잡음에 강인한 시청각 음성인식

잡음에 의한 인식율 저하를 막고자 하는 노력들 중 음성의 시각적 측면을 이용하고자 하는 시청각 음성인식(audio-visual speech recognition) 기술이 최근 주목을 받고 있다. 사람의 말은 조음기관의 움직임을 통해 생성되며 이 중 입술이나 혀, 이빨 등과 같이 사람이 눈으로 볼 수 있는 조음기관의 움직임은 음성을 인식하는데 중요한 정보로 쓰일 수 있다.

■ 이종석, 박철훈
(한국과학기술원 전자전산학부)

1. 머리말

음성은 사람과 사람 사이의 가장 자연스러운 의사소통 방법 중 하나로써 인간과 컴퓨터의 인터페이스에서도 음성을 이용한 의사소통을 가능케 하고자 하는 많은 노력이 진행되고 있다. 컴퓨터를 이용한 자동 음성인식 기술은 현재 상당한 수준에 이르렀지만 전반적으로 사람의 음성인식 능력에 비하면 아직은 낮은 인식 성능을 보인다. 특히 잡음이 존재하는 환경에서 강인하게 음성을 인식하는 문제는 사람과 기계의 성능 차이가 현저한 대표적인 상황이다. 우리의 일상을 보면 컴퓨터 팬 소음, 자동차 소음, 기계 소음 등 잡음이 존재하지 않는 경우를 찾기가 쉽지 않다. 조용한 환경에서 높은 인식율을 보이는 음성인식 시스템이라 하더라도 이러한 잡음 환경에서는 인식율이 크게 저하될 수 있다.

잡음에 의한 인식율 저하를 막고자 하는 노력들 중 음성의 시각적 측면을 이용하고자 하는 시청각 음성인식 (audio-visual speech recognition) 기술이 최근 주목을 받고 있다. 사람의 말은 조음기관의 움직임을 통해 생성되며 이 중 입술이나 혀, 이빨 등과 같이 사람이 눈으로 볼 수 있는 조음기관의 움직임은 음성을 인식하는데 중요한 정보로 쓰일 수 있다. 이와 같은 시각정보는 소리잡음에 영향을 받지 않기 때문에 청각정보를 보조하여 강인한 음성인식 성능을 얻는데 유용하게 사용될 수 있다.

음성의 시각적 및 청각적 측면이 모두 중요함은 일찍이 McGurk 효과에 의해 보여진 바 있다[1]. '바' 라는 발음을 귀로 들으면서 '가' 발음에 해당하는 입술 움직임의 비디오를 동시에 보는 경우 '다' 로 착각하여 인식할 수 있음이 보여졌으며, 이는 음성의 시각정보가 인식에 큰 영향을 미친다는 것을 입증하는 것이다. 또한 청각정보에서 혼동을 일으키기 쉬운 음소가 시각정보를 이용하면 쉽게 구분될 수 있음이 보여진 바 있다[2]. 청각에 장애가 있는 사람들의 경우 입술 움직임을 보는 것만으로도 음성을 어느정도까지는 인식할 수 있다는 보고나, 보통의 사람들도 주변에 소음으로 인해 상대방의 말소리가 잘 들리지 않을 때 입술 움직임을 관찰함으로써 보다 잘 음성을 이해할 수 있다는 보고[3] 모두 기존의 음성인식에서 고려하지 않았던 음성의 시각적 측면이 인식에 중요하게 쓰일 수 있음을 보여준다.

그림 1은 일반적인 시청각 음성인식 시스템의 구성을 보여준다. 시청각 음성인식 시스템에서 다루는 문제는 그림에서 보듯이 크게 세 가지로 요약할 수 있다. 첫째, 청각정보에서 특징을 추출하는 문제, 둘째, 시각정보에서 특징을 추출하는 문제, 그리고 마지막으로 청각정보와 시각정보를 통합하여 최종 인식 결과를 얻는 문제이다. 이 중 첫번째인 청각특징 추출은 기존의 음성인식에서 많이 연구되었기 때문에 다른 두 가지가 시청각

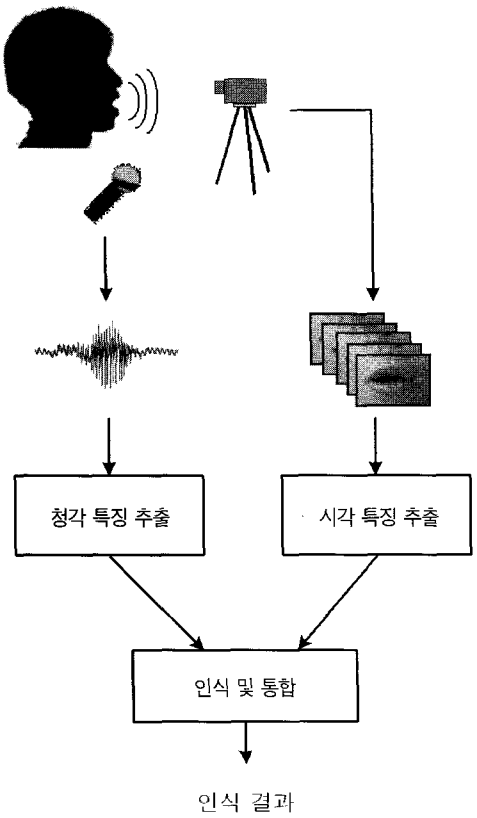


그림 1. 시청각 음성인식 시스템의 구성

음성인식 시스템 구현에서 중요하게 다뤄진다. 이 글에서는 시청각 음성인식 시스템의 구현에 관련한 연구들을 종합적으로 요약 및 개괄하고자 한다.

2. 청각특징 추출

청각신호로부터 특징을 추출하는 것은 신호를 일정크기의 프레임으로 자른 후 각 프레임 단위로 수행한다. 일반적으로 약 20~30 ms 크기의 윈도우 함수가 약 10 ms 씩 진행하며 신호를 프레임으로 나눈다. 각 프레임별로 적절한 특징을 추출하여 낮은 차원의 특징벡터를 얻는다. 이때 인식에 중요한 음성 정보를 유지하는 동시에 배경잡음, 채널잡음 및 화자간 변이와 같이 불필요한 변화가 억제된 특징을 얻는 것이 중요하다. 각 프레임별 특징 추출에서는 켑스트럼 (cepstrum) 영역을 많이 이용하며, 그 예로 멜 주파수 켑스트럼 계수 (Mel-frequency cepstral coefficient) 이나 선형예측 부호화 켑스트럼 (linear predictive coding cepstrum) 등을 들 수 있다. 잡음에 강인한 인식 성능을 위해 켑스트럼 평균차감(cepstrum mean subtraction)[4]이나 RASTA (relative spectra)[5] 등의 특징추출 기법이 제안되기도 하였다.

시간에 따른 특징벡터의 변화에 대한 정보를 함께 고려하기 위해 특징벡터의 시간미분으로 얻는 동적 특징을 함께 사용하는 경우가 많다. 1차미분 또는 2차미분항까지 함께 사용함으로써 잡음이 없는 경우나 있는 경우 모두 인식 성능이 향상되는 것으로 알려져 있다[6].

3. 시각특징 추출

시각적 측면에서 음성의 정보는 입술과 혀의 움직임, 턱의 움직임, 눈의 움직임, 표정, 손짓 등 여러 가지 형태로 나타나지만 가장 많은 정보를 포함하는 것은 입술, 혀, 이빨 등의 조음기관을 포함하는 입 영역이라 할 수 있다. 따라서 현재까지 개발된 대부분의 시청각 음성인식은 입술영역으로부터 시각특징을 추출하여 인식에 사용한다.

일반적으로 화자가 말하는 것을 담은 동영상에는 화자의 얼굴이 포함될 수 있으며 때로는 배경이나 화자의 상반신 내지는 하반신까지도 포함될 수 있다. 따라서 시각특징을 추출하기 위해서는 우선 기록된 동영상에서 화자의 입술 영역을 추출하는 것이 필요하다. 입술영역 추출을 어렵게 하는 요인으로는 배경, 화자의 피부색, 자세, 조명 조건 등의 변화를 들 수 있다. 초기 연구에서는 화자의 입술 주변에 특정한 색의 표식을 부착하여 입술영역 추출 문제를 제외하였으나 이는 자연스러운 상황이 아니다. 기록된 영상에서 특정 표식의 도움 없이 얼굴이나 얼굴 내의 특정한 부분을 추출하는 것에 관해서는 여러 연구가 진행되어 왔는데, 컬러 정보를 이용한 분할, 경계 검출, 문턱값 적용, 템플릿 정합, 움직임 정보를 이용한 방법 등의 기법들을 들 수 있다[7].

2장에서 서술한 것과 같이 인식에 좋은 청각특징을 추출하는 것에 대해서는 상당한 연구가 진행되어 연구자들간의 견해가 어느정도 일치되었지만, 시각특징의 추출에 대해서는 아직 다양한 시도가 이루어지고 있는 상태이며, 이들간의 체계적인 비교는 부족한 편이다. 지금까지의 연구에서 시각특징을 추출하는 방법은 크게 윤곽선 기반 방식, 픽셀값 기반 방식, 그리고 움직임 기반 방식으로 분류할 수 있다.

윤곽선 기반 방식은 입술의 윤곽선으로부터 특징을 추출하는 기법을 말한다. 입술의 높이나 너비, 구강의 넓이 같은 기본적이고 간단한 기하학적 정보만으로도 인식이 가능함이 보여졌다[8]. 변형가능한 템플릿 (deformable template) 기법에서는 다항식으로 주어지는 입술 모양의 템플릿의 파라미터를 변화시

키면서 입술윤곽선을 찾고 그 파라미터를 특징으로 사용한다[9]. 또한 입술 윤곽선상의 점들의 좌표에 주성분분석을 적용하여 특징을 추출하는 동적 모양모델 (active shape model) 기법도 제안된 바 있다[10]. 이상과 같은 윤곽선 기반 방식은 조명이나 화자의 피부색과 무관한 정보를 특징으로 얻을 수 있는 장점이 있는 반면, 혀나 이빨의 유무와 같은 입 안쪽의 변화나 입술의 돌출과 같은 정보를 잃어버리는 단점이 있다.

픽셀값 기반 방식에서는 입술영역 영상에 대해 영상변환을 적용하고 그 일부를 특징을 사용한다. 많이 쓰이는 영상변환으로는 주성분분석 (principal component analysis), 이산코사인변환 (discrete cosine transform), 이산웨이블릿변환 (discrete wavelet transform), 선형분별분석 (linear discriminant analysis) 등이 있다[11]. 픽셀값 기반 방식에서는 윤곽선 기반 방식과 달리 입 안쪽의 변화나 입술의 돌출과 같은 정보가 포함되어 이용될 수 있다. 반면, 조명이나 화자의 피부색의 변화가 특징에 반영되지 않도록 처리하는 것이 중요하다.

움직임 기반 방식은 입술의 움직임이 중요한 정보를 포함하는 것으로 간주하고 이와 관련된 특징을 추출하는 것이다. 이를 위해 광학흐름 (optical flow)와 같은 기법이 이용될 수 있다[12].

이상의 세 가지 특징추출 방식에서 둘 이상의 기법을 결합하여 사용하는 것도 가능하다[10,13].

4. 시청각 정보 통합 및 인식

4.1 음성 모델링

시청각 음성인식의 인식기로는 기존의 음성인식에서와 같이 은닉 마르코프 모델 (HMM: hidden Markov model)이 가장 많이 사용된다[6]. HMM은 이중의 확률 모델로써, 시간에 대해 탄력적인 음성의 동적특성을 잘 모델링할 수 있고, 학습 및 인식 과정에 필요한 알고리즘에 대한 연구가 잘 되어 있다. HMM의 학습은 Baum-Welch 알고리즘으로 수행할 수 있으며 이를 통해 해당 클래스의 발음에 대한 모델을 만든다. HMM 외에도 인공신경회로망이나 동적 시간정합 등이 인식기로 이용되기도 한다.

HMM에 의한 음성 (추출된 음성 특징)의 모델은 인식하는어휘의 수나 시스템의 응용분야에 따라 단어 또는 음절이나 음소와 같이 보다 작은 말소리 단위를 기반으로 할 수 있다. 청각음성의 음소에 대응하는 시각음성의 단위를 시소 (視素, viseme)라 한다. 일부 음소들은 시각음성으로는 구분할 수 없으며 이들은 하나의 시소로 묶여진다. 따라서 시소의 수는 음소의 수보다 적은 것이 일반적이다. 표 1은 영어에 대한 음소와 시소의 사상

표 1. 영어에 대한 음소와 시소의 관계[14]

시소기호	해당 음소
OV	/ax/, /ih/, /iy/, /dx/
BV	/ah/, /aa/
FV	/ae/, /eh/, /ay/, /ey/, /hh/
RV	/aw/, /uh/, /uw/, /ow/, /ao/, /w/, /oy/
L	/el/, /l/
R	/er/, /axr/, /r/
Y	/y/
LB	/b/, /p/
LCl	/bcl/, /pcl/, /m/, /em/
AlCl	/s/, /z/, /epi/, /tcl/, /dcl/, /n/, /en/
Pal	/ch/, /jh/, sh/, /zh/
SB	/t/, /d/, /th/, dh/, /g/, /k/
LFr	/f/, /v/
VlCl	/gcl/, /kcl/, /ng/

관계를 보여준다.

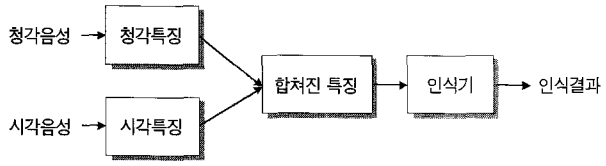
4.2 시청각 음성의 통합 모델

시각정보와 청각정보를 통합하는 것은 어느 시점에서 어떤 방식으로 두 정보를 합쳐 최종 인식 결과를 낼 것인가 하는 문제이다. 이는 시청각 음성인식의 목표인 잡음에 강인한 인식 성능을 얻기 위해 매우 중요한 문제이다. 여러 잡음상황에 대해서 통합에 의해 얻을 수 있는 인식 성능은 어느 하나의 정보만을 사용해서 얻는 인식 성능과 비교할 때 다음의 세 경우로 구분할 수 있다.

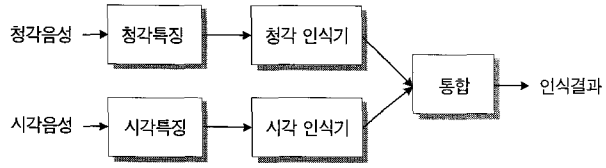
1) 단일정보에 의한 인식 결과보다 낮은 성능을 보이는 경우
2) 모든 잡음상황에 대해 두 단일정보의 인식 결과 중 좋은 성능과 같은 성능을 보이는 경우
3) 어느 한 정보만을 사용했을 때보다 항상 향상된 인식 성능을 보이는 경우

두번째 경우와 같은 성능을 얻는 것으로도 단일정보만을이용했을 때보다는 강인한 성능을 얻는 것이므로 시청각 정보를 모두 이용하는 의미가 있다고 할 수 있다. 그러나 시청각 음성인식의 최종 목표는 세번째 경우처럼 두 정보의 통합에 의한 상승효과 (synergy)를 얻는 것이며, 이는 두 정보의 상호보완적인 특성을 충분히 잘 활용할 때에 얻을 수 있다.

시각정보와 청각정보를 통합하는 방법은 그림 2에서 보는 것



(a) 초기통합 모델



(b) 후기통합 모델

그림 2. 시청각 정보의 두 가지 통합 모델

과 같이 크게 초기통합 모델과 후기통합 모델의 두 가지로 나눌 수 있다. 초기통합 모델에서는 청각특징과 시각특징을 하나의 특징벡터로 합쳐 하나의 인식기에 입력하여 인식 결과를 얻는다. 일반적으로 두 정보의 프레임 비율이 다르기 때문에 내삽의 과정을 거쳐 프레임 비율을 같도록 해야 한다. 후기통합 모델에서는 각각의 정보가 별개의 인식기를 통해 인식에 사용된 후 두 인식기의 출력을 합치는 방식으로 통합을 수행한다.

어느 모델이 더 적합한지에 대해서는 논쟁의 여지가 있지만 많은 경우 후기통합 모델을 선호하는 경향이 있다. 그 이유는 다음의 네 가지로 요약할 수 있다.

첫째, 후기통합에서는 각 정보가 독립적으로 인식된 후 합쳐지기 때문에 각 인식기의 출력을 이용하여 최종 인식 결과에 대한 두 정보의 기여도를 조절하기 쉽다. 이는 후기통합을 선호하는 가장 중요한 이유로 꼽을 수 있다. 주어진 음성 데이터에 포함된 잡음의 종류와 양에 따라서 각 정보의 가중치를 변화시킴으로써 두 정보의 상호보완적 특성을 활용하여 다양한 잡음 조건에서 강인한 인식 성능을 얻는 시스템을 구현할 수 있다.

둘째, 초기통합은 두 정보 사이에 완벽한 동기화를 가정하지만 후기통합은 동기화 문제에 있어 유연하다. 음성의 두 측면, 즉 말소리와 시각적으로 관찰할 수 있는 조음기관의 변화는 완전히 동기화 되어있지는 않다. 어떤 발음의 경우에는 입술과 혀가 실제 소리보다 수백 ms 앞서 움직이기도 한다[15]. 또한 사람이 음성을 시각과 청각을 통해 인식할 때 두 정보간에 어느 정도 시차가 있는 경우에도 인식 성능의 저하가 일어나지 않는다는 연구 결과가 있다[16]. 이러한 사실들은 두 정보의 완벽한 동기화보다는 유연한 시간적 관계가 필요함을 입증한다.

셋째, 초기통합에서는 두 특징벡터를 합쳐 높은 차원의 특징을 인식기에 통과시키며 이에 따라 학습해야 하는 인식기의 파라미터 수도 늘어난다. 일반적으로 학습해야 할 파라미터가 늘어남에 따라 필요한 데이터 역시 늘어나야 한다. 따라서 초기통합의 인식기를 안정적으로 학습하기 위해서는 그만큼 많은 양의 학습 데이터를 필요로 한다. 이를 해결하기 위한 방법의 하나로 합쳐진 특징벡터에 주성분분석이나 선형분별분석과 같은 기법을 사용하여 특징벡터의 차원을 줄이는 과정을 거치기도 한다.

넷째, 후기통합 모델에 기반한 시스템을 구현할 때에는 기존에 존재하는 단일 정보 기반 인식 시스템을 그대로 이용할 수 있다. 청각정보를 이용한 인식 시스템이 많이 개발되어 있는 것을 감안할 때 이는 시스템 구현에 드는 비용과 시간을 크게 절약할 수 있는 것이다. 반면 초기통합 모델은 완전히 새로운 시스템을 구현하고 인식기를 새로이 학습시켜야 한다.

4.3 가중 통합 기법

통합 과정에서 강인한 성능을 보장하기 위해서는 최종 인식 결과에 대한 두 정보의 상대적 기여도를 가중치를 통해 조정하는 작업이 필요하다. 후기통합 모델에서는 주어진 시청각음성 데이터 x에 대해 각 정보의 단독 인식 결과를 얻은 후 여기에 가중치를 주어 최종 인식결과 u^*를 얻는다. 이는 다음의 식으로 표현할 수 있다.

$$u^* = \arg\max\left[w y_u^A(x) + (1-w) y_u^V(x)\right] \qquad (1)$$

여기서 $y_u^A(x)$와 $y_u^V(x)$는 각각 청각정보 및 시각정보에 대한 인식기의 u번째 클래스에 대한 출력이며 가중치 w는 0과 1 사이의 값을 가진다. HMM의 경우 인식기 출력은 우도 (likelihood)로 주어지며 사후확률을 이용할 수도 있다. 청각신호에 잡음이 적게 포함되어 있는 경우 청각정보의 인식기가 시각정보의 인식기에 비해 대개의 경우 더 좋은 성능을 보이기 때문에 큰 가중치값을 써서 청각정보 인식기에 더 의존하여 최종 결과를 내도록 해야 한다. 반대로 잡음이 많이 포함되어 있는 경우 작은 가중치값을 통해 최종 인식결과가 시각정보 인식기에 더 의존하도록 해야 한다.

가중치의 결정을 위한 가장 단순한 방법으로써, 어떤 조건에서도 항상 같은 값의 가중치를 이용하는 방법이 이용되기도 하였다[14]. 보다 유연한 가중치 결정의 한 예로 청각신호에 포함된 잡음의 수준을 신호대잡음비 (SNR: signal-to-noise ratio)로 추

정하고, 사전에 정의된 SNR과 적정 가중치의 함수관계에 의해 가중치를 얻는 방법을 들 수 있다[10]. 또한 주어진 잡음 환경에서 기록된 적응 데이터를 이용하여 적정 가중치를 추정하는 적응 방법도 있다[17].

많은 연구에서 고려된 통합 가중치의 결정 방법은 각 인식기의 출력의 분포를 바탕으로 하는 것이다. 이 방법의 장점은 잡음 수준을 추정하는 과정이나 추가의 적응 데이터를 필요로 하지 않는다는 것이다. 주어진 청각신호에 잡음이 많이 포함되어 있는 경우 각 클래스에 대한 인식기 출력은 잡음이 적게 포함된 경우에 비해 그 차이가 작아지는 경향이 있다[18]. 이는 잡음의 존재로 인해 청각 인식기의 신뢰도가 낮아짐을 의미한다. 따라서 주어진 시청각음성에 대해 청각정보와 시각정보에 대한 인식기 출력의 분포를 비교함으로써 두 정보의 상대적인 신뢰도를 추정할 수 있다. 신뢰도와 최적 가중치 사이의 함수관계를 선형적 또는 비선형적으로 모델링한 후, 잡음조건을 알 수 없는 데이터가 주어졌을 때 그에 대한 인식기 출력으로부터 신뢰도를 측정하고 적정 가중치를 얻어 사용할 수 있다. 인식기 출력으로부터 어떠한 방식으로 신뢰도를 정의하는 것이 강인한 인식에 도움이 되는가에 대한 연구도 찾아볼 수 있다[18].

4.4 확장된 시청각 음성 모델

시각정보와 청각정보의 보다 섬세한 모델링을 위해서는 기존의 HMM에서 확장된 모델들이 사용되기도 한다. 이러한 모델로는 multistream HMM[10], product HMM[19], coupled HMM[20], asynchronous HMM[21] 및 동적 베이시안 네트워크[22] 등이 있다. 이들은 두 정보의 상호작용과 시간적 동기화 관계를 구체적으로 모델링할 수 있다.

5. 시청각 음성 데이터베이스

시청각 음성인식의 연구에서 인식 시스템을 설계하고 그 성능을 평가하기 위해서 반드시 필요한 것은 시청각 음성의 데이터베이스이다.

외국의 경우 영어, 불어 등의 언어로 된 데이터베이스가 존재한다. IBM의 AV-ViaVoice 데이터베이스는 290명의 화자의 연속음성 발음을 기록한 것으로써 약 50시간 길이의 데이터베이스이다[23]. MIT에서는 최근 AV-TIMIT 데이터베이스를 작성하고 발표하였는데, 여기에는 223명의 화자에 대한 450개의 문장 발음이 포함되어 있다[14]. 이 두 데이터베이스는 현재 공개되어 있지 않다. 공개된 데이터베이스로는 M2VTS[24]와 그 후속격인 XM2VTS[25]가 대표적이다. 이들은 각각 불어의 고립숫자 및 영어 연결숫자음을 기록한 것이다. M2VTS는 37명의 화자, XM2VTS는 295명의 화자로부터 얻은 데이터이다.

우리나라의 경우 외국에 비해 데이터베이스가 부족한 편이며 각 연구그룹에서 자체적으로 제작한 데이터베이스를 쓰는 경우가 많다. 최근 전자통신연구원 (ETRI)에서는 디지털 홈 개발용으로 수집된 300명 화자에 대한 멀티모달 음성명령 및 정보검색용 대화체 문장을 포함하는 시청각 음성 데이터베이스를 유상으로 배포하고 있다[26].

기존의 청각음성에 대한 데이터베이스와 비교할 때 국내나 국외를 막론하고 질적, 양적 측면에서 충분한 조건을 갖춘 시청각 데이터베이스는 드문 실정이다. 이는 시청각음성 연구의 역사가 짧다는 점 외에, 시청각 음성이 상당히 많은 저장용량을 필요로 하여 유지비가 많이 든다는 점, 시청각 음성 기록시 마이크 외에 카메라도 동시에 신경써야 하기 때문에 많은 노력이 든다는 점, 영상 기록시 생길 수 있는 다양한 변이를 통제하기가 쉽지 않다는 점 등에 기인한다.

6. 응용분야

시청각 음성인식 기술은 여러 분야에서 응용되어 쓰일 수 있으며 몇 가지 예를 들어보고자 한다.

먼저, 시청각 정보를 이용한 자연스러운 인간과 컴퓨터의 인터페이스를 들 수 있다. 시각정보와 청각정보를 함께 사용하여 향상된 인식 성능을 얻음으로써 인터페이스에서 효율을 높이고 자연스러움을 극대화할 수 있다. 또한 음성 외의 신호, 즉 눈의 움직임이나 손짓과 같은 다른 정보원을 함께 통합하는 멀티모달 인터페이스 기술로 확장하면 궁극적으로 자연스러운 인간과 컴퓨터의 상호작용을 구현할 수 있는 기반이 될 것이다.

시청각 음성인식이 잡음 상황에서 강인한 성능을 보일 수 있다는 점은 기계설비가 있는 공장, 자동차 및 항공기 등과 같이 주변에 잡음이 많이 포함되어 있어 강인한 인식 성능이 요구되는 동시에 인터페이스를 위해 손을 사용하기 어려운 상황에서 유용하게 쓰일 수 있을 것이다.

청각장애 시청자를 위한 텔레비전 방송 서비스의 하나로써 시청각 음성인식을 통한 자막 서비스를 생각할 수 있다. 특히 야외에서 뉴스를 중계하는 상황 같이 잡음이 많은 경우 음성을 인식하여 자막을 생성하고자 할 때 시각정보를 이용하는 것이 유용할 것이다.

방대한 동영상 자료로부터 사용자가 원하는 부분을 검색하

고 발췌하는 멀티미디어 검색 엔진 기술에 시청각 음성인식 기술을 접목함으로써 보다 정확한 검색이 가능할 수 있다.

화상회의와 같은 상황에서도 시청각 음성인식을 응용할 수 있다. 회의실의 반향이나 주변 잡음이 존재하는 상황에서 시청각 음성 인식 기술과 합성 기술을 함께 사용함으로써 회의와 관련된 시청각 데이터를 효과적으로 전송할 수 있을 것이다.

시청각 음성인식이 시청각 화자인식 기술로 응용 및 결합되면 보안 시스템으로 사용될 수 있다. 목소리와 더불어 각 화자의 고유한 입술의 움직임을 함께 사용함으로써 사칭자에 대한 보안성을 높이고 주변 잡음에 강인한 성능을 나타낼 수 있다.

7. 맺음말

이상에서 살펴본 바와 같이 시청각 음성인식은 일반적인 음성인식 환경인 잡음환경에서 강인한 인식 성능을 얻을 수 있는 새로운 기법으로써 앞으로 그 발전 가능성이 높고 응용분야가 많다. 하지만 충분한 데이터베이스 구축, 시각정보로부터 인식에 도움이 되는 특징을 추출하는 연구, 청각정보와 시각정보의 시간 동기 및 상호작용을 규명하고 모델링하는 방법에 대한 연구, 주어진 인식환경에서 최적의 통합을 수행하는 연구 등 해결해야 할 문제가 많이 남아 있다 할 수 있다. 이러한 문제 해결 과정에서는 공학적 접근뿐 아니라 심리학적, 생물학적, 언어학적 접근이 함께 이루어질 필요가 있다.

또한 각 언어별로 고유한 특성의 차이가 존재하기 때문에 외국의 연구 결과를 이용하는 것뿐 아니라 우리나라 자체적인 연구를 통해 우리말에 적합한 기술을 개발하는 것이 필요하다. 미국, 영국, 프랑스, 일본 등 선진국들이 자국의 언어를 대상으로 한 연구에 힘을 쏟고 있으며, 언어마다 존재하는 고유한 특징을 활용한 연구가 진행중이다. 다른 언어에 대한 연구에서 응용할 수 있는 공통적인 요소 외에 우리말의 고유한 특징을 고려한 인식 기술을 개발하는 연구가 앞으로 있어야겠다.

참고문헌

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746-748, Dec., 1976.

[2] A. Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception, in B. Dodd and R. Campbell, eds., *Hearing by Eye: The Psychology of Lip-reading*, pp. 3-51, Lawrence Erlbarum, London, 1987.

[3] L. A. Ross, D. Saint-Amour, V. M. Leavitt, D. C. Javitt, and J. J. Foxe, "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," *Cerebral Cortex*, vol. 17, no. 5, pp. 1147-1153, 2007.

[4] R. M. Stern, B. Raj, and P. J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition using Unknown Communication Channels*, Pont-a-mousson, France, pp. 33-42, Apr. 1997.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, 1994.

[6] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[7] H. P. Graf, E. Cosatto, and G. Potamianos, "Robust recognition of faces and facial features with a multi-modal system," in *Proc. Int. Conf. Systems, Man and Cybernetics*, pp. 2034-2039, 1997.

[8] M. N. Kaynak, Q. Zhi, A. D. Cheok, K. Sengupta, Z. Jian, and K. C. Chung, "Lip geometric features for human-computer interaction using bimodal speech recognition: comparison and analysis," *Speech Communication*, vol. 43, no. 1-2, pp. 1-16, Jan. 2004.

[9] T. Coianiz, L. Torresani, and B. Capril, "2D deformable models for visual speech analysis," in D. G. Stork and M. E. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems and Applications*, pp. 391-398, Springer-Verlag, Berlin, German, 1996.

[10] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141-151, Sept. 2000.

[11] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. Int. Conf. Image Processing*, vol. 3, Chicago, pp. 173-177, 1998.

[12] M. S. Gray, J. R. Movellan, and T. J. Sejnowski, "Dynamic features for visual speechreading: a systematic comparison," *Advances in Neural Information Processing Systems*, vol. 9, pp. 751-757, 1997

[13] 이종석, 심선희, 김소영, 박철훈, "제어되지 않은 조명 조건 하에서 입술움직임의 강인한 특징추출을 이용한 바이모달 음성인식," *Telecommunications Review*, 제 14권 1호, pp. 123-

134, 2004년 2월.

[14] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 3, pp. 1082-1089, May 2006.

[15] C. Benoit, "The intrinsic bimodality of speech communication and the synthesis of talking faces," in M. M. Taylor, F. Nel, and D. Bouwhuis, eds., *The Structure of Multimodal Dialogue II*, John Benjamins, Amsterdam, The Netherlands, pp. 485-502, 2000.

[16] B. Conrey and D. B. Pisoni, "Auditory-visual speech perception and synchrony detection for speech and nonspeech signals," *Journal of Acoustical Society of America*, vol. 119, no. 6, pp. 4065-4073, June, 2006.

[17] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 469-472, 2005.

[18] 이종석, 박철훈, "시청각 음성인식을 위한 정보통합: 신뢰도 측정방식의 비교와 신경회로망을 이용한 통합 기법," *Telecommunications Review*, 제 17권 3호, pp. 538-550, 2007년 6월.

[19] S. Nakamura, "Statistical multimodal integration for audio-visual speech processing," *IEEE Trans. Neural Networks*, vol. 13, no. 4, pp. 854-866, Jul. 2002.

[20] S. M. Chu and T. S. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 2, Orlando, FL, pp. 2009-2012, May 2002.

[21] S. Bengio, "Multimodal speech processing using asynchronous hidden Markov models," *Information Fusion*, vol. 5, pp. 81-89, 2004.

[22] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Applied Signal Processing*, vol. 11, pp. 1-15, 2002.

[23] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306-1326, Sept. 2003.

[24] S. Pigeon and L. Vandendorpe, "The M2VTS multimodal face database," in *Proc. Int. Conf. Audio- and Video-based Biometric Person Authentication*, pp. 403-409, 1997.

[25] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTS: the extended M2VTS database," in *Proc. Int. Conf. Audio and Video-based Biometric Person Authentication*, pp. 72-76, 1999.

[26] http://voice.etri.re.kr/DBSearch/Voice.asp

저자약력

이 종 석

· 1999년 한국과학기술원 전기전자공학과 학사.
· 2001년 한국과학기술원 전자전산학과 석사.
· 2006년 한국과학기술원 전자전산학과 박사.
· 현재 한국과학기술원 전자전산학부 연수연구원.
· 관심분야 : 시청각 음성인식, 멀티모달 인터페이스, 패턴인식.

박 철 훈

· 1984년 서울대학교 전자공학과 학사.
· 1985년 Caltech 전자공학과 석사.
· 1990년 Caltech 전자공학과 박사.
· 현재 한국과학기술원 전자전산학부 교수.
· 관심분야 : 지능시스템, 신경회로망, 최적화, 지능제어.