DOI: 10.3745/KIPSTB.2008.15-B.3.229

컬러 입술영상과 주성분분석을 이용한 자동 독순

이 종 석[†] · 박 철 훈[††]

요 약

본 논문은 화자의 입술 움직임으로부터 음성을 인식하는 자동 독순에서 회색조 영상 대신 컬러 영상을 사용하는 것의 유용성에 대해 고찰한다. 먼저 인간의 독순 실험을 통해 컬러 정보가 인식 성능에 어떠한 영향을 미치는지 확인한다. 다음으로 주성분분석을 이용한 자동 독순에서 회색조 또는 컬러 입술영상을 사용하는 경우에 대해 인식 성능을 비교한다. 다양한 컬러 좌표계에 대한 실험을 통해 컬러 영상의 사용으로 인식율이 향상됨을 보인다. 특히 RGB 좌표계를 사용했을 때 가장 좋은 성능을 얻으며, 회색조의 경우에 비해 잡음이 없는 환경에서는 4.7%, 잡음이 있는 경우 평균 13.0%의 상대적 오인식율 감소를 얻을 수 있음을 확인한다.

키워드 : 독순, 컬러, 특징 추출, 강인함

Automatic Lipreading Using Color Lip Images and Principal Component Analysis

Jong-Seok Lee[†] · Cheol Hoon Park[††]

ABSTRACT

This paper examines effectiveness of using color images instead of grayscale ones for automatic lipreading. First, we show the effect of color information for performance of humans' lipreading. Then, we compare the performance of automatic lipreading using features obtained by applying principal component analysis to grayscale and color images. From the experiments for various color representations, it is shown that color information is useful for improving performance of automatic lipreading; the best performance is obtained by using the RGB color components, where the average relative error reductions for clean and noisy conditions are 4.7% and 13.0%, respectively.

Key Words : Lipreading, Color, Feature Extraction, Noise-robustness

1. 서 론

자동 독순(automatic lipreading)은 입술 움직임이 기록된 영상을 통해 사람의 말을 인식하는 기술이다. 이는 기존의 청각정보만을 이용한 음성인식이 주변 잡음에 의해 성능이 크게 저하되는 단점을 보완하기 위한 방법으로써 최근 각광받고 있다[1]. 영상정보는 소리잡음에 영향을 받지 않기 때문에 자동 독순을 통해 잡음 환경에서 음성인식의 성능 저하를 보완하고 전체 인식 시스템의 성능 향상을 얻을 수 있다.

자동 독순은 카메라로 기록된 영상에서 입술을 추적하고 인식에 적절한 특징을 추출하여 인식기에 입력함으로써 이루어진다. 좋은 인식 성능을 얻기 위한 중요한 문제 중 하나는 인식에 효과적인 특징을 어떻게 정의하는가 하는 것이다. 여기에는 크게 두 가지 접근 방식이 있다. 첫번째는 윤곽선 기반 방식으로, 입술의 기하학적인 특징, 즉 입술의 높이나 너비를 사용하거나 입술의 윤곽선을 모델링한 후 그 모델의 파라미터를 특징으로 사용하는 방식이다. 두번째는 화소 기반 방식으로, 입술 영역의 영상에 적절한 영상변환을 적용하여 특징을 얻는 방식이다. 일반적으로 윤곽선 기반 방식은 입술의 윤곽선을 추적하는 과정에서 오차가 생길 수 있고, 혀나 이빨과 같은 입 안쪽의 변화나 입술의 돌출 정도와 같은 인식에 유용한 정보를 잃기 때문에 화소 기반 방식이 더 널리 사용되고 있다[2].

자동 독순을 위한 영상은 기록장치에 따라 회색조(grayscale)나 컬러 중 한가지이다. 최근에는 관련 산업의 발전으로 컬러 정보를 처리하는 것이 더욱 용이해졌다. 일반적으로 컬러의 경우가 회색조보다 더 많은 정보를 포함하고 있음은 분명하다. 기존의 연구에서는 입술이 피부에 비해 붉은색을 더 많이 포함한다는 관찰에 근거하여 입술 영역을 추적하거나 추출하는데 컬러 정보를 이용한 바 있다

* 본 연구는 2007년 한국과학기술원 BK21 정보기술사업단에 의하여 지원되었음.
† 정 회 원 : 한국과학기술원 전자전산학부 연수연구원
†† 정 회 원 : 한국과학기술원 전자전산학부 교수
논문접수 : 2007년 12월 7일
수 정 일 : 2008년 1월 22일
심사완료 : 2008년 2월 4일

[3,4]. 그러나 인식을 위한 특징을 추출하는데 컬러정보를 사용한 연구는 거의 없는 실정이다. Chiou와 Hwang이 입술영역의 컬러 화소값에 Karhunen-Loeve 변환을 적용하여 특징을 추출하였으나, 그러한 기법에 대한 근거나 회색조 정보에 의한 특징과의 성능비교는 이루어지지 않았다[5].

본 논문에서는 자동 독순을 위한 특징추출에서 회색조 정보 대신 컬러 정보를 이용하는 경우의 성능 향상 가능성을 알아보고자 한다. 이를 위한 배경 연구로써, 다수의 피험자를 대상으로 한 실험을 통해 인간의 독순에서 컬러 정보의 유용성을 알아본다. 다음으로, 컬러 정보에서 추출한 특징이 자동 독순에 유용한지 실험으로 알아본다. 컬러영상은 다양한 색좌표계를 이용하여 표현될 수 있는데, 기록된 영상을 여러 색좌표계를 이용하여 변환하고 그로부터 얻는 특징들에 의한 인식 성능을 서로 비교한다. 본 논문에서 이용하는 특징 추출 기법은 주성분분석(PCA: principal component analysis)을 이용하는 것이다. 기록된 영상에서 입술영역 영상을 추출하고 PCA를 통해 학습용 영상들 내의 주요한 변화를 특징으로 추출한다. PCA는 다른 기법들에 비해 기저벡터를 계산하는데 있어 많은 계산량이 요구되지만 가장 좋은 정보의 압축율을 보이는 장점이 있다. 또한 기존의 연구에서는 PCA를 이용한 특징 추출이 다른 변환 기법과 비교했을 때 비슷하거나 더 우수한 성능을 보이는 것으로 알려져 있다[6].

본 논문에서는 깨끗한 영상을 이용하는 경우 이외에도 영상에 잡음이 존재하는 환경에 대해서도 다룬다. 실제 인식 환경에서는 영상의 획득, 기록, 처리 및 전송 과정에서 영상에 잡음이 포함될 수 있으나 자동독순에서의 잡음은 최근에서야 다루어지기 시작했다[7]. 이러한 잡음 환경에서도 강인한 인식 성능을 얻는 것이 중요한데, 본 논문에서는 회색조와 다양한 컬러 특징의 잡음에 대한 강인성을 비교한다. 우리말 숫자로 구성된 데이터베이스에 대한 실험을 통해 인간이나 컴퓨터의 경우 모두 컬러 정보를 사용함으로써 잡음에 대한 강인함의 향상을 얻을 수 있음을 확인한다.

이하 논문의 구성은 다음과 같다. 2장에서는 인간의 독순에 대한 실험 과정과 그 결과를 보인다. 3장에서는 자동 독순의 과정 및 실험을 통한 비교 결과를 보인다. 4장에서는 인간과 컴퓨터의 인식 결과를 구체적으로 비교한다. 마지막으로 5장에서 결론과 추후과제로 논문을 맺는다.

2. 컬러 정보를 이용한 인간의 독순

인간은 음성을 인식할 때 귀로 듣는 말소리 뿐 아니라 눈으로 관찰되는 입술의 움직임을 함께 사용한다. 청각 장애가 있는 사람들의 경우 입술 움직임을 보는 것만으로도 음성을 어느 정도 인식할 수 있으며, 일반인들도 주변 소음으로 인해 상대방의 말소리가 잘 들리지 않을 때 입술 움직임을 관찰함으로써 보다 잘 음성을 이해할 수 있다[8].

회색조 영상과 컬러 영상에 대해 인간의 독순 및 시청각 음성인식 성능을 비교한 기존 연구에서는 두 경우 인식 결과의 차이가 거의 없는 것으로 나타났으며 이는 밝기 정보가 인간의 독순에서 컬러보다 더 중요한 역할을 함을 의미한다[9]. 그러나 이 결과는 영상에 잡음이 없는 경우에 국한된 것이며 잡음이 존재하는 경우에 대한 비교는 보고된 바가 없다. 본 장에서는 잡음이 없는 경우 및 있는 경우 회색조와 컬러 영상에 대한 인간의 인식 성능을 알아본다.

2.1. 실험 자료

우리말 숫자 '일'부터 '구', 그리고 '공'과 '영' 등의 11개 단어가 고립단어 형식으로 기록된 데이터베이스를 사용하였다[10]. 이 데이터베이스는 발음하는 입술 주변 얼굴 부분을 720x480 화소 크기, 30Hz 프레임 비율의 컬러 동영상으로 기록한 것이다. 56명의 화자(남자 37명, 여자 19명)가 각 단어를 세 번씩 발음하였는데, 본 장의 실험에서는 그 중 한 번의 발음을 사용한다.

각 동영상은 잡음 없는 회색조, 잡음 없는 컬러, 잡음 섞인 회색조, 잡음 섞인 컬러 등의 네 가지로 변환된다. 영상 잡음으로는 백색잡음을 사용하였다. 채널 잡음, 영상의 기록과정에서 생기는 잡음, 필름입자, CCD(charge-coupled device) 카메라의 잡음 등 많은 잡음이 신호와 무관한 백색 가우시안 잡음으로 모델링되기 때문이다[11]. 즉,

$$K(m,n) = I(m,n) + \eta(m,n) \tag{1}$$

여기서 (m,n)은 영상좌표, η는 가우시안 분포를 따르는 잡음신호, I와 K는 각각 잡음이 없는 경우와 있는 경우의 영상을 의미한다. 각 영상에 이처럼 잡음을 섞어 최대 신호대 잡음비(PSNR: peak-signal-to-noise ratio) 기준으로 3dB의 영상을 만든다. PSNR은 다음의 식으로 정의된다.

$$PSNR = 10\log_{10}\left\{\frac{(\text{최대화소값})^2 \times MN}{\sum_{m=1}^{M}\sum_{n=1}^{N}(K(m,n)-I(m,n))^2}\right\} \tag{2}$$

여기서 최대화소값은 화소가 가질 수 있는 최대값을 의미하며 본 논문에서 이 값은 255이다. M과 N은 영상의 가로와 세로의 크기, 즉 720과 480이다. 잡음 수준을 3dB로 한 것은 잡음이 없는 경우와 비교했을 때 인간의 인식 성능의 차이가 현저히 나타나는 값으로 정한 것이다. 컬러 영상의 경우 R, G, B 각 성분별 잡음 효과가 독립인 것으로 가정하고 각 성분별로 잡음효과를 준다[12].

2.2. 실험과정

16명의 참가자가 실험에 참가하였다. 이들의 나이는 21세부터 33세까지 분포하며 평균은 25세이다. 모두 정상 수준의 시력을 가진 한국인이다.

각 참가자마다 두 번의 세션에 걸쳐 실험을 수행한다. 첫 세션은 잡음 없는 동영상에 대한 인식 실험이고 두 번째 세션은 잡음 섞인 동영상에 대한 인식 실험이다. 두 세션은 약 2주의 격차를 두고 수행되었다. 각 세션에서 참가자는 32~80개의 동영상 파일에 담긴 입술 움직임을 통해 단어를 인식한다. 회색조와 컬러 동영상이 무작위 순서로 주어지며

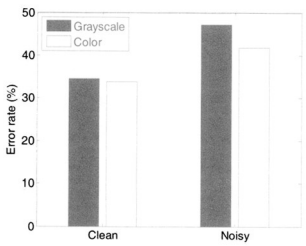
한 세션에서의 회색조와 컬러 동영상의 수는 같다. 동영상에 포함된 단어 역시 무작위 순서로 주어진다.

실험은 조용한 연구실 환경에서 이루어졌다. 참가자가 모니터가 있는 책상 앞에 앉아 준비를 마치면 각 동영상이 한 번씩 소리 없이 보여진다. 전체 단어의 목록은 미리 주어지며, 참가자는 동영상 속의 화자의 입술 움직임을 본 후 해당 단어를 인식하여 적는다. 전 실험 과정에서 모든 참가자에게 실험의 목적은 비밀로 하였다.

인식 성능은 전체 단어들 중 오인식된 단어의 수의 백분율로 정의되는 오인식율(%)로 나타낸다.

2.3. 결 과

(그림 1)은 잡음이 없는 경우와 있는 경우에 대한 참가자들의 독순 성능을 나타낸다. 잡음이 없는 경우 컬러 정보를 사용함으로써 회색조 정보에 비해 약간의 성능 향상이 있는 것으로 나타나지만 쌍체 t검증(paired t-test) 결과 이는 통계적으로 유의하지 않은 것으로 나타났다. 이는 앞서 소개한 이전의 연구[9]와 일치하는 결과이다. 잡음이 있는 경우 컬러 영상에 대한 인식율은 회색조 영상보다 눈에 띄게 높은 것을 볼 수 있으며, 이러한 성능 차이는 쌍체 t검증 결과 신뢰수준 0.05에서 통계적으로 유의한 것으로 나타났다. 따라서 잡음 환경에서 컬러 정보가 인간의 독순에 도움이 되는 것으로 결론지을 수 있다.

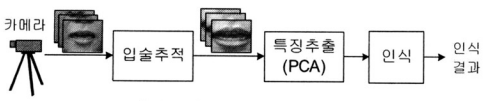


(그림 1) 회색조 및 컬러 동영상에 대한 인간의 독순 성능

3. 컬러 정보를 이용한 자동 독순

본 논문에서 자동 독순은 (그림 2)의 과정으로 수행된다. 먼저 데이터베이스의 기록된 영상에서 일정 크기의 입술 부분 영상을 추출한다. 이 과정에는 영상별 조명 조건의 차이나 피부색의 차이를 보상하는 정규화 과정을 포함한다. 다음으로 각 화소별로 평균이 제거된 입술영역 영상에 PCA를 적용하여 특징을 추출한다. 추출된 특징을 사전에 학습된 인식기에 입력하여 인식 결과를 얻는다.

이하에서 데이터베이스, 입술영역 추출, 회색조 및 컬러 영상에 대한 특징추출, 인식기 및 실험 결과를 설명한다.



(그림 2) 자동 독순의 과정

3.1. 데이터베이스

앞서 2.1절에서 설명한 것과 같은 데이터베이스를 사용하였으며, 화자별 세 번의 발음을 모두 사용하였다. 각 데이터는 모두 다른 시간적 길이를 가지며, 평균 길이는 23.8 프레임(0.79초)이다. 인식 실험은 화자 독립 방식으로 하였으며, 결과의 신뢰성을 높이기 위해 56명의 화자를 총 네 모듬으로 나누고 그 중 세 모듬을 학습에, 나머지 한 모듬을 테스트에 사용하는 과정을 돌아가며 총 네 번 수행한다. 이러한 과정을 통해 56명에 대한 모든 데이터가 인식 테스트에 사용된다.

잡음이 있는 경우에 대한 실험을 위해 영상에 가우시안 잡음을 섞어 10dB, 5dB 및 0dB의 PSNR을 갖는 영상을 생성한다. 그 과정은 2.1절에서 설명한 것과 같다.

3.2. 입술 영역 추출

독순을 위한 첫 단계는 기록된 영상에서 입술 영역을 추출하는 것이다. 이 과정은 화자별 피부색이나 영상 내, 혹은 영상간의 조명 조건의 차이를 보상하는 과정을 포함한다[10]. 먼저, 영상의 좌우의 밝기 차이를 줄여 입술 양 끝점을 정확하게 찾도록 한다. 다음, 히스토그램 명세화(histogram specification)를 통해 각 영상의 화소값 분포를 정규화하여 모든 영상이 같은 화소값 분포를 갖도록 한다. 마지막으로 문턱값(threshold)을 적용하여 입술의 양 끝점을 찾는다. 양 입술 사이는 입술의 그림자나 구강에 의해 항상 어둡게 나타나기 때문에, 문턱값을 적용해서 얻는 검은 부분의 양 끝이 입의 양 끝점이 된다. 찾은 끝점들을 바탕으로 입술 영역을 잘라내고 그 크기를 정규화하여 회전과 크기 변화에 불변하는 높이 44, 너비 50화소의 입술 영역을 얻는다. 입을 벌린 경우에도 입술 영역을 모두 포함하도록 높이를 정했으며, 더 큰 영상을 사용해도 성능의 향상이 없는 최소한의 크기로 정한 것인데 이는 성능 저하가 없는 한도 내에서 특징 추출시의 계산량을 줄이기 위한 것이다.

위의 전처리 과정은 컬러 영상에서도 RGB 각 성분에 대해 동일하게 적용된다. 공정한 비교를 위해 회색조 특징과 컬러 특징은 같은 입술영역에서 추출한다.

3.3. 회색조 특징 추출

입술 영역 영상을 얻은 후 각 화소마다 발음 전체에 대한 평균을 제거함으로써 발음마다 다르게 나타나는 화자의 피부밝기 및 조명의 차이를 줄인다. 다음, PCA를 회색조 영상에 적용하여 최종 특징을 얻는다. 입술영역 영상의 화소값을 n_0차원 열벡터로 만든 것을 $\mathbf{x}$, 학습 데이터 전체에 대한 평균 영상의 열벡터를 $\bar{\mathbf{x}}$라 할 때, $\mathbf{x}$에 대한 n차원 정적(static) 특징벡터 $\mathbf{s}$는 다음과 같이 얻는다.

$$\mathbf{s} = P^T(\mathbf{x} - \bar{\mathbf{x}}) \qquad (3)$$

여기서 행렬 P의 각 열은 학습 데이터의 모든 $\mathbf{x}$에 대한 공분산 행렬에서 얻은 고유벡터(eigenvector) 중 고유치(eigenvalue)가 큰 n개에 대한 고유벡터이다. n은 n_0에 비해

훨씬 작은 값을 가지며 위의 변환을 통해 영상에서 낮은 차원의 특징벡터를 얻는다. 본 논문에서는 n=12로 한다[10].

인식율의 향상을 위해 정적 특징벡터와 더불어 다음 식과 같이 정적 특징의 시간미분으로 정의되는 동적(dynamic) 특징을 함께 사용한다[13].

$$\Delta \mathbf{s}(t) = \left(\sum_{k=-2}^{2} k \cdot \mathbf{s}(t+k) \right) / \left(\sum_{k=-2}^{2} k^2 \right) \quad (4)$$

결과적으로 각 프레임별로 총 24차원의 특징벡터를 얻는다.

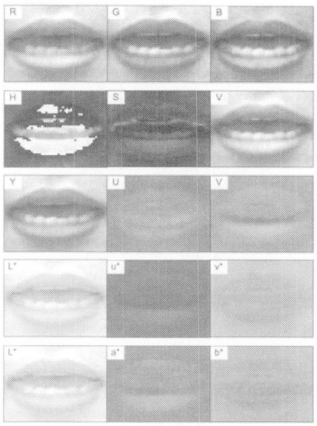
3.4. 컬러 특징 추출

컬러 영상에서의 특징 추출은 앞 절의 과정과 마찬가지로 평균제거와 PCA를 거쳐 이루어진다. 단, 회색조 영상과 달리 컬러 영상은 다수의 컬러성분을 가지기 때문에 각 성분별로 평균을 제거하고 여러 성분에 해당하는 화소값을 벡터로 하여 식 (3)과 같이 PCA를 거쳐 최종 정적 특징벡터를 얻는다.

컬러 영상은 다양한 색좌표의 형태로 표현될 수 있다[14]. 본 논문에서는 기록된 영상의 원래 색좌표인 RGB와 그로부터 유도되는 네 가지 좌표들, 즉 HSV, YUV, CIE L*u*v*, 및 CIE L*a*b* 좌표를 사용한다. RGB 색좌표는 빨간색, 녹색, 파란색의 세 가지 조합으로 구성되는 가산 모델로써 3차원 데카르트 좌표계에 기반하고 있다. 모니터나 비디오기기 등 많은 영상기기에서 사용된다. HSV 색좌표는 인간이 색깔을 인지하는 방식과 유사한 모델이다. 색상(hue) 성분은 색깔의 지배적인 빛의 파장을 규정하며 색깔의 종류를 결정한다. 채도(saturation) 성분은 얼마나 색깔이 순수한가를, 명도(value)는 색깔의 밝기를 뜻한다. YUV 색좌표는 휘도(luminance)를 나타내는 Y 성분과 색도(chrominance)를 나타내는 U와 V로 정의되며, TV 시스템에서 많이 쓰인다. 세 성분은 RGB 성분에 선형변환을 적용하여 얻을 수 있으며, R, G, B 성분 간의 상관관계를 일부 제거한다. CIE L*u*v*와 L*a*b* 색좌표는 화소값을 같은 양만큼 변화시킬 때 같은 시각적 중요도의 변화를 얻는, 즉 인지적으로 선형적인(perceptually linear) 모델이다. 또한, 영상기기에 상관없이 항상 같은 값을 갖는 특징이 있다. 이들 모델은 R, G, B의 선형변환으로 얻는 X. Y, Z 성분의 비선형 변환으로 계산되며 색깔과 밝기 정보를 분리해서 나타낸다. 결과적으로 L*는 밝기를, u*와 v* 그리고 a*와 b*는 색도 성분을 나타낸다.

(그림 3)은 입술영역 영상을 각 색좌표로 표현했을 때 각 성분별 영상의 예를 보인다.

입술영역 영상을 이상과 같은 색좌표의 성분들로 변환하고, 각 성분별로 시간에 대한 평균을 제거한 후, 이들 전부 또는 일부 성분을 열벡터로 하여 PCA를 거쳐 특징벡터를 얻는다. 따라서 이 경우 식 (3)의 $\mathbf{x}$는 사용되는 컬러성분의 개수에 따라 길이가 n_0, $2n_0$ 또는 $3n_0$로 주어진다. 그리고 회색조 특징에서와 마찬가지로 동적 특징을 함께 사용한다.



(그림 3) 여러 색좌표에 의한 입술영역 영상의 표현

3.5. 인식기

인식기로는 음성인식에서 가장 많이 사용되는 은닉 마르코프 모델(HMM: hidden Markov model)을 사용한다. HMM은 음성의 시간적 유연성을 모델링하기에 적합하며 데이터마다 다른 시간적 길이를 정규화할 필요가 없다. 사용하는 HMM은 전형적인 전단어(whole-word) 모델 좌우 연속밀도 HMM이다. 상태(state)의 수는 각 발음의 음소 수에 비례하게 하고, 각 상태의 가우시안 함수는 3개로 하였다. 이는 여러 실험을 통해 가장 좋은 성능을 내는 것으로 결정한 것이다. HMM의 학습은 잘 알려진 Baum-Welch 알고리즘을 사용한다[13]. 학습된 HMM을 이용한 인식 과정에서는 클래스를 알 수 없는 데이터가 주어졌을 때 이를 모든 클래스에 대한 HMM에 입력하여 가장 높은 확률값을 보이는 HMM을 선택함으로써 인식 결과를 얻는다. 인식 성능은 테스트 데이터의 단어들 중 오인식된 단어의 수의 백분율로 정의되는 오인식율(%)로 나타낸다. 인식기의 학습은 기본적으로 잡음 없는 학습 데이터로 수행하며 이를 잡음 없는 데이터 및 잡음이 포함된 데이터에 대한 인식에 사용하고 그 결과를 아래에서 나타낸다.

3.6. 잡음이 없는 상황에 대한 실험

<표 1>은 잡음이 없는 경우 회색조 및 컬러 영상 에서 얻은 특징에 의한 인식 성능을 비교한다. 일부 컬러 특징이 회색조 특징보다 좋은 성능을 내는 것을 알 수 있다(굵은 글씨로 표시).

RGB 각 성분을 단독으로 사용한 경우에는 서로 비슷한 성능을 보인다. 이는 (그림 3)에서 보는 것처럼 각 성분간 상관관계가 높으면서 밝기 정보를 어느 정도 충분히 포함하고 있기 때문이다. 이들 성분의 조합은 회색조의 경우보다 좋은 성능을 보이는데, 그 중 세 성분 모두를 이용한 특징이 가장 우수하며 이 때 상대적 오인식율 감소는 4.7%로 나타났다.

HSV 좌표의 경우, SV의 경우에만 회색조 특징보다 좋은 성능을 보인다. H 성분을 사용하는 것은 인식에 좋지 않은

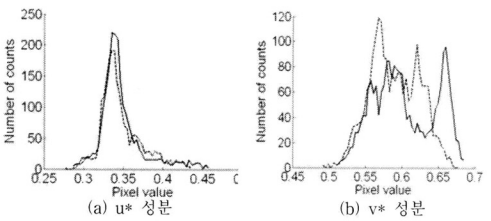
〈표 1〉 잡음이 없는 경우 색좌표의 성분 조합에 따른 자동 독순 성능

색좌표	사용된 성분	오인식율 (%)
회색조		36.1
RGB	R	36.6
	G	37.0
	B	36.1
	RG	**35.0**
	RB	**34.8**
	GB	**35.8**
	RGB	**34.4**
HSV	H	47.0
	S	42.5
	HS	46.8
	HV	43.9
	SV	**34.3**
	HSV	43.5
YUV	Y	**35.7**
	U	45.2
	V	46.4
	YU	**35.3**
	YV	36.4
	UV	44.5
	YUV	**35.3**
L*u*v*	L*	**34.5**
	u*	44.6
	v*	48.9
	L*u*	**33.7**
	L*v*	**35.2**
	u*v*	41.8
	L*u*v*	**35.2**
L*a*b*	a*	44.6
	b*	42.4
	L*a*	**34.7**
	L*b*	**35.1**
	a*b*	40.5
	L*a*b*	**34.7**

영향을 미치는데, 이는 H 영상이 작은 값과 큰 값이 실제로는 유사한 색깔을 의미하는 특성이 있어 매우 noisy하며 S 값이 작은 경우 그 신뢰도가 떨어지기 때문이다[15]. V 성분은 회색조 영상과 거의 같기 때문에 표에 보이지 않았다.

YUV 색좌표의 각 성분은 성능에서 다소 차이를 보이는데, 이는 성분간 상관관계가 RGB 성분에 비해 많이 줄었기 때문이다. Y 성분을 단독으로나 다른 성분과의 조합으로 사용하는 경우에만 회색조 특징보다 좋은 성능을 보이는데, 이를 통해 Y 성분이 나타내는 휘도 정보가 인식에 중요함을 알 수 있다.

CIE L*u*v* 좌표의 경우 회색조 특징보다 좋은 결과를 보이는 경우가 몇몇 나타난다. L* 성분의 성능이 회색조보다 더 좋은 결과를 보이는데, L* 영상이 회색조 영상과 거의 같은 의미임을 생각할 때 이는 주목할 만한 것이다. 이는 L*u*v* 좌표의 인지적으로 선형적인 특성이 입술 움직임에 대한 정보를 표현하는데 도움이 되는 것으로 추측할 수 있다. 얼굴에서 입술 영역을 추출하는 것과 관련한 기존 연구에서 이와 비슷한 맥락에서 L*u*v* 좌표와 L*a*b* 좌표가 사용된 바 있다[4]. L*u* 특징의 성능은 표에서 가장 좋다. 반면, v* 성분은 L* 성분과 같이 사용될 때 성능향상



(a) u* 성분 (b) v* 성분

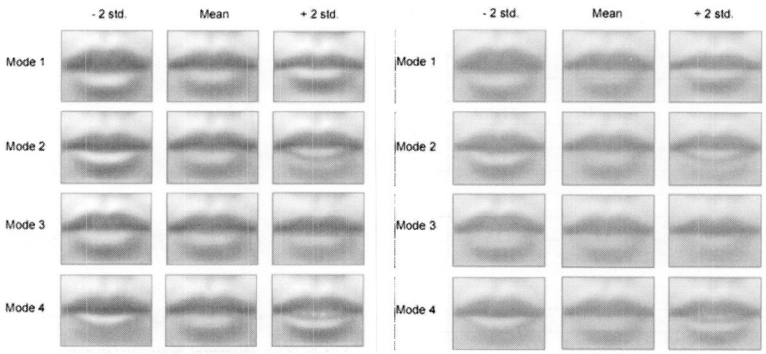
(그림 4) u*와 v* 성분의 두 화자에 대한 화소값 분포

에 도움이 되지 않는다. (그림 4)는 왜 u* 성분과 v* 성분이 성능에 대해 다른 영향을 미치는지 보여주기 위한 것이다. 그림은 같은 발음을 하는 두 화자에 대한 입술 영상에서 각 성분의 분포를 나타낸다. u* 성분의 경우 두 화자에 대한 분포가 매우 비슷하지만 v*의 경우는 그렇지 않은 것을 볼 수 있다. 이처럼 v* 성분은 화자에 따라 변화가 크기 때문에 인식에 좋은 영향을 주지 못한다.

L*a*b* 좌표계의 경우, 그 결과는 L*u*v*와 유사한 경향을 보인다. L*a*, L*b*, 및 L*a*b* 특징은 회색조 특징보다 좋은 성능을 보이지만 L*만을 사용한 경우보다는 좋지 않다. 이는 a*와 b* 성분이 나타내는 색도 정보가 도움이 되지 않음을 의미한다. 실제로, 이 두 성분에서 v* 성분처럼 화자에 따른 분포 차이가 큰 것을 관찰하였다.

이상의 결과에서 일부 컬러 특징은 회색조 특징에 비해 인식 성능이 향상되는 것을 볼 수 있다. 인식에 가장 중요한 것은 밝기(또는 휘도) 정보이며 몇몇 색도 정보는 인식율 향상에 도움이 되는 것으로 나타났다. 밝기 정보와 색도 정보가 분리되어 나타나는 색좌표의 경우 색도 성분의 기여도는 밝기 성분보다는 낮게 나타났다. 색도만을 사용한 경우 인식율은 좋지 않으며, 일부 색도 성분은 밝기 성분과 함께 사용될 때 성능에 도움이 되기도 한다. 전체적으로 L*u*, SV, RGB, L* 및 L*a*b* 성분에 의한 특징이 좋은 결과를 보인다. McNemar 검증[16] 결과, L*u*, RGB 및 L*의 경우 신뢰수준 0.05에서, 그리고 SV와 L*a*b*의 경우 신뢰수준 0.1에서 성능 향상이 통계적으로 유의한 것으로 나타났다.

(그림 5)는 입술영역의 회색조 및 RGB 영상에 대해 학습 데이터에서 얻은 평균 영상과 상위 4개의 주성분(principal component)에 대한 변화를 나타낸 것이다. 각 주성분마다 입의 개폐, 입술의 돌출여부, 이빨의 유무, 그림자의 변화 등과 같은 발음에 의한 변화를 나타내는 것을 볼 수 있다. 회색조 영상과 RGB 영상에 대한 주성분은 전반적으로 큰 차이 없이 유사한 변화를 주로 나타내고 있다. 그러나 회색조의 경우보다 RGB의 경우에 입술의 움직임 정보를 확연히 구분하는 것을 볼 수 있다. 예를 들어, 두 번째 또는 세 번째 주성분의 -2 std를 비교하면 이빨과 입술의 구분이 RGB 경우에서 잘 드러나는 것을 볼 수 있고, 세 번째 주성분의 +2 std를 비교해보면 입술과 피부와의 구분이 회색조에서는 다소 모호한 반면, RGB에서는 확연한 것을 볼 수 있다. 이러한 차이가 컬러 정보를 사용하였을 때의 성능의 향상으로 이어지는 것이라 할 수 있다.



(a) 회색조 영상 (b) RGB 영상

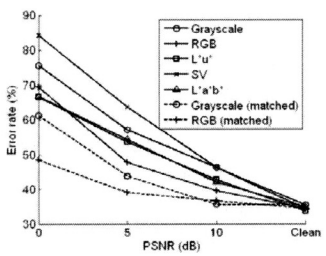
(그림 5) PCA에 의한 입술영상의 주요 변화분석

3.7. 잡음이 있는 상황에 대한 실험

앞 절에서 좋은 성능을 보인 성분조합인 RGB, L*u*, SV 및 L*a*b 특징에 대해 잡음 환경에서의 성능을 회색조 특징과 비교하여 (그림 6)에 나타내었다. 인식기는 잡음 없는 영상에서 추출한 특징으로 학습을 시킨 것이다. 회색조와 RGB의 경우 학습 데이터와 인식테스트 데이터가 같은 양의 잡음을 포함하는 동일 조건(matched condition)에 대한 결과도 함께 나타내었다.

결과에서 RGB, L*u* 및 L*a*b 특징은 회색조 특징에 비해 잡음에 대한 강인함이 크게 향상된 것을 볼 수 있다. 가장 좋은 결과는 RGB 특징에 의해 얻어지며, 이 경우 0dB에서 10dB까지의 잡음수준에 대한 평균 상대적 오인식율 감소는 13.0%로 나타났다. L*u* 성분의 경우 평균 8.5%, L*a*b*의 경우 평균 8.6%의 상대적 오인식율 감소를 보이며, 이 두 성분의 결과는 거의 유사하다. McNemar 검증 결과 이들 세 특징에 의한 성능 향상은 각 잡음 수준마다 0.001의 신뢰수준에서 통계적으로 유의함을 확인하였다. 반면, SV 특징은 회색조 특징보다 좋지 못한 결과를 보이는데 이는 RGB 좌표에서 HSV 좌표로의 비선형 변환을 하는 과정에서 S 영상이 원래 영상에 비해 더 잡음 수준이 높아지기 때문이다. 실제로 S 영상의 PSNR은 잡음 섞인 R, G 또는 B 영상의 약 절반밖에 되지 않음을 확인했다.

컬러 특징의 잡음에 대한 강인함은 동일조건에 대한 결과에서도 확인할 수 있다. RGB 특징은 회색조 특징에 비해



(그림 6) 잡음 섞인 영상에 대한 컬러 특징의 성능

평균 9.5%의 상대적 오인식율 감소를 나타낸다. 동일조건은 학습에 쓰인 영상과 테스트에 쓰인 영상의 잡음 조건이 같기 때문에 학습과 테스트 사이의 차이가 줄어들어 실선으로 나타낸 비동일조건에 비해 오인식율이 낮다.

컬러 특징에 의한 강인함은 각 컬러 성분이 한 화소값을 나타내는데 있어서 상호보완적이기 때문이다. 회색조 영상에서 한 화소가 잡음에 의해 오염되면 그 화소에 포함된 정보는 완전히 사라진다. 그러나 컬러 영상에서는 한 컬러 성분이 오염되어도 다른 성분들에 정보가 부분적으로 남아 있어서 정보를 완전히 잃지는 않는다. 다음과 같은 간단한 실험을 통해 이를 확인할 수 있다. 5dB의 잡음 섞인 영상을 만드는 잡음 패턴을 만들고 이를 각 R, G, B 영상에 똑같이 더해서 세 성분의 같은 화소가 같은 정도로 오염되도록 한다. 이 경우 RGB 특징에 의해 44.3%의 인식율을 얻는데, 이는 회색조 특징에 의한 42.8%의 인식율에 비해 큰 차이가 없는 것이다.

3.8. 계산량의 비교

컬러 정보를 사용함에 있어서 한가지 고려할 것은 계산량과 메모리 사용량의 증가이다. 회색조 정보에 비해 그만큼 더 많은 화소값을 기억하고 처리해야 하기 때문에 사용되는 컬러 성분의 개수에 따라 최소 1~3배의 계산량과 메모리 사용량 증가가 필연적이다. 특히 PCA에서 기저벡터를 얻는 것은 학습 데이터의 공분산 행렬을 고유치 분해하는 과정을 포함하며 이것이 독순 시스템 설계에서 가장 많은 시간을 소요하는데, 컬러의 경우 공분산 행렬의 크기가 커짐에 따라 고유치분해의 계산량도 많아진다. 이를 AMD 4400+ dual core의 CPU와 4GHz의 메모리를 가지는 컴퓨터에서 Matlab 환경으로 실험한 결과, 회색조 정보의 경우 364초의 시간이, RGB와 같이 세 컬러 성분을 모두 사용하는 경우 10919초의 시간이 소요되었다. 그러나 이러한 소요시간의 증가는 시스템의 설계에서만 필요한 것으로써 사전에 충분한 시간을 가지고 할 수 있다. 또한 실제 인식 과정에서 회색조와 컬러 정보를 사용하는 차이는 입술영역 영상으로부터 특징을 추출하는 과정에서만 나타나는데, 여기에 걸리는 시간은 두 경우 모두 측정이 불가능할 만큼 작게 나타났다. 특징벡터

의 크기는 두 경우가 같기 때문에 특징벡터로부터 인식하는 과정에서는 시간과 메모리 사용량의 차이가 없다.

4. 인간과 컴퓨터의 성능 비교

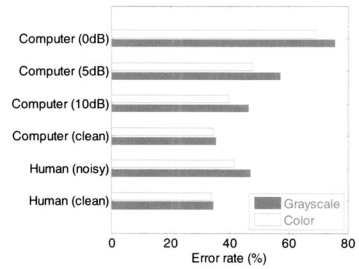
본 장에서는 앞의 두 장에서 각각 알아본 인간과 컴퓨터의 성능을 비교한다. 컴퓨터의 경우 앞의 결과에서 가장 좋은 성능을 보이는 RGB 특징을 대상으로 한다.

앞서 2.3절에서는 잡음이 없는 경우 통계적 측면에서 컬러 정보가 인간의 독순 성능에 그다지 도움이 되지 않는다는 결과를 얻었다. 그러나 3.6절의 자동 독순 결과에서는 잡음이 없는 경우에도 컬러 정보가 도움이 되는 것으로 나타났다. 인간과 컴퓨터의 독순에서 이처럼 다른 결과를 얻은 것은 영상에서 추출하여 이용하는 정보의 차이에서 기인하는 것이라 여겨진다. 컴퓨터에 의한 독순에서는 영상내의 화소값의 변화와 그 값들의 시간적 변화가 고려되는 반면, 인간은 입술 경계선이나 피부/입술의 질감과 같이 보다 높은 차원의 정보들을 추가로 이용할 수 있다. 이러한 고차원 정보는 회색조 영상이나 컬러 영상 모두에서 추출할 수 있는 것이다. 예를 들어, (그림 5)에서 보인 것과 같이 회색조 영상에서는 입술과 이빨 또는 입술과 피부의 밝기가 비슷하게 나타날 수 있다. 이는 자동 독순의 경우 회색조 영상에서는 구분되지 않지만 컬러영상에서는 구분될 수 있다. 하지만 인간은 입술의 생김새에 대한 사전 지식이 있기 때문에 회색조 영상에서도 입술과 다른 부분을 쉽게 구분할 수 있는 것이다.

(그림 7)은 인간과 컴퓨터의 인식 성능을 종합하여 비교한다. 잡음이 없는 경우 인간과 컴퓨터의 인식 성능은 큰 차이를 보이지는 않는다. 그러나 잡음이 있는 경우를 보면 3dB 잡음환경에서 인간은 5dB에서의 컴퓨터의 성능보다도 우수하며 컴퓨터에 비해 잡음에 대해 더 성능 저하가 적다. 앞서 기술한 바와 같이 인간은 고차원적 정보를 추출하는 능력이 우수하기 때문에 잡음 환경에서도 강인한 성능을 보인다.

잡음 환경에서 인식 성능의 비교를 위해 인식 결과에 대한 혼동 행렬(confusion matrix)을 (그림 8)에 나타내었다.

컴퓨터의 경우 5dB 환경을 대상으로 하였다. 각 행렬에서 세로축은 데이터의 원래 소속 클래스를, 가로축은 인식 결과를 나타낸다. 행렬의 숫자는 11개의 클래스 중 어느 클래스로 인식되었는가를 백분율로 나타낸 것이며 빈 칸은 0%를 의미한다. 인간이나 컴퓨터 모두 '일'과 '이'와 '칠', '오'와 '공', '구'와 '육' 등과 같은 혼동하기 쉬운 발음 사이에서 많은 오인식이 일어나며, '삼', '사', '영', '팔'과 같이 독특한 발음은 상대적으로 오인식이 적은 것을 볼 수 있다. 인간의 경우 컬러를 이용할 때의 성능 향상은 주로 혼동하기 쉬운 발음간의 오인식율의 감소에 기인한다. 반면, 회색조를 이용한 자동독순의 경우 많은 단어가 '이'로 오인식된 것으로 나



(그림 7) 인간과 컴퓨터의 인식 성능 비교

	일	이	삼	사	오	육	칠	팔	구	공	영
일	30	23	4	5			29	4			5
이	30	38	4				25	2			2
삼	2		88				9			2	
사	2		18	66			2	9			4
오					52	5			7	32	4
육	2	4	2		7	61	4		18		4
칠	11	38	4	2		2	43	2			
팔	5	2		2			4	86			2
구		4		2	5	61	2		23	4	
공				0	59	2		2	5	30	2
영	2	4	5	5	2	4	5	5	2		66

(a) 회색조 영상에 대한 사람의 인식 결과

	일	이	삼	사	오	육	칠	팔	구	공	영
일	43	29	4	4			20	2			
이	34	39	2		2		23				
삼	2		98								
사			25	66			2	5			2
오					75		2			21	2
육					5	68	2		23		2
칠	14	54	2	7			23				
팔	4		2	7				84			4
구					5	55	2		36	2	
공					55		5			36	4
영	4		7	11			4	2			73

(b) 컬러 영상에 대한 사람의 인식 결과

	일	이	삼	사	오	육	칠	팔	구	공	영
일	22	50	1	3		1	4	2			18
이	17	60	1	2		2	12				7
삼	1	7	90	1							1
사	7	5		70				2			15
오		42	1	1	7	27			1	7	14
육	1	35		1		45			1	1	17
칠	18	49	1	1		2	17	1			12
팔	1	11		7				73			8
구	1	51		1	1	38			3	2	4
공		36			7	24				11	21
영	4	11	1	11				1	1		73

(c) 회색조 영상에 대한 자동독순 결과

	일	이	삼	사	오	육	칠	팔	구	공	영
일	45	27		1	1	2	16				9
이	35	32	1		1	2	26		1		3
삼	5	4	82	1							8
사	13	2		61			4	1			19
오	4				30	29			7	29	1
육	4	4			4	70			15	2	2
칠	35	20				2	39				4
팔	16	3	1	1		2	2	67			8
구	7	4			2	42			43	3	
공	1	1			23	24	1		10	36	4
영	11	4		5	1	8	1			1	69

(d) 컬러 영상에 대한 자동독순 결과

(그림 8) 잡음 환경에서 사람과 컴퓨터의 인식 결과에 대한 혼동 행렬

타나며, 컬러를 이용했을 때에는 이러한 쏠림 현상이 완화됨으로써 전체적인 오인식율이 감소된다.

5. 결 론

본 논문에서는 회색조 정보와 컬러 정보가 사람과 컴퓨터의 독순에 미치는 영향에 대해 실험하고 분석하였다. 컴퓨터의 경우 컬러 정보에 의한 특징이 인식 성능 향상에 도움이 되며 잡음환경에서 향상의 정도가 두드러졌다. 특히 RGB 성분으로부터 얻은 특징을 사용한 경우가 전반적으로 가장 좋은 성능을 보였다. 이러한 결과의 심리학적 근거로써 사람의 경우도 잡음 환경에서 컬러 정보가 독순에 도움이 됨을 보였다.

본 논문에서는 기록된 영상에서 입술을 추적하는 문제는 배제하였다. 그러나 때때로 기록된 영상에는 보다 넓은 얼굴 영역이나 배경도 포함될 수 있는데, 기존 연구에서는 컬러정보가 입술을 정확히 추적하는데 도움이 됨을 보였다[3-5]. 따라서 화자 추적과 독순을 모두 수행하는 시스템을 설계하는 경우, 두 작업에서 모두 컬러 정보가 도움이 된다는 사실은 시스템에서 컬러 정보를 이용하는 강한 동기가 될 것이다. 추후 연구에서는 이와 같은 전체적 독순 시스템에서 컬러 정보를 이용하는 것에 대한 평가가 필요할 것이다.

또한 본 논문에서는 기존에 존재하는 색좌표에 대한 성능 비교를 하였는데, 독순을 위해 보다 적합한 새로운 색좌표를 만들어내는 것 또한 추후 연구를 통해 가능하리라 기대한다.

참 고 문 헌

[1] 이종석, 박철훈, "잡음에 강인한 시청각 음성인식," iCROS, 제13권 제3호, pp.28-34, 2007년 9월.

[2] P. Scanlon and R. Reilly, "Feature Analysis for Automatic Speechreading," Proc. Int. Conf. Multimedia and Expo, Tokyo, Japan, pp.625-630, Apr., 2001.

[3] P. Daubias, "Is Color Information Really Useful for Lip-reading? (Or What Is Lost When Color Is Not Used)," Proc. Interspeech, Lisbon, Portugal, pp.1193-1196, 2005.

[4] S. L. Wang, W. H. Lau, and S. H. Leung, "Automatic Lip Contour Extraction from Color Images," Pattern Recognit., Vol.37, No.12, pp.2375-2387, 2004.

[5] G. I. Chiou and J. N. Hwang, "Lipreading from Color Video", IEEE Trans. Image Processing, Vol.6, No.8, pp.1192-1195, 1997.

[6] S. Lucey, "An Evaluation of Visual Speech Features for the Tasks of Speech and Speaker Recognition," Proc. Int. Conf. Audio-Video-based Biometric Person Authentication, Guildford, UK, pp. 260-267, 2003.

[7] K. Saenko, T. Darrell, and J. Glass, "Articulatory Features for Robust Visual Speech Recognition," Proc. Int. Conf. Multimodal Interfaces, State College, PA, USA, pp. 152-158, Oct., 2004.

[8] 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.

[9] M. V. McCotter and T. R. Jordan, "The Role of Facial Colour and Luminance in Visual and Audiovisual Speech Perception," Perception, Vol.32, No.8, pp.921-936, 2003.

[10] J.-S. Lee and C. H. Park, "Training Hidden Markov Models by Hybrid Simulated Annealing for Visual Speech Recognition," Proc. Int. Conf. Systems, Man, Cybernetics, Taipei, Taiwan, pp.198-202, Oct., 2006.

[11] A. R. Weeks Jr., 'Fundamentals of Electronic Image Processing,' SPIE/IEEE Press, 1995.

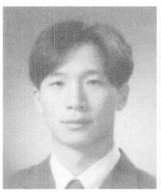
[12] H. Park, L. Gopishankar, and Y. Kim, "Adaptive Filtering for Noise Reduction in Hue Saturation Intensity Color Space," Opt. Eng., Vol.41, No.6, pp.1232-1239, 2002.

[13] L. Rabiner and B.-H. Juang, 'Fundamentals of Speech Recognition,' Prentice-Hall, 1993.

[14] R. C. Gonzalez and R. E. Woods, 'Digital Image Processing,' Prentice-Hall, 2002.

[15] N. Evano, A. Caplier, and P.-Y. Coulon, "A New Color Transformation for Lips Segmentation," Proc. Multimedia Signal Processing, Cannes, France, pp. 3-8, 2001.

[16] L. Gillick and S. J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," Proc. Int. Conf. Acoustics, Speech, Signal Processing, Glasgow, UK, pp.532-535, 1989.



이 종 석

e-mail : jslee@nnmi.kaist.ac.kr

1999년 한국과학기술원 전기및전자공학과 (학사)

2001년 한국과학기술원 전자전산학과 (공학석사)

2006년 한국과학기술원 전자전산학과 (공학박사)

2006년~현 재 한국과학기술원 전자전산학부 연수연구원

관심분야 : 시청각 음성인식, 멀티모달 인터페이스

박 철 훈

e-mail : chpark@kaist.ac.kr

1984년 서울대학교 전자공학과(학사)

1985년 Caltech 전자공학과(공학석사)

1990년 Caltech 전자공학과(공학박사)

1991년~현 재 한국과학기술원 전자전산학부 교수

관심분야 : 지능시스템, 신경회로망, 최적화, 지능제어