Lipreading과 음성인식에 의한 향상된 화자 인증 시스템

°지 승 남*, 이 종 수**

* 홍익대학교 전기제어공학과(Tel : 82-02-325-7514; Fax : 82-02-320-1110 ; E-mail: jitwins@hitel.net)
** 홍익대학교 전기제어공학과 부교수(Tel : 82-02-320-1669; Fax : 82-02-320-1110 ; E-mail: leejs@wow.hongik.ac.kr)

Abstract : In the future, the convenient speech command system will become an widely-using interface in automation systems. But the previous research in speech recognition didn't give satisfactory recognition results for the practical realization in the noise environment. The purpose of this research is the development of a practical system, which reliably recognizes the speech command of the registered users, by complementing an existing research which used the image information with the speech signal.

For the lip-reading feature extraction from a image, we used the DWT(Discrete Wavelet Transform), which reduces the size and gives useful characteristics of the original image. And to enhance the robustness to the environmental changes of speakers, we acquired the speech signal by stereo method. We designed an economic stand-alone system, which adopted a Bt829 and an AD1819B with a TMS320C31 DSP based add-on board.

Keywords : lipreading, speech recognition, robot sensing, discrete wavelet transform(DWT)

1. 서론

편리한 음성 명령 기능은 향후 자동화 시스템에서 널리 쓰이는 사용자 인터페이스가 될 것이다. 기존의 음성 인식에 관한 연구는 잡음이 많은 환경이나 여러 가지 문제점이 있어서, 만족할만한 인식 결과를 보여 주지 못하는 경우가 많았다. 본 연구에서는 음성신호와 영상 정보를 동시에 이용한 기존의 연구를 보완하여, 등록된 화자의 음성 명령에 반응하는 신뢰도가 높은 실용적인 시스템의 구현을 목적으로 한다.

영상 정보와 음성 정보를 통합하여 음성 인식의 성능을 높이기 위한 이전의 연구 사례는 잡음에 강인한 음성 인식기를 구현하려는 목적에서 비롯되었다[1][2].

한국어 이외의 언어에 대하여는 비교적 많은 연구가 있었다. Petatjan[3]은 입력 동영상으로부터 인덱스 벡터를 생성하여 코드북과 가장 가까운 영상의 인덱스로써 대응하는 입 모양 영상을 결정하기 위한 벡터 양자화의 샘플벡터와 각 모델간의 거리를 구하여 인식하였고, Finn과 Montgomery[4]는 화자의 입 주위에 12개의 특징 점을 정하고, 각 영상에서 구한 이 특징 점들의 이동 변위를 인식에 대한 안면 벡터로서 사용하였다. Mase와 Pentland[5]는 시간에 대한 입술 움직임을 열림 정도(opening level)와 벌어짐(elongation)의 두 형태로 표현하기 위해 광류(optical flow)를 사용하였다. 또한 Bregler[6]는 modular MS-TDNN(multi-state time delay neural network)이라고 하는 음성인식 시스템을 사용하여 청각적이고 시각적인 음성 데이터를 첫, 끝 음절에서 미리 분류하여 인식하는 방법을 제시하였다.

반면에 한국어에 관한 연구로는 3차원 모델을 사용하여 입술 영상 정보에서 음성 정보를 추정[7]하거나, 음성 정보와 영상 정보를 결합하는 방식에 따른 평가[8], 음성 정보의 잡음 비율에 따른 평가[9]에 관한 연구 등, 아직까지는 입술 영상 정보를 추가적으로 활용하여 실용적인 시스템을 구현하는 사례는 많이 부족한 실정이다.

본 논문에서는 위에서 언급된 연구 결과를 토대로, 실용적인 시스템에서 입술 영상 정보를 활용하는 방안을 모색하고 실험 결과를 통하여 그 유용성을 검증하고자 한다.

이를 위하여 음성 정보를 이용하는 방법으로는 스테레오 방식을 사용하여, 정보 추출의 전처리 단계에서 잡음이나 음원의 위치 등, 환경적인 성능 저해 요인에 대한 보상을 고려하였다. 또한 입술 영상 정보를 이용하는 방법에는 일반적으로 세 가지 방법이 사용되고 있는데, 첫 번째로 이미지에 근거해서(Pixel-based method) 입술 파라미터를 추출하는 방법[10][11], 두 번째로 입술 모양을 모델화하여 모델에 근거(Model-based method)한 파라미터 추출 방법[12][13] 그리고 빛의 밝기의 변화를 벡터 화(Optical flow method)하여 파라미터로 추출하는 방법[14]이 있다. 이중 본 논문에서는 마지막 방법과 유사성이 있는 웨이블렛 변환을 통해 영상 정보를 이용하였다.

또한 시스템의 소형, 경량화를 보조적인 목표(subgoal)로 잡고, 이를 위해서 기존 연구에서 사용된 고가의 영상 처리 장비와 A/D Converter Data Acquisition System을 배제하고, 실제 사용처에서 응용할 수 있을 만한 시스템을 구현할 수 있도록 설계하였다.

2. 시스템 구성과 전처리

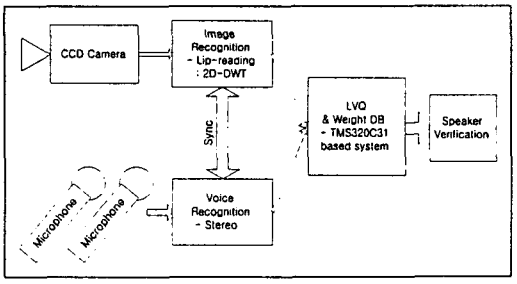

그림 1. 전체 시스템 개요
Fig. 1. System Architecture

본 논문에서는 Texas Instrument의 TMS320C31 DSP를 기반으로 Conexant(BrookTree)의 Bt829 Video Stream II Decoder와 Analog Device의 AD1819B AC97 SoundPort Codec를 채용한 저