음성인식 기반 컨텐츠 네비게이션 시스템

김기백*, 최종호*

요 약

최근 들어 인간의 의지를 각종의 전자시스템에 전달하기 위한 수단으로 음성인식 기술을 이용하고자 하는 연구가 늘어 진행되고 있다. 음성인식 인터페이스에서 가장 중요한 이슈는 처리시간의 감소 및 범용 인터페이스의 개발이다. 이러한 문제점을 해결하기 위하여 본 연구에서는 하드웨어 기반의 음성 IC로 생산되고 있는 음성인식 프로세서인 RSC-4128이 내장된 음성인식 모듈 VR-STAMP을 사용하였다. 본 연구에서 새롭게 개발한 시스템은 T2SI(Text To Speaker Independent) 기반의 화자의(말하는)목적 방식으로 음성인식 신호를 컨텐츠 네비게이션 시스템의 제어신호로 활용하여 입력터멀 시스템 및 PC 등에 설치된 와이드오즈 기반의 음성 소프트웨어를 제어할 수 있는 시스템이다. 필드 테스트를 통해 그 유용성을 확인한 결과, 본 연구에서 개발한 시스템은 컨텐츠 네비게이션은 물론 가전기기 제어 및 홈 네트워크 등에 널리 응용될 수 있을 것으로 판단된다.

Keyword: 화자독립, 음성인식, 입력터멀 시스템, PC, 컨텐츠 네비게이션

I. 서 론

최근 들어 인간의 의지를 직접적으로 기계에 전달하기 위한 수단으로 음성을 이용하고자 하는 연구가 늘어 진행되고 있다. 일상의 대화에서 자 ($.5cm)값과 사용되고 있는 음성을 이해하는 인터페이스를 구현할 수 있다면 음성은 의사전달에 대응하는 자연스러운 입력수단으로 널리 이용될 수 있을 것이다.

음성인식/Voice Recognition)은 사람의 음성신호를 전자장치에 입력하여 이를 처리하는 방식을 통합하여 표현하는 것으로 그 구현 방법(1,2,3)으로는 크게 화자중속 방식 (Speaker Dependent)과 화자독립 방식 (Speaker Independent)이 있다. 화자중속 방식으로는 화자(말하는)의 음성신호 파형을 에로리와 같은 저장장치에 저장 후 이를 인식하는 방식으로 대표적인 예로 기존의 핸드폰 전화 저장 장치를 들 수 있다. 이 방식은 사용자의 음성파형을 사전에 입력하는 방식이기 때문에 인식오류가 적고, 특정 사용자만을 위한 시스템의 구축이 가능하다는 장점이 있으나 외부감응에 크게 영향 받는다는 차별적인 단점이 있다.

* 제1저자: 김기백
* 경남대학교 전자공학전공
이와는 반대로 화자독립 방식은 화자(話者)의 음성신호와는 별도로 텍스트 형태로 저장되는 발음기호를 입력되는 음성신호와 비교하여 인식하는 방식이다. 대표적인 예로 ARS 서비스에서 제공하는 음성서비스를 들 수 있다.

본 연구에서는 빌용의 시스템을 구현하기 위하여 T2SI(Text To Speaker Independent) 기반의 화자독립 방식을 이용하여 컨텐츠 내비게이션 시스템을 개발하였다. 음성인식 인터페이스에서 가장 중요한 이슈는 처리시간의 감소 및 병용 인터페이스의 개발이다. 이러한 문제점은 해결하기 위하여 본 연구에서는 하드웨어 기반의 상용 IC로 생산되고 있는 음성인식 프로세서인 RSC-4128이 내장된 음성인식 모듈 VR-STAMP(4,5,6)을 사용하였다.

본 연구에서 새롭게 개발한 시스템은 T2SI(Text To Speaker Independent) 기반의 화자(話者)독립 방식으로 음성인식 신호를 컨텐츠 내비게이션 시스템의 제어신호로 활용하여 엽페디드 시스템 및 PC 등에 설치된 원도우즈 기반의 음성 소프트웨어를 제어할 수 있는 시스템이다. 실험을 통하여 그 유용성을 확인한 결과, 본 연구에서 개발한 시스템은 컨텐츠 내비게이션을 돕는 가전기기 제어 및 홈 네트워크 등에 널리 응용될 수 있을 것으로 판단된다.

2. 음성임력 시스템

2.1 음성인식 주요 단어

본 연구에서는 음성인식 인터페이스를 자동차용 내비게이션 시스템에 적용하는 것을 목표로 발음하기 전과, 발음의 구조가 비교적 정확한 단어를 선정하여 음성인식 단어로 활용하였다. 본 연구에서는 음성인식을 이용하여 자동차용 내비게이션 시스템에 내장된 음성소프트웨어를 제어하는 음성인식 인터페이스 시스템의 개발을 목표로 발음이 자연스러우면서 간결하면서도 두드러지게 구별되는 것에 주안점을 두고, (표 1)의 의미를 갖는 단어를 음성인식 주요 단어로 선정하였다.

<table>
<thead>
<tr>
<th>단어</th>
<th>기능</th>
<th>비교</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMB</td>
<td>DMB 실행</td>
<td>처음, 이전, 다음으로 재생설정</td>
</tr>
<tr>
<td>DVD</td>
<td>DVD 실행</td>
<td>처음, 이전, 다음으로 컨텐츠 선택</td>
</tr>
<tr>
<td>MP3</td>
<td>MP3 실행</td>
<td>처음, 이전, 다음으로 컨텐츠 선택</td>
</tr>
<tr>
<td>비</td>
<td>내비게이션 실행</td>
<td>주요 목적지 선택 기능 연계</td>
</tr>
</tbody>
</table>

2.2 음성임력 마이크 시스템

마이크 시스템에서 가장 중요한 고려해야 할 사항은 외부에서 입력된 신호에 포함되는 외부 노이즈를 최소화하는 것이다. 마이크를 회로에 마운트 할 경우 흐리와 접촉할 수 있는 공간을 제거해야만 한다. 먼저 마이크 소자를 회로 표면에 부착하여 일차적인 외부 환경으로부터의 노이즈를 제거하는 것이 필요하다.

또한 회로 표면에 장착된 마이크 앞부분에는 인식을 방해할 수 있는 모든 상황에서 자유로워야 한다는 것이다. 0.7mm 이상의 플라스틱 케이스 등이 이 부분에 부착되었을 경우, 인식률의 저하를 가져올 수 있다.

그리고 마지막으로, 하우징(Housing)을 마이크 주변에 고무와 같은 피막을 둘러서 외부 노이즈를 흡수할 수도 있도록 제작하는 것이 매우 중요하다. 이를 통해 시스템의 부조화로 인한 외부 노이즈들을 제거할 수 있기 때문이다.

한편, 마이크 제조사에서 제공하는 민감도 값(Sensitivity)을 이용하면 식 (1)을 이용하여 외부 저항값을 얻을 수 있다.[4]

\[
R_s = \frac{46 \pm 8 \text{ Sensitivity}}{20}
\]

(1)

본 연구에서 사용하고 있는 VR-STAMP의 마이크 입력 시스템의 부하회로는 (그림 1)에 나타났다.[5]
3. 응성인식 및 컨텐츠 네비게이션 시스템

T2SI(Text To Speaker Independent)는 응성 DB를 만들고 프로그램으로 테스트를 입력하여 반응하시면 SV(Speaker Voice) 템플릿을 만드는 것이다. 먼저 트리거(Trigger) 위치에 실제 응성인식에 사용될 단어를 작성한 후, 단어의 인식률을 테스트한다. 인식률이 낮은 단어는 실제 반응기록을 수정하여 높이는 과정을 거친다. 마지막으로 완성된 음성신흐를 보여주는 실제 데이터 파일과 이를 사용할 수 있도록 하는 라이브러리를 생성한다. 응성인식 DB 제작절차는 (그림 2)에 나타났다.

![그림 2. 응성인식 DB]

응성인식 DB를 기반으로 응성을 인식하는 과정은 다음과 같다. 트리거 신호가 입력될 때 입력을 기다린다. 트리거 신호 입력 후, 입력되는 음성신호와 일치하는 단어가 DB에 존재하면 응성인식을 수행하여 저장된 단어의 인덱스 값(Index)을 10진수 형태로 출력한다. 입력된 음성신호와 일치하는 단어가 DB에 존재하지 않는 경우에는 트리거 입력생태로 다시 진입한다.

응성인식 결과는 윈도우즈 기반의 시스템으로 전송하는 인터페이스 시스템은 Atmega128를 사용하여 구성하였다. 외부 인터럽트 신호의 예치되어 인터럽트 서비스루틴이 실행되면 1~10까지의 단어를 RS-232C로 통해 컨텐츠 네비게이션 대상 시스템으로 전송한다. 이 과정이 완료되면 다시 페드백 되어 신호를 대기한다. (그림 3)에 통신 인터페이스 시스템의 절차를 나타냈다.

![그림 3. 인터페이스 절차]

〈그림 3. 인터페이스 절차〉

건텐츠 네비게이션은 응성인식 모델의 출력포트를 실시간으로 감지하여 수신되는 인덱스 값을 따라 메크로를 실행시키는 응용 프로그램을 동작시키는 과정이다. 메크로를 사용하여 윈도우즈 어플리케이션(Windows Application)을 실행하는 절차는 (그림 4)에 나타났다.

![그림 4. 메크로 실행절차]

메크로 제어부분은 크게 두 가지로 분류될 수 있다. 모든 윈도우즈 어플리케이션(Windows Application)이 독립적으로 동작한다는 점을 고려하여, 먼저 프로그램을 실행시키기 위한 메크로 루틴과 이를 내부에서 동작할 수 있게 하는 메크로 루틴으로 구분하여 구현한다.

실행을 위한 메크로는 API 함수를 이용하여 실제 타겟 프로그램을 실행 시킬 수 있도록 제어한다. 이를 통해 응성인식 시스템으로부터 출력하여 최종 시스템까지 넘어온 10진수 형태의 인덱스 데이터에 따라 각 1~4까지의, 그리고 9~11까지의 문자를 사용하여 각 독립적인 프로그램으로 실행한다.

위 과정을 먼저 수행한 후, 두 번째로 5~8까지의 단어를 통해 내부 프로그램을 동작시킨다. 이는 키보드 값에 각 메크로 값으로 변환하여, 마치 키보드 입력값이 입력되는 것처럼 각 독립적인 어플리케이션을 제어하는 것이다. 예를 들어, 만약 종료루틴의 실행을 원할 경우, 종료라는 단어가 입력되었고 먼저 가장하므로, 이 후 컨텐츠 네비게이션 내부에서 동작하는 제어 어플리케이션(Application)은 윈도우즈의 종료 Function 키와 같은 Alt + F4 키가 입력되어 독립적인 프로그램을 제어할 수 있도록 한다.
또한 DVD, DMB, MP3 Player 등과 같은 다른 응용 프로그램에서도 이를 직접 제어할 수 있도록 각 상태에 맞는 메크로 키 값을 입력하여 이를 처리한다.

4. 실험 및 결과

본 연구에서 개발한 응성인식 기반 컨텐츠 네비게이션 시스템의 시제품을 (그림 5)에 나타냈다.

   (a)메인보드  (b)인터페이스

   (c)시제품

그림 5. 응성인식 기반 컨텐츠 네비게이션 시스템

본 연구에서 개발한 시제품의 성능검증을 위한 필드 테스트에서 정상적인 동작을 확인하였다. 본 연구에서 사용한 응성인식 모듈에는 약 20개의 입력 데이터를 저정할 수 있는 메모리가 내장되어 있다. 그러나 컨텐츠 네비게이션 시스템의 기능을 다양하게 구현할 목적으로 입력대상 단어를 15개 이상으로 증가한 필드 테스트에서는 인식률이 저하되는 것을 확인하였다.

5. 결론

본 연구에서는 TTSI (Text To Speaker Independent) 기반의 화자(일반)독립 방식으로 응성인식 신호를 컨텐츠 네비게이션 시스템의 제어신호로 활용하여 임베디드 시스템 및 PC 등에 설치된 윈도우즈 기반의 응용 소프트웨어를 제어할 수 있는 시스템을 개발하였다.

12개 내외의 응성인식 대상 단어를 사용하는 시제품의 성능검증을 위한 필드 테스트에서 정상적인 동작을 확인하였다. 하지만 본 연구에서는 외부환경으로부터 발생되는 각 종의 소음을 제거하는 문제는 고려하지 않았으나, 외부 소음에 강한 실시간 컨텐츠 네비게이션 시스템을 개발하기 위해서는 필수적으로 고려하여야 할 사항이다. 향후의 연구가 필요하다.

필드 테스트를 통하여 본 연구에서 개발한 시제품의 유용성을 확인한 결과, 본 시스템은 컨텐츠 네비게이션을 동결 가능한 애플리케이션 및 홈 네트워크 등에 널리 활용될 수 있을 것으로 판단된다.

참고문헌