pISSN 2005-8063
eISSN 2586-5854
2020. 12. 31.
Vol.12 No.4
pp. 81-90

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2020.12.4.081

Building a Korean conversational speech database in the emergency medical domain

Sunhee Kim[1] · Jooyoung Lee[2] · Seo Gyeong Choi[3] · Seunghun Ji[2] · Jeemin Kang[3] · Jongin Kim[4] · Dohee Kim[5] · Boryong Kim[1] · Eungi Cho[1] · Hojeong Kim[1] · Jeongmin Jang[1] · Jun Hyung Kim[6] · Bon Hyeok Ku[6] · Hyung-Min Park[6] · Minhwa Chung[2,*]

[1]*Department of French Language Education, Seoul National University, Seoul, Korea*
[2]*Department of Linguistics, Seoul National University, Seoul, Korea*
[3]*Department of English Language and Literature, Seoul National University, Seoul, Korea*
[4]*Department of Interdisciplinary Program in Cognitive Science, Seoul National University, Seoul, Korea*
[5]*Department of Foreign Language Education, Seoul National University, Seoul, Korea*
[6]*Department of Electronic Engineering, Sogang University, Seoul, Korea*

Abstract

This paper describes a method of building Korean conversational speech data in the emergency medical domain and proposes an annotation method for the collected data in order to improve speech recognition performance. To suggest future research directions, baseline speech recognition experiments were conducted by using partial data that were collected and annotated. All voices were recorded at 16-bit resolution at 16 kHz sampling rate. A total of 166 conversations were collected, amounting to 8 hours and 35 minutes. Various information was manually transcribed such as orthography, pronunciation, dialect, noise, and medical information using Praat. Baseline speech recognition experiments were used to depict problems related to speech recognition in the emergency medical domain. The Korean conversational speech data presented in this paper are first-stage data in the emergency medical domain and are expected to be used as training data for developing conversational systems for emergency medical applications.

Keywords: conversational speech, speech data, speech recognition, annotation, emergency medical domain

1. 서론

의료 영역에서 음성과 관련된 연구로는 발성기관이나 조음기관의 장애와 관련된 음성장애 영역의 연구와 음성언어처리 기술을 활용하는 연구로 나누어 볼 수 있다. 먼저 음성장애 영역의 연구는 음성 데이터를 이용하여 그 음향학적 특징을 바탕으로 음성 질환을 진단하고 분류하는 주제들을 포함한다 (Hernandez et al., 2020; Maryn et al., 2009; Seo & Seong, 2013). 최

* mchung@snu.ac.kr, Corresponding author
Received 14 November 2020; Revised 15 December 2020; Accepted 15 December 2020

근에는 이러한 발성이나 조음과는 관련이 없는 우울증(depression)이나 치매(dementia), 혹은 불안 장애(anxiety) 등과 같은 정신과적 질병이 발화에 영향을 미친다는 연구들이 보고되고 있고(Cummins et al., 2015; Laukka et al., 2008; Weiner et al., 2017), 이를 바탕으로 하여 다양한 기계학습이나 DNN(Deep Neural Network) 방법을 이용하여 환자들의 음성을 자동으로 분류하는 방법들이 제안되고 있다(Huang et al., 2020; Xezonaki et al., 2020).

의료 영역에서 음성언어처리 기술을 활용한 예로는 음성인식 기반의 딕테이션 기술을 영상의학과의 영상 판독에 직접적으로 이용한 사례가 대표적이라고 할 수 있다(Chapman et al., 2000; Mariani et al., 2006). 기존의 영상 판독은 의사들이 구두로 발화한 판독 내용을 녹음한 다음 이 녹음 파일을 전사자가 듣고 작성하는 녹취록에 의존하였으나 음성인식 기술을 이용하여 자동 녹취로 이를 대체하게 된 것이다. 최근의 음성인식 기술은 조용한 환경이나 일정한 정도의 소음이 존재하는 환경에서는 매우 높은 성능을 보여 스마트 스피커의 인공지능 시스템인 어시스턴트 등에 널리 활용되고 있다. 이러한 성능 향상에 따라 영상 판독 외의 다양한 의료 상황에서도 음성인식 기술이 진료기록을 목적으로 활용될 것이라는 기대가 높아지고 있다. 그러나 실제 의료 상황에서는 단순히 의사와 환자 간의 대화 외에도 환경 잡음을 비롯한 여러 잡음이 존재하기 때문에 신뢰할 만한 음성인식 성능을 기대하기는 쉽지는 않다.

의료 영역 가운데에서도 특히 음성인식과 관련하여 어려움이 예상되는 영역으로는 응급 의료 상황을 들 수 있다. 응급실 상황에서의 의료 대화는 주로 환자의 증상을 빠르게 파악하기 위한 의사의 질문과 증상을 설명하는 환자의 답변으로 이루어진다. 의사는 대화 내용 중 진료에 중요한 정보를 파악하여 기록하지만, 대화 내의 모든 정보를 정확하게 기록하는 것은 어려운 일이다. 또한, 진찰 후 진료 내용 전문을 확인하기도 쉽지 않다.

현재까지의 진료 기록은 의사의 진찰 메모와 기억에 크게 의존하였고 엄밀한 의미의 데이터를 집적한 예는 많지 않다. 경우에 따라 음성을 녹음하기도 하지만 이를 확인하기 위하여 사람이 일일이 다시 듣고 전사를 하는 것은 현실적으로 많은 시간이 소요될 뿐만 아니라 그 정확성을 다시 확인하는 것도 쉽지 않다. 이러한 문제를 극복하기 위해 양질의 음성 데이터를 수집하여 음성인식 기술을 통해 대화 내용을 텍스트 형태로 변환하고, 최종적으로 환자의 증상과 관련된 임상정보를 객관적인 기준에 따라 자동으로 추출하여 진료 및 임상 정보를 데이터베이스화하는 것이 필요한 실정이다(Chapman et al., 2000; Xu et al., 2010).

이와 같이 음성인식 기술을 의료 영역에 활용하기 위해서는 해당 영역의 실제 데이터를 구축하는 것이 선행되어야 한다. 실제 환경에서 수집되지 않은 데이터를 사용하는 경우에는 그 결과를 사용할 수 있을 만한 음성인식 성능을 보장할 수 없다. 따라서 음성인식 기반의 임상 정보를 추출하기 위해서는 실제 환경에서 데이터를 수집하여 그 음향적, 언어적 특성을 파악하는 것이 필수적이다. 최근 음성인식은 DNN 기술의 적용과 이를 기반으로 한 종단간(end-to-end) 모델의 적용과 함께 높은 성능을 보이고 있는데(Wang et al., 2019), 이러한 성능은 모델과 함께 사용할 수 있는 대용량의 데이터에 기인하는 것으로 현재 모델에서 데이터의 중요성이 더 커지고 있는 상황이다.

한국어 자유 발화 음성데이터로는 최근에 한국정보화진흥원의 주도로 구축된 AIHub 데이터(https://aihub.or.kr/) 가운데 조용한 환경에서 2,000여 명이 발성한 한국어 대화음성 969시간 분량의 데이터(Bang et al., 2020)와 방송 콘텐츠로 구성된 2,000시간 분량의 데이터가 있다. 이 두 데이터의 경우는 비교적 조용한 환경에서 수집된 데이터인데 반하여 본 연구는 의료 분야에서도 가장 분주하고 소음이 많은 환경에서 일어나는 환자와 의사 간의 대화 수집을 대상으로 한다. 본 논문은 응급의료 환경에서 음성인식 성능을 향상시키기 위하여 실제 환경에서 데이터 수집 방법을 정의하고 정의된 환경에서 수집된 데이터를 전사하는 방법을 제시한다. 그리고 제안된 방법으로 수집되고 전사된 데이터를 이용하여 기본 음성인식 실험을 진행함으로써 제안한 수집 및 전사 방법을 평가하고 향후 연구 방향을 제시하고자 한다.

이러한 연구는 응급영역에서의 음성대화 데이터 구축을 바탕으로 응급의료 영역의 음성인식 성능을 향상시켜 진료기록 자동화 및 임상정보 자동 추출 도구를 개발하는 것을 목표로 하고 있다. 나아가 최종적으로는 진찰에 필요한 정보를 객관적인 기준에 의해 저장하고 관리하는 음성 데이터베이스를 구축하여 의료정보화에 기여하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 응급 의료 영역에서의 음성 데이터를 수집하고 전사하는 방법을 기술한다. 3장에서는 기본 음성인식기를 개발하는 방법을 소개하고, 이를 이용하여 수집된 데이터의 일부에 대한 음성인식 실험과 그 결과를 설명한다. 마지막으로 4장에서는 연구의 한계와 향후 진행할 연구를 소개하고 마무리한다.

2. 데이터 수집 및 전사

2.1. 수집 환경

근거리 마이크를 이용한 음성인식의 경우와는 달리 1 m 이상 원거리에 마이크가 설치되어 있는 경우는 음성이 마이크까지 도달하는 동안 왜곡이 일어나고 이러한 왜곡으로 인하여 음성인식 성능이 크게 저하된다. 주변의 잡음에 의한 왜곡과 음성이 마이크에 도달하는 동안 벽이나 다른 사물 등에 부딪히는 반향 성분으로 인한 왜곡이 대표적이다. 이와 같은 음성인식의 성능 저하를 보상하기 위하여 다채널 기반의 선형 필터를 이용하여 잡음과 반향 성분을 제거하는 전처리 기술이 필요하다. 일반적으로 음성인식에서 전처리는 다채널의 마이크 어레이를 이용하여 유입되는 음성에 대하여 기본적으로 다음과 같은 세 단계의 방법으로 진행된다.

- 반향 제거(Yoshioka & Nakatani, 2012, 2013)
- 음원 위치 추정(Grondin & Glass, 2019; Higuchi et al., 2016)

· 빔포밍(Cho et al., 2019; Kubo et al., 2019)

최근에는 언급한 각각의 전처리 방법을 통합하는 알고리즘 연구도 진행되고 있다(Boeddeker et al., 2020).

응급 의료 상황은 의사나 환자가 마이크를 근거리에 두고 대화를 하는 상황이 아니라 원거리 마이크를 이용하여 대화가 수집되는 상황이다. 의사와 환자의 대화를 실제 환경에서 수집할 때에는 이와 같은 원거리 상황을 상정한 마이크 어레이를 통한 데이터 수집이 필요하다. 본 연구에서 응급의료 상황의 데이터는 실제로 환자에 대한 의사의 진료가 이루어지는 다음의 두 장소에서 수집하기로 하였다.

먼저 그림 1과 같이 응급실 내 환자의 침대 구역에서 환자와 의사 사이에서 일어나는 대화를 6개의 마이크를 이용하여 수집한다. 그림에서 보는 바와 같이 환자 위치에 5개로 구성된 마이크 어레이와 의사 위치에 1개의 마이크를 설치하였다.[1]

그림 1. 응급실 내 침대 전경 및 마이크 설치 현황
(상) 응급실 내 침대 전경, (하좌측) 환자 쪽 마이크 위치(1 – 5번), (하우측) 의사 쪽 마이크 위치(6번)
Figure 1. Overall view of a bed in the emergency room and microphone installation. (top) Overall view of a bed in the emergency room, (bottom left) Microphones on the patient's side (1–5), (bottom right) Microphone on the doctor's side (6)

응급실 내에서 데이터가 수집되고 있는 또 다른 장소는 환자 분류소로, 여기에서는 그림 2에서 보는 바와 같이 의사와 환자 사이에 16개 채널 마이크 어레이를 설치하였다. 그러나 마이크 어레이가 설치되기 전인 초기에는 기본적으로 환자 위치와 의사 위치에 각각 하나씩의 마이크가 설치되어 2개의 채널로 수집되었다. 그림 2와 같이 의사와 환자/보호자 사이에 거리가 있고, 서로 마주 보고 있는 상황에서 각 대화 참여자의 소리를 효과적으로 녹음하기 위해서는 여러 위치에서 녹음이 가능한 16채널 마이크가 적합하다. 데이터 전사에는 16채널 가운데 음도가 강하게 들어오는 1번이나 2번, 그리고 15번이나 16번 채널을 통하여 수집된 발화를 추출하여 사용하였다. 모든 음성은 16비트의 해상도와 16 kHz 샘플링으로 저장되었다.

그림 2. 응급실 내 환자 분류소
(상) 환자 분류소 전경, (하) 의사 쪽 마이크 어레이 배치
Figure 2. Triage station in the emergency room. (top) Overall view of the triage station, (bottom) Microphone array on the doctor's side

마이크 어레이는 다양한 빔포밍 방식을 이용하여 주어진 지역 내에서 음원의 위치를 검출하고 음원이 아닌 다른 여러 방향에서 입력되는 잡음을 제거함과 동시에 원하는 음성 성분을 강화하는 데 필요한 방법론이다. 이는 응급실과 같이 다양한 잡음 환경에서의 음성인식 성능을 향상시키기 위하여 필수적인 연구라고 할 수 있다. 그림 3은 16채널 마이크 어레이인데, 이 가운데 1번 혹은 2번이 환자, 혹은 환자와 보호자 위치의 음성이 주로 수집되는 마이크이고, 15번과 16번이 의사 위치의 음성이 주로 수집되는 마이크이다.

16채널 마이크는 그림 3과 같이 원형 어레이 6개의 마이크(7-12)와 선형 어레이 10개의 마이크(1-6, 13-16)로 구성하였다. 원형 어레이는 중심으로부터 4 cm에 육각형 모양으로 6개가 배치되었다. 선형 어레이는 원형 어레이의 중심을 통과하는 10개의 마이크로 4 cm 간격으로 일직선 모양을 이룬다. 이때, 15-16번 간격만 포트 연결을 위하여 2 cm로 구성하였다.

1 현재는 COVID-19 상황으로 인하여 응급실 환자 침실 구역에 설치된 마이크 등의 설치 시설이 폐쇄되어 잠정적으로 데이터 수집이 중단된 상태이다.

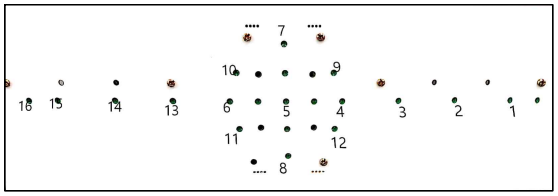

그림 3. 16채널 마이크 어레이
Figure 3. Sixteen-channel microphone array

2.2. 수집 데이터

다음 표 1은 위에서 언급한 방법대로 수집하고 전사가 완료된 데이터 현황이다. AV-CV-2 데이터는 환자 침대 구역에서 환자와 의사의 대화를 의사와 환자 양쪽에 각각 1개의 마이크의 마이크를 이용하여 2채널로 수집된 초기 데이터로서 총 39건이 수집되었다. AV-TR-2 데이터는 환자 분류소에서 환자와 의사의 대화를 수집한 데이터로 초기에는 AC-CV-2 데이터와 마찬가지로 환자 위치와 의사 위치에 각각 설치된 2개의 마이크로 수집되었고, 이 환경에서 수집된 데이터는 총 19건이다. 이후 동일한 장소에 16채널 마이크를 설치하여 수집한 AV-TR-16의 경우는 수집된 데이터 가운데 현재까지 108건으로 전체 대화는 총 166건이다.

표 1. 수집 데이터 요약
Table 1. Summary of collected data

	AV-CV-2	AV-TR-2	AV-TR-16
진료 대화 건수	39건	19건	108건
	총 166건		
음성 재생 시간 (휴지 구간 포함)	3시간 31분	1시간 0분	4시간 4분
	총 8시간 35분		

환자 침대 구역의 대화인 AV-CV-2의 경우 평균 대화 시간은 5분 4초이고, 환자 분류소 대화인 AV-TR 데이터의 경우 평균 대화 시간은 2분 6초로 환자 침대 구역이 대화가 더 길게 진행된 것을 알 수 있다.

2.3. 전사 방법

데이터 전사를 위하여 전사 규칙 초안을 만들고 일부의 데이터에 대하여 전사자들이 다 같이 전사를 수행하여 전사 규칙을 수정 및 보완하였다. 두 개의 마이크를 사용하여 2채널로 수집된 초기 데이터의 경우 AV-CV-2 데이터와 AV-TR-2 데이터의 전사는 기본적으로 의사 위치의 채널을 통하여 수집된 음성을 사용하였고, 환자 위치의 채널 데이터를 참고하였다. 16채널로 수집된 AV-TR-16의 경우도 마찬가지로 기본적으로 의사 위치의 마이크 16번 채널을 통하여 수집된 음성을 사용하였고, 환자 위치의 마이크 1번 채널을 통하여 수집된 음성을 참고하였다.

대화음성의 경우 인위적으로 녹음된 낭독체 발화와는 달리 발화의 경계를 정하기 어렵다. 따라서 문장을 단위로 분절하지 않고 일정 길이의 휴지를 기준으로 발화 단위를 분절한다. 현재 전체 수동으로 전사를 진행하고 있으므로 전사자가 음성을 직접 들으면서 휴지가 느껴지거나 Praat(Boersma & Weenink, 2018)에서 보았을 때 육안으로 보이는 휴지 구간을 경계로 구분하였다. 전체 음성 데이터의 전사는 Praat의 Annotation 기능을 이용하여 수행하였다. 다음 그림 4는 Praat를 이용한 전사 예이다.

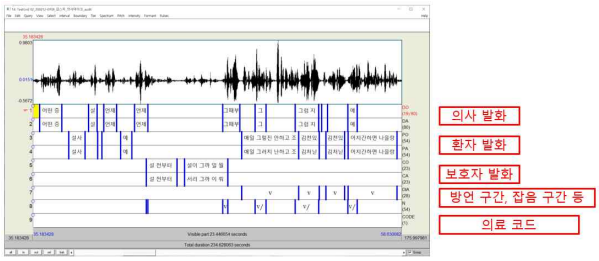

그림 4. Praat를 이용한 음성 전사 예
Figure 4. A sample of transcription using Praat

우선 의사와 환자/보호자의 발화를 구분하여 각각의 층위(Tier)로 표시하였다. 이때, 각 발화자의 발화는 표기법대로 전사하는 철자 전사(orthographic transcription)와 발음 나는 대로 전사하는 음소 전사(phonemic transcription)로 구분하여 전사하였다. 본 데이터는 대화체 발화 특성상 두 명 이상의 화자가 동시에 발화하는 발화 겹침이 발생하는데, 별도의 잡음 층위를 통해 발화 겹침 구간을 표시하였다. 또한, 추가적인 정보 파악을 위해 환자 발화에 한하여 방언 발화 구간을 표시한 방언 층위와 대화 내 각종 잡음을 표시한 잡음 층위도 함께 기록하였다. 마지막으로, 진단에 있어서 중요하다고 판단되는 단어를 검출하여 이에 해당하는 의료코드를 추가한다. 주요 전사 지침은 아래에서 설명한다.

2.3.1. 철자 전사

철자 전사는 기본적으로 표준어 맞춤법에 맞게 전사하는 것이다. 이와 같은 철자 전사는 음성인식기의 언어모델 학습을 위한 텍스트 데이터의 기본 형태이다. 철자 전사에는 발음을 잘못하고 다시 말을 제대로 고치는 경우(발화 수정)나 단어 일부를 번복하여 발화하는 경우(발화 반복)도 모두 포함된다. 그리고 숫자, 외래어, 기호의 경우도 모두 한글 철자로 표기하고 문장부호는 전사하지 않는다. 다만, 경우에 따라 의미 파악이 어려운 발화는 들리는 발음으로 전사한다.

대화체 발화 특성상 '어...', '음...' 등의 간투어 발화가 다수 나타나는데, 이러한 경우는 철자 전사 열에서 해당 단어 뒤에 기호 '/'를 붙인다. 간투어의 예는 다음과 같다.

- 아/, 그/, 어/, 음/, 저/, 저기/, 에/, 으/, 응/, 등

2.3.2. 음소 전사

음소 전사는 철자 전사에 해당하는 발음을 표기하는 부분으로 기본적으로 음성인식의 음향모델에 관여하게 된다. 한국어의 음소 전사는 기본적으로 음운현상에 따라 변동된 음소를 반영하게 되는데, 본 연구에서는 일관된 음소 전사를 위하여 발음

열 생성기(KoG2P: Lee et al., 2018)를 사용하여 철자 전사로부터 음소를 자동으로 추출한 다음 그 결과를 수동으로 검토하였다.

음소 전사에 있어서 주의할 점으로는 기본적으로 철자 전사와 음절수를 동일하게 유지하는 것인데, 경우에 따라 음절수가 맞지 않는 경우는 철자 전사 열이나 음소 전사 열에 기호 '@'를 추가하였다.

2.3.3. 방언 전사

방언 전사는 발화 가운데 방언으로 판단되는 부분의 구간을 표시하는 방식으로 진행한다. 전사 업무자의 판단 하에 방언이 들렸다고 할 때 그 구간을 표시하도록 한다. 억양이나 강세는 판단하기 어렵기 때문에 어휘 차원에서 방언을 사용했을 경우에 대해서만 방언 구간으로 표시하기로 하였다. 이는 다양한 방언 정보가 포함된 자유 발화를 텍스트로 기록하여 이후 방언의 특징이 객관적으로 표시하기 위함이다.

2.3.4. 잡음 전사

잡음 전사는 발화 가운데 말소리가 아닌 부분의 구간을 구분하여 전사하는 것으로 음성인식에서 제외되거나 따로 모델링 할 필요가 있는 부분에 해당한다. 잡음은 그 특성에 따라 다시 분류하는데, 일단 아래와 같이 4종류로 분류하였고, 필요하면 기타로 분류된 잡음 가운데 새롭게 분류 기호를 추가할 수 있다.

- h /: (human), 사람한테서 발생하는 잡음. 예) 웃음소리, 기침 등
- b/: (background), 배경에서 발생하는 잡음. 예) 자동차 소리, 다른 사람의 목소리 등
- k/: 기타 모든 일시적으로 발생한 소리. 예) 틱, 톡, 탁, 삐삐삐 등
- v/: (overlap), 두 사람 이상이 동시에 발화하여 발화가 겹치는 경우

본 연구에서는 두 사람의 발화가 겹치는 중복 발화의 경우도 잡음과 같이 그 구간을 표시한다. 그림 4에서와 같이 두 사람의 발화가 중복되는 부분은 기본적으로는 두 발화자의 철자 전사와 음소 전사 층위에 각각 구분하여 전사하는데, 겹치는 구간의 경우 음성인식에서 문제가 되므로 이 구간을 따로 표시해 주는 것이 필요하다.

2.3.5. 의료 코드 전사

의료 환경에서 의사는 진단을 위하여 환자에게 증상을 물어보고 환자가 대답을 하였을 때, 의사가 다시 환자의 대답을 따라 말하거나 대답에서의 증상을 반복하여 말하기도 한다. 따라서 많은 경우에 의사의 발화와 환자의 발화에 진단의 실마리가 되는 단어들이 포함된다. 의사나 환자의 발화에 대한 철자 전사 결과 가운데 이러한 의료 코드에 해당하는 단어들을 검출하고 이에 해당하는 진단명을 의료코드 층위에 전사한다. 이때, 철자 전사 층위에서 의료 코드와 관련 있는 단어 앞에 #를 각각 붙이고, 의료 코드 층위에 해당 의료 코드를 추가한다. 예를 들면, 의사 발화의 철자 전사에서 저혈압에 해당하는 단어를 다음과 같이 표시한다.

- 그러니까 너무 #혈압이 #떨어지는 거죠 일어날 때

2.4. 전사 데이터 통계

표 1에서 전사 대상인 데이터는 총 166건인데 전사 결과 이 가운데 환자와 의사 2인 간의 대화는 103건이었고, 환자와 의사, 그리고 보호자, 3인의 대화는 63건이었다. 대화에 참여한 의사는 남성 4명과 여성 4명, 총 8명이었고, 대화에 참여한 환자는 남성이 77명, 여성이 89명으로 166명이었다. 그리고 대화에 보호자가 등장한 63건에서 남성은 35명, 여성 보호자는 28명이 포함되어 있었다. 따라서 데이터에는 남성 116명, 여성 121명으로 총 237명의 화자가 포함되어 있다.

다음 표 2는 수집 데이터에 대한 전사 결과를 요약한 표이다. 발화 수를 보면, 의사의 발화 수는 6,977개, 환자의 발화 수는 4,779개, 그리고 보호자의 발화 수는 853개로 총 발화 수는 12,609개이다. 전체 166개의 대화 가운데 방언 전사 구간은 108건으로 비교적 그 빈도가 높지 않았다. 잡음 전사 구간은 총 4,086건으로 대화 당 평균 24.6건의 잡음이 포함되어 있었다. 이 가운데 배경 잡음(/b/)이 2,594건(평균 15.7건)으로 가장 빈번하였고, 두 사람 이상이 동시에 말하는 구간도 1,234건(평균 7.4건)이 전사되었다.

표 2. 수집 데이터에 대한 전사 결과 요약
Table 2. Summary of transcription of collected data

		AV-CV-2	AV-TR-2	AV-TR-16
발화 수	의사	3,825	523	2,629
	환자	2,553	375	1,851
	보호자	0	128	725
	총 12,609 발화			
방언 전사 결과		85건	23건	0건
		총 108건		
잡음 전사 결과	h/	258		
	b/	2,594		
	k/	0		
	v/	1,234		
	총 4,086건			

표 3은 대화 참여자별 발화 수 및 어절 수를 나타낸다. 의사의 발화 수는 6,977개, 환자의 발화 수는 4,779개, 그리고 보호자 의 발화 수는 853개이다. 각 대화 당 의사의 평균 발화 수는 42.0발화, 환자는 28.8발화, 보호자는 13.5발화로 의사의 발화 수가 환자나 보호자보다는 많으나 실제로 각 대화에서 환자와 보호자의 발화를 합친 수와 의사의 발화 수가 비슷한 것을 알 수 있다. 어절 수를 보면, 의사 발화의 어절 수는 26,288개, 환자 발화의 어절 수는 17,984개, 그리고 보호자 발화의 어절 수는 4,052개로 총 어절 수는 48,324개이다. 대화 당 평균 어절 수는 의사는 158.4어절, 환자는 108.3어절, 그리고 보호자는 64.3어절로 전체

어절 수와 비슷한 양상을 보인다. 발화 당 평균 어절 수는 의사와 환자, 보호자가 각각 3.8어절, 3.8어절, 4.8어절로 보호자의 발화가 의사나 환자에 비하여 평균 1어절이 길다는 것을 볼 수 있다.

표 3. 대화 참여자별 발화 수 및 어절 수
Table 3. Number of utterances and number of Eojeols by speaker groups

		의사	환자	보호자
발화 수	전체	6,977	4,779	853
	평균	42.0	28.8	13.5
어절 수	전체	26,288	17,984	4,052
	대화 평균	158.4	108.3	64.3
	발화 평균	3.8	3.8	4.8

3. 음성인식 실험

3.1. 음성인식기 구축

현재 서울대학교 언어학과에서 보유 중인 음성인식기는 Kaldi 기반의 음성인식기이다(Povey et al., 2011). 음향모델은 체인(Chain) 기반 딥러닝 구조로, 언어모델은 Sentencepiece Model (Kudo & Richardson, 2018)을 거친 서브워드(subword)를 학습한 HMM(Hidden Markov Model) 구조로 이루어져 있다.

음향모델은 단계별 GMM(Gaussian mixture model) 학습과 정렬(alignment) 과정을 통해 DNN 기반 음향 모델의 초기 값을 생성하였다. 25 ms 길이의 음성 프레임에서 추출한 40개의 MFCC (mel frequency cepstral coefficients) 음향 특징을 차례로 모노폰과 트라이폰 단위 GMM에서 학습하고 LDA(linear discriminant analysis)와 MLLT(maximum likelihood linear transform)를 통해 정렬 과정을 거쳐 각 유사음성단위(phone-like-unit)의 음향 특징 구간을 학습하였다. 이때 2,400시간 분량의 원본 데이터와 이를 기반으로 음성 열화(speech perturbation) 기법을 이용하여 생성한 0.9배속과 1.1배속 데이터를 합친 총 7,200시간 분량의 데이터를 GMM 학습에 사용하였다(Park et al., 2019).

학습이 완료된 GMM 모델의 파라미터는 Kaldi에서 지원하는 체인 모델이라는 DNN 기반 음향모델 학습 알고리즘의 초기 값으로 사용하였다. GMM 학습에서 유사음성단위당 3개의 HMM state를 사용하는 대신에 유사음성단위당 1개의 HMM state만 사용하는 방식으로 변경하였고, 이를 토대로 만들어진 토폴로지(topology)는 체인 기반 트리(tree)를 생성하는 데 사용하였다.

언어모델은 Google의 Sentencepiece Model(Kudo & Richardson, 2018)을 통해 인공신경망이 분절한 40만 개의 서브워드 사전을 만든 후, 이를 사용하여 HMM 기반 언어모델을 학습하였다. 이때 Sentencepiece Model이 생성한 서브워드는 단어의 형태소 경계와 무관한, 단어 분절 빈도수를 계산하여 만든 단위이다. 언어모델에 이용한 데이터의 양은 뉴스, SNS, 대화체 음성 등 낭독발화와 자유 발화를 모두 포함한 약 1억 8천만 개의 문장이다.

3.2. 실험 결과

3.2.1. 음성인식기의 기본 성능 실험

성능 실험을 위하여 비교 대상으로 대화체 자유 발화인 AIHub 문장 5,000개와 10,000개를 각각 사용하여 실험하였다. 다음 표는 이에 대한 단어 오류율(word error rate, WER)로 나타낸 실험 결과이다. 오류 유형으로는 삭제나 삽입보다는 두 경우 모두 단어 대체 오류가 많은 것을 볼 수 있었다. 실험 결과 단어 오류율은 각각 23.80%와 23.47%로 비슷한 결과를 보였다(표 4).

표 4. 음성인식기 기본 성능
Table 4. Baseline performance of speech recognition

발화 수	5,000개	10,000개
WER	23.80%	23.47%

WER, word error rate.

3.2.2. 응급의료 영역 데이터에 대한 기본 성능 실험

본 실험에서는 수집된 AV-CV-2, AV-TR-2, AV-TR-16 데이터에서 의사, 환자, 보호자 각각의 발화를 분절하여 개별적으로 저장한 다음 기본 인식 성능을 실험하였다. 실험에는 AC-CV-2 유형은 6,378발화, AV-TR-2와 AV-TR-16 유형은 1,973 발화를 사용하였고, 그 결과는 다음 표 5와 같다. 응급실 내 침대 위치에서의 단어 오류율은 71.24%, 환자 분류소에서는 81.06%로 환자 분류소에서의 인식률이 낮게 나타난 것을 볼 수 있다.

표 5. 응급의료 영역 음성인식 기본 성능
Table 5. Baseline performance of speech recognition in emergency medical domain

데이터 유형	AV-CV-2	AV-TR-2 AV-TR-16
발화 수	6,378	1,973
WER	71.24%	81.06%

3.3. 논의

본 연구를 위하여 개발한 대어휘 음성인식기는 전통적인 방식의 음성인식기로서 단어오류율 23% 정도의 성능을 보였는데, 이는 최근 AIHub 데이터 1,000시간 분량을 이용하여 종단간(end-to-end) 방식으로 개발한 음성인식기(Bang et al., 2020)에 비하여 좀 더 높은 성능을 보였다. 그러나 응급의료 영역의 데이터를 사용한 경우에 크게 성능이 저하된 것을 볼 수 있었다.

현재 인식기의 학습 데이터는 대부분은 근거리(close-talk) 마이크를 사용하여 수집된 데이터이고, 기본 성능 실험 데이터인 AIHub 데이터도 마찬가지로 근거리 마이크로 수집된 데이터이다. 이에 반하여 응급의료 영역의 데이터는 모두 고정된 마이크를 통해 수집하고 의사/환자와의 거리가 1 m 이상 되는 원거리에 마이크가 설치되어 있어 마이크에 유입되는 음성의 특성이 기존의 학습 데이터 및 AIHub 데이터와는 많이 다르다고 할 수 있다. 뿐만 아니라 배경에서 들리는 각종 잡음이 음성인식 성능

을 저하시키는 또 다른 요인이 될 수 있다. 또한, 환자가 고령이거나 질환으로 인하여 음성이 뚜렷하지 않은 경우 환자의 발화는 특히 인식이 잘 안 되는 것을 볼 수 있다.

그리고 AIHub의 데이터는 본 데이터와 같이 대화음성이긴 하지만 잡음 없는 환경에서 일정한 주제에 대하여 인위적으로 대화를 유도하여 수집된 데이터인데 반하여, 본 데이터는 실제 상황에서 수집된 데이터로서 기계에서 나오는 배경 잡음과 대화 미참여자들의 배경 대화 등이 포함되는 등 그 특성이 많이 다를 수 있다. 즉, 본 데이터는 응급의료 환경이라는 실제 상황에서의 대화 음성으로서 심리적인 요인을 포함한 다양한 요인으로 인하여 일반적인 대화 음성과는 다른 특성을 보인다. 기본적으로 빔포밍 방법을 이용한 음원 강화와 잡음 처리를 포함하는 전처리 기법이 전혀 적용되지 않은 것이 낮은 성능의 주요한 요인이라고 할 수 있다. 아직은 데이터의 분량도 적은 편으로 기본적인 성능 향상을 위해서는 이러한 모든 문제들을 고려한 데이터 수집이 계속되어야 할 것이다.

4. 결론

본 연구에서는 응급환자의 임상정보 추출을 목적으로 의료 대화 166건을 2채널과 16채널 마이크로 수집하였다. 수집된 데이터는 Praat를 이용하여 철자 전사, 음소 전사, 방언 전사, 잡음 전사, 그리고 의료 코드 전사를 수행하여 다양한 정보를 포함한 텍스트 데이터를 구축하였다. 본 연구에서는 대화음성에 대한 전사 방법을 제안하고 있는데, 이는 응급의료 영역에서의 음성 발화 정보를 텍스트로 기록하고 전사에 대한 체계적이고 표준화된 기준을 만들기 위함이다. 또한, 본 논문에서 제안한 방법에 따라 수집한 데이터는 응급의료 영역의 1단계 데이터로서 향후 의료 영역에서의 음성인식 모델의 학습 데이터로 활용될 수 있으며, 환자의 주요 임상정보의 종류와 분포를 구축된 텍스트 데이터를 통해 확인할 수 있다는 점에서 중요하다 할 수 있다.

본 논문에서는 기본 베이스라인 실험을 통하여 응급의료 영역에서의 음성인식 문제를 실제로 확인할 수 있었다. 현재 기본 베이스라인 실험으로 Kaldi를 이용하였으나 이후 기본 성능을 높이기 위하여 종단간(end-to-end) 모델을 이용한 실험들을 진행할 계획이다. 또한, 응급의료 영역에서의 데이터 증강이나 기본 인식 단위 등 관련 이슈들에 대해서도 이후 연구가 필요한 실정이다. 특히, 다양한 잡음이 존재하는 환경에서 수집된 데이터를 이용하여 음성인식 성능을 향상시키기 위한 전처리 연구도 계속 진행될 계획이다. 뿐만 아니라, 환자와 의사간의 대화의 특성에 대한 언어학적 연구들도 진행될 예정이다.

결론적으로, 본 연구는 실제 의료 상황에서 의사와 환자 간에 이루어진 대화를 수집한 데이터를 구축하는 구체적인 방법을 제안하고, 이러한 데이터를 기반으로 개발한 음성인식 시스템을 통하여 의료 대화를 전사하고 의료 정보를 추출하는 기초 연구라고 할 수 있다. 다른 한편으로는 이와 같은 의사와 환자 간에 실제로 일어나는 대화들을 이용하여 의사의 일부 역할을 기계가 대신하는 인공지능 의료 도우미 시스템을 개발하는 데 있어서도 본 연구가 중요한 자료로 이용될 수 있을 것으로 기대한다.

감사의 글

본 연구는 정보통신기획평가원 대학ICT연구센터지원사업 과제 “의료 빅데이터 융합 전문가 인력 양성을 위한 비정형 빅데이터의 정형화 기술 및 분석 플랫폼 개발” (IITP-2020-2018-0-01833)의 연구지원비 지원에 의해 수행되었습니다.

References

Bang, J. U., Yun, S., Kim, S. H., Choi, M. Y., Lee, M. K., Kim, Y. J., Kim, D. H., & Kim, S. H. (2020). KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences, 10*(19), 6936.

Boeddeker, C., Nakatani, T., Kinoshita, K., & Haeb-Umbach, R. (2020, May). Jointly optimal dereverberation and beamforming. *Proceedings of the 2020－2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 216-220). Barcelona, Spain.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer (version 6.0.37) [Computer program]. Retrieved from http://www.praat.org/

Chapman, W. W., Aronsky, D., Fiszman, M., & Haug, P. J. (2000). Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department. *Proceedings of the AMIA Symposium* (p. 131).

Cho, B. J., Lee, J. M., & Park, H. M. (2019). A beamforming algorithm based on maximum likelihood of a complex Gaussian distribution with time-varying variances for robust speech recognition. *IEEE Signal Processing Letters, 26*(9), 1398-1402.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication, 71,* 10-49.

Grondin, F., & Glass, J. (2019, May). SVD-PHAT: A fast sound source localization method. *Proceedings of the 2019~2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4140-4144). Brighton, UK.

Hernandez, A., Kim, S., & Chung, M. (2020). Prosody-based measures for automatic severity assessment of dysarthric speech. *Applied Sciences, 10*(19), 6999.

Higuchi, T., Ito, N., Yoshioka, T., & Nakatani, T. (2016, March). Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5210-5214). Shanghai, China.

Huang, Z., Epps, J., Joachim, D., Stasak, B., Williamson, J. R., &

Quatieri, T. F. (2020). Domain adaptation for enhancing Speech-based depression detection in natural environmental conditions using dilated CNNs. *Interspeech 2020* (pp. 4561-4565). Shanghai, China.

Kudo, T., & Richardson, J. (2018, August). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66-71).

Kubo, Y., Nakatani, T., Delcroix, M., Kinoshita, K., & Araki, S. (2019). Mask-based MVDR beamformer for noisy multisource environments: introduction of time-varying spatial covariance model. *Proceedings of the 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (pp. 6855-6859). Brighton, UK.

Laukka, P., Linnman, C., Åhs, F., Pissiota, A., Frans, Ö., Faria, V., Palmquist, Å. M., & Furmark, T. (2008). In a nervous voice: Acoustic analysis and perception of anxiety in social phobics' speech. *Journal of Nonverbal Behavior, 32*(4), 195.

Lee, Y., Shon, S., & Kim, T. (2018). Learning pronunciation from a foreign language in speech synthesis network. *arXiv*. Retrieved from https://arxiv.org/abs/1811.09364

Mariani, C., Tronchi, A., Oncini, L., Pirani, O., & Murri, R. (2006). Analysis of the X-ray work flow in two diagnostic imaging departments with and without a RIS/PACS system. *Journal of Digital Imaging, 19*(1), 18-28.

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., & Corthals, P. (2009). Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America, 126*(5), 2619-2634.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019* (pp. 2613-2617). Graz, Austria.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., ... Vesely, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* Big Island, HI.

Seo, I., & Seong, C. (2013). Voice quality of dysarthric speakers in connected speech. *Journal of the Korean Society of Speech Sciences, 5*(4), 33-41.

Wang, D., Wang, X., & Lv, S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry, 11*(8), 1018.

Weiner, J., Engelbart, M., & Schultz, T. (2017). Manual and automatic transcriptions in dementia detection from speech. *Interspeech 2017* (pp. 3117-3121). Stockholm, Sweden.

Xezonaki, D., Paraskevopoulos, G., Potamianos, A., & Narayanan, S. (2020). Affective conditioning on hierarchical attention networks applied to depression detection from transcribed clinical interviews. *Interspeech 2020* (pp. 4556-4560). Shanghai, China.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: A medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association, 17*(1), 19-24.

Yoshioka, T., & Nakatani, T. (2012). Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(10), 2707-2720.

Yoshioka, T., & Nakatani, T. (2013, September). Dereverberation for reverberation-robust microphone arrays. *21st European Signal Processing Conference (EUSIPCO 2013)* (pp. 1-5). Marrakech, Morocco.

• **김선희(Sunhee Kim)**
서울대학교 불어교육과 교수
서울시 관악구 관악로 1
Tel: 02-880-7693
Email: sunhkim@snu.ac.kr
관심분야: 프랑스어 음성학, 음성언어처리

• **이주영(Jooyoung Lee)**
서울대학교 언어학과 박사과정
서울시 관악구 관악로 1
Tel: 02-880-6162
Email: excalibur12@snu.ac.kr
관심분야: 음성인식, 방언학, 딥러닝

• **최서경(Seo Gyeong Choi)**
서울대학교 영어영문학과 석사과정
서울시 관악구 관악로 1
Tel: 02-880-6162
Email: csganna@snu.ac.kr
관심분야: 음성인식, 제2언어습득, 사회언어학

• **지승훈(Seunghun Ji)**
서울대학교 언어학과 석사과정
서울시 관악구 관악로 1
Tel: 02-880-9039
Email: seunghun.ji@snu.ac.kr
관심분야: 음성언어처리, 음성학, 기계학습

• **강지민(Jeemin Kang)**
서울대학교 영어영문학과 석사과정
서울시 관악구 관악로 1
Tel: 02-880-6162

Email: bling1104@snu.ac.kr
관심분야: 음성인식, 발음평가, 딥러닝

• **김종인(Jongin Kim)**
서울대학교 인지과학 협동과정 박사과정
서울시 관악구 관악로 1
Tel: 02-880-6162
Email: prows12@gmail.com
관심 분야: 음성인식, 자연언어이해, 대화모델, 패턴인식

• **김도희(Dohee Kim)**
서울대학교 외국어교육과 불어전공 석사과정
서울시 관악구 관악로 1
Tel: 02-880-7693
Email: dohee826@snu.ac.kr
관심분야: 음성학, 외국어교육

• **김보령(Boryoung Kim)**
서울대학교 불어교육과 석사과정
서울시 관악구 관악로1
Tel: 02-880-7693
Email: jadebr@snu.ac.kr
관심분야: 음성인식, 교육 평가, 외국어교육

• **조은기(Eungi Jo)**
서울대학교 불어교육과 석사과정
서울시 관악구 관악로 1
Tel: 02-880-7693
Email: eungi78@snu.ac.kr
관심분야: 음성 데이터 분석, 음성 합성

• **김호정(Hojeong Kim)**
서울대학교 불어교육과 석사과정
서울시 관악구 관악로 1
Tel: 02-880-7693
Email: hojeong43@snu.ac.kr
관심분야: 음성학, 통사론, 불어교수법

• **장정민(Jungmin Jang)**
서울대학교 불어교육과 석사과정
서울시 관악구 관악로 1
Tel: 02-880-7693
Email: jjungmini@snu.ac.kr
관심분야: 음성학, 불어교수법

• **김준형(Jun Hyung Kim)**
서강대학교 전자공학과 박사과정
서울시 마포구 백범로 35
Tel: 02-711-8916
Email: imalbert@naver.com
관심분야: 음성인식, 음성신호처리

• **구본혁(Bon Hyeok Ku)**
서강대학교 전자공학과 석사과정
서울시 마포구 백범로 35
Tel: 02-711-8916
Email: k01032633280@gmail.com
관심분야: 음성인식, 음성신호처리

• **박형민(Hyung-Min Park)**
서강대학교 전자공학과 교수
서울시 마포구 백범로 35
Tel: 02-711-8916
Email: hpark@sogang.ac.kr
관심분야: 음성인식, 음성신호처리

• **정민화 (Minhwa Chung)** 교신저자
서울대학교 언어학과 교수
서울시 관악구 관악로 1
Tel: 02-880-9195 Fax: 02-882-2451
Email: mchung@snu.ac.kr
관심분야: 음성인식, 음성언어처리, 컴퓨터 기반 언어교육

응급의료 영역 한국어 음성대화 데이터베이스 구축*

김 선 희[1] · 이 주 영[2] · 최 서 경[3] · 지 승 훈[2] · 강 지 민[3] · 김 종 인[4] · 김 도 희[5] · 김 보 령[1] ·
조 은 기[1] · 김 호 정[1] · 장 정 민[1] · 김 준 형[6] · 구 본 혁[6] · 박 형 민[6] · 정 민 화[2]

[1]서울대학교 불어교육과, [2]서울대학교 언어학과, [3]서울대학교 영어영문학과,
[4]서울대학교 인지과학협동과정, [5]서울대학교 외국어교육과, [6]서강대학교 전자공학과

국문초록

본 논문은 응급의료 환경에서 음성인식 성능을 향상시키기 위하여 실제 환경에서 데이터 수집 방법을 정의하고 정의된 환경에서 수집된 데이터를 전사하는 방법을 제안한다. 그리고 제안된 방법으로 수집되고 전사된 데이터를 이용하여 기본 음성인식 실험을 진행함으로써 제안한 수집 및 전사 방법을 평가하고 향후 연구 방향을 제시하고자 한다. 모든 음성은 기본적으로 16비트 해상도와 16 kHz 샘플링으로 저장되었다. 수집된 데이터는 총 166건의 대화로서 8시간 35분의 분량이다. 수집된 데이터는 Praat를 이용하여 철자 전사, 음소 전사, 방언 전사, 잡음 전사, 그리고 의료 코드 전사를 수행하여 다양한 정보를 포함한 텍스트 데이터를 구축하였다. 이와 같이 수집된 데이터를 이용하여 기본 베이스라인 실험을 통하여 응급의료 영역에서의 음성인식 문제를 실제로 확인할 수 있었다. 본 논문에서 제시한 데이터는 응급의료 영역의 1단계 데이터로서 향후 의료 영역에서의 음성인식 모델의 학습 데이터로 활용되고, 나아가 이 분야의 음성기반 시스템 개발에 기여할 수 있을 것으로 기대된다.

핵심어: 음성대화, 음성 데이터, 음성인식, 전사, 응급의료 영역

참고문헌

서인효, 성철재(2013). 연결발화에서 마비말화자의 음질 특성. *말소리와 음성과학*, *5*(4), 33-41.

* 본 논문의 일부 내용은 한국음성학회 2020년 가을학술대회에서 발표함.