공학학사학위논문

디지털 음성의 대역폭 제한에 따른 감정 인지 능력 비교

Comparison of Recognizing Emotions depending on Bandwidth-limitation of Digital Speech

2022년 06월

서울대학교 공과대학

산업공학과

황 보 진 경

디지털 음성의 대역폭 제한에 따른 감정 인지 능력 비교

Comparison of Recognizing Emotions depending on Bandwidth-limitation of Digital Speech

지도교수 박 우 진

이 논문을 공학학사 학위논문으로 제출함

2022년 06월

서울대학교 공과대학

산업공학과

황 보 진 경

황보진경의 공학학사 학위논문을 인준함

2022년 06월

위 원 장 _____(인)

부위원장 _____(인)

위 원 _____(인)

초록

대화 중 상대방의 감정을 정확히 인식하는 것은 매우 중요하다. 최근에 코로나 바이러스의 대유행 및 메타버스의 출현은 디지털 의사소통을 가속화하였다. 음성이 디지털 신호로 전환되어 네트워크를 통해 송수신되는 과정에서 대역폭 제한은 빈번하게 발생한다. 본 연구에서는 대역폭 제한이 음성의 감정 전달에 어떤 영향을 미치는지 파악하고자 한다. 피실험자는 대역폭이 NB, WB, SWB, FB로 제한된 음성들을 듣고 발화자의 감정을 맞추는 실험을 수행하였다. 정답 여부와 응답 시간에 대해 통계적 분석을 수행한 결과, 음성의 주파수 대역폭 제한은 발화자의 감정의 전달에 통계적으로 유의미한 영향을 미치지 않았다. 그러나 발화자가 의도한 감정의 종류에 따라서는 전달력이 달라졌다. 따라서 음성의 컨텐츠를 전달하기에 충분한 대역폭은 음성의 반언어적인 부분도 전달하기에 충분함을 보였다. 또한, 발화자가 전달하고자 하는 감정의 종류에 따라 전달 효과를 높이기 위해서는 음성 외에도 다른 보조적인 채널이 필요하다.

주요어: 감정 인식, 대역폭 제한, 인간공학, 산업공학
학번: 2018-16729

목차

표 목차

그림 목차

제 1장 서론

1.1. 연구 목적

본 논문은 음성의 대역폭이 제한된 상황에서 음성에 담긴 감정의 전달력이 어떤 차이를 보이는지 파악하고자 한다. 음성의 중요한 요소들은 대부분 100~5,000Hz 주파수 대역에 존재하지만, 대역폭이 제한되는 상황에서 음성의 내용 외에 분위기, 감정 등의 요소가 전달되기에 충분한지에 대한 연구는 부족하다. 대역폭이 제한되는 상황은 디지털로 음성을 송수신하는 과정에서 종종 발생한다. 네트워크 및 통신 규격의 불균일성, 송수신 기기에서 사용하는 코덱 호환성 등으로 인해 음성 신호의 대역폭을 제한한 후, 송수신이 이루어진다. 본 논문에서는 한국어 음성에 대해 대역폭이 제한되는 다양한 상황에서 음성을 통한 감정의 전달력이 어떤 차이를 보이는지 알아보고자 한다. 본 논문의 연구 결과를 통해 대역폭 확장 기술의 필요성을 부각하고 오디오 방송, 오디오북, ASMR 등 다양한 음성 콘텐츠를 제공할 때, 전송 효율성과 감정 전달력 사이의 trade-off를 결정하는 데 도움이 될 것이라 기대한다.

1.2. 연구 동기 및 공헌

최근 오디오 방송, 오디오북, ASMR, 메타버스 등 음성을 활용한 콘텐츠들이 많은 인기를 얻고 있다. 음성 콘텐츠들은 텍스트와 달리 내용 전달의 측면 외에도 분위기, 감정 등의 반언어적 요소를 담고 있다. 음절의 빠르기, 높이, 억양 등에 따라 희로애락의 감정들이 담길 수 있고, 이에 따라 같은 텍스트이더라도 다른 의미로 전달될 수 있다.

스마트 기기 및 음향 기기들의 발전으로 뛰어난 품질의 소리를 재생할 수 있는 능력을 갖추었지만 네트워크 송수신 과정을 거치며 음질의 저하가 발생한다. 네트워크를 통해 음성이 송수신 되는 과정에서 오류를 줄이고 전송 효율을 높이기 위해서 코덱을 이용한 부호화·복호화 과정이 필수적이다. 이로 인해 코덱의 규격에 맞추어 음성의 대역폭이 제한되는 상황이 발생한다. 네트워크 송수신 과정에서는 이동통신 표준화기술 협력기구(3GPP)에서 표준으로 채택한 EVS 코덱을 사용하기 때문에 대역폭은 SWB(Super Wideband; ~16,000Hz)로 제한된다. 최근에는 무선 이어버즈를 사용하여 통화하는 상황이 증가하면서 다양한 송수신 과정을 거치는 음성 신호의 품질이 저하된다. 블루투스 통신의 경우 mSBC(modified SBC) 코덱을 표준으로 사용하기 때문에 음성 신호의 대역폭은 WB(Wide Band; 50~7000Hz)로 제한된다. 음성 콘텐츠들을 온전히 이용하기 위해서는 음질의 저하에도 불구하고 감정이나 분위기 등의 반언어적 요소의 전달력이 유지되는지 확인하는 것이 매우 중요하다.

본 연구는 음성 신호의 품질 저하가 감정의 전달에 미치는 영향을 파악한다. 음성에서 나타나는 억양과 감정의 표현 방식은 발화자의 문화와 사회를 반영하기 때문에 서구권 음성에 대한 연구 결과를 그대로 반영하는 것은 무리가 있다. 본 연구는 한국어의 감정 표현을 의미하는 음성에 대해 주파수 대역폭 제한이 미치는 영향을 분석한다는 의의가 있다.

코로나 바이러스와 그의 변이 바이러스들의 전염이 지속되면서 소통의 매개체가 대면에서 비대면으로 변화하였다. 이 외에도 메타버스의 출현, OTT 시장의 확대 등으로 인해 디지털 음성 신호를 청취하는 상황이 다분하다. 본 논문의 결과를 통해 콘텐츠 청취나 디지털 의사소통 상에서 대역폭 제한으로 인해 감정이 온전히 전달되고 있는지 점검할 수 있으며, 더 나아가 신호 처리의 효율성과 감정의 전달력 사이의 상충점을 제시하는 근거로 활용할 수 있을 것이라 기대한다.

1.3. 논문구성

본 논문은 5 장으로 구성된다. 제 2장에서는 배경이론을 살펴본다. 제 3장에서는 실험 과정을 설명한다. 제 4장에서는 실험 결과를 살펴본다. 마지막으로 제 5장에서는 결론과 향후 연구방향을 제시한다.

제 2장 배경 이론

2.1. 음성에 담긴 감정

감정이란 특정 현상에 대해 느끼는 기분이다. 의사소통 상황에서는 사람은 자신의 감정을 드러내기도, 타인의 감정을 받아들이기도 한다. 서로의 감정을 정확히 인식하는 것은 원활한 대화를 진행할 때 중요한 역할을 한다. 특히, 목소리는 감정과 관련된 다수의 정보를 포함한다. 이는 화자의 감정의 상태에 따라 발성과 관련된 생리학적 변수들이 변하기 때문이다. 예를 들어, 감정적 각성이 증가하면 심장 박동수, 혈류, 근육의 긴장 등의 변화를 매개할 뿐만 아니라 후두의 긴장과 성문하 압력이 증가하기 때문에 화자의 발성 강도를 증가시킨다.

감정 외에도 화자의 의도에 따라서도 목소리는 다르게 나타난다. 예를 들어, 분노한 화자의 목소리는 적에게 공포감을 주기 위해 거칠고 불쾌하게 들린다. 이처럼 발성 체계의 생리학적 조절이나 화자의 의도에 따라 말하는 방식이 달라지기 때문에 청자는 목소리만으로 화자의 감정 상태를 추론할 수 있다. (Schrirmer & Kotz, 2006)

Figure 1은 인간이 음성의 감정을 처리하는 과정을 모식화한 것이다. 먼저 청각 처리 영역을 통해 감각을 처리한다. 음향학적 정보를 다른 정보들과 통합 한 후, 인지의 과정을 거치게 된다. 이 때, 맥락적 혹은 개인적 중요도에 따라 각 과정을 촉진시킬 수 있다.

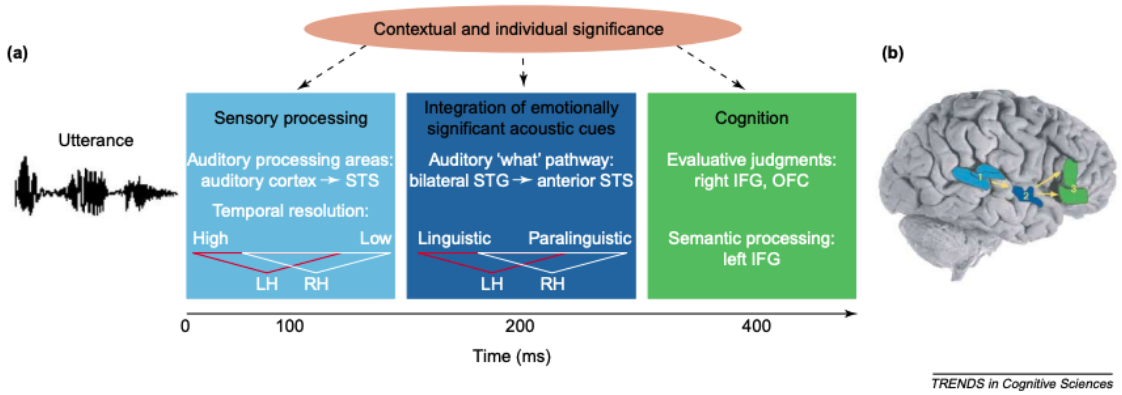

Figure 1 인간의 감정 처리 과정 (Schrirmer & Kotz, 2006)

2.2. 선행 연구

Gallardo et al.(2012)에서는 독일어 음성에 대해 주파수 대역폭을 NB와 WB로 제한하였을 때, 발화자 인식률을 비교하였다. 서로 잘 알고 있는 그룹 내에서 발화자들의 음성에 주파수 대역폭을 제한한 후 발화자의 신원을 맞추는 실험을 진행하였다. 정답률과 응답시간을 비교한 결과, NB로 주파수 대역폭을 제한한 경우 정답률이 낮고 응답시간이 길어졌다.

Labelle et al.(2016)에서는 영어 음성에 대해 6.6과 12.65kbps AMR-WB 코덱을 적용하였을 때, 청취자가 느끼는 감정의 인식률을 비교하였다. 총 다섯 가지 감정(중립, 슬픔, 기쁨, 분노, 공포)에 대해 각각 6개의 발화를 피험자들에게 들려준 후 객관식 형태의 테스트를 진행한 후 각 감정의 정답률을 원본 음성의 정답률과 비교하였다. 그 결과, 6.6kbps 코덱을 적용한 음성의 경우 슬픔을 제외한 모든 감정들은 99% 신뢰도로 청취자의 감정 인지 능력을 저하시켰으며, 12.65kbps 코덱의 경우 99%의 신뢰도로 중립과 행복을, 95%의 신뢰도로 슬픔과 분노를 인지하는 능력을 저하시켰다.

본 논문에서는 총 일곱 가지 감정(행복, 분노, 불쾌, 공포, 중립, 슬픔, 놀람)에 대해 코덱의 활용을 제외하고 음성의 주파수 대역(NB, WB, SWB, FB)에 따라 청취자의 감정 인지 능력의 변화를 알아보고자 한다.

제 3장 실험 과정

3.1. 실험 설명

본 실험은 한국어를 모국어로 사용하는 20대 남녀를 대상으로 진행하였다. 모든 피실험자들은 유선 이어폰을 착용하고, 듣기 편안한 정도의 볼륨 상태로 실험을 진행하였다.

주파수 대역의 제한에 따라 청취자의 감정 인지 능력의 변화를 탐구하고자 다음의 실험을 진행하였다. 피실험자들은 주어진 음성을 듣고, 발화자의 감정으로 가장 가까운 것을 Happiness(행복), Anger(분노), Disgust(불쾌), Fear(공포), Neutral(중립), Sadness(슬픔), Surprise(놀람) 중 하나 선택한다. 실제 대화 상황을 반영하기 위해 음성은 최대 두 번까지 재생할 수 있다. 피실험자들은 총 56개의 음성을 랜덤한 순서로 듣게 된다.

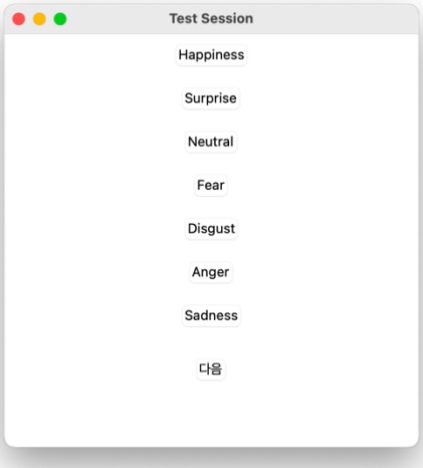

Figure 2 실험에서 사용되는 감정의 예시

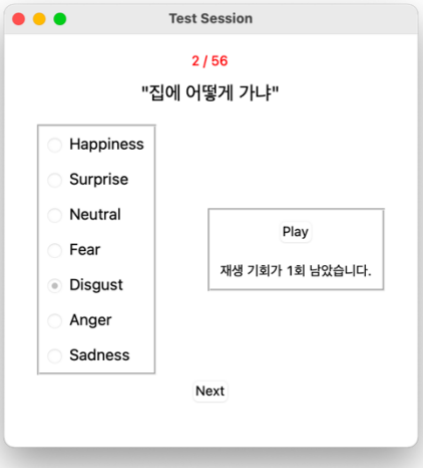

Figure 3 실험 세션의 GUI

먼저 본 실험에 앞서서 실험에서 사용하는 감정들을 익히고자 각 감정에 해당하는 음성의 예시를 들어볼 수 있다. 각 감정의 예시를 들어볼 수 있는 GUI(Graphic User Interface)는 Figure 2과 같다.

피실험자들이 실험 방법을 제대로 숙지할 수 있도록 돕기 위해 실제 실험 환경과 동일한 연습 세션을 3회 실시한다. 피실험자들은 연습 세션을 통해 GUI 화면의 구성, 조작

법 등에 충분히 적응한 뒤 본 실험을 실시하였다.

실험 세션의 GUI는 Figure 3와 같다. 화면 상단에는 실험의 진행도, 현재 음성의 대사가 표기되어 있다. 우측에는 Play 버튼과 남은 재생 기회가 표기되어 있다. 좌측에는 발화자의 감정을 선택할 수 있는 Radio button들이 나열되어 있다. 하단에는 다음 샘플로 넘어갈 수 있는 Next 버튼이 위치하고 있다.

3.2. 실험 변수

본 실험의 독립 변수는 음성의 대역폭 제한과 감정의 종류이다. 대역폭을 제한하는 최대 주파수는 총 네 가지로, NB(Narrow Band, ~4,000Hz), WB(Wide Band, ~8,000Hz), SWB(Super Wide Band, ~16,000Hz), FB(Full Band, ~24,000Hz)이다. 해당 주파수 대역은 EVS에서 제공하는 설계 기준을 따른 것이며 가장 대중적으로 사용하는 주파수 대역폭이다. Nyquist Frequency에 의해 디지털 신호는 그 신호에 포함된 가장 높은 진동수의 두 배에 해당하는 Samping rate를 가질 때, 완벽히 복원된다. 따라서 Sampling rate가 48,000Hz인 FB의 디지털 음성 신호의 Sampling rate를 각각 32,000Hz, 16,000Hz, 8,000Hz로 조절하여 독립변수에 해당하는 대역폭 제한을 구현하였다.

본 실험에서 사용하는 감정의 종류는 총 일곱 가지이다. 행복, 분노, 불쾌, 공포, 슬픔, 놀람은 감정 인식 분야에서 흔히 사용되는 분류로, 전세계인의 보편적인 감정으로 알려져 있다. (Ekman, 1994) 여기에 별다른 감정을 담지 않는 중립을 추가하였다. (박지인 et al., 2010)

본 실험에서 사용하는 종속 변수는 정답률과 응답 시간이다. 정답률이 높을수록 발화자의 감정이 명확히 전달됨을 의미한다. 그리고 응답 시간이 짧을수록 청취자가 발화자의 감정을 예측하는 데 필요한 mental load가 적다는 것을 의미한다.

독립 변수가 종속 변수에 미치는 영향을 알아보기 위해 이 외의 변수들을 통제하였다. 실험 환경은 조용하고 밀폐된 방이다. 모든 피실험자들은 Macbook Pro 13 inch (2018)에 연결한 SAMSUNG AKG C타입 이어폰을 착용하고 실험을 진행하였다.

3.3. 실험 데이터

실험 데이터는 KETI R&D에서 제공하는 감정 분류용 데이터셋[11]을 사용하였다. 이것은 100명의 연기 지망생 및 전문가들이 일곱 가지 감정에 대해 100번씩 발화 및 연기한 동영상 데이터셋이다. 각 감정별로 2명의 화자(남1, 여1)를 선정한 후, 16bit, 48000Hz 음성을 wav format으로 추출하여 사용하였다. 본 논문에서 사용한 음성들은 Table 1과 같다.

Table 1 실험에 사용한 음성 정보

감정	발화자	대사	맥락
행복	남	정말 믿어지지가 않네요	영화 시상식. 나의 첫 데뷔작으로 신인상을 수상 받게 되었다. 떨리는 마음을 가까스로 다독이며 마이크를 잡았다
	여	저 왔어요	오랜만의 연휴에 고향에 계신 부모님을 뵈러 내려왔다. 문이 열리고 언제나처럼 반갑게 맞이해 주시는 부모님의 얼굴엔 웃음꽃이 가득하다
분노	남	왜 이제 말해	친구와 만나기로 약속한 날. 만나기 30분 전에 오늘 못 나갈 것 같아 라고 전화가 왔다
	여	아 어디갔어	방을 온통 헤집으며 찾아도 필요한 물건이 나오질 않는다
불쾌	남	집에 어떻게 가냐	한창 퇴근시간으로 붐비는 지하철역. 저 멀리 힘겹게 다가오는 지하철 안에는 이미 발 디딜 틈 없이 사람들로 가득 차 있다
	여	인생 참 재밌게 사네	한 편의점 안. 일이 능숙해진 점원이 창밖을 보며 한숨을 쉬고 있다. 1주일에 꼭 한 번씩 와서 이상한 소리를 늘어놓고 잔소리를 하는 진상손님이 가게 문을 열고 있다
공포	남	야 어디갔어	친구와 배낭 하나 들고 멀리 떠나온 여행. 길가에서 텐트를 치고 자고 일어났는데 짐은 그대로고 친구만 감쪽같이 없어졌다
	여	빨리 집으로 와	늦은 저녁 반지하 원룸. 외출을 하려고 화장을 하는데 화장대 거울에 비친 창문 사이 낯 선 두 눈동자. 힘겹게 손을 뻗어 친구에게 전화를 건다
중립	남	안녕히 주무세요	잠자리에 들 저녁. 나는 옆방에 계신 부모님 방문 앞에서 말한다
	여	시청자 여러분 안녕하십니까	정각을 알리는 TV 소리와 함께 시작되는 뉴스. 무미건조한 앵커의 목소리가 들린다
슬픔	남	다들 기뻐보이네	시린 겨울. 너무 춥고 추운데 오가는 사람들은 모두 행복해 보인다

	여	갑자기 왜 그래	마주 보며 앉아있는 남녀. 차갑게 식은 커피보다 날 보는 눈빛이 차갑다. 나는 힘겹게 입을 뗀다
놀람	남	저거 다 몇 층이야	높은 건물들이 잔뜩 있는 잠실역. 그 중 독보적으로 하늘을 향해 우뚝 솟은 건물이 눈에 띈다
	여	아저씨 정신 좀 차려보세요	하루의 피로를 풀기 위해 찾은 찜질방. 아무도 없는 황토방 문을 열었는데 바닥에 쓰러져있는 한 사람. 황급히 다가선다

3.4. 실험 가설

피실험자들로부터 얻은 실험 결과를 바탕으로 아래 두 가지 가설의 통계적 유의성을 검정하고자 한다.

(a) 대역폭 제한의 최대 주파수 값이 낮아짐에 따라 청취자들의 감정 인지 능력은 저하될 것이다.

Labelle et al.(2016)의 연구 결과에 따르면 대역폭이 낮아지고 코덱을 적용함에 따라 청취자들의 정답률이 낮아졌다. 본 실험에서도 유사하게 최대 주파수 값과 감정 인지 능력이 비례하는 경향성을 얻을 수 있을 것이라 추측한다.

(b) 대역폭 제한과 감정의 종류에 따라 교호 작용이 발생할 것이다.

Labelle et al.(2016)의 연구 결과에 따르면 음성의 압축률과 코덱이 동일한 상황이더라도 감정의 종류에 따라 정답률이 차이를 보였다. 이는 장인창 et al.(2004)에 따르면 감정의 변화에 따라 모음 부분에서의 피치, 포먼트 등의 음성 정보 값이 변하기 때문으로 보인다. 본 실험에서도 대역폭을 제한함에 따라 음성에 포함된 감정 정보의 소실 정도가 다를 것이라 추측한다.

제 4장 실험 결과

피실험자들의 평균 나이는 22.9세 이며, 남자는 10명, 여자는 8명이다. 모든 피실험자들은 청력 및 정보 파악 등의 인지 능력에 이상이 없는 건강한 상태이다. 3.4에서 세운 가설의 성립여부를 보이기 위해 정답여부와 응답시간에 대해 통계적 검정을 실시하였다.

4.1. 실험 결과

Figure 4~Figure 7은 차례로 NB, WB, SWB, FB로 주파수 대역폭을 제한했을 때의 confusion matrix를 시각화한 것이다. 가로축은 피실험자들이 선택한 감정을 의미하고, 세로축은 발화자가 의도한 감정을 의미한다. 모든 주파수 대역폭을 통틀어 가장 많은 오답이 발생한 경우는 Disgust이다.

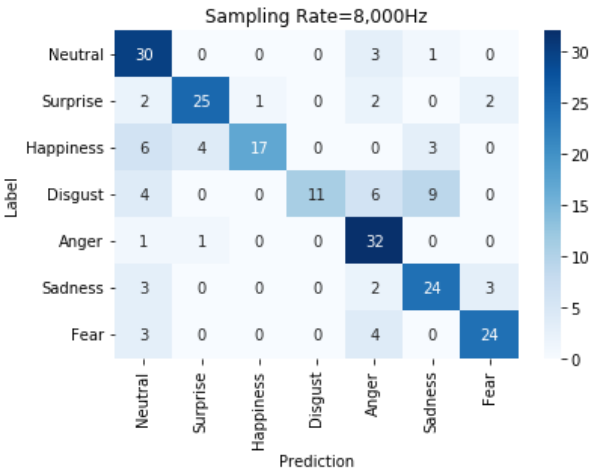

Figure 4 NB의 Confusion Matrix

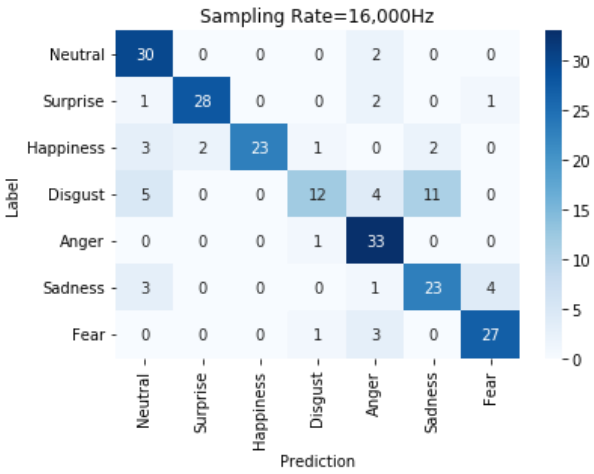

Figure 5 WB의 Confusion Matrix

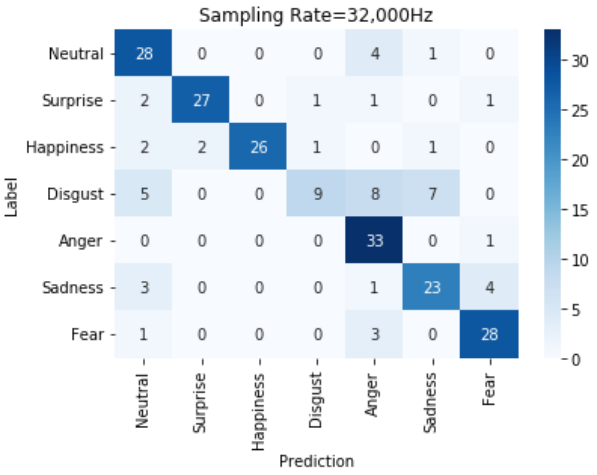

Figure 6 SWB의 Confusion Matrix

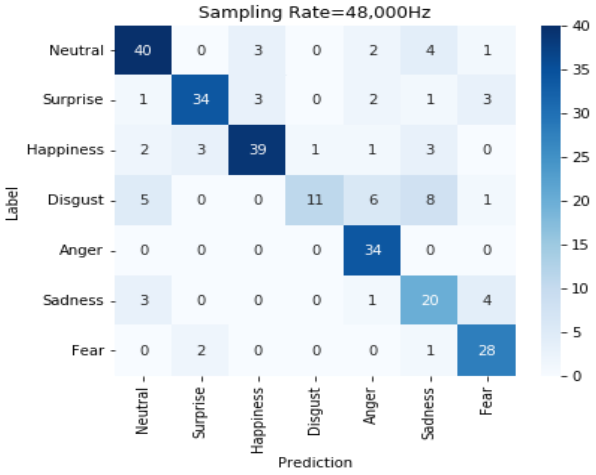

Figure 7 FB의 Confusion Matrix

Figure 8은 위 결과를 막대 그래프로 시각화한 것이다. x축은 발화자가 의도한 감정을 나타내고, y축은 피실험자들의 정답률을 의미한다. 각각의 색깔은 차례로 NB, WB, SWB, FB를 나타낸다. 회색 선은 95% 신뢰구간을 의미한다. 위와 마찬가지로, Disgust의 정답률이 가장 낮은 것을 확인할 수 있다.

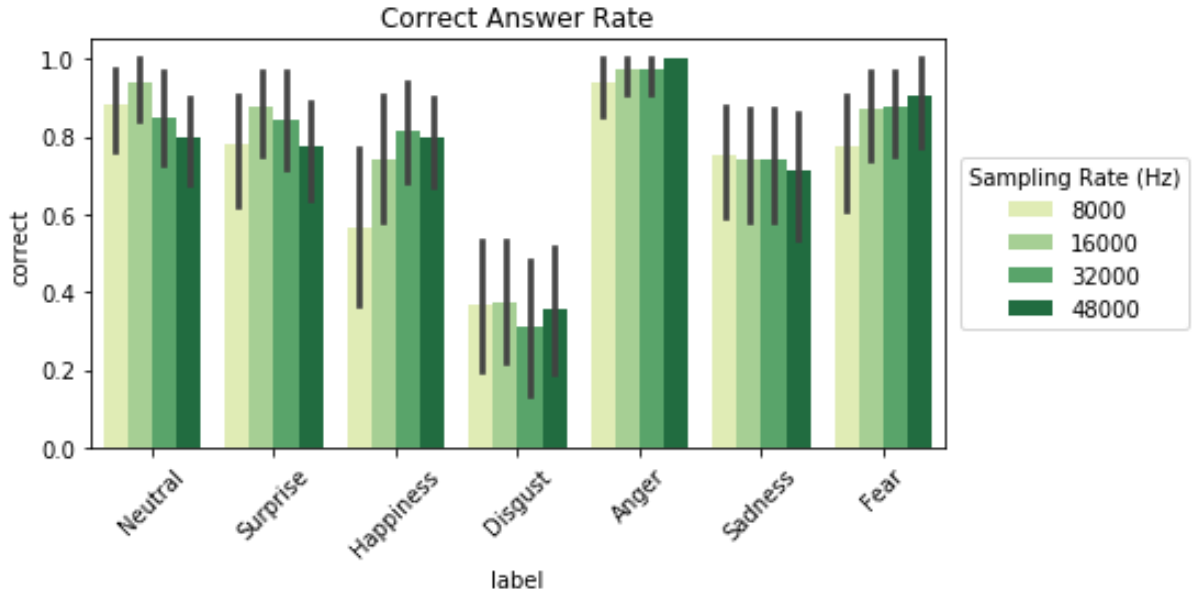

Figure 8 대역폭 제한과 감정의 종류에 따른 정답률

Figure 9는 대역폭 제한과 감정의 종류에 따른 피실험자의 응답시간을 막대 그래프로 시각화한 것이다. 응답시간은 발화자의 음성이 재생되기 시작하는 순간부터 피실험자가 Next 버튼을 누르는 순간 까지를 의미한다. x축은 발화자가 의도한 감정을 나타내고, y축

은 피실험자들의 응답시간의 평균값을 초로 나타낸 것이다. 위와 동일하게 각각의 색깔은 차례로 NB, WB, SWB, FB를 나타내고, 회색 선은 95% 신뢰구간을 의미한다.

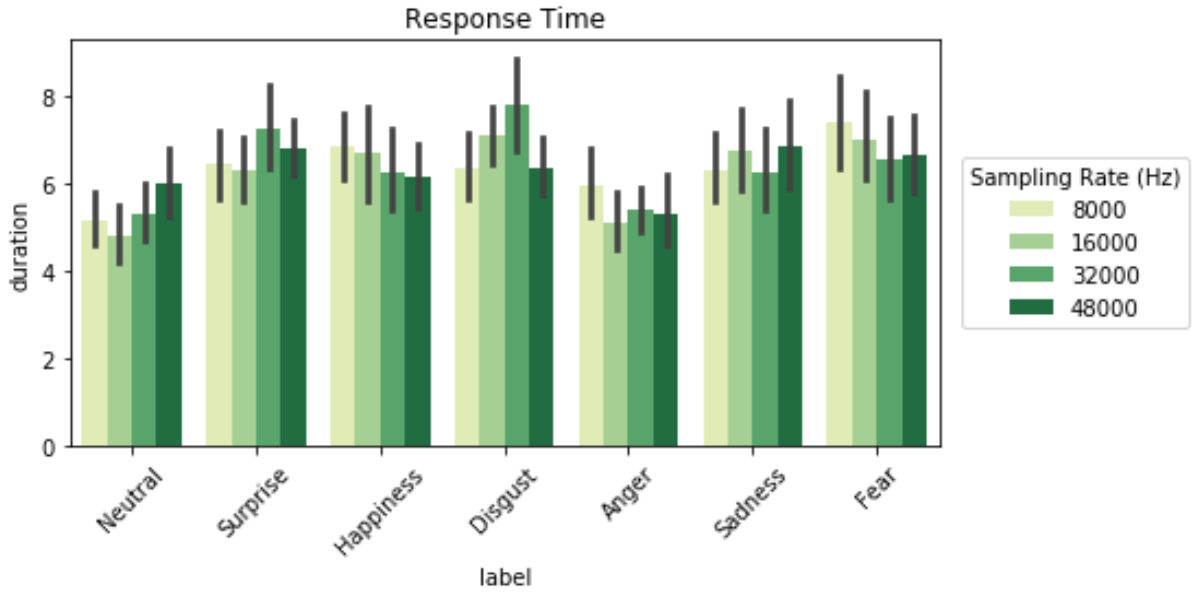

Figure 9 대역폭 제한과 감정의 종류에 따른 응답시간(단위: 초)

4.2. 통계적 분석

먼저, 정답 여부를 예측하는 다중 로지스틱 회귀분석(Multivariate Logistic Regression Analysis)을 수행하였다. 독립 변수는 대역폭 제한과 발화자가 의도한 감정이며, 종속 변수는 정답여부(0: 오답, 1: 정답)이다. 대역폭 제한의 경우 순서를 고려하며, 0과 1사이의 값을 가지도록 하기 위해 NB, WB, SWB, FB를 차례로 0, 0.25, 0.5, 1.0으로 매핑하였다. 발화자가 의도한 감정의 경우 범주형 변수이기 때문에 one-hot encoding을 진행하였다. 로지스틱 회귀분석을 수행한 결과는 Table 2와 같다.

Bandwidth의 회귀계수가 0에 가깝고 p값이 유의수준 0.05보다 크기 때문에 대역폭 제한은 정답 여부에 영향을 미치지 않는다고 할 수 있다. 즉, 3.4의 가설 (a)는 통계적으로 유의하지 않다. 음성의 대역폭 제한이 감정의 전달에 있어 유의미한 영향을 끼치지 않는다는 결론을 내릴 수 있다.

발화자가 의도한 감정의 경우 p값이 모두 유의수준 0.05보다 작으므로 정답 여부에 유의미한 영향을 미친다. 4.1의 실험 결과 그래프에서 확인하였듯이 Disgust의 회귀 계수가 음수이다. 이것은 발화자의 의도가 Disgust인 경우 청취자들의 감정 인식 정확도가 낮아짐을 의미한다. Disgust라는 감정의 경우 음성만으로 구분하는 것은 다소 어려운 과업임을 짐작할 수 있다.

Anger의 경우 가장 큰 회귀계수를 가지고 있으며, Odds ratio는 31.1563으로 종속 변수와의 관계가 매우 강하다. 발화자의 의도가 Anger라는 감정인 경우, 청취자들의 감정 인식 정확도가 매우 높음을 의미한다. **Error! Reference source not found.**에서 언급했듯이 Anger의 경우 감정적 각성 및 발화자의 의도에 따라 발성이 달라지기 때문에 음성을 통한 감정의 인식이 잘 되는 것으로 보인다.

Table 2 정답 여부에 대한 로지스틱 회귀분석 결과

	Coef.	Std. Err.	p-value	Odds ratio
Bandwidth	0.1340	0.2140	0.5311	1.1434
Anger	3.4390	0.5154	0.0000	31.1563
Disgust	-0.6331	0.2022	0.0017	0.5309
Fear	1.7600	0.2635	0.0000	5.8123
Happiness	0.8090	0.2063	0.0001	2.2457
Neutral	1.7181	0.2528	0.0000	5.5740
Sadness	0.8530	0.2108	0.0001	2.3468

Surprise	1.4743	0.2363	0.0000	4.3681

다음으로, 응답시간에 대한 이원분산분석(two-way ANOVA)을 수행하였다. 대역폭 제한의 p-value는 유의수준 0.05보다 크기 때문에 귀무가설을 기각할 수 없다. 따라서 대역폭 제한에 따른 응답 시간의 차이는 없다. 3.4의 가설 (a)는 통계적으로 유의하지 않다. 음성의 대역폭 제한이 감정의 인식에 있어 유의미한 영향을 끼치지 않는다는 결론을 내릴 수 있다.

발화자가 의도하는 감정의 p-value는 유의수준 0.05보다 작기 때문에 귀무가설을 기각할 수 있다. 따라서 발화자가 의도하는 감정에 따라 응답시간의 차이가 있다.

대역폭 제한과 발화자가 의도하는 감정 간의 교호작용의 경우 p-value가 0.05보다 크기 때문에 귀무가설을 기각할 수 없다. 따라서 유의한 교호작용은 발생하지 않는다. 3.4의 가설 (b)는 통계적으로 유의하지 않다. 발화자가 의도한 감정과 대역폭 제한의 조합에 따라 다른 경향성이 나타나지 않는다.

	Df	**Sum of Squared**	**Mean of Squared**	**p-value**
Bandwidth	3	64.6370	21.5457	0.4895
Label	6	1280.6821	213.4470	0.0000
Bandwidth:Label	18	570.7862	31.7103	0.2621
Residual	908	25995.4771	26.6620	-

제 5장 결론

5.1. 결론

본 논문은 디지털 음성의 대역폭 제한이 발화자의 감정의 전달에 미치는 영향을 확인하기 위해 NB, WB, SWB, FB 대역폭의 음성에 대해 감정 인지 능력의 차이를 검정하였다. 피실험자들이 음성을 듣고 발화자의 감정으로 추정한 응답의 정답률과 응답 시간을 측정하였으며, 음성의 주파수 대역폭 제한은 발화자의 감정의 전달에 통계적으로 유의미한 영향을 미치지 않았다.

반면에 발화자가 의도한 감정의 종류에 따라 전달력이 달랐다. Disgust의 경우 정답 여부에 부정적 영향을 끼치는 것이 통계적으로 유의하였으며, 응답 시간도 유의미하게 길었다. 즉, 발화자가 Disgust라는 감정을 음성만으로 전달하는 것은 쉽지 않으며, 청취자도 이 감정을 구분하는 과업을 수행할 때 피로함을 느낄 수 있다. 반면에 Anger의 경우 음성의 크기, 높낮이 등의 발성 방법이 다른 감정들과 확연히 다르기 때문에 정답률이 가장 높고, 응답 시간 또한 짧았다. 따라서 발화자가 전달하고자 하는 감정의 종류에 따라 전달력을 높이기 위해서는 반언어적 표현인 음성 외에도 몸짓이나 표정 등 비언어적 표현을 함께 전달할 수 있는 Multimodal 방법을 사용할 필요가 있다.

5.2. 토의

본 연구의 결과는 선행 연구들을 바탕으로 세운 가설에 상반된다. 기존의 선행 연구들은 서구권 음성에 대해 수행한 결과인 반면, 본 연구는 한국어 음성에 대해 수행한 결과이다. 영어권 음성의 경우 /s/, /z/와 같은 마찰음이나 /f/, /S/, /Z/의 경우 NB 주파수 대역폭 만으로는 완벽히 표현하기 힘들다. 에너지의 상당 부분이 고주파수 대역에 위치하기 때문에 저주파 대역만으로는 위 소리들을 구분하는 것이 혼동될 수 있다. (Gajjar et al., 2012) 반면 한국어 음성에서는 마찰음 사용이 영어권 음성보다는 적기 때문에 대역폭 제한에 의한 영향이 적은 것으로 보인다. 또한, 코덱의 유무도 실험 결과에 영향을 미쳤을 것이다. 기존의 선행 연구들은 대역폭 제한 뿐만 아니라 코덱을 사용했기 때문에, 코덱을 부호화·복호화하는 과정 중에 음성이 일부 손실되거나 의도치 않은 아티팩트가 발생하는 등 음성의 품질에 영향을 미쳤다. 반면에 본 연구에서는 대역폭 제한만을 독립 변수로 설정하고, 이 외의 요인들을 통제했기 때문에 코덱에 의한 영향이 포함되지 않았다.

이 외에도 본 연구에는 몇 가지 실험상의 한계가 있다. 먼저, 피실험자 집단이 적고 편향되었다. 충분히 다양한 피실험자들을 확보하지 못하였고, 피실험자의 피로도를 고려하여 10분 내의 실험을 구상하기 위해 샘플의 수를 상당히 제한하였다. 둘째로, 데이터의 한계가 있다. 발화자의 감정 표현 능력이 충분하지 못했을 가능성이 있다. 그리고 발화 이전의 맥락에 대한 정보를 제공하지 않았고, 중립적인 텍스트를 선별하였기 때문에 감정을 명확히 구분하는 데 어려움이 있을 수 있다.

선행연구와 본 연구의 결과 사이의 차이를 바탕으로 반언어적인 부분의 전달에 영향을 미치는 요인으로는 노이즈를 지목할 수 있을 것이다. 복수의 화자가 존재하는 공간에서 그 화자들이 동시에 발화하는 경우, 인간의 청각 기관에는 모든 소리가 혼합되어 들리지만 인간은 자신이 집중하고자 하는 화자의 소리만을 들을 수 있다. 이는 선택적 주의의 결과로, 칵테일 파티 효과라고 칭한다.(이보원, 2015) 배경 소음이 존재하는 상황에서 사람은 배경 소음을 없애는 데 선택적 주의를 사용하기 때문에 감정을 인식하는 데에 많은 주의를 사용할 수 없을 것이다. 향후 과제로 SNR(Signal Noise Ratio)에 따라 반언어적 부분의 전달력이 어떤 차이를 보이는지 검정하는 연구가 진행되기를 기대한다.

참고문헌

[1] 박지인, 강지인, 조현상, 이 은 & 안석균, 2010, 한국인에서의 Ekman 얼굴 표정 사진의 신뢰도와 타당도, 대한우울·조울병학회, 8(2), pp.145-151.

[2] 오세진, 2022, 한국인과 중국인 고급 학습자의 감정 발화에 대한 음성학적 특성 비교, PhD Thesis, 한양대학교.

[3] 이보원, 2015, 칵테일 파티 효과 – 복수 화자의 검출 및 지역화 방법. 한국소통학회 2015 가을철 정기학술대회, pp.200-210.

[4] 장인창, 박미경, 김태수, & 박면웅, 2004, 감정변화에 따른 음성정보 분석에 관한 연구, 한국정밀공학회 학술발표대회 논문집, pp.25-28.

[5] Ekman, P, 1994, *Strong evidence for universals in facial expressions*: a reply to Russell's mistaken critique.

[6] Gajjar, P., Bhatt, N., & Kosta, Y, 2012, *Artificial bandwidth extension of speech & its applications in wireless communication systems: A review*, In 2012 International Conference on Communication Systems and Network Technologies, IEEE, pp. 563-568.

[7] Gallardo, L. F., Möller, S., & Wagner, M., 2012, *Comparison of human speaker identification of known voices transmitted through narrowband and wideband communication systems*, ITG Symposium, VDE, pp.1-4.

[8] Labelle, F., Lefebvre, R., & Gournay, P, 2016, *A subjective evaluation of the effects of speech coding on the perception of emotions*, ISPACS, IEEE, pp.1-6.

[9] Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A., 200, *Recognizing emotions in a foreign language,* Journal of Nonverbal Behavior, 33(2), pp.107-120.

[10] Schirmer, A., & Kotz, S. A., 2006, *Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing*, Trends in cognitive sciences, 10(1), 24-30.

[11] https://aihub.or.kr/opendata/keti-data/recognition-visual/KETI-01-001

Abstract

Comparison of Recognizing Emotions depending on Bandwidth-limitation of Digital Speech

Jinkyoung Hwangbo
Department of Industrial Engineering
College of Engineering
Seoul National University

It is crucial to recognize someone's emotion while engaging in conversation. Recently, digital communication has increased due to the outbreak of COVID-19 and the emergence of metaverse. When the speech is transmitting through networks, the quality of the speech is degraded because of bandwidth limitation and multiple codec artifacts. This paper is intended to identify how bandwidth limitation affects emotions in one's speech. The human subjects choose the speaker's emotions after listening the speech of which bandwidth is limited to NB, WB, SWB, and FB. As a result, not the bandwidth limitation of the speech, but the type of emotion the speaker intended has a statistically significant effect on recognizing speaker's emotion. Therefore, if it is enough to convey the linguistic part of the speech, it is also enough to convey the semi-verbal part of the speech. Also, auxiliary channels are needed to convey the speaker's emotion effectively depending on the type of emotion intended.

Keywords: Emotion Recognition, Bandwidth-Limitation, Ergonomics, Industrial engineering
Student Number: 2018-16729