사용자 프라이버시 보호를 위한 음성 특징 제거 기법 설계 및 구현

유병석*, 임수현**, 박미소*, 이유진*, 윤성현*
*백석대학교 정보통신학부, **고려대학교 컴퓨터 · 전파통신공학과
e-mail : byeongseok.yu@gmail.com

Design and Implementation of the Voice Feature Elimination Technique to Protect Speaker's Privacy

Byung-Seok Yu*, SuHyun Lim**, Mi-so Park*, Yoo-Jin Lee*, Sung-Hyun Yun*
*Div. of Information & Communication Engineering, Baekseok University
**Dept. of Computer and Radio Communications Engineering, Korea University

요 약

음성은 가장 익숙하고 편리한 의사 소통 수단으로 스마트폰과 같이 크기가 작은 모바일 기기의 입력 인터페이스로 적합하다. 서버 기반의 음성 인식은 서버를 방문하는 다양한 사용자들을 대상으로 음성 모델을 구축하기 때문에 음성 인식률을 높일 수 있고 상용화가 가능하다. 구글 음성 인식, 아이폰의 시리(SiRi)가 대표적인 예이며 최근 스마트폰 사용자의 증가로 이에 대한 수요가 급증하고 있다. 서버 기반 음성 인식 기법에서 음성 인식은 스마트폰과 인터넷으로 연결되어 있는 원격지 서버에서 이루어진다. 따라서, 사용자는 스마트폰에 저장된 음성 데이터를 인터넷을 통하여 음성 인식 서버로 전달해야 된다[1, 2]. 음성 데이터는 사용자 고유 정보를 가지고 있으므로 개인 인증 및 식별을 위한 용도로 사용될 수 있으며 음성의 톤, 음성 신호의 피치, 빠르기 등을 통해서 사용자의 감정까지도 판단 할 수 있다[3]. 서버 기반 음성 인식에서 네트워크로 전송되는 사용자 음성 데이터는 제 3 자에게 쉽게 노출되기 때문에 화자의 신분 및 감정이 알려지게 되어 프라이버시 침해를 받게 된다. 본 논문에서는 화자의 프라이버시를 보호하기 위하여 사용자 음성 데이터로부터 개인의 고유 특징 및 현재 상태를 파악할 수 있는 감정 정보를 제거하는 기법을 설계 및 구현하였다.

1. 서론

음성은 사람에게 가장 익숙하고 거부감이 적은 의사 소통 수단이다. 음성 인식은 컴퓨터를 이용하여 사용자의 음성을 문자로 변환하는 기술로 스마트폰 제어를 위한 사용자 인터페이스로 활용될 수 있다. 구글 음성 인식, 아이폰의 시리(SiRi) 등과 같은 서버 기반의 음성 인식은 다양하고 많은 사용자들을 대상으로 음성 모델을 구축하기 때문에 음성 인식률이 매우 높고 서버의 성능, 네트워크 지연 시간의 최적화를 통해서 음성 인식 인터페이스의 상용화를 가능하게 한다.

음성은 사용자마다 고유한 데이터이기 때문에 개인 인증 및 식별을 위한 목적으로 사용될 수 있다. 최근에는 음성의 톤, 음성 신호의 피치, 빠르기 등을 통해서 사용자의 감정을 판단하는 용도로도 활용된다[3].

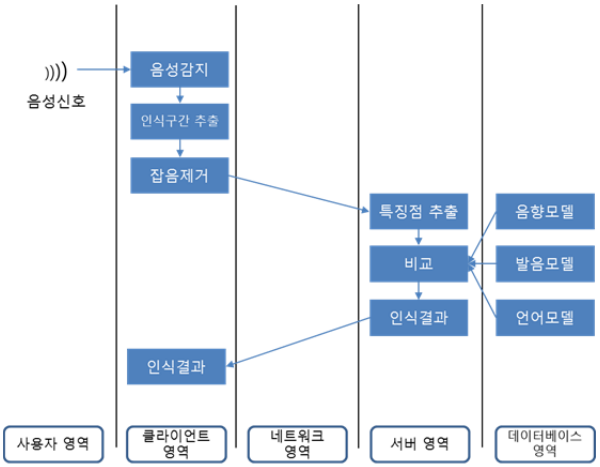
서버 기반 음성 인식 기법에서 음성 인식은 스마트폰과 인터넷으로 연결되어 있는 원격지 서버에서 이루어진다. 따라서, 사용자는 스마트폰에 저장된 음성 데이터를 인터넷을 통하여 음성 인식 서버로 전달해야 한다. 인터넷은 공중망이기 때문에, 사용자 음성 데이터가 제 3 자에게 노출될 수 있다. 음성 데이터는 고유 정보이기 때문에 사용자 신분과 톤, 피치, 빠르기 등을 분석한다면 감정 상태까지 알 수 있게 되어 사용자 프라이버시를 침해하게 된다. 음성 데이터를 암호화해서 보내더라도 이 데이터를 복원하는 서버에서 사용자 프라이버시 정보를 알 수 있게 된다.

본 논문에서는 사용자 프라이버시를 보호하는 서버 기반 음성 인식 모델을 제안한다. 프라이버시 보호를 위하여 사용자 음성 데이터로부터 톤과 피치 정보를 제거하는 모듈을 설계 및 구현하였으며 이에 대한 성능 평가를 하였다. 클라이언트(스마트폰)의 전처리 구간에 음성 변조 기능을 추가하여 화자 종속적인 특징을 제거함으로써 음성 데이터가 노출되어도 화자의 신원 및 감정을 파악할 수 없다.

2 장에서는 서버 기반의 음성 인식에서의 위험 요소를 분석하고 3 장에서는 제안한 음성 특징 제거 기법에 대해서 기술한다. 4 장에서는 변조된 음성에 대한 인식률을 분석하였다.

2. 서버 기반 음성 인식 시스템

그림 1 은 클라이언트-서버구조의 음성 인식 처리 과정을 보여준다[4].



<그림 1> 서버 기반 음성 인식 프로세스

클라이언트-서버 기반의 음성 인식에서 클라이언트는 음성 입력에 대해 음성 감지, 인식구간 추출, 잡음 제거와 같은 전 처리를 수행한다. 클라이언트 즉, 스마트폰 단말기에서는 이와 같은 전 처리가 완료되면 음성 인식을 담당하는 서버로 음성 데이터를 전송한다. 서버에서는 수신한 음성 데이터로부터 특징 값들을 추출하고 음향 모델, 발음 모델, 언어 모델을 이용하여 인식 결과를 계산하고, 결과값을 클라이언트로 보낸다.

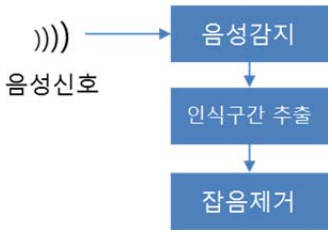
그림 1 을 보면 실제 데이터가 이동되는 영역은 클라이언트 영역, 네트워크 영역, 서버영역으로 구분된다. 클라이언트 영역에서 생성한 음성 데이터는 인터넷과 같은 공중망(네트워크 영역)을 통하여 서버로 전달되기 때문에, 제 3 자에게 유출될 수 있다. 음성 데이터를 암호화해서 보내더라도 이 데이터를 복원하는 서버에서 사용자 프라이버시 정보를 알 수 있게 된다. 음성 데이터는 화자 종속적인 특성이 반영되어 있기 때문에 화자의 신원 노출 또는 감정 상태 파악 등이 가능하다.

3. 사용자 음성 특징 제거 기법

본 논문에서 구현 하고자 하는 어플리케이션은 인터넷과 같이 보안 기능이 취약한 네트워크를 통해 전송되는 음성 데이터의 화자 종속적인 측면을 제거하여 사용자 프라이버시를 보호하는 것을 목적으로 한다.

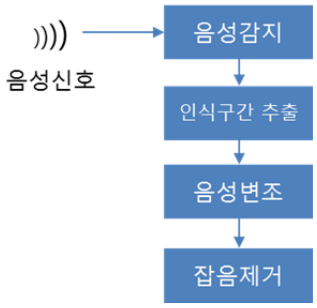
어플리케이션은 클라이언트에서 음성 신호를 변조하는 역할을 수행하며, 변조된 음성은 변조되기 이전의 음성과 의미상 동일한 내용을 전달해야 한다.

어플리케이션에서 변경할 수 있는 값은 음속, 피치, 포만트주파수(Formant Frequency)로 한정했다.



<그림 2> 기존 전처리 프로세스

그림 2 는 기존의 클라이언트 영역에서의 음성 인식을 위한 전처리 과정을 보여준다. 클라이언트 영역에서는 음성 신호가 발생하게 되면 이 신호를 감지하고 음성 인식 구간을 추출하여 잡음 제거를 한 후에 가공된 음성 데이터를 서버로 전송한다.

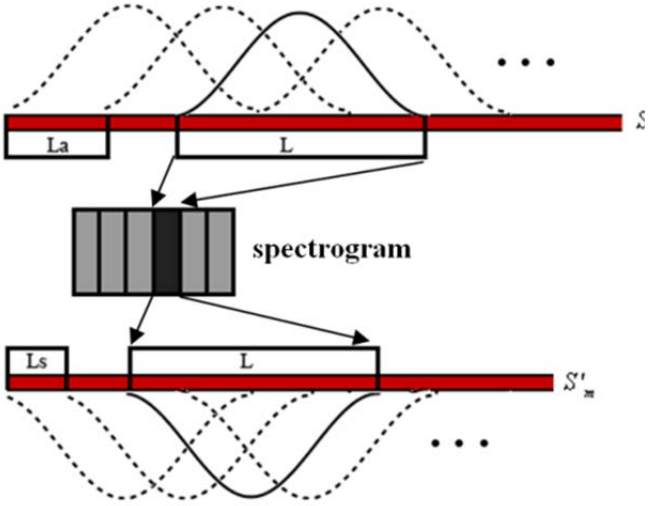


<그림 3> 제안하는 전처리 프로세스

제안하는 어플리케이션은 그림 3 과 같이 음성 신호 감지, 인식 구간 추출, 음성 변조, 잡음 제거의 순서로 전처리를 한다.

음성을 변조하기 위해서는 입력 신호에 대해 STFT (Short Time Fourier Transform)를 통하여 주파수 도메인으로 변환하게 되면 해당 주파수 도메인을 분석하여 주파수 성분의 진동수 위상 크기를 통해 소리의 특성을 파악할 수 있다. 변조 과정은 시간축의 진폭 데이터를 시간축의 주파수 스펙트럼으로 변환하기 위해서 STFT 는 입력 신호에 일련의 time window 를 부과하는 것으로 Windowing 이후 STFT 의 각 조각에 DFT (Discrete Fourier Transform)를 적용한다[5].

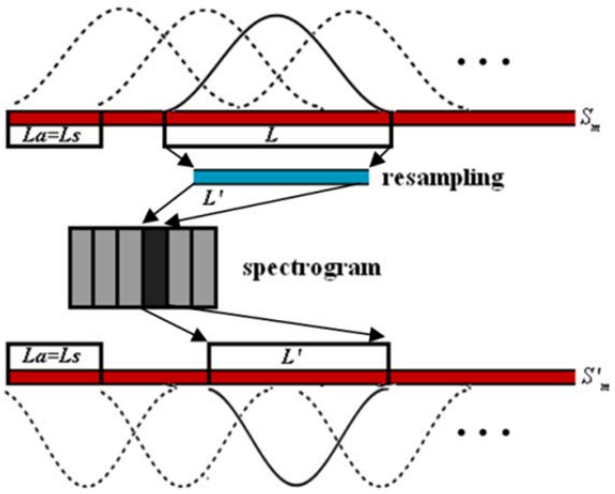
음성 신호의 템포 수정을 위해서 샘플링 속도를 조정한다. 예를 들어 44kHz 의 샘플링을 22kHz 로 수정할 경우 음의 속도는 2 배가 된다.



<그림 4> 음성신호의 템포 변환[6]

그림 4 는 소리의 템포를 변경하는 방법을 보여준다. L 은 분석 윈도우의 길이, La 는 분석 길이의 홉, Ls 는 합성된 홉의 길이를 나타내며, La 가 Ls 보다 클 경우 빠른 음성으로 변조되며, Ls 가 La 보다 클 경우 느린 음성으로 변조된다.

또한 음성신호에서 음의 높이를 피치(pitch)라고 하는데, 주파수가 높아지면 음높이는 높아진다. 피치의 수정은 입력된 음성의 주파수를 변조하여 수정한다.



<그림 5> 음성신호의 피치 변환[6]

그림 5 에서는 사운드의 피치를 변경하는 방법을 보여준다. 홉 분석 길이 La 는 홉 합성 길이 Ls 와 동일하다. 그러나, 분석 윈도우 L 은 합성 윈도우와 다른 것을 볼 수 있다. L 은 L'으로 재 샘플링 되며, L 이 L' 보다 클 경우 음 높이는 크게 변조되고, L'이 L 보다 클 경우 음 높이는 낮게 변조된다.

포만트 주파수는 음의 공진 주파수로 일반적으로 '우'는 500c/s 부근, '오'는 700c/s, '아'는 1,000c/s 부근, '에'는 2,000c/s 부근, '이'는 3,000c/s 부근에서 강한 진동을 갖는 형태를 나타낸다.

4. 성능 분석 및 구현 결과

어플리케이션은 ios 상에서 구현하였고 cocos2d 프레임워크를 이용하였다. dirac3 라이브러리를 이용하여 음성을 변조하였으며 연속음성인식 어플리케이션은 카네기멜론대학교의 스핑크스[7]를 ios 용으로 포팅한 openears[8]를 이용하여 구현했다.

변조 어플리케이션은 원본 음성이 감지되었을 경우 어플리케이션에서 설정한 값으로 변조한다. 이 값들은 음성 특징에 영향을 미치는 템포, 피치, 포만트 주파수 세 가지 조건을 가지고 변조한다.

변조된 음성에 대한 음성 인식은 단순 인식률만을 비교하기 위해 원본 음성과 변조 음성을 순차적으로 출력하여 인식된 결과가 동일한지 판단하기 위해 ios 상에서 연속음성인식 어플리케이션을 구현했다.

인식률의 분석은 원본 음성과 변조된 음성의 인식 결과 차이를 알아보기 위해 5 개의 단어를 표 1 과 같이 3 가지 변조 값을 이용하여 실험했다.

[표 1] 변조테스트를 위한 변수설정

	조건 1	조건 2	조건 3
Speed	1	0.8	1
Pitch	0.7	0.7	0.7
Formant	1	1	0.7

실험을 위한 조건은 3 가지로 한정하였으며, 조건 1 의 경우에는 피치의 값만 변경시켜 실험을 진행하였고, 조건 2 는 속도와 피치를 변경하였으며, 조건 3 은 포만트 주파수를 변경하였다.

조건 1 의 경우에는 음성 신호의 변조를 위해 음의 높낮이만을 수정한 경우, 조건 2 의 경우에는 음의 높낮이뿐만 아니라 사람마다 다를 수 있는 말하는 속도에 대한 조작, 조건 3 은 포만트 주파수를 수정하여 사람의 구강 구조에 따라 달라질 수 있는 공진 주파수 영역에 대하여 왜곡했다.

각각의 실험의 경우 조건당 100 회 반복했으며, 실험결과는 표 2 과 같다.

[표 2] 원본 음성과 변조 음성의 음성인식 비교

	조건 1	조건 2	조건 3
Go	85%	77%	48%
Right	90%	83%	62%
Left	89%	85%	69%
Forward	93%	83%	65%
Backward	95%	84%	62%

실험 내용은 원본 음성과 변조된 음성을 순차적으로 인식시켜 비교한 것으로 표 2 에서 나온 확률은 원본 음성과 변조된 음성의 일치율을 보여주고 있다.

실험 결과를 볼 경우 조건 1 의 경우 일치율에 가장 적은 영향을 주었으며, 조건 2 의 경우 원본 음성에 비해 손실되는 부분의 증가로 인해 조건 1 에 비하여 낮은 일치율을 보여주고 있다. 조건 3 의 경우 포만트 주파수를 변경하여 원본 음성에서 발성된 모음의 특

성이 변질되어 일치율이 저조한 것을 볼 수 있다.

5. 결론

본 논문에서는 프라이버시 보호를 위한 음성 특징 제거 기법을 설계 및 구현하였다. 클라이언트 전처리 구간에서 입력된 음성 신호에 대해 템포, 피치, 포만트 주파수를 각각 변조하여 변조된 음성과 원본 음성의 인식 결과에 대한 일치도 실험을 하였다. 원본 음성에 대해 변조를 할 경우 피치만 변경한 경우가 원본 음성과 인식 결과가 가장 유사했다. 음성 데이터에서 사람마다 다른 특성을 제거하는것이 목적이기 때문에 모음의 음성이 변하는 포만트 주파수의 변경은 부적합하며, 템포의 변경 또한 일정 수치 이상일 경우 음성의 의미적 파악이 힘들었다.

Acknowledgement

이 논문은 2012 년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(지역대학우수과학자사업, No. 2012-0004515)

참고문헌

[1] Eric Thelen, Stefan Besling, US6487534, Nov, 26 ,2002
[2] 이윤근, “음성인터페이스 기술 개요 및 스마트폰 환경에서의 서비스 동향”, 한국통신학회지 29(4), 2012.3, 3-9
[3] 심귀보, 박창현, “음성으로부터 감성인식 요소 분석”, 퍼지 및 지능시스템학회 논문지 2001, Vol. 11, No. 6, pp. 510-515
[4] Johan Schalkwyk, Doug Beeferman, Francoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Garret, Brian Strope, “Google Search by Voice: A case study”, Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, Springer (2010), pp. 61-90
[5] M. R. Portnoff, “Time-scale modification of speech based on short-time Fourier analysis” IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, vol. ASSP-29, pp. 374-390, June S981.
[6] http://www.codeproject.com/Articles/245646/How-to-change- the-pitch-and-tempo-of-a-sound
[7] CMU Sphinx - Speech Recognition Toolkit, http://cmusphinx.sourceforge.net/
[8] OpenEars - iPhone Voice Recognition and Text-To-Speech, http://www.politepix.com/openears/