

노인발화특성을 고려한 음성인식기술기반 재난관련 민원데이터분석 프레임워크 개발

박성원, 김민기, 장동민
한국과학기술정보연구원 데이터기반문제해결연구단
swpark@kisti.re.kr

Development of Framework on Disaster-related Civil Complaint Analysis Based on Automatic Speech Recognition Considering Elderly's Speech Features

Dongmin Jang, Minki Kim, Sung-Won Park
Dept. of Data-centric Problem Solving Research, Korea Institute of Science and Technology
Information

요약

우리나라를 비롯한 주요국가의 가장 큰 문제는 인구구조의 고령화이다. 특히 우리나라는 주요도시의 노인계층의 증가와 정보격차에 따른 디지털 소외현상이 가속화되고 있다. 노인계층은 특히 재난발생취약성이 높은 반면, 위협으로부터 벗어날 수 있는 위기대처능력은 현저히 낮고 민원요청을 통한 대응 및 복구 단계의 조치에 어려움을 겪고 있다. 또한 국가 및 지자체 단위의 재난관련 정책결정 시에도 검증이나 활용을 위한 데이터의 부재로 인하여 악순환이 반복되는 상황이다. 따라서 본 논문에서는 4차산업혁명과 정보화시대 가속화에 따라 디지털 소외계층이 스마트디바이스를 이용하여 재난피해에 대해 민원 요청하는 경우 디지털 불평등을 해소하기 위한 방안으로 노인발화 특성을 고려한 음성인식기술 개발을 위한 프레임워크를 제시하고자 하였다. 즉 기존 음성인식기술의 기본프레임워크를 기반으로 노인들의 발화특성을 고려하여 잡음 처리 및 방언해석 등의 추가적인 모듈구성하여 재난발생 취약계층의 민원데이터를 극대화하고 나아가 스마트에이징기술의 고도화를 유도할 것으로 기대한다.

주요어 : 노인계층; 디지털 소외계층; 음성인식기술; 프레임워크; 민원데이터;

1. 서론

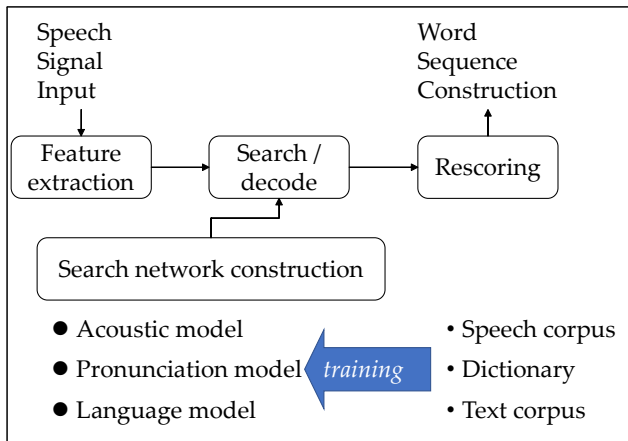
전세계적으로 선진국들의 노령인구 증가문제는 가장 큰 사회 문제 중 하나로 대두되고 있으며 또 다른 사회문제의 원인이거나 과정으로 확산되는 추세이고, 이로 인한 연령별 계층사이의 갈등과 신기술 활용능력 격차 심화, 소외계층의 확산 및 빈곤화 양극화 등의 문제는 앞으로 지속적으로 가속화될 전망이다. 우리나라의 경우, 통계청 발표에 따르면 2020년 기준으로 5,184만명에서 향후 10년간은 연평균 6만명 내외로 감소하여 2070년에는 3,766만명에 이를 것으로 전망하고 있다. 이 중에서 65세 이상 고령인구 비중이 2020년 15.7%에서 46.4%로 증가하는 것으로 집계 발표되었고, 2020년 OECD 국가들에 비해서는 낮은 수준이지만, 2070년에는 가장 높아질 것으로 예상되었다(통계청, 2022). 평균수명의 연장으로 은퇴 후 시간은 늘어나고 산업화 시대에 따라서 과학기술 발달이 가속화되지만 4차산업혁명을 필두로 하는 다양한 분야의 정보격차에 따른 소외현상은 노령인구를 더욱 고립시키면서 복지 문제로 야기될 가능성이 높다. 노령인구 증가문제와 동시에, 도시 외곽으로 주거지역을 이동한 노령인구는 재난발생 취약

계층으로 분류되어 산불이나 침수재난 등의 다양한 재난피해 발생에 대해서 능동적인 대처가 어려운 상황이다. 또한 피해 복구를 위한 민원요청에 있어서도 대처가 원활하지 않아서 어려움이 지속되는 상황이며, 재난발생 해당지자체는 물론 정부부처에서도 지속적인 모니터링을 위한 재난발생 데이터 확보에 어려움이 가중되고 있는 실정이다. 신속한 재난피해 상황 파악과 대처를 위해서 행정적 절차의 자동화와 다양한 스마트 디바이스의 보급, 키오스크 등의 편의성이 제공되는 상황에도 불구하고 노령인구의 대부분은 디지털 소외계층으로의 편입이 가중되는 상황이며 다양한 혁신기술 중 음성인식기술의 활용이 필요한 실정이다. 하지만 대부분의 음성인식기술기반 서비스는 평균적인 연령대나 청년층의 사용자들에게 맞춰져서 성능평가되고 지원되는 실정이다.

따라서 본 연구에서는 재난발생에 대해서 민원요청을 원활하게 할 수 있는 음성인식기술기반의 서비스 기초기술 중 노령인구의 발화특성을 고려한 음성인식기술 개발 프레임워크를 개발하고자 하였다. 즉 기존의 음성인식기술 아키텍처를 기반으로 하여 노인발화특성을 추가적으로 고려한 개선모형과 적용방향을 검토하였다.

2. 연구동향

음성인식(ASR: Automatic speech recognition)은 AI기술 활용에 있어서 가장 활발한 상용화 분야 중에 하나이다. 음성 인식은 사람의 말소리를 음성신호로 분리하고, 특징을 적절하게 추출하여 문자열로 변환하여 직접적으로 이해하는 기술이다. 신호입력 후 특징추출 이후에 음향모델(acoustic model), 언어모델(pronunciation model), 발음사전(language model) 등으로 구성된 탐색기와 디코딩(decoding) 과정이후 후처리(post-processing)을 통해서 자연발화자료 출력하는 구조로 구성되어 있다[그림 1].



[그림 1] 기본 ASR 아키텍처

코로나19의 세계적 유행으로 디지털 기반의 비대면 소통방법이 활발하게 사용되는 가운데 자연어 처리 및 음성인식과 AI 기술이 다양한 서비스로 적용되어 편의를 제공하고 있다. 최근 자연어 음성인식은 종단형(end-to-end) 트랜스포머 인공지능 모델과 수만시간의 음성데이터 학습으로 인식율이 개선되어 대화형 AI 서비스로 확산되고 있다.

김유빈 등(2022)은 AI기반 무인결체시스템의 인터페이스 프로토타입 개발에 관한 연구를 수행하면서 고령 사용자의 효율을 극대화 할 수 있는 개선사항에 대한 실험을 수행하여 범용성 확보를 위한 연구를 수행하였고, 고령 사용자의 발화특성에 대한 고려에 관한 가장 최근 연구이다.

박지용 등(2013)은 청년층과 노인층 음성 인식에 대한 비교 분석을 위한 전처리 기술에 관한 고찰과 중요성을 언급한 바 있으며 특히 음성신호 분석을 통해 발음의 명료성, 변이 등을 고려하였다.

권유환 등(2020)은 딥러닝 기술을 활용하여 음성신호데이터로부터 발화자 벡터를 추출하는 시스템을 발표한 바 있다. 최근들어 판별기 이론기반으로 개발된 생성적 적대 신경망 모형(GAN: Generative Adversarial Network model)을 이용하여 실험결과를 검증한 바 있으며 최근 기술과 비교 활용가능할 것으로 판단된다.

3. 연구내용

3.1 기존음성인식기술 적용성 검토

기존 기술기반음성인식 모델인 LAS (Listen-Attend-Spell, Chan et al., 2016)와 ESPnet(Watanabe et al., 2018)에 대한 적용성을 검토하고자 하였다. LAS 모델은 기존에 사용되었던 DNN-HMM (Deep Neural Network-Hidden Markov Model) 구조나 CNN(Convolutional Neural Network)-HMM 기반 구조와는 다르게 발화자로부터 발생된 음성신호에서 문장을 바로 출력하는 신경망 구조를 갖는다. 즉 모델로부터 음성신호를 입력받은 뒤 언어모델 및 발음사전 등[그림 1]을 구성하는 요소를 사용하지 않고 단순히 문자나 문장을 출력해내는 모델을 시도한다. ESPnet 모델은 end-to-end 음성인식에 기반을 둔 모델이며 이 모델의 인코더는 특정 길이의 음성 특징 벡터값을 입력받은 뒤에 pBLSTM을 통해서 음향 정보를 추출하는 방식으로 구성된다. 디코더는 RNN(Recurrent Neural Network) 모델로 구성되는데 SOS (start of sentence) 토큰과 EOS (end of sentence) 토큰을 추가하여 학습하게 된다. 본 연구에서는 기존에 제안된 음성인식기술 방식을 사용하며 세부 계층별 모델변수들은 실험을 통해서 개선해 나가기로 하였다.

3.2 노인발화특성과 학습데이터 확보

국내에서 연구 목적으로 공유 중인 음성 데이터베이스는 대부분 성인남녀를 대상으로 구성되어 있으며, 음성인식기술 기반의 서비스를 제공하는 민간기업에서도 데이터 수집의 어려움과 비용 상의 문제로 인해 성인 남녀의 음성을 중심으로 음성인식 해석엔진을 개발하는 상황이다(김준우와 정호영, 2000). 따라서 노인이나 어린이 음성인식을 위한 시스템 개발을 위해서 본 연구에서는 AI-HUB에서 제공하는 다양한 개방형데이터 중 노인남녀의 자유대화 음성데이터를 확보하여 활용하고자 하였다[표 1].

[표 1] 노인발화 학습데이터 속성(AI-HUB)

수집대상	형태
데이터속성	• 남녀 발화자 대상(60세 이상, 1,000명 이상) • 3,000여 시간 음성데이터
원천데이터	• PCM(WAV) 음성 파일 • 대상자 및 대화 시나리오 정보를 포함한 음성파일
메타데이터	• Json type • 대상자 상세정보 (성별 / 연령 / 지역) • 녹음환경 정보 (실내외 : 대중교통, 거리 등) • 대화 주제 및 상세내용
• 원천데이터(음성파일)와 메타데이터(Json)로 구분 • 원천데이터(음성파일)은 각각의 파일명으로 구분 (Ex. sample1.wav)	

3.3 아키텍처 개선방향

고령 노인층의 발화 산출 특성을 파악하기 위해서는 성별과 노화정도, 그리고 교육정도를 선제적으로 파악하는 것이 중요하다고 알려져 있다(최현주, 2012). 또한 노인의 발화특성 중 사투리와 억양 그리고 의미를 포함하지 않는 다양한 음성 표현을 적절하게 제거할 필요가 있다. 하지만 일정 나이 이상이 지난 고령 노인층의 경우에는 성별 및 교육정도가 발화특성과 상관관계가 감소한다고 알려져 있다(윤필상과 윤지혜, 2015). 따라서 노인의 성별, 노화정도, 혹은 교육정도에 대한 상세한 분리를 통한 학습데이터셋의 구성은 제외하였다.

또한 본 연구에서는 재난발생과 민원요청을 위한 음성인식 기술의 개발 아키텍처 구성을 목표로 하고 있기 때문에 대부분의 재난발생 상황은 거주 주소지를 기준으로 민원이 발생한다. 주소지에 대한 DB 활용은 제한된 학습데이터셋의 활용이 가능하며 순차모형의 적용에 있어서 자연어처리 적용에 용이함을 갖게 된다. 또한 노인의 발화-담화 사이에 발생하는 특정 파열음과 기침 등의 제거모듈을 포함할 필요가 있다. 전술한 내용을 바탕으로 아래와 같은 고려사항을 아키텍처 개선방향으로서 제시할 수 있다.

- 가. 주소지(도로명 주소 등) 기반 데이터셋 활용
- 나. 위치정보 서비스 및 플랫폼과 연동형 DB 활용
- 다. 노인발화 속도조정 및 음파 증폭 모듈 추가 결합
- 라. 불필요한 음성신호(파열음 등) 제거 모듈 추가 결합

4. 고찰 및 향후연구

본 연구는 재난발생 취약계층으로 분류되는 노령인구의 발화특성을 고려하여 민원데이터 처리의 원활함을 극대화할 수 있는 음성인식기술 개발의 아키텍처 구성을 위한 검토내용을 제공한다. 청년층 위주로 성능 평가되고, 해석모듈이 개발되는 기존의 음성인식기술의 아키텍처를 기반으로 구성한 뒤에 노인발화특성에 해당하는 다양한 문제를 극복할 수 있는 추가 모듈구성에 구상안을 제시하였다. 국내는 물론 선진국의 인구구성이 고령화됨에 따라 이러한 노령인구의 디지털 소외감을 줄일 수 있는 기술로써 활용성을 높일 수 있으며 나아가 노인은 물론 장애인이나 유아계층이 겪을 수 있는 민원요청 시 어려움을 극복할 수 있는 디지털 포용성 확대방안으로 활용을 기대할 수 있다.

감사의 글

This research was supported by the Ministry of Science and ICT (MSIT, Republic of Korea (Project Number: K-22-L04-C06-S01).

참고문헌

- [1] 통계청, 장래인구추계, 2022년.
- [2] 김유빈, 남양희, 강민정, “고령층을 위한 패스트푸드점 AI 음성인식 키오스크 인터페이스의 사용성 비교 평가”, 산업디자인학연구지, 제 16권, 1호, pp. 99-109, 2022년
- [3] 박지웅, 이승준, 권순일, “노인음성인식을 위한 전처리에 관한 연구”, 한국정보처리학회 추계학술발표대회 논문집, 제 20권, 2호, 11월, 2013년.
- [4] 권유환, 정수환, 강홍구, “화자 인식을 위한 적대학습 기반 음성 분리 프레임워크에 대한 연구”, 한국음향학회지, 제 39권, 5호, pp. 447-453, 9월, 2020년.
- [5] Chan, W., Jaitly, N., Le, Q., and Vinyals, O. “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition”, *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal (ICASSP)*, pp. 4960-4964, March, 2016.
- [6] Watanabe, S., Hori, T., Karita, S. Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., Ochiai, T. “ESPnet: End-to-End Speech Processing Toolkit”, *arXiv*. Retrieved from <https://arxiv.org/abs/1804.015>.
- [7] 김준우, 정호영, “End-to-end speech recognition models using limited training data”, 한국음성학회지, 제 12권, 4호, pp. 63-71, 12월, 2020년.
- [8] AI-HUB (<https://aihub.or.kr>).
- [9] 최현주, “노년층의 담화 산출 특성: 노화, 성별, 교육정도에 따른 차이”, 말소리와 음성과학, 제 4권, 2호, pp. 135-143, 2012년.
- [10] 윤필상, 윤지혜, “정상 노인의 절차 담화 특성”, 한국노년학, 제 35권, 4호, pp. 929-940, 2015년.