# 이동통신 PTT 서비스 음성 전화 호출을 위한 음성인식 연구

김재명, 유대승
한국전자통신연구원
jaemkim@etri.re.kr, ooseyds@etri.re.kr

# A Study on speech recognition for voice calling in mobile PTT service

JaeMyoung KIM, Dae Seung Yoo
ETRI (Electronics and Telecommunications Research Institute)

요 약

음성인식 기술은 다양한 분야에 상용화되어 적용되고 있으며, AI 음성인식 스피커 등 실생활에 활발히 적용되고 있다. 사용 환경이 열악한 산업용 이동통신 PTT(Push-to-Talk) 단말에서 음성으로 전화를 걸수 있는 음성 호출 서비스를 음성인식 기술을 적용하고자 한다. PTT 단말은 특성상 공동으로 사용 가능해야 하고 작업 중 손을 사용하지 않고 음성으로 1:1 통화 혹은 그룹 통화 호출을 위한 음성 처리 기술이 필요하나, 기존의 기술은 인식 단어가 매우 제한적이어서 호출어 인식 등의 임베디드 프로세서에 탑재되는 기능에 한정되었다.

본 연구는 명령어 인식을 하기 위한 오디오 분류 방식과 다양한 음성 데이터를 통해 학습된 하이브리드 모델을 통한 음성인식 방식을 설명하고, 각 방식의 장단점을 분석하였다. 전자의 방식은 음성인식의 결과로 사용자 디렉토리를 탐색하여 호출하여야 하는데 분류 카테고리가 제한되어 있어 사용이 곤란하다. 후자의 경우 다양한 데이터셋과 KALDI 레시피를 사용하여 모델을 훈련하고 그 인식률을 측정한 결과 27% 정도의 낮은 정확도를 보였다. 따라서, 본 논문에서는 KALDI 활용하여 이름으로 구성된 데이터셋으로 학습된 이름 음성 호출 모델을 적용한 결과 84% 정도의 정확도를 보였으며, 기존의 음성인식 방식에서 문제점 및 개선 사항을 제시하였다.

## Ⅰ. 서 론

인공지능 기술의 발달로 인해 음성인식 기술은 다양한 분야에 상용화되어 적용되고 있으며, AI 음성인식 스피커 등 실생활에 활발히 적용되고 있다.

인공지능 음성인식 기술의 활용은 우리 생활에 이미 많이 사용되고 있는 AI 음성인식 스피커에 기본 기술로 사용될 뿐아니라 실시간 음성인식 자동 자막 방송 등의 방송 분야, 차량 제어, 카투홈(Car To Home) 서비스를 위한 자동차 분야, 본인 인증 등을 위한 콜센터 분야, 독거노인 케어, 안심 화장실, 안심 보안 등의 서비스 등 공공안전 분야, 심지어는 AI 무인 로봇 제어, 자동 번역 등의 국방 분야에도 적용되고 있다.

PTT 단말은 특성상 공동으로 사용 가능해야 하고 작업 중 손을 사용할 수 없는 경우가 많으므로 음성으로 1:1 통화 혹은 그룹 통화를 할 필요가 있으나 인식 단어가 매우 제한적이어서 음성 전화 호출 등에는 사용되지 못하고 호출어 인식 등의 임베디드 프로세서에서 동작하는 기능에 한정되었다.

본 연구는 명령어 인식을 하기 위한 오디오 분류 방식과 다양한 음성 데이터를 통해 학습된 하이브리드 모델을 통한 음성인식 방식을 설명한다.

전자의 방식은 v1 시험셋의 정확도는 v1, v2 훈련셋에 대해 각각 85.4%, 89.7%이며, v2 시험셋의 정확도는 v1, v2 훈련셋에 대해 각각 82.7%, 88.2%로 높은 편이나. 사용자 디렉토리를 서치하여 호출하여야 하는데 분류 카테고리가 제한되어 있어 사용이 곤란하다[1].

후자의 경우 다양한 일반 음성 데이터셋과 KALDI 레시피를 사용하여 훈련하고 이름으로 구성된 문장의 정확도를 측정한 결과, 27% 정도의 아주 낮은 결과가 도출되었다. 이를 개선하기 위해 이름으로 구성된 데이터셋을 만들고 학습한 결과 84%의 정확도를 나타내었다.

본 논문은 음성인식을 하기 위한 오디오 분류 방식과 하이브리드 모델을 설명하고, 각 방식의 결과와 문제점을 살펴본다. 또한 이동통신 PTT 단말에 사용 가능한 이름 호출 음성 모델을 훈련하여 결과를 살펴보고, 개선 방향을 제시하고자 한다.

## Ⅱ. 음성 인식 방식 사전 연구

### 1. 오디오 분류 방식

음성인식에서 오디오 분류 방식은 클라우드 기반으로 오디오 데이터셋을 생성하고 임베디드 시스템에서 주로 사용하는 Keyword Spotting 시스템을 훈련하고 평가하는데 사용하기 위해 개발되었다[1].

이 방식은 35개 카테고리의 오디오 파일을 스펙트로그램 이미지로 변환하고 간단한 CNN(컨볼루션 신경망)을 사용하여 분류하였으며, 2017년 수집 데이터 v1과 2018년 수집 데이터 v2의 평가 결과, v1 시험셋의 정확도는 v1, v2 훈련셋에 대해 각각 85.4%, 89.7%이며, v2 시험셋의 정확도는 v1, v2 훈련셋에 대해 각각 82.7%, 88.2%이다. v2 데이터셋이 v1보다 높은 정확도를 나타내고 있다.

상기 데이터셋을 기반으로 최신의 신경망 처리 기법인 MHAtt-RNN(Multihead attention RNN)을 사용하여 v1 시험셋의 정확도 97.2, v2 시험셋의 정확도 98%를 달성하였다[2].

이 방식의 경우 카테고리가 이미 정해진 관계로 음성 이름 인식하여 통화 연결하는 이동 휴대전화의 PTT 전화기에서는 사용이 곤란하다.

### 2. 하이브리드 모델 방식

음성인식은 입력된 음성이 가장 높은 확률을 가지 단어 조합을 찾는 것이며, 일정 시간 T 동안 입력된 음성열 X에 대해 인식기가 표현할 수 있는 모든 단어 조합 중 확률적으로 가장 높은 단어 열 W를 찾는 것이며, 다음과 같은 수식으로 요약이 가능하다.

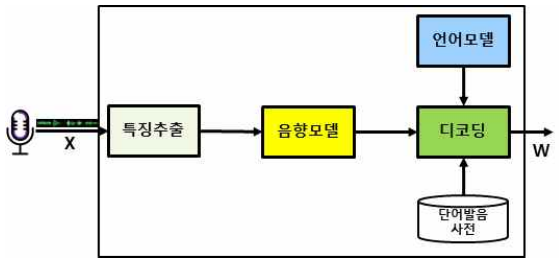
$$\arg\max_{W} P(W|X) = \arg\max_{W} P(X|W)P(W)$$

X : T시간의 발화된 음성
W : N개의 단어들로 이루어진 문장

상기 수식은 베이즈룰에 따라 정리된 것이며 P(X|W)은 음향모델의 확

률로 W라는 단어를 발성했을 때 X라는 신호 특성이 발생될 확률이며, P(W)는 언어모델의 확률로 단어 W가 말해질 확률이며 이는 음성신호와 무관하다.

따라서 음성인식 과정은 음성분석, 음향모델 계산, 언어모델 계산, 디코딩의 4단계로 구분가능하며 다음 그림과 같은 과정을 거친다[3, 4].




(그림 1) 음성인식 과정

본 논문에서는 실험을 위해 KALDI 도구[5]의 zeroth 레시피[6]와 krs 레시피[7]를 사용하였으며,. 각 레시피는 Monophone, Triphone1 (DELTA), Triphone2(LDA+MLLT), Triphone3(LDA+MLLT+SAT), Triphone4(DNN) 음향 모델을 생성하며, 데이터셋의 볼륨에 따라 DNN 모델은 생성이 불가능할 수도 있다. DELTA, LDA(Llinear Discriminant Analysis), MLLT(Maximum Likelihood Linear Transform) 등의 방식은 음성 입력을 feature로 만들 때 잡음, 채널 등 다양한 특성을 고려한 처리 방식이다.

음성인식을 위해 사용된 데이터셋으로는 zeroth 레시피에 내장된 말뭉치[6], 서울말 낭독체 발화 말뭉치[8], AI허브, 명령어 음성(일반남녀) 데이터셋[9]을 사용하였으며, 각각의 훈련 데이터의 시간은 51시간 40분, 75시간 22분, 248시간 35분이다.

음성인식 평가는 WER(Word Error Rate; 단어 에러 비율)로 계산되며, 음성인식 평가 결과는 krs 레시피와 서울말 낭독체 발화 말뭉치, 명령어 음성(일반남녀) 데이터셋을 사용한 모델의 종류와 상관없이 아주 나쁜 결과를 보였다. 가장 좋은 경우의 결과는 WER 137.38 [ 577 / 420, 157 ins, 0 del, 420 sub ] 로 Monophone 모델이며, 데이터 많고 적음에 관계없이 나쁜 결과를 보였다.

zeroth 레시피의 경우 krs 레시피 보다는 좋은 결과를 보이나, 음성 인식 기반 전화 호출을 하기에는 사용할 수 없는 수준이다. 예로 들어 보면,

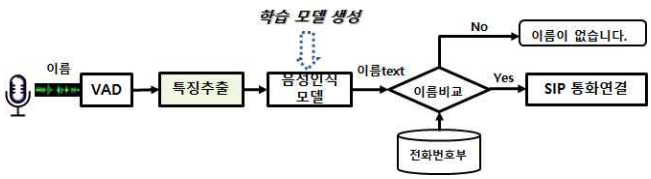
| |
|---|
| WER 0.82 [ 9 / 11, 1 ins, 0 del, 8 sub ] |
| 정답: 이은우 이정우 박시윤 박승우 이승현 이준혁 차지환 이윤우 박승민 서유찬 최지우 |
| 예측: 이윤호 이정우 박시온 박성호 이승연 이준혁 차지헌 이윤호 박승민 사육 되지 요 |

문장상으로는 비슷하게 예측하였으나 전화 번호부를 찾기 위한 정답은 3건으로 정확도는 27%밖에 되지 않는다.

따라서, 음성 호출에 적합한 이름을 예측할 수 있도록 데이터셋 보완이 필요하며, 평가도 문장이 아닌 이름 호출에 대한 정확도로 예측할 필요가 있다.

## Ⅲ. 이름 음성 호출 모델

우리가 만드는 모델은 (그림 2)와 같은 과정을 통해 사용된다.




(그림 2) PTT 단말 음성 인식 호출 과정

마이크를 통해 발화된 음성은 VAD(Voice Activity Detection), 특징추출을 거쳐, 음성 인식 모델의 결과가 전화번호부에 존재하면 SIP(Session Initiation Protocol) 프로토콜을 통해 호출하게 된다. SIP 프로토콜은 VoIP(Voice over IP), 멀티미디어 통신에 있어 세션이나 호(Call)을 관리하는 프로토콜이다.

이름 음성 호출 모델을 생성하기 위해 기존의 데이터, 이름 음성을 6명이 녹음한 음성 데이터와 이름과 성, 이름과 성을 랜덤 조합한 이름을 파파고와 구글번역기를 통해 얻은 음성데이터 49분의 데이터를 krs 레시피를 사용하여 학습한 결과, 특정 음성의 경우 다음의 결과를 보였다.

WER 15.91 [ 7 / 44, 0 ins, 0 del, 7 sub ]

이름의 정확도를 계산하기 위해서는 WER로는 부적절하나 한 문장을 하나의 이름으로 하여 계산한 결과이며, 상기 7개의 문자 변경이 발생하였으므로 84%의 정확도를 가진다고 할 수 있다.

이름 음성 호출 모델을 효율적으로 구성하기 위해서는 일반적인 음성인식 모델과는 다르게 다음의 문제점이 고려되고 해결되어야 한다.

- 언어모델의 n-gram 기술 등을 이용할 수 없다
- 이름은 고유명사이므로 일반적인 음성 데이터로는 학습이 곤란하다
- PTT 단말 사용자의 음성이 학습되어야 높은 인식률을 보일 수 있다
- 문장으로 되어 있지 않기 때문에 WER로 평가할 수 없다

## Ⅳ. 결론

본 논문에서는 이동통신 PTT 단말기에서 이름 음성 호출을 통해 통화 연결하기 위한 모델을 연구하였다. 그 결과 여러 가지 문제가 있지만, 확보된 데이터 통해 KALDI 도구를 활용한 krs 레시피를 사용하여 정확도를 측정한 결과 84%의 정확도를 보였다.

이름 음성 호출 모델의 정확도를 개선하기 위해서는 응용이 적합한 데이터의 확보를 통해 학습되어야 하며, 언어 모델을 이용할 수 없기 때문에 문자 단위의 연관성을 가진 문자 상관 모델의 개발이 필요하다.

## ACKNOWLEDGMENT

본 연구는(or 본 논문은) 울산광역시-ETRI (2차) 공동협력사업 일환으로 수행되었음. [22AS1600, 제조 혁신을 위한 주력산업 지능화 기술 개발 및 산업현장에서의 사람-이동체-공간 자율협업지능 기술 개발]

## 참 고 문 헌

[1] Pete Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," Apr 2018, https://arxiv.org/abs/1804.03209

[2] Oleg Rybakov외 4인, "Streaming keyword spotting on mobile devices," Jul. 2020, arXiv:2005.06720.

[3] 김훈, 음성인식 방법과 카카오i의 음성형엔진, 카카오AI리포트, Vol.06, 2017.06, https://brunch.co.kr/@kakao-it/105.

[4] 윤상범, AI 보안 음성인식 - 1강 음성인식의 이해, 2021.1학기, http://www.kocw.net/home/cview.do?mty=p&kemId=1379711.

[5] 천궈궈 외3, 칼디로 배우는 음성인식, 지니북스, 2022년 04월.

[6] 해커의 개발일기, "#4 음성인식 KALDI 툴을 이용한 한국어 음성인식 (zeroth project)", 2019. 8. , https://bourbonkk.tistory.com/28.

[7] 양형원, "칼디로 한국어 음성인식 구현하기 파트 1~4", 2017-07, https://hyungwonsnotebook.blogspot.com/2017/07/kaldi-tutorial-for-korean-model-part-[1234].html

[8] 국립국어원, 서울말 낭독체 발화 말뭉치, https://www.korean.go.kr/front/board/boardStandardView.do?board_id=4&mn_id=17&b_seq=464.

[9] AI허브, 명령어 음성(일반남녀) 데이터셋, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=96.

[10] 한국어 STT, CER과 WER 계산, 2022.1, https://mingchin.tistory.com/240.