청유형 문장의 종결어미 음성신호 감정분류

5조

제출일	2023. 12. 3	전공	글로벌미디어학부
과목	멀티미디어론	담당교수	유 광 복
이름	20192694 신승훈, 20213004 김혜령, 20212012 이혜원		

[목차]

1. 서론.....오류! 책갈피가 정의되어 있지 않습니다.

1.1. 연구의 배경 4

1.2. 연구의 필요성 5

2. 실험.....오류! 책갈피가 정의되어 있지 않습니다.

2.1. 데이터 수집 6

2.2. 실험 내용 7

2.3. 실험 결과 7

3. 결론.....오류! 책갈피가 정의되어 있지 않습니다.

3.1. 남녀 별 음성 신호 분석 10

3.2. 감정별 음성 신호 분석 12

3.3. 결론 12

3.4. 참고 문헌 ..... 1**오류! 책갈피가 정의되어 있지 않습니다.**

[표 목차]

1. 표 1. Evolution of AI speakers technology 4
2. 표 2. 선정된 종결어미의 서법적 의미(감정) 6
3. 표 3. "요"의 남녀별 Spectrogram 11
4. 표 4. "요"의 남녀별 AMDF 11
5. 표 5. "요"의 감정별 Spectrogram ..... 1**오류! 책갈피가 정의되어 있지 않습니다.**
6. 표 6. "요"의 감정별 AMDF ..... 1**오류! 책갈피가 정의되어 있지 않습니다.**
7. 표 7. "요"의 감정별 LPC ..... 1**오류! 책갈피가 정의되어 있지 않습니다.**

[그림 목차]

1. 그림 1. Market of Artificial Inteligence 5
2. 그림 2. 유성음 "요"의 Spectrogram 8
3. 그림 3. 유성음 "요"의 ZCR 8
4. 그림 4. 유성음 "요"의 ACF 9
5. 그림 5. 유성음 "요"의 LPC 스펙트럼 10

1. 서론

1.1 연구의 배경

AI 스피커는 많은 부분에서 유용하게 사용된다. 기존의 인간들의 음성을 이용한 소통방식을 AI가 이해하고 학습하여 간단한 명령을 이해하고 우리의 삶을 보다 더 편리하게 만들었다. 'AI 스피커'란 기존 스피커에 인공지능(Artificial Intelligence) 기능이 더해진 스피커를 의미하며, 스피커의 기능에서 나아가 음성 비서로서 사용자가 가전제품을 제어할 수 있도록 '스마트 홈 지휘자' 역할을 한다. AI 스피커의 주된 인터페이스가 '음성'이라는 점에서 인간과 컴퓨터 간의 상호작용에 있어 가장 이상적이며 다양한 서비스 플랫폼으로 확장될 가능성이 높다. 또한, AI 스피커가 스마트 홈 허브가 되기까지의 기술적인 진화 배경을 살펴보면 그 중요성을 이해할 수 있다.

Generation	Technology
1st	The degree to which people follow.
2st	not practicable, simple voice command or voice control and low accuracy.
3st	To improve the speech recognition rate, the discriminative learning technique such as MCE is used, the N-GRAM-based language model technology, and the development of the Sphinx system of Carnegie Mellon University.
4st	Techniques to identify users and feelings or recognize the situation using big data and deep learning technology in real time.
5st	

표 1. Evolution of AI speakers technology

<표 1>은 AI 스피커 기술의 발전 과정이다. <표 1>에서 확인할 수 있듯이 과거 OS 기반 플랫폼에서는 정보의 생산, 처리, 저장, 분석이 스피커 기기 자체에 한정되었다. 하지만 최근 실시간 클라우딩 서비스가 활성화되고 빅 데이터 딥러닝 기술 등 ICT 환경이 개선되며 대용량 데이터 분석이 가능해졌다. 이를 통해 음성인식 정확도가 높아졌을 뿐 아니라 스피커가 사용자의 감정을 인지할 수 있게 되었다. 이는 AI 스피커의 파급력 확산에 일조했으며 현재 국내외 IT기업들이 AI 스피커에 심혈을 기울이는 이유이다[1].

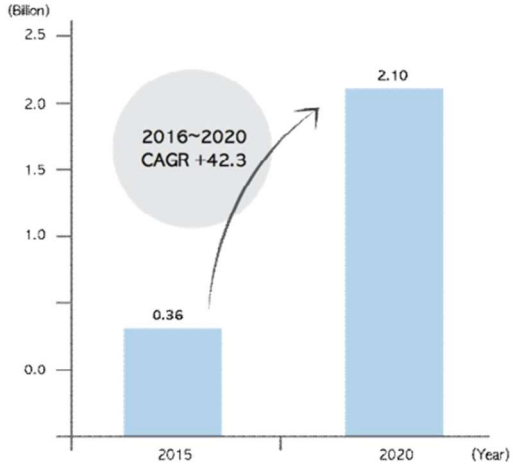


그림 1. Market of Artificial Inteligence

AI 스피커의 시장 성장률과 관련된 미국의 사례를 확인할 경우 2016년 AI 스피커 판매량은 570만대로 집계되며 1년뒤인 2017년에는 전년 대비 +329.8% 성장한 2450만대를 기록했다[2]. 위의 지표에서 알 수 있듯이 AI 스피커는 폭발적인 시장 성장률을 보이는 것을 확인할 수 있었다[2].

1.2 연구의 필요성

인간의 의사소통에는 언어의 의미 전달 이외에도 어조, 억양, 말의 빠르기, 목소리의 크기와 음색 등 반언어적 요소가 있다. 준언어적 의사소통에 관여하는 요소들 중 목소리의 정보는 사람의 미묘한 감정을 전달하기도 하며, 어떤 경우에는 단어에 있는 의미보다 더 큰 신뢰를 갖기도 한다. 더불어 준언어적 요소는 화자에 대한 청자의 인식하는데 큰 영향을 줌으로 인해 이미지 형성에도 많은 영향을 주기도 한다[3].

AI 스피커가 정말 인간처럼 상호작용하기 위해서는 감정이라는 요소의 학습이 추가적으로 진행되어야 한다. 한 연구에서는 인공지능이 마음을 인식할 때 느끼는 사람의 감정에 대해 연구하였다. 사람들에게 인공지능과의 개인적인 상호작용에서 인공지능이 마음 또는 정신 능력을 가지고 있다고 인식하는지에 대해 질문함으로써 경험적 접근 방식을 취한다. 다양한 유형의 인공지능에 대한 다양한 경험에도 불구하고 사람들의 반응에서 분명한 패턴이 나타났는데, 바로 인공지능에서 마음을 인식할 때 사람들은 감정을 경험한다는 것이다[4]. 따라서 AI 스피커의 사용자 경험의 질을 향상시키기 위해서 본 프로젝트에서는 음성신호를 분석하고 감정분류를 진행하고자 한다.

국내 초창기 AI 스피커 누구의 최대 한계점으로 한국어 인식 능력이 떨어진다. 이

에 SK텔레콤은 IBM의 음성인식 시스템 왓슨과 파트너십을 가져 보완점을 찾으며 꾸준한 성장을 시도했다. 하지만 여전히 실제 AI 스피커를 사용하고 있는 사용자들을 살펴보았을 때 아직 AI 스피커 이용에 한계를 느끼는 사람들이 많다. 음성 인터페이스를 통해 컴퓨터와 인간이 상호작용이 쉽다는 AI 스피커의 최대장점이지만 한국어 인식률이 비교적 떨어지기에 음악 스트리밍, 날씨, 알람과 같은 기본 기능 이상을 기대하기 어렵다는 단점이 있다[1]. 한국어 기반의 4,5세대 AI 스피커를 위한 한국어의 감정 요인들을 분석하고 이를 직접 실험하여 의미 있는 결과를 얻고자 한다.

2. 실험

2.1 데이터 수집

이는 서법과도 밀접한 연관이 있는데 동사의 굴절 또는 교착으로 문장에 대한 화자의 태도를 나타내는 형태론적 수단을 의미한다. 위의 문장은 행위 서법의 '제안'에 해당된다.

	서법적 의미	종결 어미	예문
감정	감탄/놀라움	-(으)ㄴ/는데, -는(군), -(ㄴ/는)다니, -구려	① 그 옷이 너한테 아주 잘 어울리**는데**. ② 세월이 참 빨리 가**는군**. ③ 이런 더러운 의자에 누가 앉**는다니**? ④ 어릴 적에 살던 곳인데 그새 많이 변했**구려**.
	아쉬움	-지(요), -(으)ㄹ걸	① 불꽃놀이는 벌써 끝났는데 조금만 일찍 오시**지요**. ② 가: 민준이가 고시에 합격했대. 나: 아, 이럴 줄 알았으면 나도 계속 공부**할걸**.
	못마땅	-(ㄴ/는)대, -기는	① 가: 이 많은 책을 언제 다 읽**는대**? 나: 정말. 숙제가 많아도 너무 많아. ② 아는 것도 없으면서 잘난 척하**기는**.
	친근함	-지(요)	① 가: 여기 가까운 화장실이 어디**지요**? 나: 저기 복도 끝으로 가시면 왼쪽에 있어요.

표 2. 선정된 종결어미의 서법적 의미(감정)

<표 2>는 종결어미의 서법적 의미를 감정으로 분류한 표다. <표 2>에서 보면 알

수 있듯이 한 종결어미가 여러 개의 감정 서법적 의미와 대응할 수 있다는 것을 알 수 있다. 화자가 청자에 대한 태도를 나타내는 서법을 '의향서법'이라 하는데 평서형, 의문형, 명령형, 청유형의 네 가지로 나뉠 수 있다[5]. Korean Emotional Speech Dataset(KESDy18) 데이터셋에서 청유문의 문장 '오늘 신문에 난 기사를 읽어봐요' 라는 문장을 선별했다. 기쁨, 슬픔, 화남의 감정 서법들이 들어갈 수 있는 청유문의 문장이라 위의 데이터셋을 추출한다.

2.2 실험 방법

위의 데이터셋으로 진행된 실험은 선행 연구에서 참고했다[6]. '읽어봐요' 부분을 추출한 뒤 Energy와 Spectrogram을 검출한다. 이후에 반복되는 파형을 제일 잘 확인할 수 있는 유성음을 추출하였다. 유성음은 청유문의 종결어미 '-요'를 512샘플로 프레이밍했다. 본 고에서 사용한 데이터는 외부 요인들을 최대한 배제하기 위해 동일인물의 성별, 감정을 변수로 설정하였다. 남자 3명과 여자 3명의 음성을 기쁨, 슬픔, 화남의 음성 데이터를 16kHz의 주파수로 샘플링하여 실험을 진행하였다. 여러 개의 감정 서법적 의미와 대응할 수 있는 청유형 문장인 "오늘 신문에 난 기사를 읽어봐요"를 사용하였고. 이 중에서 종결어미인 "요"를 집중적으로 분석하였다. 실험의 진행은 Pitch Detection Algorithms을 사용하였으며 자세한 내용은 아래와 같다.

2.3 실험 내용

실험 방법에 기반하여 실험을 진행하였다. 음성 신호는 시간에 따른 공기 압축의 변화로 나타난다. 음성 신호는 말하는 사람의 음성 파형을 기록하며, 각종 주파수 및 시간 도메인에서 다양한 특징을 가지고 있다. 여기에는 일반적으로 사용되는 몇 가지 음성 신호의 특징이 있다. 음성 파형의 높낮이는 주파수와 관련이 있다. 고주파수는 높은 음을 나타내며, 저주파수는 낮은 음을 나타난다. 각 발음의 지속시간은 음성 신호의 시간적 특징 중 하나다. 발음이나 음절 간 간격도 중요한 정보를 제공한다. 또한 음성 신호의 에너지 레벨을 통해서 음성 신호의 강도를 나타낸다.

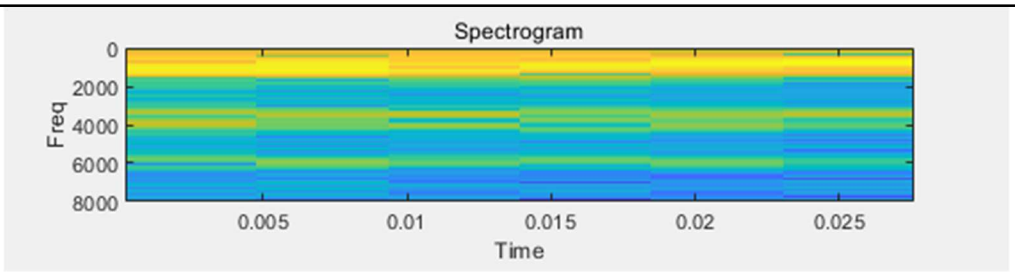


그림 2. 유성음 "요"의 Spectrogram

<그림 2>는 Spectrogram의 예시이다. x축은 시간, y축은 주파수를 의미하므로 피치나 톤으로 분석되며, 가장 낮은 주파수는 하단에, 가장 높은 주파수는 상단에 해당한다. 특정 시간에 특정 주파수의 진폭(에너지 또는 소리의 크기)은 색상으로 표시되며, 낮은 진폭에 해당하는 어두운 파란색과 점진적으로 더 큰 진폭에 해당하는 노란색을 통해 더 밝은 색상으로 표시된다. 특정 감정 상태에서 어떻게 바뀌는지 언어적인 특성을 파악할 수 있다. 음성 파형은 시간에 따른 음압의 변화를 시각적으로 나타낸 데이터로, 소리의 압력이 상승과 감소하는 부분을 포착한다. 이는 양극성과 음극성으로 나눌 수 있다. 양극성은 파형이 양의 값을 가지는 부분으로, 압력이 상승하는 구간이며, 음극성은 파형이 음의 값을 가지는 부분으로, 압력이 감소하는 구간이다. ZCR은 이 양극성과 음극성이 교차하는 지점이다.

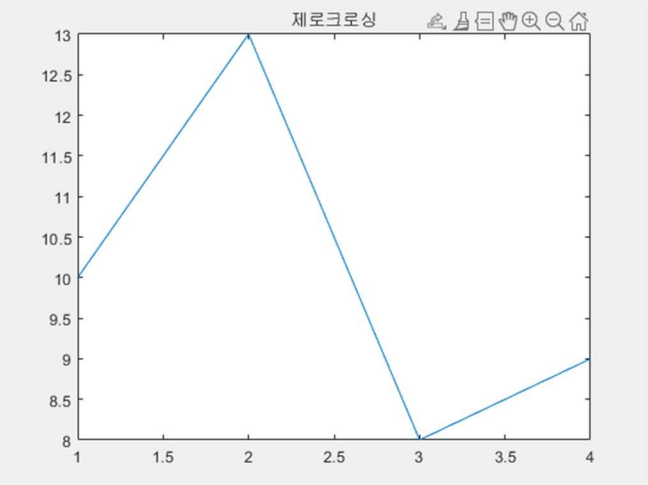


그림 3. 유성음 "요"의 ZCR

본 연구에서는 512개의 샘플로 이루어진 음성 파형을 100개씩 5개의 프레임으로 나누어 각 구간에서 ZCR의 빈도를 계산한다. ZCR의 빈도가 높을수록 음성 파형이 급격하게 변하는 소리로, 에너지가 높고, 엣지가 뚜렷하다는 특징이 있다. 이는 주로 고주파수 성분에서 나타나며, 주파수가 높을수록 음성 파형이 빠르게 변하므로 음과 양 사

이의 전환 수가 증가한다. 반면, ZCR의 빈도가 낮으면 주로 저주파수 성분이 많은 신호에서 나타나며, 파형이 부드럽고 엣지가 뚜렷하지 않다.

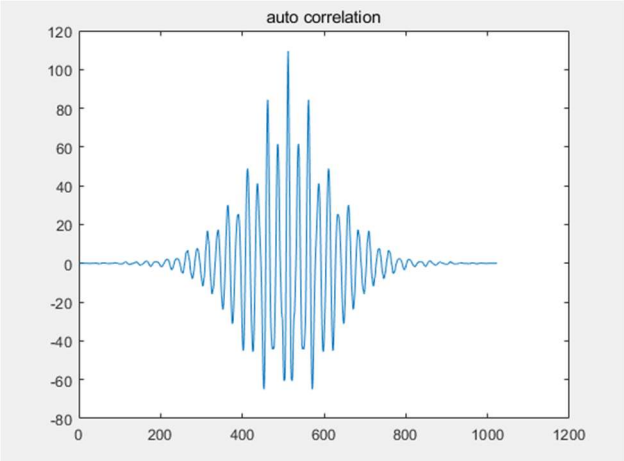


그림 4. 유성음 "요"의 ACF

<그림 3>은 유성음 '요'의 ACF를 구한 것이다. ACF는 좌우 대칭의 우함수로 음성신호의 피치 (pitch 혹은 fundamental frequency), 즉 기본주파수를 측정한다. 음성신호는 시간에 따라 변하는 준주기적인 신호이기에 그 주기, 피치를 추정하는 것이 매우 어렵다고 알려져 있다. 허나 ACF를 사용할 경우 위상 왜곡이 발생할 수 있는 신호의 피치를 탐지하는데 좋은 성능을 보인다.

ACF가 음성신호의 유사성을 기반으로 피치를 검출하는 것에 반해 AMDF는 신호의 차이로 피치를 검출한다. AMDF는 신호의 차이와 크기를 계산하므로, 그 연산량이 ACF에 비해 적고 동작시간이 빠르다는 장점이 있다.

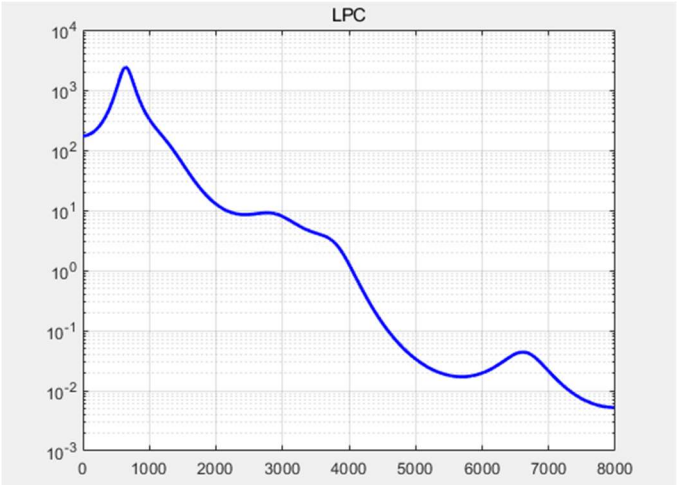


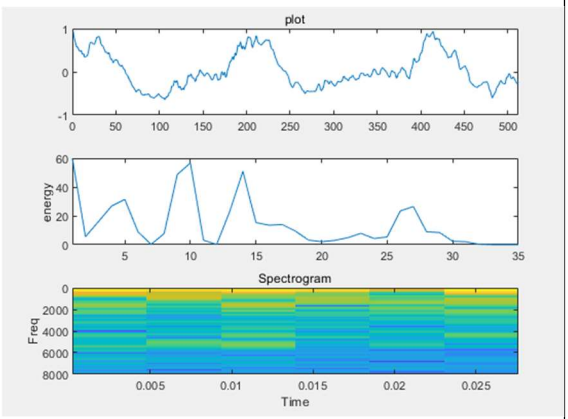
그림 3. 유성음 "요"의 LPC 스펙트럼

선형 예측 코딩(LPC)는 음성 파형을 선형 예측 모델로 분해하여 음성의 특징을 추출한다. 추정된 선형 예측 모델을 사용하여 다양한 목적에 맞게 음성의 특성을 변환할 수 있다. 본 고에서는 음성신호의 생성에서 중요한 기관인 성도 (vocal tract)를 선형화하여 음성신호를 모델링하고자 한다. 성도의 반응을 시간에 따라 변하는 시스템으로 보고 이의 필터 계수를 음성 생성의 파라미터로 볼 수 있다[6].

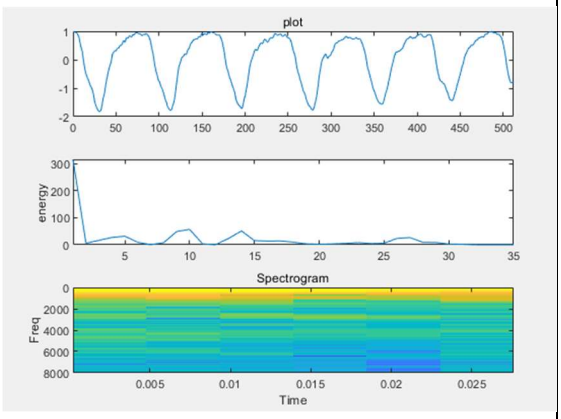
결론

3.1 남녀별 음성 신호 분석

남녀별 음성 신호 분석을 위해 같은 감정인 '슬픔' 데이터셋 간의 여성과 남성의 데이터를 비교하였다. 남성의 경우 ZCR의 평균은 3.5, 여성의 경우는 9.5로 여성이 비교적 높은 것을 확인할 수 있었다. ACF를 확인한 결과 남성의 피치 주기 샘플 간은 평균 96로 주파수 간격은 약 167이고 여성의 피치 주기 샘플 간격은 69로 주파수 간격이 232Hz였다. 즉 남성은 1초간 167번의 주기를 반복하고, 여성은 1초간 232번의 주기를 반복한다. 따라서 여성이 피치 주기가 더 짧은 것으로 보였다.



남자 우는 소리 "요"의 Spectrogram



여자 우는 소리 "요"의 Spectrogram

표 3. "요"의 남녀별 Spectrogram

남녀별 Spectrogram을 비교한 결과, 여자의 경우는 비교적으로 가장 밝은 색인 노란색을 띠고 있는 것을 볼 수 있고, 반대로 남자의 경우 녹색을 띠고 있는 것을 볼 수 있다. 이는 여자의 소리가 크고 더 강한 에너지를 가지고 있는 것으로 볼 수 있다.

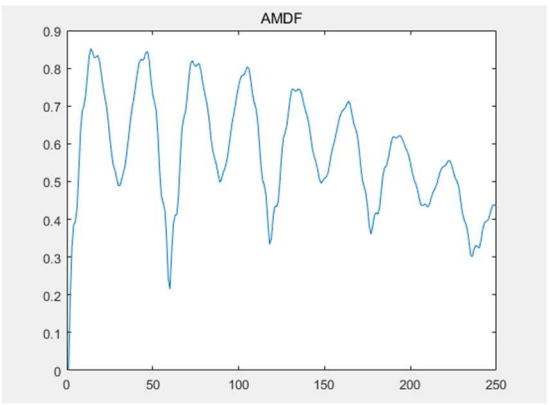
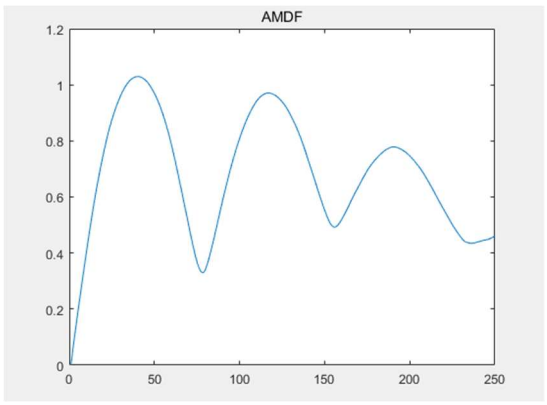
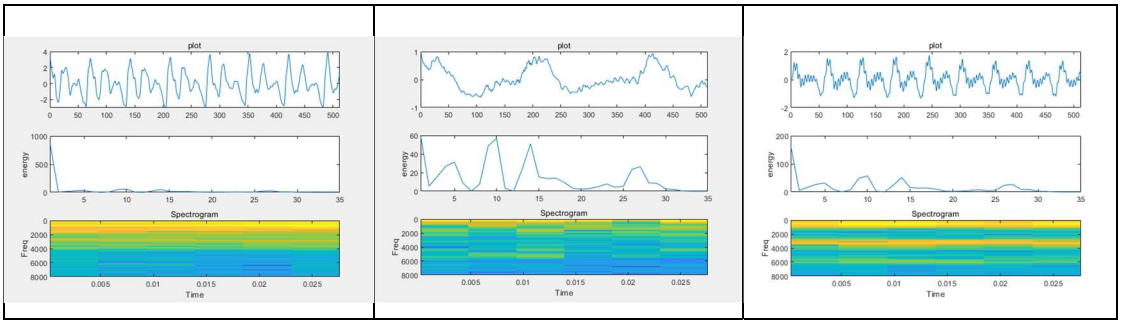
	
남자 우는 소리 “요”의 AMDF	**여자 우는 소리 “요”의 AMDF**

표 4. “요”의 남녀별 AMDF

AMDF를 확인했을 때, ACF의 결과와 마찬가지로 남성의 경우 AMDF 주기가 길고 여성의 경우 AMDF 주기가 보다 짧은 것을 확인할 수 있었다. 이는 주기가 긴 남성 음성에서 저주파수 성분이 강조되었고, 반대로 주기가 짧은 여성 음성에서 고주파수 성분이 강조되었음을 의미한다. 남녀별 LPC의 차이는 크게 보이지 않았다.

3.2 감정별 음성 신호 분석

감정별 음성 신호 분석을 위해 같은 성별인 ‘남성’의 데이터셋만 선별해 ‘기쁨’, ‘슬픔’, ‘화남’의 감정 데이터를 비교하였다. ‘기쁨’의 경우 ZCR의 평균은 10, ‘슬픔’의 경우 3.5, ‘화남’의 경우 18.7로 화남-기쁨-슬픔 순으로 ZCR 수치가 높은 것을 확인할 수 있었다. ACF를 확인한 결과 ‘기쁨’의 피치 주기 샘플 간격은 평균 81였고, ‘슬픔’은 평균 167, ‘화남’은 68만큼의 간격을 가졌다. 즉 ‘기쁨’은 1초간 평균 198번의 주기를 반복하고, ‘슬픔’은 1초간 96번의 주기, ‘화남’은 1초간 235번 반복한다. 따라서 ‘화남’과 ‘기쁨’ 감정의 피치 주기는 슬픔의 주기와 비교적 더 긴 것을 확인할 수 있었다.



남자 기쁜 소리 "요"의 Spectrogram	남자 슬픈 소리 "요"의Spectrogram	남자 화난 소리의 "요"의 Spectrogram

표 5. "요"의 감정별 Spectrogram

감정별 Spectrogram을 비교한 결과, 슬픈 소리, 기쁜 소리, 화난 소리의 순서로 Spectrogram이 전체적으로 밝아지고 에너지가 높아지는 것을 확인할 수 있다. 이를 통해 슬픈 소리는 다른 감정들에 비해 소리가 작고 낮은 에너지를 가지고 있는 편이며, 화난 소리는 소리의 크기가 크고 강한 에너지를 가지고 있다.

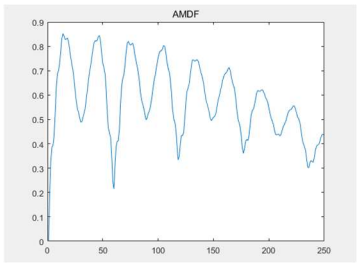
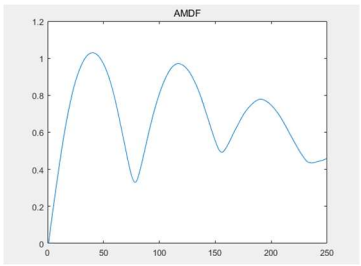
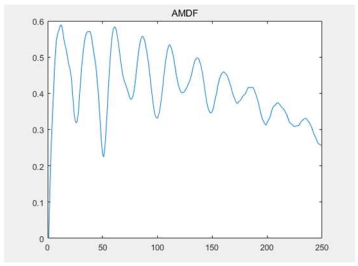
		
남자 기쁜 소리의 AMDF	남자 슬픈 소리의 AMDF	남자 화난 소리의 AMDF

표 6. "요"의 감정별 AMDF

AMDF를 확인했을 때, ACF의 결과와 마찬가지로 '슬픔'의 경우 AMDF 주기가 길고 '기쁨'과 '화남'의 경우 AMDF 주기가 확연히 짧은 것을 확인할 수 있었다. 이는 주기가 긴 '슬픔' 감정에서 저주파수 성분이 강조되었고, 반대로 주기가 짧은 '기쁨', '화남' 감정에서 고주파수 성분이 강조되었음을 의미한다.

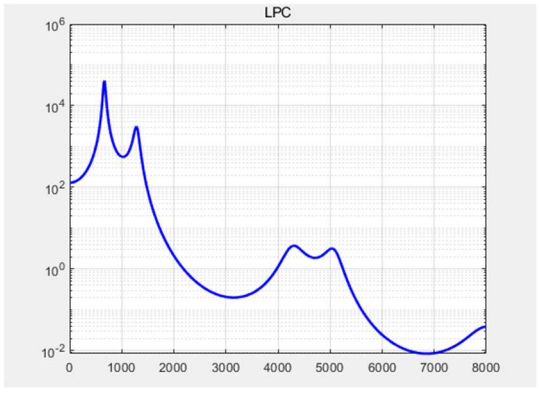
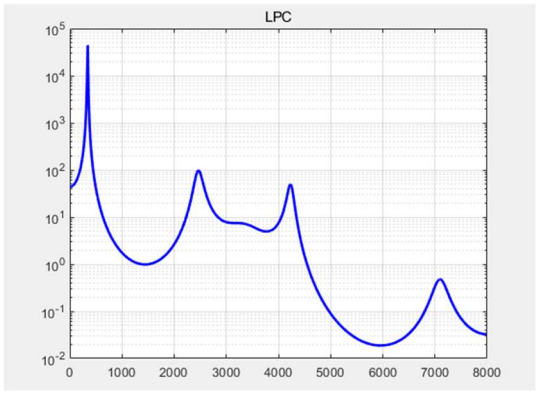
	
여자 화난 소리 "요"의 LPC	여자 우는 소리 "요"의 LPC

표 7. "요"의 감정별 LPC

성도는 음성 신호의 주파수 특성과 밀접한 관련이 있다. 음성 신호의 주파수 대역

은 성도의 상태에 따라 변화한다. 예를 들어, 성대가 짧고 두꺼우면 주파수 대역이 낮아지고, 성대가 길고 얇으면 주파수 대역이 높아진다. 화남 상태에서는 에피네프린이라는 호르몬이 분비되는데 이는 혈관을 수축시킨다. 신주영 의사의 말에 따르면 이러한 상태에서 성대는 긴장하게 되며 이는 곧 성대가 길어지고 얇아질 수 있다. 이에 따라 주파수 대역이 높아지는 것을 확인할 수 있다. 제 1음형대 주파수 상승과 제 2음형대의 주파수의 하강은 인두강이 좁아지는 경우와 관계된다[7]. 제 1음형대가 상승하고 제 2음형대가 하강하면 두 피크의 간격 차이가 작아지기 때문에, 슬픔 같은 경우 제 1음형대와 제 2음형대의 피크 차이가 2.125차이가 나나 화남 같은 경우는 전체 다 1.300 이하로 나오는 것을 알 수 있었다.

3.3 결론

감정 분류를 통한 AI 스피커의 성능 향상은 사용자 경험의 측면에서 중요한 발전을 가져올 수 있다. 사용자의 감정 상태를 파악하여 상호작용이 조절되어 자연스러운 대화가 가능해지며, 이는 사용자와의 감정적 연결을 강화할 수 있다. 이를 위해 본 고에서는 종결어미의 유성음 파형을 분석하여 음성신호에서 감정을 구분하려는 시도를 하였다. 대체적으로 화남-기쁨-슬픔 순으로 높은 에너지를 갖고 있는 것을 확인했고, 피치 주기를 확인하였을 때, '기쁨'과 '화남'은 각각 1초간 198번, 235번의 주기를 반복하고 '슬픔'은 96번의 주기를 반복하는 것을 보며 확연한 차이를 보였다. '기쁨'과 '화남'의 감정이 둘 다 격양된 목소리 톤을 갖고 있기 때문에 '슬픔'의 감정보다 주기가 높게 나온 것이라 예측된다. 실험 결과에서는 '기쁨'과 '화남'의 종결어미를 분석했을 때, 드러나는 차이는 화남이 비교적 주기가 높다는 것이다. 정확한 결과는 실험해본 데이터셋이 부족해서 명확히 결론을 내릴 수 없다는 한계점도 있습니다.

각 대비점에 있는 '기쁨'과 '화남'을 AI가 잘못 인식하여 답변을 준다면 오히려 사용자에게 부정적인 영향을 끼칠 수 있다. 여전히 감정분류는 커져가는 AI 스피커의 극복해야할 과제로 남아있다. 본 실험에서는 한국어의 종결어미에 표현되는 비언어적 요소를 분석하려는 의의가 있다.

3.4 참고 문헌

[1] 조규은. (2018). A study on User Experience of Artificial Intelligence speaker. Journal of the Korea Convergence Society Vol. 9. No. 8, pp. 127-133.

[2] Y. D. Kim. (2017). Artificial Intelligece speaking speakers brought to life. Zoom in

[3] Knapp, M. L. (1980). Essential of nonverbal communication. New York: Holt Rinehart and Winston Inc.

[4] Shank, D. B. (2019). Feeling Our Way to Machine Minds: People's Emotions When Perceiving Mind in Artificial Intelligence. Computers in Human Behavior, 98, 256-266.

[5] 왕주. (2019). A Study on the Significance of Korean Ending Suffix Narrattion Education Program for Chinese Learners(Master's Thesis). Inha University Graduate School.

[6] 염정석, 유광복. (2021). A Study on the Emotion Classification of the Speech Signal using Support Vector Machine. 한국통신학회논문지 제46권 제10호, pp. 1741-1749.

[7] 서경식, 김재영, 김영기. (1994). Relationship between Formants and Constriction Areas of Vocal Tract in 9 Korean Standard Vowels. 대한음성언어의학회지 제5권 제1호. pp. 56