 나타나는 움직임 영상을 대상으로 수집하여 초기값으로 활용하였다. 영상은 조명과 같은 영상 획득환경에 따라 특징값의 편차가 많으나, 상대적인 유사도 산출에 활용 가능하다. 입술움직임 영상의 주요 특징벡터는 7가지로 요약된다. 각 파라미터와 특징값을 [표 1]에 요약했다.

표 1. 입술움직임 영상 특징요소별 데이터
Table 1. Lip movement image features.

Lip Feature Vecter	(1) length	(2) width	(3) width/ leng.	(4) dimension rate	(5) pixel value	(6) length location	(7) width location
average	5	20	3.95	0.37	16	1.03	0.68
stand. dev.	2.18	6.88	0.87	0.07	3.92	0.18	0.1

위에 표시된 숫자는 입술움직임 영상을 구성하는 픽셀의 개수 또는 이들의 비율을 나타내고 각 요소값의 표준편차를 나타낸다. 면적율 (dimension rate)은 입술움직임 영상의 면적을 외접사각형의 넓이로 나눈 것이다. 평균 픽셀값 (pixel value)은 프레임간의 비교를 통하여 움직임 영상을 검출하는 과정에서 산출된 명암값 차이의 평균이다. 가로 및 세로위치 (length, width location)는 다수의 움직임 영상들의 중심좌표로부터 떨어진 정도이다. 구현에서는 이외에도 좌우상하 특성으로서, 해당 방향으로 분포한 움직임 영상의 분포정도를 부가하여 적용하였다. 예를 들어, 눈의 움직임 영상의 경우에는 좌우에 최소한 다른 눈의 움직임 영상이 있다. 입술의 경우에는 아래 부분에 턱의 움직임 영상이 있다. 입술움직임과 여타의 움직임을 차별화할 수 있는 특징 파라미터의 지속적인 발굴이 필요하다.

입술움직임 영상 특징값은 다수의 움직임 영상들의 특징값과 비교하기 위한 기준값으로서 적용된다. 그러나 개별 특징요소별로는 입술움직임과 영상획득 환경에 따라서 데이터의 변화가 많으므로, 특징벡터들의 종합에 의해 유사성의 척도로 활용하였다. 많은 움직임 영상 (M_i)들을 대상으로 입술움직임 영상 (L_i) 특징벡터에 대비시켜 유사성 척도 ($M_i sim$)를 다음 식 (2)와 같이 산출하였다.

$$(M_i)sim = \sum_{j=1}^{k}\left(-0.1 * \frac{|M_{i(j)} - L_i avg_{(j)}|}{L_i std_{(j)}} + 1\right) * w_{(j)} \qquad (2)$$

(식 2)에서는 움직임 영상 (M_i)의 특징요소 (j)의 특징값과 입술움직임 영상 특징요소 (j)의 기준값 ($L_i avg_{(j)}$)간의 차이의 절대값을 입술움직임 영상 특징요소 (j)의 표준편차 ($L_i std_{(j)}$)로 나눈 것이다. 움직임 영상의 유사도가 1.0 이하의 값을 갖도록 하였고 입술움직임 영상 특징과의 편차가 클수록 낮은 값을 갖도록 하기 위하여, 기울기 (-0.1)와 절편 (1)을 설정하였다. 특징요소별 가중값 ($w_{(j)}$)은 별도로 설정하지 않고 균등 ($\frac{1}{k}$)하게 적용하였다. 움직임 영상 프레임마다 다수의 움직임 영상을 대상으로 입술움직임 영상 특징벡터와의 유사도를 산출하였고, 이 결과를 토대로 각각의 움직임 영상 프레임별로 유사도가 높은 순으로 6개씩의 움직임 영상을 선정하였다. 프레임별로 유사도가 높은 순으로 선정한 6개씩의 움직임 영상의 예를 [표 2]에 나타냈다.

표 2. 입술움직임 영상과의 유사도 (예)
Table 2. Similarity with the lip movement image features.

	M_1	M_2	M_3	M_4	M_5	M_6	...
F_1	0.38	0.06	0	0	0	0	
F_2	0.54	0.40	0.38	0.28	0.18	0	...
F_3	0.68	0.58	0.58	0.45	0.29	0.28	...
F_4	0.93	0.75	0.66	0.56	0.31	0.29	...
F_5	0.94	0.84	0.8	0.55	0.51	0.40	...
F_6	0.94	0.74	0.64	0.51	0.40	0.29	...
F_7	0.84	0.76	0.72	0.56	0.29	0.23	...
F_8	0.93	0.82	0.77	0.77	0.70	0.33	...
F_9	0.82	0.72	0.50	0.30	0.28	0	...
F_{10}	0.47	0.39	0.32	0.29	0.29	0.27	...
...	...	...	...	...	...	...	...

위에서 F_i는 움직임 영상 프레임을 나타내고, M_i는 움직임 영상 프레임 내에 있는 움직임 영역들을 나타낸다. 프레임 내에는 6개 미만의 움직임 영역이 있는 경우 (F_1, F_2, F_9)도 있는데, 6개 이상인 경우가 대부분이다. 여기에서는 최대 6개까지 선정된 움직임 영역을 대상으로 입술 움직임 영상특징과의 유사도의 예를 나타냈다. 입술움직임 영상특징과의 유사도가 가장 높은 영역 (M1)부터 가장 낮은 영역 (M6)까지 차례로 보였다. 프레임에 따라서는 유사도가 모두 낮아 입술움직임 영상이 포함되지 않았다고 판단되는 경우도 있고, 0.90 이상의 유사도를 갖는 영역 (F_4, F_5, F_6, F_8)도 있음을 보이고 있다. 또한 입술움직임 영역의 포함여부를 판단하기 어려운 경우 (F_2, F_3)도 있다. 따라서 입술움직임 영상 여부를 판단하기 위해서는 유사도 외에도 추가적인 분석이 필요하다.

3.4. 템플릿정합 및 입술움직임 영상 검출

움직임 영상 중에서 입술움직임 영상특징과의 유사도가 상대적으로 높은 경우에도 실제 입술움직임 영상인지의 여부를 판별하는 절차가 필요하다. 이는 입술움직임 외에도 움직임 요소가 많기 때문이다. 또한 입술을 움직이는 정도와 경우에 따라서 입술움직임 영상의 크기와 모양은 다양하게 변화한다. 이런 점들을 감안하면, 입술움직임 영상특징 외에 입술움직임 영상 자체와의 직접적인 비교가 가장 확실한 방안이 될 수 있다. 그러나 영상간의 직접적인 비교는 많은 계산량이 요구되는 점과, 입술영상 자체는 움직임에 따라 변화가 많다는 점이 감안되어야 한다.

이에 따라, 입술움직임 영상특징과의 유사도가 높은 순으로 3개의 움직임 영상만을 대상으로, 입술 주위의 일부 영역으로 비교대상 템플릿을 정하여 움직임 영상과의 적합도를 분석하는 방법을 택하였다. 3개의 입술움직

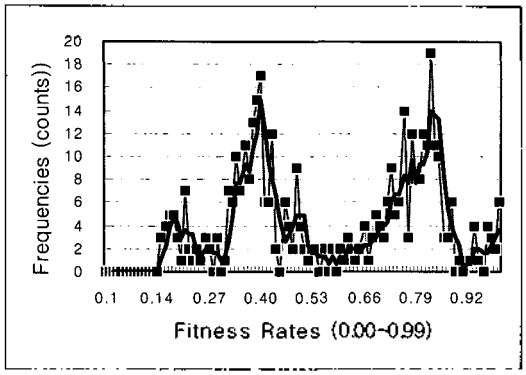

그림 3. 움직임 영상들의 템플릿 적합도 분포
Fig. 3. Template matching rate distribution.

임 영상들에 대한 적합도를 산출하였고, 가장 높은 적합도를 보이는 움직임 영역을 대상으로 입술움직임 영상여부를 판별하도록 하였다. 이를 위해서는 적합도에 대한 임계값 설정이 필요하다.

적합도 산출을 위한 템플릿은 그 크기와 모양의 변화가 거의 없고 명암의 대비가 뚜렷한 코의 일부로 하였다. 코를 포함하는 얼굴영상 파일을 저장해 놓고, 적합도 산출과정에서 이를 읽어 들인 다음 코의 일부 템플릿 영역과, 입력되는 움직임영상의 중심으로부터 위쪽으로의 일정부분과 비교되도록 하였다. 즉 움직임 영역이 입술일 경우에는 상단 일정부분에는 코의 영상이 있을 것이고 코의 템플릿과의 비교결과 (적합도)도 높게 나오게 된다. 적합도 데이터가 축적됨에 따라 입술움직임 영상과의 적합도 분포범위가 여타 움직임 요소의 적합도와 구분되도록 하였다. (그림 3)은 움직임 영상의 템플릿 적합도와 그 빈도수의 분포의 예를 보여준다.

템플릿 적합도에 따른 빈도 수의 편차가 많으므로 3구간 이동평균선 (검은 곡선)을 구하여 분석에 활용하였다. 우측의 볼록곡선은 입술움직임 영상의 적합도이고 좌측의 볼록곡선들은 여타 얼굴요소의 움직임 영상에 대한 것임을 실험을 통해 확인하였다. 여기서 입술움직임 영상과 여타의 움직임 영상의 적합도 사이에 최저 빈도를 보이는 적합도 (여기에서는 0.55)가 이들을 구분하는 임계값이다. 임계값이 자동으로 도출되는 과정에서 관련

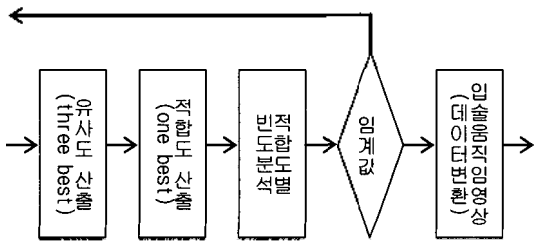

그림 4. 입술움직임 영상 검출절차 요약도
Fig. 4. Lip movement image detection procedure.

데이터 축적을 위해 실행 초기에 몇 초 정도 소요될 수 있다. 이상의 절차를 (그림 4)에 요약했다.

3.5. 2단계 입술움직임 영상 검출

입술움직임의 크기와 모양이 다양하게 변하는 점에 기인하여, 실제 입술움직임 영상인 경우에도 입술움직임 영상특징과의 유사도가 다소 낮아질 수 있다. 이러한 영상들도 가능한 찾아내야 한다. 따라서 임계값 이상의 적합도를 보이는 움직임 영상이 검출되지 않는 경우에는 2단계로 재검출하는 방법을 취하였다. 이전의 검출과정에서 높은 적합도를 보였던 위치를 중심으로 움직임 영상 존재여부를 확인했다.

3.3 절의 유사도 산출결과 중에서 활용하지 않았던 움직임 영상을 포함하여 모든 움직임 영역 ($M_1 \sim M_6$)을 확인한다. 다만, 2단계 입술움직임 영상 검출이 가능하기 위해서는, 말하면서 얼굴 전체를 과다하게 움직이지 않는다는 것이 전제되었다.

3.6. 명암값 보상

움직임 영상과 템플릿 영상을 픽셀 단위로 비교함에 있어서 주변의 조명에 의하여 적합도 산출에 큰 영향을 받게 된다. 이를 보상하기 위하여, 주변의 조명에 무관하게 명암값이 0 ~ 255 사이에 골고루 오도록 명암분포를 확장시켰다 [6]. 템플릿 영상과 움직임 영상에서 정합대상 부분에 한정하여 적용하였다. 명암 대비 스트레칭은 다음과 같은 매핑함수로 표현된다.

$$V_{new}(x,y) = \frac{V_{old}(x,y) - V_{\min}}{V_{\max} - V_{\min}} * 255 \qquad (3)$$

위에서 $V_{old}(x,y)$는 템플릿 영상을 포함하여 실시간으로 입력되는 움직임 영상의 픽셀값을 나타내고, $V_{new}(x,y)$는 변환된 값을 나타낸다. $V_{\min}$은 입력 영상의 픽셀값 중에서 최소값을 나타내고 $V_{\max}$는 최대값을 나타낸다.

3.7. 입술움직임 영상신호 변환

입술움직임 영상 특징과의 유사도가 높고 임계값 이상의 템플릿 적합도를 가진 입술움직임 영상이 검출되면 이를 음성구간 검출과정에서 활용할 수 있도록 데이터로 변환해야 한다. 영상 신호값은 입술움직임 영상 특징과의 유사도와 템플릿 적합도를 합하여 구했다. 영상 신호

값은 0.0 ~ 1.0 사이의 값을 갖도록 하였고, 음성인식의 음성구간 검출모듈에서는 그 값이 양의 값을 가지면 입술움직임이 있는 것으로 판단할 수 있도록 하였다.

IV. 시스템 연동

4.1. 데이터 공유

영상처리와 음성인식은 각각 별개의 윈도우로 나뉘어져서 독립적으로 실행된다. 그러므로 영상처리와 음성인식 프로세스 간에 데이터를 공유하기 위해서는 IPC (Inter Process Communication) 통신이 필요하다. 이를 위해 공유메모리 방식을 통하여 데이터를 공유하도록 하였다 [7]. 입술움직임 영상신호는, 음성인식 과정에서의 확인여부에 구애되지 않고, 공유메모리에 기록하는 것이고, 음성인식의 음성구간 검출 모듈에서는 독립적으로 공유메모리를 확인한다.

4.2. 음성인식의 음성구간 검출기능 보완

음성구간 검출 모듈에 영상 신호값을 읽어 들일 수 있는 기능을 부가하였다. 이는 음향에너지 분석 외에 입술움직임이 있는 지를 추가로 확인하도록 한 것이다. 또한, 영상처리가 처음부터 실행되지 않는 경우나 영상처리 실행을 중도에 멈추는 경우에도 음성인식 기능은 기존의 방식대로 정상적으로 동작될 수 있도록 구현하였다.

음성구간 검출모듈에 영상신호 확인을 위한 추가기능을 구현함에 있어서는 프레임 처리 시간차를 감안하는 것이 필요하다. 음성처리는 보통 10msec 단위로 음성 프레임을 처리하는데 비해 영상은 30msec단위로 처리하기 때문이다. 따라서 영상검출 과정에서의 유실이 음성인식에 미치는 영향을 최소화하기 위하여 이를 보상할 수 있는 기능을 추가하였다. 이를 위해 영상검출에 성공한 프레임의 데이터가 다음 프레임까지 지속될 수 있도록 조정하였다. 일단 입술움직임 영상이 검출되면 그 결과가 일정 횟수의 프레임까지 지속되도록 한 것이다.

4.3. 영상처리와 음성인식 연동실행

영상처리와 음성인식은 독립적으로 실행될 수 있다. 그러나 일단 모두 실행되는 상황에서는 연동기능을 수행한다. 영상처리부에서는 입술움직임 영상신호가 있는 경우에 이를 공유메모리에 기록한다. 음성인식기는 음성구간을 검출하면서 영상신호가 있는 지를 확인한다.

V. 실험결과

영상신호 추출과정에서는 입술움직임 영상을 정확히 찾아내는 것이 무엇보다도 중요하다. 영상처리를 실행하면서 추출할 수 있는 입술움직임 영상검출 결과, 적합도, 임계값 등의 데이터를 시각적으로 실시간 확인해 볼 수 있도록 하였다. 또한, 영상과 음성인식 연동실험 환경을 구축하여, 입술움직임 영상신호를 고려한 음성인식이 실행되는 지를 확인하였다. 실험은 PC Pentium IV, 3.6GHz 컴퓨터에서 수행한 결과를 중심으로 서술하였다.

5.1. 입술움직임 영상 검출

일상의 조명환경 하에서 얼굴 전체의 움직임이 거의 없는 상태에서 입술움직임을 어느 정도 정확하게 검출하는 지를 확인하였다. 프로그램 구동부터 입술움직임 영상 초기 검출까지의 시간은 무시할 정도였으며, 얼굴요소 중에서 가장 많은 검출오류 요소는 눈의 움직임이었다. 입술움직임 영상 검출 성공률은 95%에 이르렀다. 4.2절에서와 같이 입술움직임 검출결과의 지속기능을 추가함으로써 검출오류가 영상/음성 연동으로까지 줄 수 있는 영향을 최소화하였다. 이상의 결과를 요약하면 다음 표와 같다.

표 3. 입술움직임 영상 검출실험 및 결과
Table 3. Lip movement image detection test and results.

조명환경	일반가정/ 사무실 환경 (300~500lx, 100lx 이상 가능)
화자와 화상카메라 간격	약 50cm
초기검출시작 소요시간	약 3초
오인식 다발 얼굴요소	눈, 얼굴/배경 경계면
검출 오류 보상기능	명암값 보상, 2단계 검출
입술움직임 영상 검출률	95%

5.2. 영상과 음성인식 연동실험 결과

영상과 음성을 연동시킨 실험환경 하에서 영상신호의 유무에 따른 음성인식 결과의 출력여부를 확인하였다. 실험은 에러가 발생할 때까지의 결과를 확인한 것이다. 또한 영상기능을 실행하지 않는 경우와 중도에서 멈추는 경우에도 음성인식 기능은 정상적으로 수행함을 확인하였다.

아래 [표 4]에서 Not input 에러는 발성과 함께 입술움직임이 있었는데도 입술움직임 영상신호를 확인하지 못하여 음성인식 진행을 하지 않은 경우이다. Input 에러란 카메라로 하여금 입술움직임을 검출하지 못하도록 한 상태에서 발성한 경우로서, 음성인식을 진행하는 경

표 4. 영상/음성 연동 실험결과
Table 4. Linked test results.

발성횟수	에러내역			성공률 (%)
	Not input	Input	계	
72	1	0	1	98.6
103	1	0	1	99.0
193	1	0	1	99.5
199	1	0	1	99.5
567	4	0	4	99.3

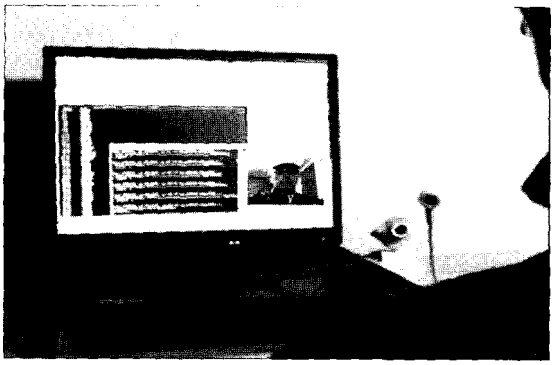

그림 5. 영상/음성 연동 테스트 화면
Fig. 5. Linked test environment.

우는 없었다. (그림 5)는 발성목록에 따라 발성하고 음성인식 결과를 확인하는 모습이다.

VI. 결 론

본 논문에서는 음성인식 과정에서 유입될 수 있는 동적 음향잡음을 입술움직임 영상신호 확인을 통하여 효과적으로 방지할 수 있는 방안과 실험결과를 제시하였다. 영상획득에서부터 영상신호 추출과정을 살펴보았고, 음성인식기와의 연동 하에서 음성인식 기능의 정상적인 진행여부를 확인하였다. 입술움직임 영상신호 추출과정에서의 부하가 음성인식 과정으로 전이되지 않도록 실행되는 구조로서, 연속음성인식 과정에서의 동적잡음 처리에 적극 활용될 수 있을 것으로 기대한다.

향후에는 다양한 영상획득 환경 하에서 강인한 입술움직임 영상검출을 실현하고자 한다.

참 고 문 헌

1. G. Potaminanos, H.P. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading, Image Processing", 1988. ICIP 98, Proceeding, 173-177, Oct. 1998.
2. M.T. Chan, Y. Zhang, and T.S. Huang, "Real-Time Lip Tracking and Bimodal Continuous Speech Recognition", IEEE Second Workshop on Multimedia Signal Proceeding, 65-70, 7-9 Dec. 1998.
3. Rafael C. Gonzalez, Richard E. Woods, *Digital Image Processing*, (Second Edition, 2002) pp 567-642.
4. 이수종, 박 준, 이영직, 김응규, "연속 영상 프레임으로부터 입술움직임 영상신호 검출", 한국음향학회 2006하계학술발표대회, 2006.8.26. 281-284.
5. F. Leymarie and M.D. Levine, "Simulating the Grassfire Transform Using and Active Contour Model", Trans. IEEE Pattern Analysis and Machine Intelligence, 14 (1):56-75, 1992.
6. Z.Q.Wu, J.A.Ware, W.R.Stewart, and J.Jiang, "The Removal of Blocking Effects Caused by Partially Overlapped Sub-Block Contrast Enhancement", Journal of Electronic Imaging--July-September 2005--Volume 14, Issue 3, 033006(8 pages).
7. 김상형, *Windows API 정복*, (가 남사 2005.3.10) pp 1019-1153.

저자 약력

• 이 수 종 (Soo-jong Lee)

1984년 2월: 충남대학교 경제학과 (학사)
1990년 9월: 건국대학교 경제학과 (석사)
2003년 3월~현재: 한밭대학교 정보통신공학과 (박사과정)
1984년 3월~현재: 한국전자통신연구원 음성인터페이스연구팀, 책임연구원

• 박 준 (Jun Park)

1981년 2월: 서울대학교 전자공학과 (학사)
1983년 2월: 서울대학교 전자공학과 (석사)
1994년 8월: Univ. of Southern California 전기과 (박사)
1983년 3월~현재: 한국전자통신연구원 음성인터페이스연구팀장, 책임연구원

• 이 영 직 (Youngjik Lee)

1979년 2월: 서울대학교 전자공학과 (학사)
1981년 2월: 한국과학기술원산업전자공학과 (석사)
1989년 1월: Polytechnic University 전기 및 전산과 (박사)
1981년~1984년: 삼성전자(주) 컴퓨터개발실
1989년 1월~현재: 한국전자통신연구원 음성/언어정보연구센터장, 책임연구원

• 김 응 규 (Eung-Kyeu Kim)

1976년2월: 충남대학교 공업교육학과(학사)
1978년2월: 충남대학교 공업교육학과(석사)
1993년9월: 오사카대학 기초공학연구과 정보공학전공 (박사)
1982년 6월~1987년 9월: 충남대학교 공과대학 조교
1987년 10월~1989년 9월: 교오토대학 공학부 연구원
1990년 2월~1992년 3월: 오사카대학 기초공학부 문부교관 (조수)
1993년 9월~1994년 2월: 충남대학교 공과대학 시간 강사
1994년~현재: 한밭대학교 정보통신, 컴퓨터공학부 교수, 한밭대학교 정보통신전문대학원장