ICT 챌린지 2020 사업계획서

Attention 기반 딥러닝 알고리즘을 활용한 방언 핵심 억양 모델링

주관대학: **서울대학교**

센터명: **의료빅데이터연구센터**

팀명: **Attention to Dialect**

팀장: **이주영 (언어학과 박사과정)**

목차

- 연구 개발 필요성 및 목표
- 연구 개발 내용 (계획)
  - 개요
  - 데이터 선정
  - 데이터 수집 및 가공
  - 방언 식별 딥러닝 아키텍쳐
- 기대 성과 및 활용
- 진행 일정

연구 개발 필요성

- 과학수사를 위한 화자 프로파일링
  - 지역 미확인 화자의 음성의 방언을 자동으로 추적
  - 방언과 관련한 화자 신원 파악 측면에서 수사 비용(시간 등) 절감
  - 음성 감정관의 보조 도구로서 활용

- 의사-환자의 진료 과정에서 대화 참여자의 방언 발화 인식
  - 현재 음성 인식은 표준어 사용에 대해서는 매우 높은 인식률을 보이나, 방언 인식은 저조
  - 방언 식별을 우선 진행한 후 해당 방언에 특화된 인식 시스템을 통해
    정확하 진료 데이터 구축 가능

연구 개발 목표

1. 연구를 위한 음성 데이터 수집 및 가공
   - 음성과 텍스트만 주어진 원시 데이터에서 연구에 필요한 음성 구간 추출
2. 한국어 방언학에서 기술한 방언 표지 정리
   - 한국어 방언학에서 정리한 방언별 음운, 문법, 어휘, 운율 등 정리
3. 방언학 표지로 활용할 수 있는 음향 특징 선정
   - 실험 음성학 측면에서 분석한 음향 특징 중 억양과 관련한 특징
4. Attention을 활용한 딥러닝 아키텍쳐 구현 및 실험
   - 타 분야 연구까지 참고하여 억양 특징을 파악할 수 있는 아키텍쳐 제안 및 성능 고도화
5. 여러 방언 데이터셋을 비교하여 방언 농도에 따른 방언 식별율 확인
   - 방언 농도가 짙은 발화부터 방언 특징이 억제된 발화까지 서로 비교하며 모델 성능 확인

연구 개발 내용 (개요)

데이터

대검찰청 DB
낭독발화

대검찰청 DB
자유발화

국립국어원 DB
자유발화

한국어 방언학 연구 내 방언 표지

분절음적 요소

음운 – 음운변동, 모음
문법 – 종결어미
어휘 – 고유어휘

초분절음적 요소

운율 – 억양

계산 모델링에의 적용

분절음적 요소

음운변동 → 인식 결과 어휘
모음 → 포먼트
종결어미 → 인식 결과 어휘
어휘 → 인식 결과 어휘

초분절음적 요소

억양 → 음향 특징 + 딥러닝

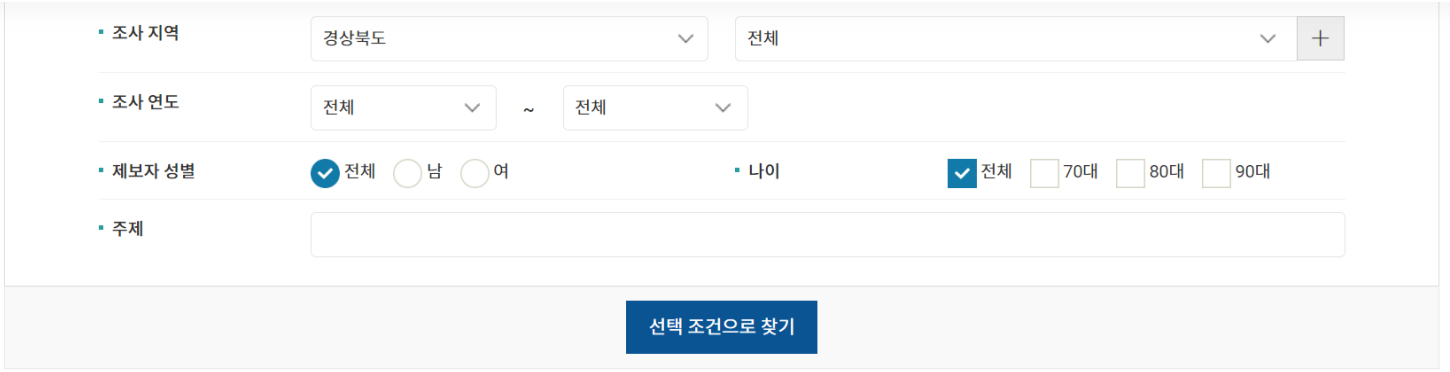
분류 모델

Attention 기반
Bidirectional LSTM

Support
Vector
Machine

연구 개발 내용 – 방언 데이터 선정

- 국립국어원 DB (https://dialect.korean.go.kr)
  - 전국 각지의 70대 이상 자료제공인의 옛 이야기를 녹음한 자료
  - 약 150개 자료
  - 음성과 전사 텍스트 존재
  - 조사자와 제보자의 대화
    - 주로 제보자가 대부분 발화



검색 가능한 주제 분류 목록은 자료실 에서 확인할 수 있습니다.

음성 듣기 기능은 Chrome 브라우저에서만 재생이 가능합니다

총 21개의 검색 결과가 있습니다.

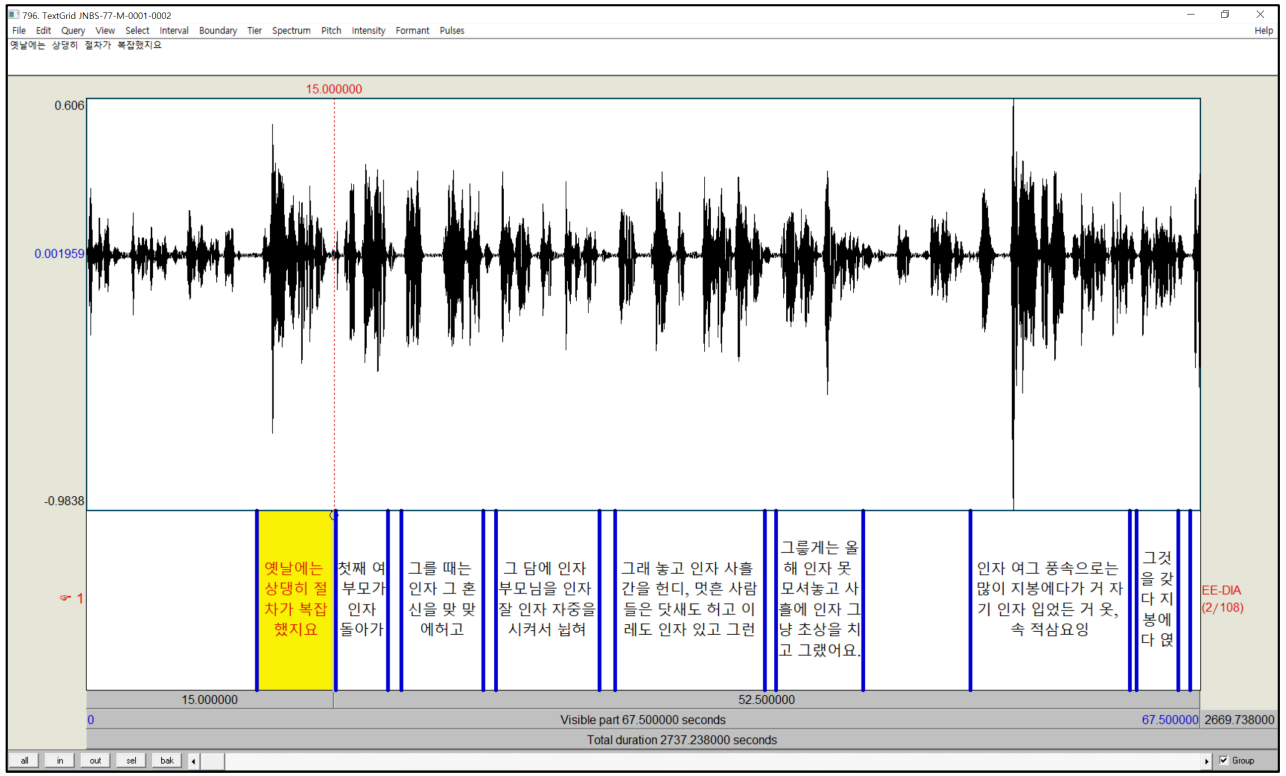
번호	지역	주제	제보자	조사 연도	음성	전사
1	경상북도 고령군	채소 재배와 요리	남 86세	2007		
2	경상북도 고령군	나물 채취와 요리	남 86세	2007		
3	경상북도 고령군	밑반찬의 조리	남 86세	2007		
4	경상북도 고령군	집짓기	남 86세	2007		
5	경상북도 고령군	가신과 조상 숭배 신앙	남 86세	2007		

연구 개발 내용 – 방언 데이터 수집 및 가공

- 전사 텍스트 태깅 작업
  - 자료제공인의 음성 구간을 추출하기 위해 태깅 작업 진행

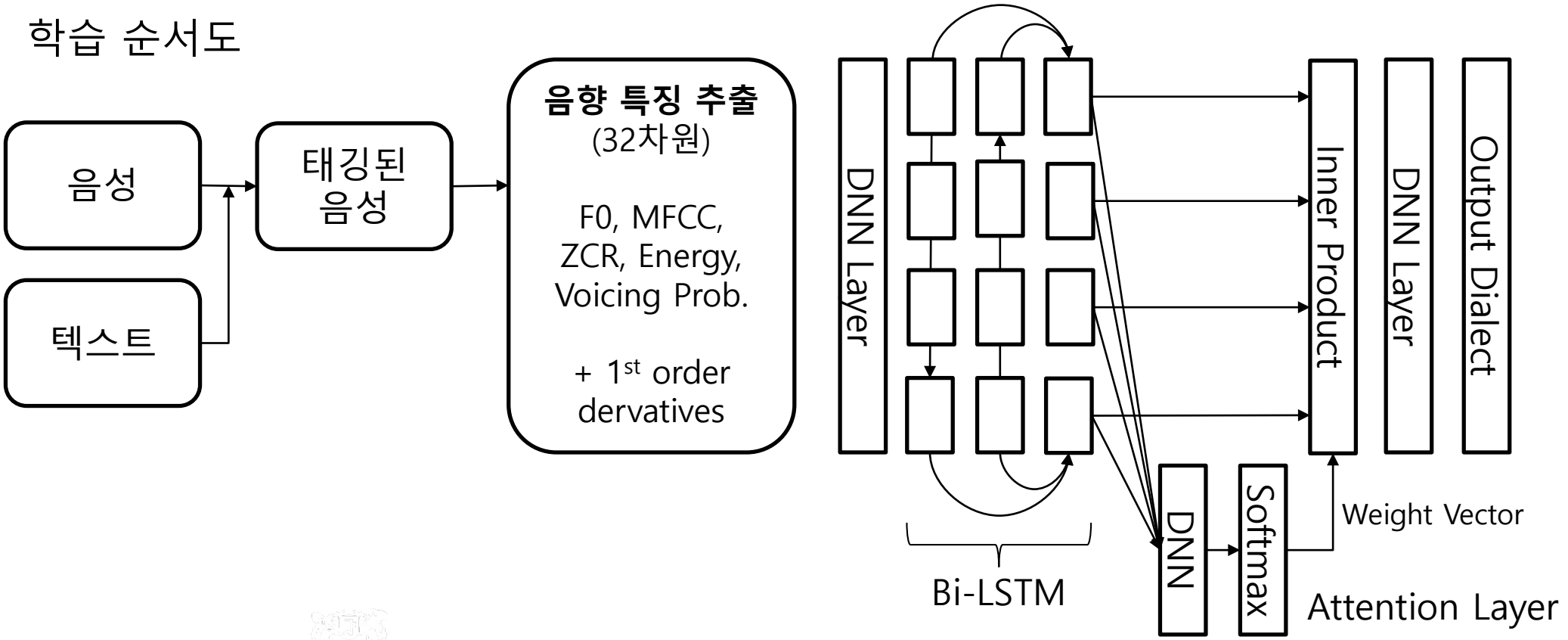
조사 지역: 전라남도 보성군
조사 연도: 2013
제보자: 남 77세

조사자:사람 죽을 때, 사람 죽을 때 어트게 초상 치루는가, 고 이야기를 조금 자세하게 해 주실랍니까? 옛날에? 사람이?
(사람 죽을 때 사람 죽을 때 어떻게 초상 치르는지 그 이야기를 조금 자세하게 해 주시겠습니까? 옛날에? 사람이?)
제보자:옛날에는 상댕히 절차가 복잡했지요
(옛날에는 상당히 절차가 복잡했지요.)
조사자:예
(예)
제보자:첫째 여 부모가 인자 돌아가셨다,
(첫째 부모가 이제 돌아가셨다,)
조사자:예
(예)
제보자:그를 때는 인자 그 혼신을 맞 맞에허고
(그럴 때는 이제 그 혼신을 맞이하고)
조사자:에
(에)
제보자:그 담에 인자 부모님을 인자 잘 인자 자중을 시켜서 넙혀 놓고
(그 다음에 이제 부모님을 이제 잘 이제 자중을 시켜서 눕혀 놓고)
조사자:음
(음)
제보자:그래 놓고 인자 사흘 간을 헌디, 멋흔 사람들은 닷새도 허고 이레도 인자 있고 그런디
(그래 놓고 이제 사흘 간을 하는데 어떤 사람들은 닷새도 하고 이레도 이제 있고 그런데)
조사자:아
(아)
제보자:그릉게는 올해 인자 못 모셔놓고 사흘에 인자 그냥 초상을 치고 그랬어요.
(그렇게는 오래 이제 못 모셔 놓고 사흘에 이제 그냥 초상을 치르고 그랬어요.)
조사자:예.
(예.)
제보자:예.
(예.)
조사자:그러먼언 혼신을 불러올라먼 어뜽게 합니까?
(그러면은 혼신을 불러오려면 어떻게 합니까?)
제보자:인자 여그 풍속으로는 많이 지붕에다가 거 자기 인자 입었든 거 옷, 속 적삼요잉
(이제 여기 풍속으로는 많이 지붕에다가 그 자기 이제 입었던 것, 옷, 속적삼이요응,)
조사자:예
(예)
제보자:그것을 갖다 지붕에다 엱어 놓든지,
(그것을 가져다 지붕에다 얹어 놓든지,)
조사자:예
(예)
제보자:그렇지 않으먼언 지붕 몬멩이로 올라가서 용머리를 타고
(그렇지 않으면은 지붕 꼭대기로('몬멩이'는 '꼭대기'의 뜻) 올라가서 용마루를('용머리'는 '용마루'의 방언형) 타고)



연구 개발 내용 – 방언 식별 딥러닝 아키텍쳐

- 핵심 억양 표지 학습
  - 억양과 관련한 음향 특징을 딥러닝 모델을 통해 학습하여 방언 자동 식별
  - 방언과 관련한 핵심 억양을 학습하도록 Attention 기법 사용
- 학습 순서도



기대 성과 및 활용

- 한국어 방언을 정의하는 데 사용되는 음성학적 특징 확인
- 과학 수사에서 화자의 신원 확인 과정에서 화자의 사용 방언 예측
  - 수사망을 크게 좁힐 수 있음
- 향후 데이터 추가를 통해 모델 성능 고도화로 세분화된 방언 지역까지 예측
- 다양한 배경의 환자가 존재하는 의료 대화에서 방언 발화에도 대응 가능
  - 방언 식별 과정을 통해 해당 지역 방언 발화에 특화된 의사/환자 방언 발화를 정확히 인식 가능
  - 진료 후 대화 메시지를 살펴보는 과정에서 정확히 인식된 텍스트로 의사는 정확한 진료 판단 가능

진행 일정

월별 진행 내용	10월	11월	12월
한국어 방언학 연구 정리 및 계산 모델링 적용 가능성 평가	**객관적인 평가 위해 외부 연구자와 회의** →		
국어원 DB 음성-텍스트 태깅 작업	**태깅 진행하면서 모델 업데이트** →	→	
Attention 억양 검출 모델 학습 진행 및 중간 테스트	**모델 학습** →	**성능 중간 평가** →	
데이터셋 간 교차 검증 (대검 DB & 국어원 DB)		**대검 자유/낭독 DB로 테스트하여 성능 비교** →	→