3차원 모델을 이용한 입모양 인식 알고리즘에 관한 연구

김동수* · 남기환* · 한준희* · 배철수* · 나상동**

*관동대학교 전자통신공학과

**조선대학교 컴퓨터공학과

A study on the lip shape recognition algorithm using 3-D Model

Dong-soo Kim* · Kee-hwan Nam* · Jun-hee Han* · Cheol-soo Bae*

Sang-dong Ra**

*Department of Electronic Communication Eng. Kwandong University

**Department of Computer Eng. Chosun University

E-mail : baecs@kdccs.kwandong.ac.kr

Abstract

Recently, research and developmental direction of communication system is concurrent adopting voice data and face image in speaking to provide more higher recognition rate then in the case of only voice data. Therefore, we present a method of lipreading in speech image sequence by using the 3-D facial shape model. The method use a feature information of the face image such as the opening-level of lip, the movement of jaw, and the projection height of lip. At first, we adjust the 3-D face model to speeching face image sequence. Then, to get a feature information we compute variance quantity from adjusted 3-D shape model of image sequence and use the variance quality of the adjusted 3-D model as recognition parameters. We use the intensity inclination values which obtaining from the variance in 3-D feature points as the separation of recognition units from the sequential image. After then, we use discrete HMM algorithm at recognition process, depending on multiple observation sequence which considers the variance of 3-D feature point fully. As a result of recognition experiment with the 8 Korean vowels and 2 Korean consonants, we have about 80% of recognition rate for the plosives and vowels.

요 약

최근 통신 시스템의 연구와 발전 방향은 목소리의 음성 정보와 말하는 얼굴 영상의 화상 정보를 함께 적용하므로서 음성 정보만을 제공하는 경우보다 높은 인식율을 제공한다. 따라서 본 연구는 청각장애자들의 언어 대체수단 중 하나인 구화(speechreading)에서 가장 시각적 변별력이 높은 독순(lipreading)을 PC에서 구현하고자 한다. 본 논문은 기존의 방법과 달리 말하는 영상 시퀀스에서 독순(lipreading)을 행하기 위해 3차원 모델을 사용하여 입의 벌어진 정도, 턱의 움직임, 입술의 들출과 같은 3차원 특징 정보를 제공하였다. 이와같은 특징 정보를 얻기 위해 3차원 형상 모델을 입력 동영상에 정합시키고 정합된 3차원모델에서 각 특징점의 변화량을 인식파라미터로 사용하였다. 그리고, 인식 단위로 동영상을 분리하는 방법은 3차원 특징점 변화량에서 얻어지는 강도의 기울기에 의한다. 인식은 다차원(multi-dimensional), 다단계 라벨링방법을 사용하여 3차원 특징벡터를 입력으로 한 이산 HMM을 사용하였다.

I. 서 론

언어의 자동 인식에 관한 연구는 1950년대부터 활발히 진행되어져 왔다. 초기에는 음성 정보만을 이용하여 처리되었지만 90년대에 이르러 처리할 수 있는 하드웨어의 급속한 발전에 힘입어 마이크로 폰을 이용한 음성 정보와 CCD카메라 등과 하드웨어의 급속한 발전에 힘입어 마이크로 폰을 이용한 음성정보와 CCD 카메라 등과 같은 영상입력장비의 영상정보를 이용하여 특수한 실험환경이 아닌 일반적인 환경에서 음성 정보와 영상 정보를 동시에 처리하여 화자의 언어 인식률을 향상시켰다. 따라서, 독순(lipreading)은 구화에서 변별이 가장 높은 시각정보로서 컴퓨터로 구현하기 위해서는 세가지 문제점을 해결해야 한다. 첫째, 입력영상에서 정확한 특징점을 추출하는 문제이고, 둘째, 입술의 움직임을 충실히 표현할 수 있는 인식파라미터를 사용하는 것, 마지막으로 입력동영상에서 인식단위로 영상의 프레임을 분리하는 문제이다. 이러한 문제점들에 대한 기존의 연구들[1~4]은 주로 2차원적인 형태와 특정점의 움직임 변위에 대해서만 고려하였기 때문에 입력영상에서 약간의 두부의 움직임만 있어도 인식파라미터가 변하게 된다. 또한, 동영상에서 인식단위로 분리하는 과정은 거치지 않아 자동화하기가 어렵다. 한글 발성에 대한 고찰은 전무한 상태이어서 이에 대한 연구가 필요하다.[5~6]

본 논문에서는 말하는 동영상에서 입술의 움직임에 부합하는 글자를 인식하는 방법과 음절단위로 입모양을 인식하기 위해서 동영상에 음절을 효과적으로 분리할 수 있는 간단한 방법을 제안한다. 그림 1은 동영상 입력에서부터 특징벡터의 추출까지의 과정이다. 인식에서는 각 입력특징벡터에 대한 출력확률을 평균벡터로 하는 정규분포로 가정하고, 다차원(multi-dimensional), 다단계 라벨링방법을 사용하여 3차원 특징벡터를 입력벡터로 한 이산 HMM을 사용하였다.[7~8]

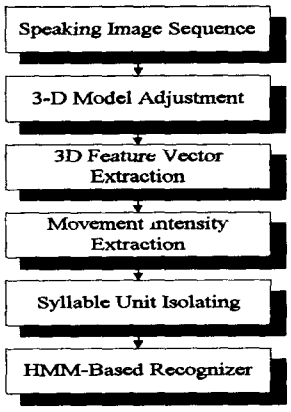


그림 1. 전체 영상처리 과정
Fig 1. Image Processing

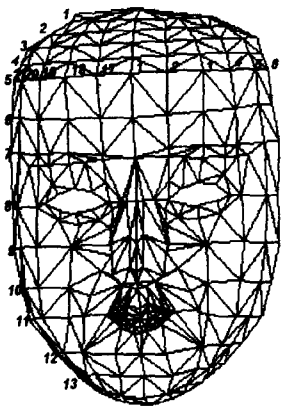


그림 2. Wire Frame Model
Fig 2. Wire Frame Model

II. 입력 영상의 3차원 모델 정합

2차원영상은 3차원영상의 투영이라는 관점에서 입술의 움직임은 3차원적으로 추정하여야 한다. 입력동영상으로부터 입술움직임에 대한 3차원 특징벡터를 얻기 위해서는 얼굴형상을 충실히 표현할 수 있는 3차원모델이 필요하다. 따라서 본 논문에서는 얼굴형상을 점과 선으로 근사한 삼각형으로 구성된 Wire Frame Model(그림. 2)을 사용한다.

WFM을 영상의 각 프레임에 정합하여 입력영상을 표현하므로 각 프레임당 입술움직임에 대한 WFM 정점의 이동변위를 구할 수 있다.

본 연구에서는 특징점 검출을 위해 영상의 휘도치 분포를 다단계로 임계화하는 방법을 사용한다. 임의의 대상 영상들로 부터 획득된 확률·통계적 분석에 의해 다단계 임계화된 구간을 수직·수평으로 투영하여 얼굴부위에 해당하는 휘도치분포의 경계값을 결정한다. 이 결정된 임계값에 따라 얼굴이외의 성분으로부터 얼굴을 분할하고, 분할된 얼굴부위를 바탕으로 안면 요소 특징점들(눈, 코, 입등)을 추출한다.

III. 3차원 정보의 특징 파라메터 검출 알고리즘

독순은 화자의 입술의 움직임을 관찰하여 음성언어를 이해하는 보조수단으로서 음성의 지각단위인 음소와 같이 시각적 지각 요소를 가질 수 있다. 이러한 시각적 지각요소를 시각소(visual

phoneme)라고 하고 이것들은 음소에 대한 입술의 독특한 움직임을 가지고 있다. 대표적인 시각소는 입술의 움직임, 입의 벌어진 정도, 턱의 움직임, 입술의 돌출, 뺨의 움직임 등과 같은 복합적인 움직임으로 이루어진다. 그러므로 이들의 복합적인 움직임 고려하여야 하지만은 입술의 돌출과 뺨의 움직임 등은 2차원적인 기존의 접근방법으로는 곤란하다. 따라서 여기에서는 입력된 동영상에 정합된 3차원모델로부터 시각소를 표현할 수 있는 3차원 특징 벡터를 얻을 수 있다. 본 연구에서는 그림 3과 같이 12개의 특징점을 사용하여 4개의 특징벡터를 추출한다. 4개의 특징벡터는 다음과 같다.

- 입술점 특징 벡터
  윗 입술 움직임 벡터(UP) :

$$UL = P_2 - P_0$$

  아랫 입술 움직임 벡터(Down) :

$$DL = P_3 - P_0$$

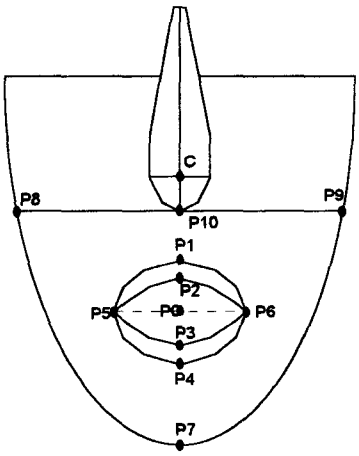
-입의 벌어진 정도 :

$$ML = P_5 - P_6$$

- 턱의 상하움직임

$$JM = P_7 - P_{10}$$

이와같은 움직임벡터를 정량적으로 취급하기 위해서는 개인차를 흡수할 수 있는 기본량이 필요하다. 여기에서는 입의 종횡폭을 이용하여 기준량으로 정한다. 움직임벡터의 프레임당 이동변위는 특징점의 이동량을 기준량으로 정규화한 것이다.



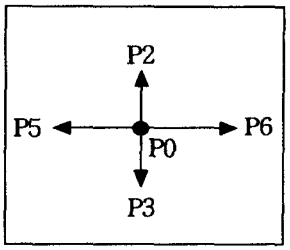
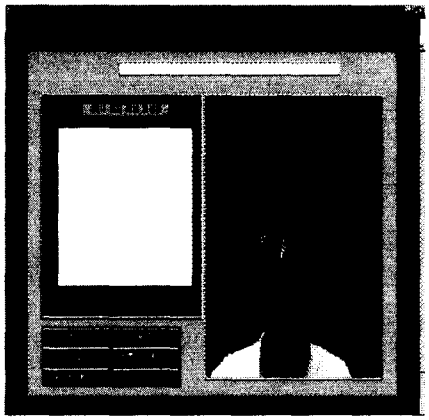
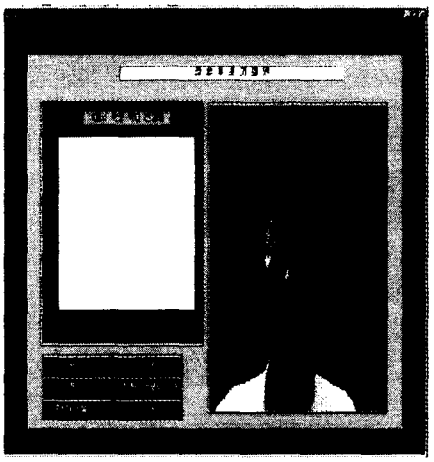


그림 3. 특징점
Fig 3. Feature Points

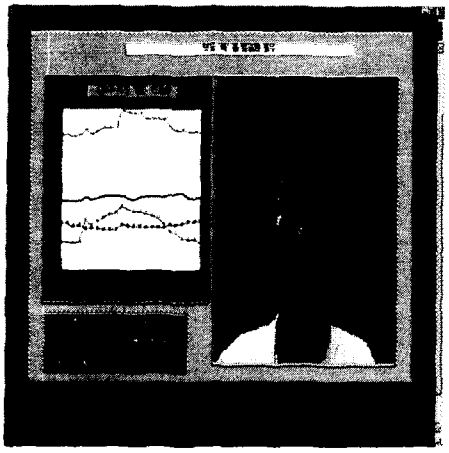
그림 4는 본연구의 입모양 인식 알고리즘을 실행한 예이다. 그림 4(a)는 입력 영상의 초기화면이고 그림 4(b)는 입력 영상에 3차원 형상 모델을 정합한 후를 나타낸다. 또한, 그림 4(c)는 모음 '아'를 발음하였을 때 각 4개의 특징 벡터 변화율과 영상 정보를 함께 나타내고 있다.



(a) 정합 전
(a) Before Adjustment



(b) 정합 후
(b) After Adjustment



(c) 모음 '아'
(c) Vowel /a/

그림 4. 3차원 모델 정합 예 '/a/'
Fig 4. A examples of 3-D model's adjustment

Ⅳ. 입력 동영상의 음절분리

입모양의 인식은 음절단위로 행해진다. 따라서 말하는 영상 시퀀스를 음절단위로 구분할 필요가 있다. 본 연구에서 입력 동영상의 음절분리는 특징벡터의 기울기의 굴곡점을 검출하여 분리할 수 있다. 3차원모델로 부터 얻어지는 특징벡터의 강도는 그림. 5에 나타낸 것처럼 선형적으로 증가하다가 다음 음절의 발성으로 갈 때 입모양 패턴의 끝점으로 연결된다. 이때 특징벡터의 시간에 대한 기울기는 변하게 된다. 이 기울기가 변하는 점을 음절의 끝부분으로보고 분리한다. 이 분리는 시간 t의 모든 구간에 대해 좌미분계수와 우미분계수를 구하여 이를수 있다.

시간 t_1에서 미분 계수는

$$\lim_{h\to 0,\, h<0} \frac{f(t_1+h)-f(t_1)}{h} \neq \lim_{h\to 0,\, h>0} \frac{f(t_1+h)-f(t_1)}{h} \quad \text{- 식 (1)}$$

과 같이 되어 좌미분계수와 우미분계수가 다르게 된다. 이때, 시간 t_1은 음절 구분점이다. 실제 이 방법은 음절분리에 좋은 결과를 주었다. 그러나 이 방법은 동일한 모음이 계속될때 좋은 결과를 얻을수 없다. 그것은 같은 패턴이 계속될 때 입모양의 특징벡터의 기울기는 변하지 않고 같은 입모양형상이 계속되기 때문이다. 예를 들면 [아가가]라고 발성하였을 때 자음에 대한 입모양의 영향은 없어지고 [ㅏ]의 입모양패턴이 세음절의 발성시간동안 계속되기 때문이다. 따라서 동일한 입모양패턴이 계속되는 경우 음절발성시간의 평균시간에 따라 음절분리를 행한다.

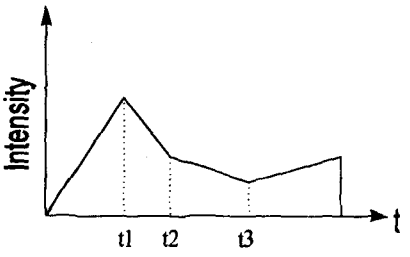


그림 5. 시간에 따른 특징벡터의 강도변화
Fig 5. The change of the feature vector intensity for the time

Ⅴ. 이산분포 HMM의 인식 알고리즘

*Speaking image sequence*로부터 추출된 각 특징점의 변화량은 multiple observation sequence로 HMM의 입력 파라메타로 사용한다. 그림 3에서 나타낸 4개의 특징벡터는 모든 모델 파라미터의 신뢰성있는 평가를 충분히 수행한다.

먼저, 다음과 같이 4개의 관측 sequence의 집합을 정의한다.

$$O=[O^{UL},\ O^{LL},\ O^{ML},\ O^{JM}] \text{ - 식 (2)}$$

우리는 각각의 관측 sequence가 매 다른 관측과 독립으로 가정하고, 우리의 목표는 모델 λ의 파라미터를 식(3)과 같이 최대로 하기위해 조정하는 것이다.

$$P(O\mid\lambda)=\prod_{K=1}^{K} P(O^{(K)}\mid\lambda) = \prod_{K=1}^{K} P_K \quad \text{- 식 (3)}$$

재산정 수식은 다양한 사건의 빈번한 발생에 기초하기 때문에, 다중 관측 sequence의 재산정 수식은 각각의 sequence에 대한 개개의 빈번한 발생과 함께 더해져서 수정된다.

이러한 multiple observation sequence를 특정 벡터로 이용하여 이산 HMM의 입력벡터로 사용하였다. 그림 6에 HMM인식기의 구성도를 나타내었다.

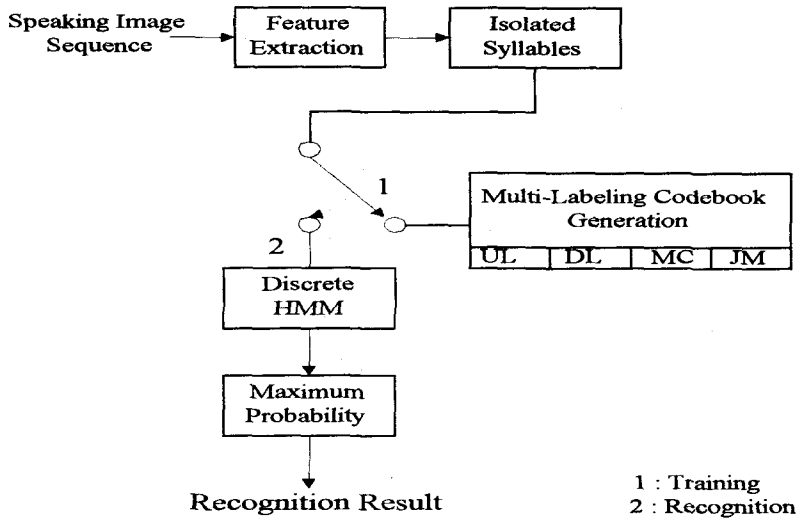


그림 6. 이산HMM 인식기

Fig 6. Discrete HMM Recognizer

Ⅵ. 결 론

본 논문에서는 3차원모델을 사용한 lipreading 방법을 제안하여 실험한 결과, 다음과 같다. 입력동영상에서 음절단위를 효과적으로 분리하였으며, 3차원모델의 움직임변위에 따른 특징벡터를 얻을 수 있었다. 인식은 입력특징벡터를 멀티라벨링하여 이산HMM의 입력으로 하였다.

인식실험은 한국어 8개모음과 2개의 자음의 조합으로 이루어진 음절단위로 행하였다. [파열음(ㅁ,ㅂ,ㅍ)+모음]의 경우는 80%이상의 인식율을 보였으며, 모음(8개)의 경우는 평상적인 조건하에서 55%의 인식율을 보였다. 그러나, 3차원모델을 정확히 정합하기 위한 특징점 검출알고리즘의 개선과 두부움직임에 따른 움직임 추정에 따른 정확한 움직임 변위를 추출하는 문제가 남아있다. 따라서 이러한 문제점을 해결하게 되면 크게 인식율을 향상시킬 수 있을 것으로 기대된다.

Ⅶ. 참 고 문 헌

[1] E.Petajan, B.Bischoff, D.Bodoff, and N. M. Brooke, " An Improved Automatic Lipreading System to enhance Speech Recognition." In ACM SIGCHI, 1988.

[2] Mase and A.Pentland."LIP Reading. Automatic Visual Recognition of Spoken Word". Proc. Image Understanding and Machin Vision, Optical of America, June. 1989

[3] K. E. Finn and A. A. Montgomery. Automatic Optically-Based Recognition of Speech. *Pattern Recognition Letters*, 8:159 - 164, 1988.

[4] K. Mase and A. Pentland. Lip Reading: Automatic Visual *Recognition* of Spoken Words. Technical Report 117, M.I.T. Media Lab Vision Science, 1989.

[5] Danial Reisfeld and Yehezkel Yeshurun, "Robust Detection of Facial Features by Generalized Symmetry" , Proc. ICPR, pp.117-120 , 1992 .

[6] Young Dong Lee, Chong Seak Choi, Kap Seak Choi, " Lip Shape Synthesis of Korean Syllable for Human Interface ". Korea Institut Comunication, vol 19, pp.614-623.

[7] L.R.Raider, "Mathematical Foundations of Hidden Markov Models", Recent *Advances in speech understanding* and Digital systems,

[8] L.R.Raider and B.H.Juang, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine Vol. 3, No.1, 99.4 - 16, Jan 1986.