ISSN 1975-8359(Print) / ISSN 2287-4364(Online)
The Transactions of the Korean Institute of Electrical Engineers Vol. 63, No. 6, pp. 813~819, 2014
http://dx.doi.org/10.5370/KIEE.2014.63.6.813

안면 움직임 분석을 통한 단음절 음성인식

Monosyllable Speech Recognition through Facial Movement Analysis

강 동 원* · 서 정 우* · 최 진 승* · 최 재 봉** · 탁 계 래†
(Dong-Won Kang · Jeong-Woo Seo · Jin-Seung Choi · Jae-Bong Choi · Gye-Rae Tack)

Abstract - The purpose of this study was to extract accurate parameters of facial movement features using 3-D motion capture system in speech recognition technology through lip-reading. Instead of using the features obtained through traditional camera image, the 3-D motion system was used to obtain quantitative data for actual facial movements, and to analyze 11 variables that exhibit particular patterns such as nose, lip, jaw and cheek movements in monosyllable vocalizations. Fourteen subjects, all in 20s of age, were asked to vocalize 11 types of Korean vowel monosyllables for three times with 36 reflective markers on their faces. The obtained facial movement data were then calculated into 11 parameters and presented as patterns for each monosyllable vocalization. The parameter patterns were performed through learning and recognizing process for each monosyllable with speech recognition algorithms with Hidden Markov Model (HMM) and Viterbi algorithm. The accuracy rate of 11 monosyllables recognition was 97.2%, which suggests the possibility of voice recognition of Korean language through quantitative facial movement analysis.

Key Words : 3-D motion capture system, Facial motion, Hidden Markov model, Speech recognition

1. 서 론

최근 음성인식기술에 있어 음성정보 이외에 잡음의 영향을 받지 않는 영상정보를 접목시킨 립리딩에 대한 연구가 활발히 진행되고 있다[1-3]. 립리딩은 화자의 입술 움직임을 분석하여 발성 단어를 인식하는 기술로 실생활에서 사용되는 TV, 휴대폰, 게임 등에 활용되고 있으며, 잡음 환경에서 현저하게 떨어지는 음성인식 정확도를 높이기 위한 보상 방법으로 많은 연구가 수행되었다[4-6]. 특히, 입술모양 및 안면근육 패턴의 변화를 통해 음성인식을 수행하는 기술은 언어장애인에 있어 의사를 인지하고 주변 재활 및 보조 기기에 연결하는 유용한 의사전달 수단으로써 사용이 가능하며, 이러한 기술은 언어학습의 교육적 콘텐츠 분야에도 활용되고 있다[7]. 국내에서도 영상의 시각정보만을 이용한 립리딩에 대한 연구가 진행 중이며, 발성 시의 입모양 변화에 큰 영향을 미치는 모음을 중심으로 수행되었다. Lee 등[8]은 카메라를 통해 '아/에/이/오/우'의 5가지 모음 발성 시의 입술 영역 영상데이터를 획득하고 6개의 관찰점을 설정하여 관찰점 간의 거리 변화를 계수화하다. 이를 통해 신경망 음성인식 알고리즘을 구현하였으며, 87.44%의 인식 정확도를 보였다. 후속연구에서는 '아/에/애/이/어/으/우/오'의 8가지 모음인식으로 다양화하고 93.7%의 보다 개선된 정확도를 나타냄으로써 음성인식 알고리즘의 효용성을 나타내었다[9]. 또한 Kim 등[10]과 Nam 등[11]의 연구에서는 영상의 2차원적인 데이터를 이용하여 3차원 안면 움직임 파라미터를 추출하여 음성인식을 수행하였다. Kim 등은 입술영역의 영상으로부터 입술에 관한 정보를 보다 쉽게 획득하기 위해 입술과 주요 부분에 마커를 부착하여 촬영하였으며, 거울을 이용하여 카메라의 영상에 얼굴의 옆모습이 투영되도록 하여 입술 파라미터의 깊이 방향에 대한 정보도 획득하였다. 마커인식을 통해 '아/에/이/오/우'의 5가지 모음의 입술 파라미터를 추출하였으며, 약 84%의 인식 정확도를 나타내었다. Nam 등은 영상의 이미지를 3차원 모델로 구현하고 입의 벌어진 정도, 턱의 움직임, 입술의 돌출과 같은 3차원 특징 정보를 이용하여 Hidden Markov Model(HMM) 인식기의 인식 파라미터로 사용하였으며, '아/어/오/우/으/이/에/외'의 8가지 모음을 대상으로 수행한 인식정확도는 55%를 나타내었다.

기존 한글 음성인식을 위한 립리딩의 연구는 카메라를 이용한 영상데이터를 통해 5가지 또는 8가지 모음에 대한 음성인식 알고리즘을 구현하였으며, 55~93%의 다양한 인식결과를 나타내었다. 그러나 기존 립리딩의 연구는 공간 및 조명 변화 등에 따라 입술 움직임의 검출에 오류가 발생될 수 있는 카메라 영상기법을 사용하고 있다. 입술 검출의 오류를 최소화하기 위해 대부분 환경변화가 없는 실내에서의 실험을 통해 영상을 획득 하고 있지만 다양한 조건 하의 경우 상당한 추정오차를 갖게 되며, 실제사용 환경에서는 이러한 오차를 피할 수는 없다. Kim 등은 영상을 통한 입술 움직임의 파라미터를 추출하는데 있어 10%의 인위적인 추정오차가 첨가될 경우, 인식률이 84%에서 19%로 크게 저하됨을 나타내었다[10]. 이러한 입술 움직임 검출의 오류는 측정변인 계산에 오차가 발생될 수 있어 인식률을 저해하는 요

† Corresponding Author : Dept. of Biomedical Engineering, BK21+ Research Institute of Biomedical Engineering, Konkuk University, Korea
E-mail : grtack@kku.ac.kr
* Dept. of Biomedical Engineering, Konkuk University, Korea
** Department of Mechanical Systems Engineering, Hansung University, Korea
Received : April 10, 2014; Accepted : May 27, 2014

소가 될 뿐만 아니라 각 음절에 대한 정량적인 패턴을 나타내는데 어려움이 있다. 또한 카메라 영상기법을 통해 획득한 2차원 영상은 3차원 공간상의 전체적인 안면 움직임의 특정패턴을 추출하는데 있어 제한점을 가진다. 얼굴 정면에서 촬영된 2차원 영상은 관찰점들 간의 거리를 이용하여 발성 시의 특정패턴을 도출함으로써, 측면 및 깊이의 공간상의 움직임과 코, 입술, 턱, 볼의 전체적인 안면 움직임의 다양한 파라미터를 나타내는데 한계가 있다. 현재 이러한 한계점을 극복하기 위한 방법으로 스테레오 카메라[12,13] 및 깊이 측정이 가능한 RGB-D(Red-Green-Blue Depth) 카메라센서[14]를 이용하여 3차원 공간상의 입술 움직임 추적과 좌표측정이 가능한 시스템개발이 수행되고 있지만, 미세한 안면 움직임의 변화를 나타나는데 있어 영상에 따른 추정오차를 배제할 수는 없다. 특히 초성, 중성, 종성으로 이루어진 방대한 양의 한글 음성인식에 있어 보다 정량적인 패턴분석과 다양한 파라미터를 통한 분석이 요구되어 진다. 따라서 영상기법을 통한 음성인식 시스템 및 알고리즘 개발에 앞서 보다 정확한 안면 움직임의 측정을 통해 정량적인 특징들을 추출할 필요가 있다.

이에 본 연구에서는 여러 가지 모음형태의 단음절 발성에 따른 보다 정확한 안면 움직임을 측정하기 위해 카메라 기반이 아닌 3차원 동작분석기를 사용하였으며, 음성인식을 수행하기 위한 다양한 인식 파라미터들을 보다 정략적으로 추출하고자 하였다. 3차원 동작분석기는 일반 카메라에 비해 촬영 공간, 마커부착 및 보정(Calibration)작업에 대한 영상획득 용이성이 낮고 데이터 후처리과정의 단점을 가지지만 공간상의 측정과 실험환경에 따른 오차가 적으며, 전체적인 안면 움직임의 획득을 통해 정량적이고 다양한 패턴분석을 수행할 수 있다는 점에서 본 연구에 사용되었다. 3차원 모션장비를 사용하여 모음형태의 단음절 발성 시의 안면 움직임을 획득하였으며, 최종적으로 데이터를 통해 추출된 파라미터들을 이용하여 음성인식 알고리즘을 구현하고 인식 정확도를 살펴보았다.

2. 본 론

2.1 단음절의 3차원 영상 획득

한글은 낱소리 문자에 속하며, 자음 19개와 모음 21개의 조합으로 구성된 문자이고 '초성 + 중성', '초성 + 중성 + 종성'과 같이 2가지 형태로 문자를 표현하고 있다. 자음은 초성과 종성 부분에 들어가고 모음은 중성 부분에 들어간다. 본 연구에서는 발음 시 입술 모양에 큰 영향을 미치는 11가지의 모음을 선택하여 단음절을 구성하였다. 표 1은 11가지의 인식대상 단음절의 한글표기와 발음표기를 나타낸다.

단음절 발성 시의 안면 움직임의 데이터 획득은 6대의 적외선카메라로 구성된 3차원 동작분석기(Motion Analysis Systems Inc., USA)를 사용하였다. 실험은 피험자의 얼굴에 직경 6mm의 36개 반사마커를 부착하여 120Hz의 샘플링으로 단음절을 표현하기 위한 안면의 움직임 및 궤적 데이터를 획득하였다. 피험자는 입을 다문 무표정 상태에서 단음절을 발성하여 안면 움직임의 변화를 측정하였다. 언어장애가 없는 20대 성인 14명을 대상으로 실험을 실시하였으며, 각 피험자는 보통속도로 11가지 단음절을 각 3회씩 수행하였다. 그림 1은 마커의 부착위치를 나타내며, MPEG-4 (Moving Picture Experts Group) 표준에 의해 정의되는 얼굴 구성 요소의 특징들을 참고하여 36개의 마커를 부착하였다[15].

표 1 기본 단음절

Table 1 Basic Monosyllables

단음절 (한글표기)	발음표기
아	/a/
야	/j/+/a/
어	/v/
여	/j/+/v/
에	/e/
오	/o/
요	/j/+/o/
우	/u/
유	/j/+/u/
으	/U/
이	/i/

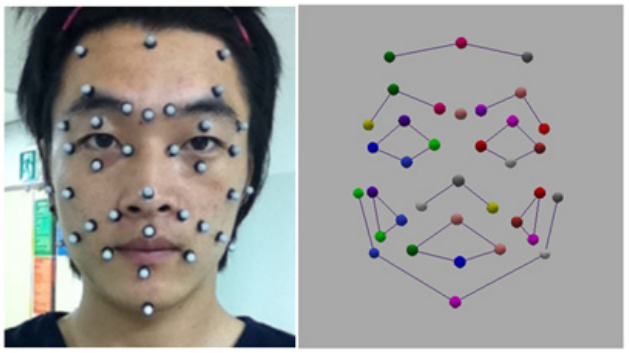

그림 1 안면 마커 셋과 3차원 동작 분석 시스템을 이용한 모션캡쳐 데이터

Fig. 1 Facial marker set and motion capture data using 3D motion analysis system.

2.2 단음절의 특징 패턴 추출

본 연구에서는 3차원 동작분석기의 데이터를 활용하여 안면움직임의 특징을 살펴보기 위해 파라미터 기반 접근법을 사용하였다. 기존 카메라 영상기법에서 사용되는 입술의 높이와 폭, 코끝과 입술의 거리, 입술 하단과 턱의 거리, 코끝과 턱의 거리의 5가지 파라미터를 바탕으로 3차원 공간상의 코, 입술, 턱, 볼의 전체적인 안면 움직임의 변화패턴을 나타내기 위한 보다 세분화된 11가지 파라미터를 선정하였다. 그림 2는 동작분석 데이터를 이용한 11가지 파라미터 추출 변인을 나타낸다.

특히 모음 발성 시에 큰 영향을 미치는 입술구조(폭, 높이, 깊이, 돌출, 면적, 각도)의 파라미터들을 계산하였으며, 각 조음기관(코, 입술, 턱)의 거리와 턴관절 각도 및 볼 면적

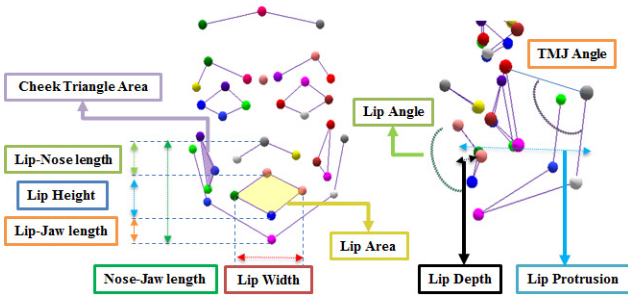

그림 2 3차원 동작 데이터를 이용한 패턴 추출 파라미터

Fig. 2 Pattern sample parameter using 3D facial motion data

의 각 파라미터를 이용하여 단음절의 패턴을 추출하기 위한 변인으로 사용하였다. 여기서 거리와 면적을 나타내는 입술의 폭(Lip Width), 높이(Lip Height), 깊이(Lip Depth), 돌출(Lip Protrusion), 면적(Lip Area), 볼의 면적(Cheek Triangle Area), 코끝과 입술의 거리(Lip-Nose length), 코끝과 턱의 거리(Nose-Jaw length), 입술 하단과 턱의 거리(Lip-Jaw length)의 9가지 파라미터는 입을 다문 정지 상태를 기본으로 거리와 면적의 변화 비율을 통해 파라미터를 계산하였다. 이는 개개인의 구강 구조가 서로 차이를 나타냄으로써, 측정변인에 대한 일반화된 데이터베이스 획득을 위해 수행되었다. 또한 입술의 벌어진 각도(Lip Angle) 및 턱관절 각도(Temporomandibular joint angle, TMJ Angle)는 정지 상태에서의 변화각도 차이를 계산하였다. 그림 3은 "아" 발성 시의 입술구조 변화에 따른 11가지 파라미터의 패턴 변화의 예를 나타낸다. 피험자를 통해 측정된 모든 데이터는 각각의 11가지 파라미터 패턴을 추출하여 단음절의 학습 및 인지를 위한 알고리즘 개발에 사용되었다.

2.3 파라미터 기반 음성인식을 위한 알고리즘 적용

음성인식 알고리즘 개발에 있어 기존 연구들에서는 퍼지 로직, Neural Network 및 Hidden Markov Model(HMM) 방법들을 주로 사용하고 있으며, 인식률을 높이기 위한 다양한 방법 및 알고리즘 개발이 수행되고 있다. 특히, Hidden Markov Model 은 시공간적인 정보를 통한 모델링과 학습 및 인식을 위한 효과적이고 우수한 알고리즘을 가지고 있어 여러 분야에서 응용되고 있으며[16], 음성인식에 있어 가장 널리 사용되어지고 있다[17]. 이에 본 연구에서는 시공간적 데이터인 11가지 파라미터의 패턴들을 적용하여 모델링, 학습 및 음성인식을 수행하기 위한 방법으로 HMM을 사용하였다. 또한 입력과 근사치의 출력을 생성하는 은닉 상태집합의 경로를 검출하는 Viterbi 알고리즘을 사용하여 최적 확률의 음절인식을 수행하였다[18]. Viterbi 알고리즘은 "초기 상태와 초기의 결정이 어떻든 간에 남아 있는 결정들은 첫 번째 결정으로부터 나온 상태에 대해 최적이어야 한다"는 최적의 원리(Principle of optimality)에 근거를 두고 있다. 그림 4는 HMM과 Viterbi 알고리즘이 적용된 음성인식 알고리즘을 통해 단음절의 패턴을 학습하고 인식하는 과정을 나타낸다.

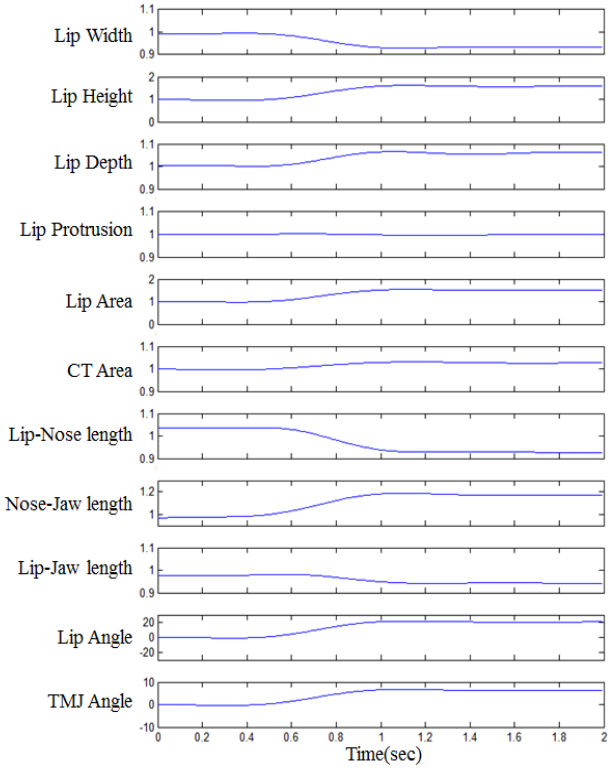

그림 3 단음절 발성 시의 11가지 파라미터 패턴

Fig. 3 Appearance of 11 parameters in monosyllable speech.

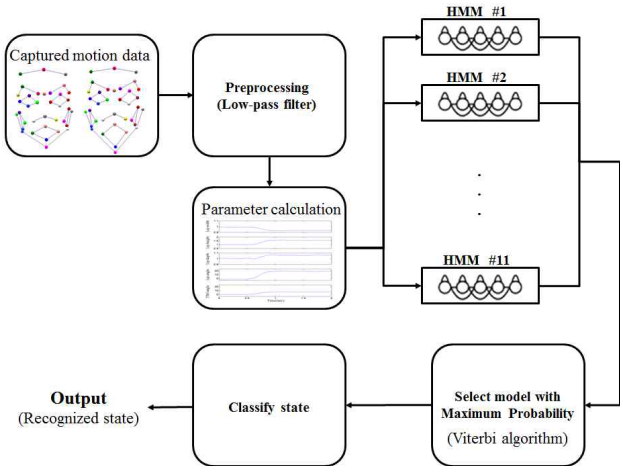

그림 4 HMM과 Viterbi 알고리즘이 적용된 음성인식 알고리즘의 전개도

Fig. 4 Block diagram of speech recognition algorithm which HMM and Viterbi algorithms are applied

음성인식 알고리즘은 먼저 획득된 안면 움직임 데이터를 6Hz의 저역통과 필터링을 거친 후에 11가지 파라미터에 대한 패턴을 추출한다. 추출된 패턴들은 기존 데이터를 통해 모델링된 HMM에 적용되어 파라미터 패턴에 대한 11가지 단음절의 상태 확률을 계산하며, 최종적으로 Viterbi 알고리즘을 이용하여 최적 확률의 단음절을 인식하게 된다. 구축

표 2 각 파라미터의 개별적 HMM을 통한 음성인식 검출 정확도

Table 2 The detection rate of speech recognition detection through individual HMM of each parameter

파라미터	Lip Width	Lip Height	Lip Depth	Lip Protrusion	Lip Area	CT Area	Lip-Nose length	Nose-Jaw length	Lip-Jaw length	Lip Angle	TMJ angle
인식률(%)	41.1	44.4	58.4	31.8	51.3	35.7	42.5	49.5	44.1	49.7	47.0

된 음성인식 알고리즘의 정확도는 3차원 안면 움직임의 각 데이터를 통해 추출된 파라미터들을 이용하였으며, 11가지 인식 파라미터의 음성인식에 대한 주요변인들을 파악하기 위해 개별적인 HMM의 정확도를 살펴보았다. 또한 선행연구 결과와의 비교를 위해 '아/에/이/오/우'의 인식을 위한 기존 5가지 파라미터의 HMM과 본 연구에서 사용된 11가지 파라미터의 HMM의 음성인식 정확도를 살펴보았다.

3. 실험결과

구축된 음성인식 알고리즘의 정확도를 살펴보기 위해 11가지 단음절의 각 42개 데이터로 총 462개를 사용하였으며, 3차원 안면 움직임의 각 데이터를 통해 추출된 파라미터들을 이용하였다. 표 2는 추출된 파라미터를 11가지 인식 파라미터의 개별적인 HMM에 적용하여 나타낸 음성인식 결과를 나타낸다. 여기서 개별적 파라미터의 HMM을 통한 11가지 단음절에 대한 음성인식 정확도는 입의 깊이(Lip Depth)에서 58.4%로 가장 높게 나타났으며, 입의 돌출(Lip Protrusion)에서 31.8%로 가장 낮은 정확도를 보였다. 선행연구에서 사용된 5가지의 파라미터(Lip Width, Lip Height, Lip-Nose length, Nose-Jaw length, Lip-Jaw length) 중에는 코끝과 턱의 거리(Lip-Jaw length)가 49.5%로 가장 높게 나타났다.

기존 선행연구의 5가지 파라미터의 HMM과 이를 바탕으로 구성된 11가지 파라미터의 HMM을 통한 음성인식 결과는 표 3에 나타내었다. 선행연구와 동일한 '아/에/이/오/우'의 단음절 인식대상의 경우, 5가지 파라미터 HMM의 인식 정확도는 87.6%로 나타났으며, 11가지 파라미터 HMM에서는 98.1%로 보다 향상된 인식 결과를 나타냈다. 또한 단음절 인식대상 '아/야/어/여/에/오/요/우/유/으/이'에서도 마찬가지로 5가지(88.1%) 보다 11가지(97.2%)의 파라미터로 구성된 HMM에서 높은 정확도를 나타내었다.

표 4는 11가지 파라미터 HMM을 통해 도출된 11가지 단음절 인식 정확도의 세부적인 결과를 나타낸다. 총 462개의 데이터 중에 449개의 단음절을 올바르게 인식하여 전체적으로 97.2%의 인식률 정확도를 나타내었다. 단음절 '야/오/우/으'는 100%로 가장 높은 인식률을 나타내었으며, '여'는 가장 낮은 85.6%로 나타났다. 또한 서로 간의 유사한 발음형태를 가지는 {아, 야, 어, 여, 에}, {오, 요, 우, 유}, {으, 이}의 3가지 그룹은 각 그룹 내에 오인식이 13개중에 11개로 그룹 간에 보다 많이 분포됨을 나타내었다. 3가지 그룹의 분류는 11가지 단음절의 모음이 입술 모양에 따라 평순모음(unrounded vowel) {ㅏ, ㅑ, ㅓ, ㅕ, ㅡ, ㅣ, ㅐ}와 원순모음(round vowel) {ㅗ, ㅛ, ㅜ, ㅠ}로 나눌 수 있으며, 평순모음은 다시 개구부의 크기의 특징에 따라 {ㅏ, ㅑ, ㅓ, ㅕ, ㅐ}, {ㅡ, ㅣ}로 구분됨을 알 수 있다. 여기서 각 그룹 내에 오인식의 발생 빈도가 높은 이유는 발성 시의 구강은 입모양뿐만 아니라 혀의 위치도 변화함으로써, 단음절 발성 시의 혀의 위치는 다르지만 유사한 입모양의 패턴을 나타낸 것으로 사료된다.

4. 고 찰

인간의 기본적인 의사소통 수단은 음성이지만, 실생활에서 인간은 표정, 몸짓, 글자 등 다양한 수단을 사용한다. 특히 말소리의 정확한 이해를 위하여 인간은 무의식적으로 영상정보를 이용한다[19]. 이러한 영상정보를 이용한 독화(speech reading)기술의 핵심은 발화자를 잘 찾아 시각적으로 보이는 안면 움직임의 변화를 정밀히 인식하여 그 변화를 파라미터로 추출하고 인식에 반영하는 것이다[20]. 영상기반 접근 방법은 영상 자체 또는 영상의 코딩값이나 영상처리 후의 영상 자체를 시각 특징으로 사용하는 방법이다. 이와 같은 영상 기반 접근 방법은 다른 접근 방법에 비해 정보량이 많아 데이터 처리 면에서 손해를 보지만 정보량을 줄이기 위한 방법들을 적용하면 극복 가능하며, 인식 성능 면에서 우수한 결과를 보인다[21]. 그러나 근본적으로 카메라 기법을 이용한 영상기법은 환경적 요인에 의한 추정오차와 입술영역을 검출하는데 있어 영상처리 기술에 따른 오차를 배제할 수 없다. 이에 반해 3차원 동작분석기는 입술 움직임의 보다 세분화된 특징들의 분석은 다양한 음성인식에

표 3 단음절 인식대상과 HMM에 따른 음성인식 정확도

Table 3 The accuracy of speech recognition through target recognition of monosyllables and HMM

단음절 인식대상	n=210	인식	오인식	인식률(%)
'아/에/이/오/우' (5가지)	5가지 파라미터 HMM	184	26	87.6
	11가지 파라미터 HMM	206	4	98.1
단음절 인식대상	n=462	인식	오인식	인식률(%)
'아/야/어/여/에/오/요/우/유/으/이' (11가지)	5가지 파라미터 HMM	407	55	88.1
	11가지 파라미터 HMM	449	13	97.2

표 4 음성인식 알고리즘을 통한 단음절 검출의 정확도

Table 4 The accuracy rate of monosyllable detection through speech recognition algorithm

인식률(%)	아 (/a/)	야 (/j/+/a/)	어 (/v/)	여 (/j/+/v/)	에 (/e/)	오 (/o/)	요 (/j/+/o/)	우 (/u/)	유 (/j/+/u/)	으 (/U/)	이 (/i/)
아 (/a/)	**95.2**		2.4	2.4	2.4						
야 (/j/+/a/)	2.4	**100**		2.4							
어 (/v/)			**97.6**	4.8							
여 (/j/+/v/)				**85.6**							
에 (/e/)				2.4	**97.6**						
오 (/o/)						**100**	2.4				
요 (/j/+/o/)							**97.6**		2.4		
우 (/u/)								**100**			
유 (/j/+/u/)									**97.6**		
으 (/U/)	2.4									**100**	2.4
이 (/i/)				2.4							**97.6**

필요한 안면움직임의 정량적 패턴을 나타낼 수 있을 뿐만 아니라 깊이와 영상을 포함하는 3차원 고해상도 카메라를 사용하는 음성인식 알고리즘의 정확도를 개선할 수 있다는 점에서 선행되어야 한다.

본 연구에서는 11가지 모음형태의 단음절 발성에 따른 보다 정확한 안면 움직임을 측정하기 위해 카메라 기반이 아닌 3차원 동작분석기를 사용하였다. 또한 음성인식을 수행하기 위한 다양한 인식 파라미터들을 보다 정략적으로 추출하고자 하였으며, 최종적으로 데이터를 통해 추출된 11가지 파라미터들을 이용하여 음성인식 알고리즘을 구현하고 인식정확도를 살펴보았다. 카메라 기반 기술의 국내 연구들과 비교하면, '아/에/이/오/우'의 5가지 모음의 인식정확도가 Lee 등[8]과 Kim 등[10]의 연구에서 각각 87.44%와 84%를 나타내었다. 본 연구에서도 동일한 파라미터를 이용하여 5가지 모음을 인식한 결과 87.6%로 나타났으며, 환경변화가 없는 카메라 기반의 실내에서의 실험을 통해 영상을 획득한 결과와 큰 차이를 나타내지는 않았다. 그러나 본 연구에서 보다 세분화된 11가지의 파라미터를 사용하였을 경우, 98.1%로 인식결과가 향상됨을 나타냈다. 이는 기존 5가지 파라미터에서 나타내지 못하는 공간상의 안면 움직임의 파라미터들이 포함되어 보다 향상된 인식결과를 나타내고 있으며, 특히 입의 깊이(Lip Depth)는 11가지 파라미터 중에 가장 높은 인식결과를 보임으로써 음성인식에 있어 주요한 변인임을 알 수 있다. Lee 등[9]과 Nam 등[11]의 연구에서는 8가지 모음을 대상으로 각각 93.7%와 55%의 인식정확도를 나타내었다. 단모음 인식대상과 인식 알고리즘의 차이로 직접적인 비교는 어려우나 보다 다양한 11가지 단모음의 인식에 대한 본 연구의 결과가 97.2%로 더 높게 나타났다. 앞서 마찬가지로 11가지 단음절에 대한 기존 파라미터의 결과(88.1%) 보다 세분화된 11가지 파라미터의 인식 정확도가 높게 나타남으로써, 음성인식을 위한 보다 정량적이고 세분화된 파라미터의 필요성을 나타내었다.

모션캡처는 안면 움직임을 포함하는 인간 동작을 측정하는데 있어 실현가능하고 강력한 접근방법으로 입증되었으며[22], 기존 많은 연구에서도 모션캡처를 이용한 정량적 측정을 통해 시각적 음성합성(Visible Speech Synthesis)[23], 애니메이션[24] 및 의사소통 프로그램 개발[25]등에 활용되고 있다. Ma 등[23]은 3차원 동작분석기를 이용하여 21가지의 단음절 발성 시, 정량적 안면움직임의 패턴을 분석하고 이를 3차원 얼굴 모델에 적용하는 시각적 음성합성에 대한 연구를 수행하였다. 개발된 시각적 음성합성 시스템은 현재 언어를 배우는 60 개 이상의 유치원에서 학습 에이전트로써 활용되고 있다. Cao 등[24]은 음성을 통해 실시간으로 그에 상응하는 사실적인 얼굴 애니메이션을 구현하는 방법을 개발하였다. 얼굴 애니메이션 모델은 3차원 모션캡처 데이터를 통해 구현하였으며, 보다 실제적인 안면 움직임을 나타내고자 하였다. 이러한 기술개발은 대화형 게임 및 가상현실 등의 적용 가능성을 제시하였다. 또한 안면 움직임의 정량적 패턴의 측정은 시청각 자동 음성인식(Audio-Visual Automatic Speech Recognition, AV-ASR)의 알고리즘에 적용하여 음성인식에 대한 정확도를 높이는데 활용될 수 있다. 음성인식을 위한 신뢰할 수 있는 영상특징들은 특히 잡음환경에서 음성만을 이용한 ASR의 수행 개선을 위해 발성의 음성과 영상을 혼합하는 AV-ASR의 성능 향상에 도움을 준다는 점[26]에서 활용도가 크다고 할 수 있다. 본 연구결과를 통해 도출된 정량적인 안면 움직임의 패턴들을 이용한 한글 음성인식 알고리즘은 기존 시청각 자동 음성인식의 성능 향상뿐만 아니라 3차원 안면 움직임의 정밀한 측정이 가능한 스테레오 및 RGB-D 형태와 같은 다양한 카메라 기법에 적용한다면, 보다 실용적인 음성인식 시스템을 구현하는데 있어 사용이 가능하겠다.

5. 결 론

본 연구에서는 3차원 동작분석 시스템을 이용하여 11가지 형태의 단음절 발성 시의 정량적인 안면 움직임을 측정하였

고 개개인의 구강구조에 따라 달라지는 파라미터들을 안면 움직임의 비율에 따라 객관적인 데이터로 나타내었다. 또한 이러한 시공간적인 파라미터 패턴들을 Hidden Markov Model과 Viterbi 알고리즘이 적용된 음성인식 알고리즘을 통해 각 단음절의 패턴을 학습하고 인식하는 과정을 수행하였다. 음성인식 알고리즘을 적용한 11가지 단음절의 인식률의 정확도는 97.2%를 나타내었으며, 정량적인 안면 움직임을 통한 한글 음성인식의 충분한 가능성을 제시하였다.

본 연구에서는 11가지 모음 형태의 단음절을 인식하는 알고리즘을 구현하였지만, 실생활에서 사용되는 보다 복잡하고 많은 형태의 음절들을 인식하기 위해서는 추후 연구가 수행되어야 하겠다. 현재 구현된 알고리즘이 수많은 형태의 음절을 인식 할 수는 없지만, 이를 바탕으로 다양한 분석 방법과 알고리즘 개발을 통해 보다 복잡한 음성인식을 수행할 수 있는 연구의 기초가 될 수 있으며, 기존 카메라를 이용한 음성인식에 적용하여 성능향상과 보다 실용적인 음성인식 시스템 구현이 가능하겠다.

감사의 글

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단-공공복지안전사업의 지원을 받아 수행된 연구임 (No. 2011-0020972)

References

[1] N. Eveno, A. Capiler, and P. Y. Coulon, “Accurate and quasi-automatic lip tracking,” IEEE Transactions of Circuits and Systems for Video Technology, vol. 14, no. 5, pp. 706-715, 2004.

[2] A. Bagai, H. Gandhi, R. Goyal, M. Kohli, and T. V. Prasad, “Lip-Reading using Neural Networks, International Journal of Computer Science and Network Security,” vol. 9, no. 4, pp. 108-111, 2009.

[3] H. Mehrotra, G. Agrawal, and M. C. Srivastava, “Automatic Lip Contour Tracking and Visual Character Recognition for Computerized Lip Reading,” International Journal of Computer Science, vol. 4, no. 1, pp. 62-71, 2009.

[4] W. J. Ma, X. Zhou, L. A. Ross, J. J. Foxe, and L. C. Parra, “Lip reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space,” PLoS ONE, vol. 4, no. 3, pp. 1-14, 2009.

[5] J. J. Shin, J. Lee, and D. J. Kim, “Real-time lip reading system for isolated Korean word recognition,” Pattern Recognition, vol. 44, pp. 559 - 571, 2011.

[6] M. G. Song, T. P. Thanh, J. Y. Kim, and S.T. Hwang, “A Study on Lip Detection based on Eye Localization for Visual Speech Recognition in Mobile Environment,” Journal of Korean institute of intelligent systems, vol. 19, no. 4, pp. 478-484, 2009.

[7] Y. T. Won, H. D. Kim, M. R. Lee, B. S. Jang, and H. S. Kwak, “A Character Speech Animation System for Language Education for Each Hearing Impaired Person, Journal of digital contents society, vol. 9, no. 3, pp. 389-398, 2008.

[8] K. H. Lee, J. J. Kum, and S. B. Rhee, “Design & Implementation of Lipreading System using the Articulatory Controls Analysis of the Korean 5 Vowels,” The Journal of Korean association of computer education, vol. 8, no. 4, pp. 281-288, 2007.

[9] K. H. Lee, R. Yong, and S. O. Kim, “A study on speechreading the Korean 8 vowels,” Journal of the Korea society of computer and information, vol. 14, no. 3, pp. 173-182, 2009.

[10] J. Y. Kim, S. H. Min, and S. H. Choi, “Robustness of Bimodal Speech Recognition on Degradation of Lip Parameter Estimation Performance,” Journal of the Korean Society of Phonetic Science and Speech Technology, vol. 10, no. 2, pp. 29-33, 2003.

[11] K. H. Nam, and C. S. Bae, “A study on the lip shape recognition algorithm using 3-D Model,” The Journal of the Korean Institute of Maritime Information & Communication Sciences, vol. 6, no. 5, pp. 783-788, 2002.

[12] C. G. Lee, I. M. So, Y. U. Kim, J. R. Kim, S. K. Kang, and S. T. Jung, “Implementation of three dimension lip reading system using stereo vision,” Proceedings of Korea multimedia society conference (KMMS ‘04), pp. 489-492, 2004.

[13] H. S. Koh, S. M. Han, J. U. Chu, S. H. Park, J. B. Choi, G. W. Choi, D. S. Hwang, and I. C. Youn, “The three-dimensional lip shape tracking system using stereo camera,” Proceedings of the Korean Society of Precision Engineering (KSPE ‘11) Conference, pp. 979-980, 2011.

[14] G. Galatas, G. Potamianos, D. Kosmopoulos, C. McMurrough, and F. Makedon, “Bilingual Corpus for AVASR using Multiple Sensors and Depth Information,“ Auditory-Visual Speech Processing (AVSP ‘11), pp. 103-106, 2011.

[15] I. S. Pandzic, and R. Forchheimer, MPEG-4 Facial Animation: The Standard, Implementation, and Applications, John Wiley and Sons, Inc., New York, 2002.

[16] X. D. Huang, Y. Ariki, and M.A. Jack, Hidden Markov Models for Speech Recognition, Edinburgh Univ. Press, Edinburgh, 1990.

[17] A. Srinivasan, “Speech Recognition Using Hidden Markov Model,” Applied Mathematical Sciences, vol. 5, no. 79, pp. 3943-3948, 2011.

[18] D. A. Pierre, Optimization Theory with Applications, Dover Publications, Inc., New York, 1986.

[19] T. Chen, H. P. Graf, and K. Wang, “Speech-assisted video processing: Interpolation and low-bitrate coding,” 28th Annual Asilomar Conference on Signals, Systems, and Computers (Asilomar ‘94), pp. 957-979, 1994.
[20] G. Baily, E. Vatikiotis-Bateson, and P. Perrier, Visual and audio-visual speech processing, MIT press, 2004.
[21] P. Scanlon, and R. Reilly, “Feature analysis for automatic speech reading,” Proc. of the IEEE Int. Conf. on Multimedia Signal Processing (MMSP ‘01), pp. 625-630, 2001.
[22] R. SCOTT, “Sparking life: notes on the performance capture sessions for the lord of the rings: the two towers,” ACM SIG-GRAPH Computer Graphics, vol. 37, no. 4, pp. 17 - 21, 2003.
[23] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, “Accurate Visible Speech Synthesis Based on Concatenating Variable Length Motion Capture Data,” IEEE Transactions on Visualization and computer Graphics, vol. 12, no. 2, pp. 266-276, 2006.
[24] Y. Cao, P. Faloutsos, E. Kohler, and F. Pighin, “Real-time speech motion synthesis from recorded motions,” In Proceedings of Eurographics/SIGGRAPH Symposium on Computer Animation (SCA ’04), pp. 345-353, 2004.
[25] G. Bailly, F. Elisei, M. Odisio, D. Pelé, D. Caillière, and K. Grein-Cochard, “Talking faces for MPEG-4 compliant scalable face-to-face telecommunication,” Proceedings of the Smart Objects Conference (SOC ‘03), pp. 204-207, 2003.
[26] S. Dupont, and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” In IEEE Trans-actions on Multimedia, vol. 2, pp. 141 - 151, 2000.

저 자 소 개

강 동 원 (姜 同 院)

1982년 1월 14일생. 2009년 건국대학교 의료생명대학 의학공학과 졸업(공학석사). 2009년～현재 동대학원 박사과정 재학 중.
E-mail : dwkang00@gmail.com

서 정 우 (徐 政 佑)

1983년 9월 8일생. 2013년 건국대학교 의료생명대학 의학공학과 졸업(공학석사). 2013～2014년 보건복지부 국립재활원 재활연구소 재활로봇중개연구사업단 연구원. 2014년～현재 동 대학원 박사과정 재학 중.
E-mail : jwseo0908@gmail.com

최 진 승 (崔 珍 丞)

1979년 4월 3일생. 2012년 건국대학교 의료생명대학 의학공학과 졸업(공학박사). 2012～2013년 건국대학교 의공학실용기술연구소 박사후연구원(한국연구재단 학문후속세대양성사업지원). 2013년～현재 건국대학교 의학공학부 조교수.
E-mail : jschoi98@kku.ac.kr

최 재 봉 (崔 載 烽)

1960년 7월 30일생. 1993년 U. of Iowa 생체공학 박사. 1994년~1998년 KIST 의과학연구센터 선임연구원. 2005년 Duke 대학 방문연구교수. 1999년~현재 한성대학교 기계시스템공학과 교수.
E-mail : jbchoi@hansung.ac.kr

탁 계 래 (卓 烓 來)

1960년 7월 12일생. 1991년 U. of Iowa, Biomedical Eng. 박사. 1995～1997년 삼성SDS(주) 정보기술연구소 PACS Lab. Director. 2003～2004년 U. of Calgary, Human Performance Lab. 방문교수. 1997년～현재 건국대학교 의학공학부 교수 및 BK21+의공학실용기술연구소장.
E-mail : grtack@kku.ac.kr