# 외부응향잡음 차단을 위한 강인한 입술움직임 영상영역 추적방법

**김응규***
한밭대*

# A Tracking Method of Robust Lip Movement Image Regions for Blocking the External Acoustic Noise

Eung-Kyeu Kim*
The Dept. of Information & Communication Eng., Hanbat National University*

**Abstract** - 본 논문에서 조명 환경하에서 음성/영상 연동시스템을 통해서 외부음향잡음의 차단을 위한 강인한 입술움직임 영상영역을 추적하는 한 가지 방법을 제안한다. 조명환경하에서 강인한 입술움직임 영상영역을 추적하기 위해 온라인상에서 입술움직임 표준영상을 수집하였고 다양한 조명환경에 적응하는 입술움직임 영상의 특징들을 추출하였다. 동시에 온라인 템플릿 영상을 획득하였고, 이 영상들을 템플릿 정합을 위해 사용했다. 음성/영상처리시스템의 연동결과, 다양한 조명환경하에서 그 연동률을 99.3%까지 높일 수 있었고 음향잡음에 의한 음성인식 실행을 원천적으로 차단할 수 있었다.

## 1. 서 론

음성인식은 기본적으로 음성에너지를 분석의 대상으로 한다. 그러나, 실제 음성인식 서비스 환경에서 마이크를 통하여 음성인식 과정에 유입되는 음향에너지에는 배경잡음 및 다양한 동적 음향잡음이 존재한다[1]. 특히, 음성인식 로봇과 같이 외부 음향잡음에 노출되어 있는 상황에서 음성인식을 수행하는 경우에 음성인식 대상이 아닌 동적잡음이 예고 없이 유입될 수 있으며, 이들 잡음이 음성으로 오인식 되는 경우에는 심각한 문제를 야기시킬 수 있다. 한편, 영상의 경우에는 음향잡음과는 무관하게 획득되고 처리될 수 있어 이를 음성인식에 활용하려는 노력이 계속되고 있다[2]-[3]. 사람은 말을 할 때 입술을 움직이게 되므로, 음향잡음을 효과적으로 방지하기 위하여 음성인식 과정에서 입술움직임 영상신호를 활용하려는 것이다[4]-[5]. 음성정보와 영상정보를 보다 효율적으로 결합함으로써 기존의 음성인식 속도와 인식률을 저하시키지 않으면서 음성인식 대상이 아닌 외부 음향잡음을 효과적으로 차단하고, 음성인식 절차가 불필요하게 진행되는 것을 방지할 필요성이 제기된다.

본 논문에서는 음성인식의 전처리 단계인 음성구간 검출 과정에서 음향에너지에 대한 기존의 분석 이외에 화자의 입술움직임 영상신호를 추가 확인하는 음성/영상 연동 시스템을 구축함으로써 다양한 영상환경에서 강인한 입술움직임 영상영역을 보다 정확히 추적하는 한 방법을 제안 한다[6]-[8]. 이를 위해 온라인 하에서 입술움직임 표준 영상을 추출하였고, 이에 관한 특징벡터를 추출하여 오프라인에서 산출된 특징벡터 초기 모델을 대체하였다. 마찬가지로 템플릿 초기 모델을 대체하였다. 또한 입술움직임 영역 추적에 관한 주요 파라미터를 시각적으로 확인하고 검증하기 위하여 영상처리 테스트베드를 구축하였다.

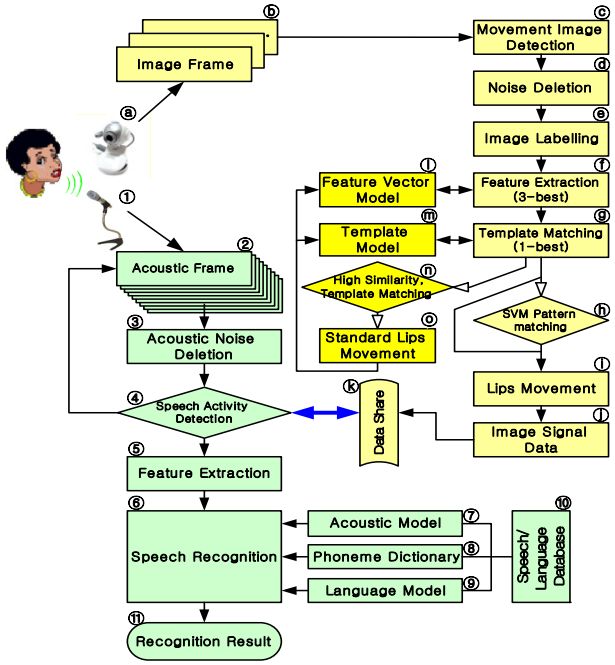
이하, 2장에서는 음성/영상 연동시스템의 주요 구조에 대해 기술한다. 3장에서는 입술움직임 표준영상 추출 및 영상환경에 대해 기술한다. 4장에서는 입술움직임 영역 추적 테스트베드 구축에 대해 기술하고, 5장에서는 실험환경 및 실험결과에 대해 기술한다.

## 2. 음성/영상 연동시스템

그림 1은 음성인식 과정의 음성구간 검출단계에서 입술움직임 영상신호를 활용하는 것을 보여주는 음성/영상 연동시스템 구조이다. 음성인식기(녹색)와 영상처리기(노란색)는 공유메모리(k)를 통하여 연동되며, 영상처리기가 실행되는 시점부터 입술움직임 영역 추적 결과가 공유메모리에 저장되기 시작하고, 음성인식의 음성구간 검출단계(4)에 반영된다.

### 2.1 음성인식절차

음성/영상 연동상태에서의 음성인식 절차를 요약하면 다음과 같다. 먼저, 음성인식을 위한 사람의 발성뿐만 아니라 다양한 형태의 외부 음향잡음을 포함한 음향에너지가 마이크(1)를 통하여 입력된다. 음향흐름 데이터를 적절한 크기의 프레임(2)별로 나누고 필터링 과정을 거침으로써 일정한 크기와 고주파를 특징으로 하는 정적인 잡음을 제거(3)한다. 이어서 음향에너지의 크기와 지속성 그리고 새로 추가된 입술움직임 영상신호 데이터(k)를 확인하여, 사람의 발성에 의한 음향에너지로 판단되면 음성구간으로 표시(4)한다. 이어서 음성구간을 대상으로 프레임별로 특징추출(5)하고 음성인식(6)을 수행한다. 음성/언어 데이터베이스(10)를 기반으로 음향모델(7), 발음사전(8), 언어모델(9)을 미리 구축하며, 이들을 통하여 음성인식을 위한 탐색공간이 형성된다. 음성인식 과정(6)에서는 탐색공간 내에서 음성특징값(5)을 비교하여 인식결과(11)를 도출한다.




**<그림 1> 음성/영상 연동시스템의 구조**

### 2.2 입술움직임 영상신호 추출

한편, 음성인식을 위해 발성하는 화자의 영상은 PC용 영상카메라(a)에 의해 영상프레임(b)으로 획득된다. 인접한 영상 프레임간의 비교에 의한 움직임 영역(c) 추출, 잡음영역 제거(d), 영역별 라벨링(e), 라벨링 영역별 특징추출(f) 및 특징벡터 모델(l)에 의한 입술움직임 후보 영역 선정, 그리고 템플릿 모델(m)에 의해 최종후보를 선정(g)한다. 템플릿 모델에 의한 템플릿 정합률이 산출되고 템플릿 정합률 분포에 의해 입술움직임 영역 또는 여타의 얼굴요소 영역을 구분하는 임계값이 형성된다. 입술움직임 최종후보는 템플릿 정합율이 임계값 이상인지에 따라 결정된다. 한편, 템플릿 정합률의 축적이 미흡한 초기단계에서는 SVM (Support Vector machine)에 의한 영역분류(h)를 거친 후 실제 입술움직임인지(i)의 여부가 판별된다. 템플릿 정합율은 영상신호

가 되어 공유메모리(k)에 저장되고, 음성구간 검출(4) 과정에서 활용된다.

#### 2.2.1 입술움직임 영역 특징

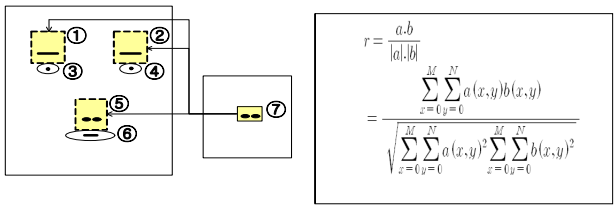
입술움직임 영역 특징벡터(l)은 다수의 움직임 중에서 입술움직임 영역과 보다 유사한 3개의 후보(f)를 선정하는 기준이 된다. 입술움직임 영역의 특징요소로는 그 폭과 높이, 이들의 비율, 넓이, 평균 픽셀값 차이, 공간 좌표상의 상대적 위치 등 7가지로 설정하였다. 입술움직임 영역 특징벡터(l) 초기화를 위해서는 오프라인에서 미리 이들 데이터를 수집하여야 하는데, 카메라로부터 약 50cm 거리에서 수집하였다.

#### 2.2.2 템플릿 선정

입술움직임 영역의 특징에 의해 선정한 3개의 입술움직임 후보(f)를 대상으로 템플릿 모델(m)과의 정합률에 의해 최종후보(g)를 선정한다. 템플릿은 코의 일부 영상으로 하였다. 왜냐하면, 입술 바로 위에 있으면서도 말하는 과정에서 그 크기와 모양의 변화가 거의 없고 명암의 구분이 뚜렷하기 때문이다. 템플릿 정합은 픽셀단위로 픽셀값 사이의 편차를 구함으로써 영상간의 상관성을 보다 정확히 측정하는데 적합한 방법이다.

#### 2.2.3 템플릿 정합률 측정

그림2는 템플릿 정합률을 측정하는 모형(a)과 산출식(b)을 나타낸다. 그림 2(b)에서 a(x,y)는 입력영상에서 취한 비교부 g(x,y)에서 평균 E(g)를 뺀 밝기값을 나타내며, b(x,y)는 템플릿 영상 t(x,y)에서 평균밝기 E(t)를 뺀 밝기값을 나타낸다. x, y는 수직 및 수평방향으로의 인덱스로서 픽셀좌표이다. 정합률 산출식(b)은 농담정규화 정합법 (Normalized Gray-level Correlation)으로서, 템플릿(그림 1(a)-7)과 이와 비교할 검색범위의 일정영역 (그림 1(a)-1,2,5)을 밝기에 대해 정규화시켜 비교함으로써 전체적인 조명의 밝기가 높아지거나 낮아지는 경우에도 패턴의 비교를 가능하게 한다. 검색범위는 각 움직임 후보(그림 1(a)-3,4,6)의 바로위에 위치한다. 템플릿 정합률이 산출되고 축적되면서 입술움직임 여부를 구분할 수 있는 임계값이 형성된다.




(a) (b)

〈그림 2〉 템플릿 정합률 산출모형(a) 및 그 산출식(b)

### 3. 입술움직임 표준영상 추출 및 영상환경 적응

3개의 입술움직임 후보를 선정하기 위한 특징벡터 초기 모델(그림1-l)과 최종후보를 선별하기 위한 템플릿 초기 모델(그림 1-m)은 오프라인 하의 특정한 영상 환경에서 산출된 것이다. 따라서 영상획득 환경이 바뀌는 실제 상황에서는 변별력이 그만큼 떨어지게 된다. 다양한 영상환경 적응을 위해, 온라인 하에서 입술움직임 표준영상을 추출한 후 이를 매개로 하여 특징벡터 초기모델과 템플릿 초기모델을 온라인에서 추출한 특징모델과 템플릿 모델로 각각 대체하였다. 이로써 다양한 영상환경에 적응할 수 있게 되었다.

#### 3.1 입술움직임 표준영상 추출

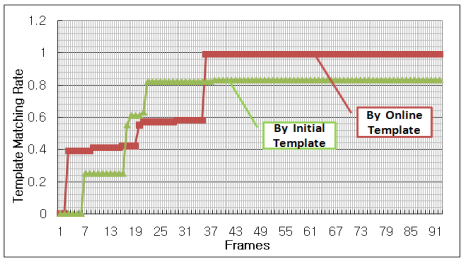
입술움직임 표준영상은 템플릿 정합률이 임계값의 일정비율 이상을 초과하는 정합률을 보이는 움직임 영역, 즉 입술움직임 영역이 확실시되는 영역만을 대상으로 특징값(그림1-o)을 추출하여 산출하였다. 입술움직임 표준영상 특징값 추출 및 학습을 위한 제한 요건은 표 1과 같다.

〈표 1〉 입술움직임 표준영상 추출 및 학습 요건(예)

| | |
|---|---|
| 템플릿 정합률 누적 최대값(A) | >0.5 |
| 템플릿 정합률(B) | >A-0.2 |
| 입술움직임 구분 임계값(C) | >0.3(minimum) |
| 템플릿 정합률 대비율(B/A) | >C * 1.2 |
| 학습시작 (위 요건 충족 횟수) | >10 |

#### 3.2 온라인 모델에 의한 대체 효과

그림 3은 초기 템플릿 모델과 온라인에서 추출된 템플릿 모델을 각각 적용한 경우에 그 정합률의 최고치가 각각 어느 수준까지 올라갈 수 있는지를 비교하여 측정한 것이다. 오프라인에서 산출한 템플릿 영상이 적용된 경우에는 그 최대값이 0.83의 수준을 넘지 못함을 볼 수 있는데 비하여 입술움직임 표준영상을 기반으로 온라인에서 추출한 템플릿 모델을 적용한 경우에는 정합률 최대값이 1.0에 수렴되고 있음을 보여주고 있다. 이것은 입술움직임 영역 추적의 변별력을 획기적으로 높였다고 볼 수 있다.




〈그림 3〉 오프라인/온라인 템플릿별 최대 정합률 비교

#### 3.2 입술움직임 영역추적 테스트베드 구축

입술움직임 영역추적 과정과 오류 여부를 시각적으로 확인하고 관련 파라미터를 실시간 분석하기 위하여 영상처리 테스트베드를 구축하였다. 인터페이스는 파라미터 확인부분, 영상 제어부분 및 영상 재생부분과 같이 크게 세 부분으로 나눌 수 있다.

### 4. 연동 실험 결과

음성/영상시스템 연동에 앞서 음성인식 대상이 아닌 음향잡음에 의해 음성인식 과정이 진행되는 것을 확인하였다. 또한, 조명환경의 적응여부를 확인하기 위하여 전등의 일부 또는 전부를 소등하는 등 조명기기의 밝기를 다르게 하면서 실험을 하였다. 그리고, 음성/영상 연동시스템에 의해 순수 음향잡음이 음성인식 과정에서 차단되는지를 확인하였다. 실험결과를 요약하면 <표 2>와 같다(생략).

### 5. 결 론

본 논문에서는 음성검출 과정에서 외부음향잡음을 차단하기위한 강인한 입술움직임 영상영역을 추적하는 한 가지 방법을 제안하였다. 이를 위해 음성인식의 음성구간 검출과정에서 입술움직임 영상신호가 있는지를 확인하도록 음성/영상 연동시스템을 구축하였다. 조명환경에서 강인한 입술움직임을 영상영역을 추적하기 위해 온라인하에서 입술움직임 표준영상을 추출하였다. 이를 기반으로 오프라인에서의 입술움직임 특징 초기모델과 템플릿 초기모델을 온라인에서 추출한 모델로 대체하였다. 그 결과, 음성/영상 연동률은 99.3%에 달했으며 음향잡음에 의한 음성인식 실행을 원천적으로 차단할 수 있었다.

### [참 고 문 헌]


[2] G. Potaminanos, H.P. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading," Image Processing, 1988. ICIP 98, Proceeding, pp.173-177, Oct. 1998.

[8] Z.Q. Wu, J.A.Ware, W.R. Stewart, and J.Jiang, "The Removal of Blocking Effects Caused by Partially Overlapped Sub-Block Contrast Enhancement," Journal of Electronic Imaging, July- Sept. 2005, Vol.14, Issue 3, 033006(8 Pages).