Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 6, pp. 549-559, December 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.6.549

Consistency Regularization을 적용한 멀티모달 한국어 감정인식

김정희 · 강필성†

고려대학교 산업경영공학부

Multi-modal Korean Emotion Recognition with Consistency Regularization

Jounghee Kim · Pilsung Kang

School of Industrial & Management Engineering, Korea University

Recently, the demand for artificial intelligence-based voice services, identifying and appropriately responding to user needs based on voice, is increasing. In particular, technology for recognizing emotions, which is non-verbal information of human voice, is receiving significant attention to improve the quality of voice services. Therefore, speech emotion recognition models based on deep learning is actively studied with rich English data, and a multi-modal emotion recognition framework with a speech recognition module has been proposed to utilize both voice and text information. However, the framework with speech recognition module has a disadvantage in an actual environment where ambient noise exists. The performance of the framework decreases along with the decrease of the speech recognition rate. In addition, it is challenging to apply deep learning-based models to Korean emotion recognition because, unlike English, emotion data is not abundant. To address the drawback of the framework, we propose a consistency regularization learning methodology that can reflect the difference between the content of speech and the text extracted from the speech recognition module in the model. We also adapt pre-trained models with self-supervised way such as Wav2vec 2.0 and HanBERT to the framework, considering limited Korean emotion data. Our experimental results show that the framework with pre-trained models yields better performance than a model trained with only speech on Korean multi-modal emotion dataset. The proposed learning methodology can minimize the performance degradation with poor performing speech recognition modules.

Keywords: Speech Emotion Recognition, Wav2vec 2.0, Multi-Modal Emotion Recognition

1. 서 론

최근 다양한 분야에서 사용자의 요구를 판별하고 이에 적절하게 응답할 수 있는 인공지능 기반 시스템에 대한 관심이 높아지고 있다(Pantic *et al.*, 2005). 특히 음성 인공지능 서비스는 한 번의 발화로 다양한 기능을 빠르게 수행할 수 있는 장점과 스마트 단말기의 폭발적인 공급으로 인해 그 수요가 증가하고 있는 추세이다(Rawat *et al.*, 2014). 현재 음성 인공지능 서비스는 항공편 예약, 텔레뱅킹(Kim *et al.*, 2015), 콜마케팅, AI스피커 등 다양한 분야에서 상용화 되었으며, 헤드업 디스플레이 등으로 확장되고 있다. 한편, 음성에는 언어적 정보와 함께 비언어적 정보가 포함되어 있으므로 사용자의 의도를 정확하게 파악하고 음성 서비스의 품질을 향상시키기 위하여 비언어적 정보인 감정을 인식하는 것이 매우 중요하다(Karlgren *et al.*,

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2021-0-00034, 파편화된 데이터의 적극 활용을 위한 시계열 기반 통합 플랫폼 기술 개발). 또한 이 연구는 한국산업기술진흥원의 산업인공지능인력양성사업의 지원을 받아 수행되었음(P0008691).

† 연락저자 : 강필성 교수, 02841 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3383, Fax : 02-929-5888,
E-mail: pilsung_kang@korea.ac.kr

2021년 9월 29일 접수; 2021년 11월 11일 게재 확정.

2012). 따라서 음성에 담긴 사용자의 감정을 파악하는 음성 감정인식(speech emotion recognition)이 활발히 연구되고 있다.

과거에는 음성에서 톤(tone), 발음속도(speech speed), 포만트 주파수(formant frequency) 등의 특징을 통계적으로 분석하여 감정을 인식하였다(Kwon *et al.*, 2003). 최근에는 데이터 처리 및 딥러닝 기술이 발달하면서 인공 신경망을 활용한 음성 감정인식 모델들(Fayek *et al.*, 2017; Han *et al.*, 2014)이 개발되었다. 또한, 음성이라는 단일 정보에만 의존하지 않고 텍스트 정보를 함께 활용하기 위해 음성인식 모듈을 적용한 감정인식 프레임워크(McDuff *et al.*, 2019; Xu *et al.*, 2019; Yoon *et al.*, 2018)가 제안되었다. 감정인식 프레임워크는 음성만 사용 가능한 환경에서 음성인식 모듈을 활용하여 음성에서 텍스트를 추출한다. 그리고 음성과 추출한 텍스트를 딥러닝 기반 멀티모달(multi-modal) 모델의 입력으로 함께 활용함으로써 감정인식률을 비약적으로 향상시켰다. <Figure 1>은 음성을 활용하여 감정을 인식하는 감정인식 프레임워크의 흐름을 나타낸 예시이다.

음성인식 모듈을 적용한 감정인식 프레임워크는 음성만 활용 가능한 환경에서 감정 인식률을 향상 시켰지만 음성에서 발화된 대화내용과 다른 텍스트를 입력으로 활용하면 감정인식 성능이 하락하는 단점을 갖고 있다(Yoon *et al.*, 2018). 일반적으로 음성에는 주변 소음 및 불특정 화자의 발음이 포함되어 있으므로 음성인식 모듈이 음성에서 대화내용을 온전히 추출하지 못하는 경우가 존재한다. 따라서 실제 사용 환경에서 감정인식 프레임워크는 음성 인식률이 하락할 뿐만 아니라 감정인식의 성능도 하락하는 상황이 빈번히 발생한다. 또한, 감정인식 프레임워크의 딥러닝 기반 멀티모달 모델은 방대한 레이블 데이터가 주어졌을 때 성능 개선이 가능하지만 공개된 한국어 감정인식 데이터는 모델을 학습시키기에 충분하지 않다. 다량의 감정인식 데이터를 생성 및 수집하는 것은 많은 비용이 소요되므로 제한된 한국어 데이터를 활용하여 딥러닝 기반 감정인식 프레임워크를 개발하는 것은 한계가 있다. 따라서, 본 연구는 한국어 감정인식의 성능을 향상시키기 위하여 선행연구의 단점을 보완한 멀티모달 감정인식 프레임워크를 제안한다.

본 연구에서 제안하는 멀티모달 감정인식 프레임워크의 특징은 크게 두 가지이다. 첫째, 주변 소음이 존재하는 실제 환경에서 음성인식 모듈이 대화내용을 온전히 추출하지 못하는 것을 고려하여, 음성의 대화내용과 음성인식 모듈을 통해 추출된 텍스트의 차이를 모델에 반영할 수 있는 Consistency Regularization 학습 방법론을 제안한다. Consistency Regularization 학습 방법론은 음성인식 모듈의 오탈자를 멀티모달 모델이 자체적으로 보정하여 인식하게 함으로써, 음성 인식률의 하락에 따른 감정인식 성능 하락을 최소화한다. 둘째, 데이터가 부족한 환경에서 감정인식 성능을 향상시키기 위하여 레이블이 없는 대량의 데이터에서 패턴을 추출하고 사전 학습이 가능한 자기지도학습 모델을 감정인식 프레임워크에 적용한다. 음성 자기지도학습 모델로는 최근 음성인식 분야에서 활발히 연구되고 있는 Wav2vec 2.0(Baevski *et al.*, 2020)을 활용하고, 텍스트 자기지도학습 모델로는 BERT 방법론(Devlin *et al.*, 2018)을 한국어 데이터에 적용한 HanBERT(Park, 2018)를 활용한다. 본 연구에서는 자기지도학습 모델을 감정인식 프레임워크에 효과적으로 적용하기 위하여 선행 연구의 다양한 멀티모달 융합 아키텍처를 적용한다.

본 연구에서 제안하는 방법론을 통하여 음성만 활용 가능한 환경에서 감정인식 프레임워크를 적용하여 한국어 감정인식 성능을 향상시킬 수 있음을 보였다. 총 여덟 가지의 감정 범주가 존재하는 한국어 영상 데이터셋을 사용하여 비교 실험을 수행한 결과, 본 연구에서 제안하는 감정인식 프레임워크를 적용한 멀티모달 모델이 음성만을 활용한 모델보다 높은 감정인식 성능을 기록하였다. 추가로, 성능이 저조한 외부 음성인식 모듈을 활용한 감정인식 프레임워크에 제안한 학습 방법론을 적용하여 실제 상황에서 불완전한 음성 인식 때문에 하락하는 감정인식 성능 중 상당 부분을 회복할 수 있음을 확인하였다. 마지막으로 자기지도학습(self-supervised learning) 모델인 Wav2vec 2.0과 HanBERT를 감정인식 프레임워크에 적용한 결과 감정 인식률이 향상됨을 확인하였다.

본 논문의 구성은 다음과 같다. 먼저 제2장에서는 음성인식 모듈을 적용한 감정인식 프레임워크와 감정인식에 활용된 자기지도학습 모델 및 융합 아키텍처에 관한 선행 연구를 소개한다. 이후 제3장에서는 제안하는 멀티모달 감정인식 프레임

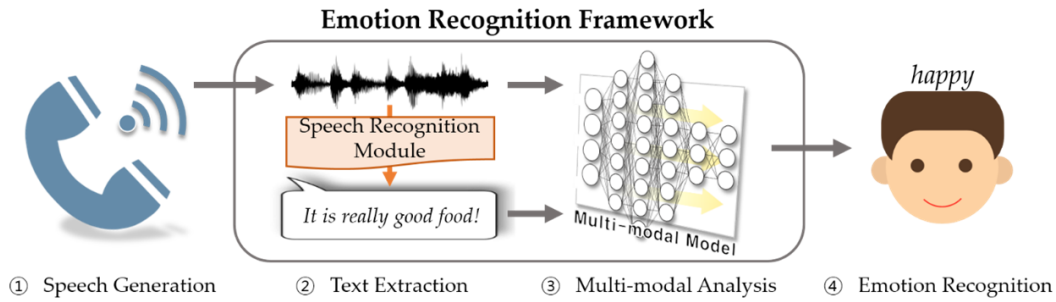

Figure 1. Multi-modal Emotion Recognition Framework

워크를 소개하면서 추가적으로 Consistency Regularization 학습 방법론 및 멀티모달 융합 아키텍처에 대해 설명한다. 제4장에서는 수행한 실험의 설계 및 결과를 소개한다. 마지막으로 제5장에서는 본 연구의 결론과 추후 연구 방향을 서술한다.

2. 관련 연구

2.1 음성인식 모듈을 적용한 감정인식 프레임워크

음성 감정인식은 음성신호의 전반적인 특징이나 Mel Frequency Cepstral Coefficient(MFCC)와 같은 스펙트럼 특징을 활용하여 감정을 분류하는 기술이다. 음성신호의 특징을 입력으로 하여 서포트 벡터 머신(Seehapoch *et al.*, 2013), 히든 마코브 모델(Nwe *et al.*, 2003) 등의 머신러닝 알고리즘이 음성 감정인식 초기 연구에 활용되었다. 최근에는 자연어처리, 컴퓨터비전과 같은 다양한 분야에서 높은 분류 성능을 가진 딥러닝 아키텍처(Fayek *et al.*, 2017; Han *et al.*, 2014)를 활용하여 음성 감정인식 성능을 개선하고자 하는 연구가 이루어지고 있다. 한편, 사람은 다양한 형태로 감정을 표현하므로 이에 영감을 받아 음성과 다른 형태의 정보를 함께 활용하여 감정을 인식하는 멀티모달 모델(Paleari *et al.*, 2010)이 제안되어 감정인식 성능을 크게 향상시켰다. 따라서 음성만 활용 가능한 환경에서 다른 형태의 정보를 추출하고 감정인식에 함께 활용하기 위하여 음성인식 모듈을 적용한 감정인식 프레임워크(Xu *et al.*, 2019; Yoon *et al.*, 2018)가 제안되었다. 감정인식 프레임워크는 음성인식 모듈을 활용하여 음성으로부터 대화내용을 텍스트 형태로 추출하고, 음성과 추출한 텍스트를 멀티모달 모델의 입력으로 활용하여 감정인식 성능을 향상시키는 전략을 사용한다. 예를 들어 구글 음성인식 모듈과 순환 신경망(recurrent neural network)을 적용한 감정인식 프레임워크(Yoon *et al.*, 2018)는 음성만을 활용한 모델보다 우수한 감정인식 성능을 기록하였으며, 음성인식 모듈에서 추출한 텍스트와 음성에서 강조된 감정을 추출하기 위해 오텐션(attention) 메커니즘을 추가한 감정인식 프레임워크(Xu *et al.*, 2019)는 감정인식 성능을 더욱 향상시켰다. 그러나 선행 연구의 감정인식 프레임워크는 음성인식 모듈의 성능에 의존적이기 때문에 주변 소음이 포함된 실제 환경에서 음성인식의 성능이 하락하면 감정인식의 성능도 하락하는 단점을 갖고 있다. 따라서 본 연구는 음성인식 모듈에서 추출된 텍스트와 실제 대화내용의 차이를 반영할 수 있는 Consistency Regularization 학습 방법론을 적용하여 선행연구의 단점을 보완한다.

2.2 자기지도학습 모델을 활용한 감정인식

자기지도학습 방법론은 레이블이 없는 대량의 데이터를 활용하여 모델을 사전 학습(pre-training)하고 과업 데이터로 미세 조정(fine-tuning)하여 모델의 예측성능을 향상시키는 연구분야이다. 자기지도 학습 방법론은 레이블이 없는 대량의 데이터에서 패턴 및 특징을 학습할 수 있으므로 과업 데이터가 충분하지 않을 때 더 효과적이다. 최근 다양한 자기지도학습 모델이 공개되었으며 자연어처리(Devlin *et al.*, 2018; Liu *et al.*, 2019), 컴퓨터 비전(Chen *et al.*, 2020; Kolesnikov *et al.*, 2019), 음성인식(Baevski *et al.*, 2019; Baevski *et al.*, 2020) 등 다양한 분야에서 모델의 예측성능을 향상시키는 것으로 나타났다. 또한, 모델을 학습하기에 데이터의 양이 충분하지 않은 감정인식 분야에서도 많은 연구들이 자기지도학습 모델을 감정인식에 활용하였다(Pepino *et al.*, 2021; Siriwardhana *et al.*, 2020a; Siriwardhana *et al.*, 2020b). 예를 들어, 다량의 영어 음성으로 사전 학습된 Wav2vec 2.0 모델(Baevski *et al.*, 2020)을 활용하여 음성으로부터 벡터를 추출하고, 추출된 벡터로 감정을 분류하는 감정인식 모델(Pepino *et al.*, 2021)이 제안되었다. 또한, 음성 자기지도학습 모델인 Wav2vec(Schneider *et al.*, 2019)과 텍스트 자기지도학습 모델인 RoBERTa(Liu *et al.*, 2019)를 활용하여 각각 음성과 텍스트의 특징을 추출하고, 멀티모달 모델의 입력으로 활용하는 방법(Siriwardhana *et al.*, 2020a)이 제안되어 감정인식 성능을 향상시켰다. 따라서, 본 연구는 한국어 감정인식 데이터가 충분하지 않은 환경을 고려하여 대량의 데이터로 사전 학습된 자기지도학습 모델을 감정인식 프레임워크에 적용한다. 음성 자기지도학습 모델로는 Wav2vec 2.0을 활용하고 텍스트 자기지도학습 모델로는 한국어 데이터로 사전 학습한 BERT 모델인 HanBERT(Park, 2018)를 활용한다.

2.3 멀티모달 융합 아키텍처

멀티모달 모델의 목적은 두 가지 이상의 형태를 갖고 있는 데이터를 함께 활용하여 한 가지 형태의 데이터만 사용하는 모델보다 우수한 정확도를 달성하는 것이다. 감정 인식 분야에서 멀티모달 모델은 사용되는 음성, 이미지, 텍스트 등 상이한 형태의 데이터의 길이가 서로 다르기 때문에 시점을 일치시킨 후 감정을 인식하거나 시간적 정보가 없는 데이터를 융합할 수 있는 멀티모달 아키텍처가 연구되었다. 전자의 경우, Gu *et al.*(2018)은 음성과 텍스트의 시간적 정보가 정렬된 데이터를 활용하여 어텐션 메커니즘을 적용한 계층적 융합 아키텍처를 제안하였다. Tsai *et al.*(2018) 연구에서는 시간 정보가 포함되어 있는 음성, 이미지, 텍스트 데이터에서 특정 시간대에 공통으로 나타난 특징정보를 추출하고, 추출된 특징벡터를 활용하여 감정을 인식하는 방법론을 제안하였다. 후자의 경우, Majumder *et al.*(2018)은 게이트 순환 유닛(gated recurrent unit; GRU)을 활용하여 서로 다른 길이의 데이터를 일정 길이의 특징 벡터로 압축한 후 정보를 계층적으로 융합할 수 있는 멀티모달 아키텍처를 제안하였다. Tsai *et al.*(2019)은 시간 정보가 정렬되지 않은 음성, 이미지, 텍스트 데이터에서 공통적으로 강조하는 감정 정보를 추출하기 위해 Cross-modal Transformer 융합 아키텍처를 제안하여 감정인식 성능을 향상시켰다. 본 연구의 감정인식 프레임워크는 음성인식 모듈을 활용하

여 텍스트를 추출하므로 음성과 텍스트의 시점 정보가 없다. 따라서 시점 정보 없이 음성과 텍스트 정보를 융합하기 위하여 다양한 선행 연구의 멀티모달 융합 아키텍처를 감정인식 프레임워크에 변형하여 적용하였다.

3. 제안 방법론

본 연구에서 제안하는 감정인식 프레임워크를 활용하여 감정을 추론하는 전반적인 구조는 <Figure 2>와 같다. 감정을 추론하는 과정은 프레임워크의 음성인식 모듈 사용여부에 따라 speech-to-text(STT) path와 Golden Path로 나뉜다. 실제 사용 환경에서는 음성만 활용 가능하므로 STT path를 활용하여 음성으로부터 대화내용을 추출하고 감정을 추론한다. 반면, 학습 환경에서는 멀티모달 데이터로부터 음성과 함께 대화내용이 텍스트 형태로 제공되므로 STT path뿐만 아니라 음성인식 모듈 없이 Golden Path를 통해 감정을 추론할 수 있다. 본 연구에서는 STT path와 Golden Path를 통해 추론한 두 확률 분포의 차이를 활용하여 멀티모델 모델을 학습하는 Consistency Regularization 학습 방법론을 제안한다. 또한, 자기지도학습 모델과 다양한 형태의 멀티모달 융합 아키텍처가 적용된 멀티모달 모델을 제안한다.

3.1 멀티모달 감정인식 프레임워크

음성인식 모듈을 활용하여 감정을 인식하는 멀티모달 감정인식 프레임워크의 상세 구조는 <Figure 3>과 같다. 먼저 음성인식 모듈을 활용하여 음성의 대화내용을 텍스트 형태로 추출한다. 이후 음성과 텍스트를 해당 자기지도학습 모델의 입력변수로 활용하여 각 정보를 일련의 벡터로 압축한다. 그리고 융합 아키텍처를 활용하여 음성 정보와 텍스트 정보를 융합하고 최종적인 감정 추정이 이루어진다. 제안 프레임워크에 대한 세부적인 절차는 다음과 같다. 입력된 일련의 음성신호를 $a = a_1, ..., a_T$ 라 할 때, 제안 프레임워크는 외부 음성인식 모듈 G을 활용하여 음성의 대화내용인 텍스트 $s = G(a)$를 추출한다. 이후 텍스트 자기지도학습 모델과 함께 학습된 토크나이저(tokenizer)를 활용하여 추출된 텍스트를 토큰 형태로 분할하면 $s = s_1, ..., s_L$과 같이 표현할 수 있다. 일반적으로 음성신호와 텍스트 토큰은 서로 다른 길이를 갖고 있다($T \neq L \in N^+$). 음성신호는 음성 자기지도학습 모델 E_{audio}의 입력 변수로 활용되어 일련의 벡터 $E_{audio}(a) = h^a = h_1^a, ..., h_n^a$로 압축된다. 본 연구에서 활용하는 음성 자기지도학습 모델인 Wav2vec 2.0은 음성신호에 있는 복잡한 패턴을 추출할 수 있도록 대량의 음성 데이터로 사전 학습되었으므로 압축된 음성벡터에는 음성의 중요정보가 포함되어 있다. 한편, 텍스트 토큰은 텍스트 자기

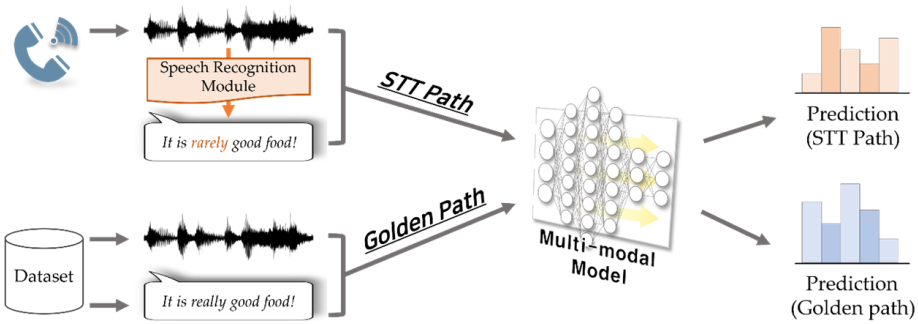

Figure 2. Emotion Recognition Processes with Multi-modal Framework

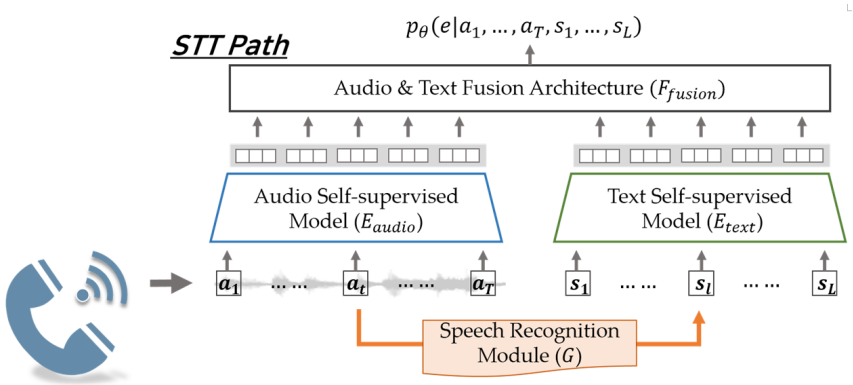

Figure 3. Overview of our Framework with Self-supervised Models

지도학습 모델 E_{text}의 입력 변수로 활용되어 일련의 벡터 $E_{text}(s) = h^s = h_1^s, \ldots, h_k^s$로 압축된다. 제안 프레임워크는 텍스트 자기지도학습 모델로 HanBERT를 적용하였다. HanBERT는 대량의 한국어 문서를 사전 학습하여 문맥과 단어의 상관관계를 고려할 수 있으므로 압축된 텍스트벡터는 감정을 유추할 수 있는 텍스트의 중요정보를 포함하고 있다. 제안 프레임워크의 마지막 단계에서는 압축된 음성벡터 h^a와 텍스트벡터 h^s를 융합 아키텍처 F_{fusion} 의 입력으로 활용하여 감정 $e \in E$ 의 확률분포 $F_{fusion}(h^a, h^s) = p_\theta(e \mid a, G(a))$를 추정한다.

3.2 Consistency Regularization 학습 방법론

감정인식 프레임워크를 적용한 선행 연구는 멀티모달 데이터에서 제공한 음성과 텍스트를 활용하여 멀티모달 모델을 학습한다. 반면, 서비스에 활용할 때에는 외부 음성인식 모듈을 활용하여 추출한 텍스트와 음성을 멀티모달 모델의 입력으로 활용한다. 실제 사용 환경에서 생성된 음성은 주변소음을 포함하고 있으므로 음성인식 모듈을 통해 추출된 텍스트는 실제 대화내용과 다를 수 있다. 따라서 선행 연구의 멀티모달 모델은 서비스 단계에서 음성인식 모듈로부터 실제 대화내용과 다른 텍스트를 입력으로 받아 감정을 추론하므로 낮은 감정인식 성능을 기록했다. 또한, 외부 음성인식 모듈을 적용한 감정인식 프레임워크는 학습과정에서 음성인식 모듈의 인식률을 향상시킬 수 없는 구조적인 문제를 갖고 있다. 이를 극복하기 위하여 본 연구는 음성인식 모듈에서 추출된 텍스트와 실제 대화내용의 차이를 학습에 반영할 수 있는 Consistency Regularization 학습 방법론을 제안한다. 제안 학습 방법론은 음성인식 모듈을 학습할 수 없는 제한된 감정인식 프레임워크의 성능을 향상시키기 위해 설계되었다. 제안 학습 방법론의 구조는 <Figure 4>과 같다.

제안 학습 방법론은 STT path와 Golden Path로 나뉘어 감정을 추론한 후 각각의 결과를 학습에 활용한다. 먼저 학습 데이터 P_{train}에서 샘플링 된 음성 $a \sim P_{train}$이 주어졌을 때, STT path는 음성인식 모듈을 활용하여 텍스트 $s^* = G(a)$를 추출하고 프레임워크의 절차에 따라 감정분포 $p_\theta(e \mid a, s^*)$을 추정한다. 반면, Golden path는 학습 데이터에서 음성과 함께 텍스트 $a, s \sim P_{train}$를 제공받아 멀티모달 모델의 입력변수로 활용하여 감정의 확률분포 $p_\theta(e \mid a, s)$를 추정한다. STT path를 통해 추정한 확률분포와 Golden path를 통해 추정한 확률분포의 차이를 Kullback-Leibler Divergence로 측정하여 Consistency Regularization Loss를 구성한다.

$$L_{consistency}(\theta) = E_{a,s \sim P_{train}}[KL(p_\theta(e \mid a, G(a)) \| p_\theta(e \mid a, s))] \quad (1)$$

두 분포의 차이는 음성인식 모듈을 통해 추출된 텍스트와 실제 대화내용이 다르기 때문에 발생한다. 따라서 Consistency Regularization Loss를 최소화하도록 모델을 학습함으로써 추출된 텍스트와 대화 내용의 차이를 멀티모달 모델이 고려하여 감정을 추정하도록 조정한다. 또한, 멀티모달 모델이 주어진 음성과 텍스트를 활용하여 감정을 예측할 수 있도록 Golden path를 통해 추정한 확률분포와 정답 라벨 $e \in E$을 활용하여 Cross Entropy Loss를 계산한다.

$$L_{cross-entropy}(\theta) = E_{a,s,e \sim P_{train}}[-\log p_\theta(e|a,s)] \quad (2)$$

본 논문에서는 Consistency Regularization Loss와 Cross Entropy Loss를 더해서 최종 목적식을 계산하고 최소화한다. 이때, 두 손실함수의 균형을 맞추기 위하여 가중 계수 $\lambda : \lambda > 0$ 를 곱하여 최종 목적식을 구성한다.

$$\min_\theta \; L_{Total}(\theta) = \lambda L_{consistency}(\theta) + L_{cross-entropy}(\theta) \quad (3)$$

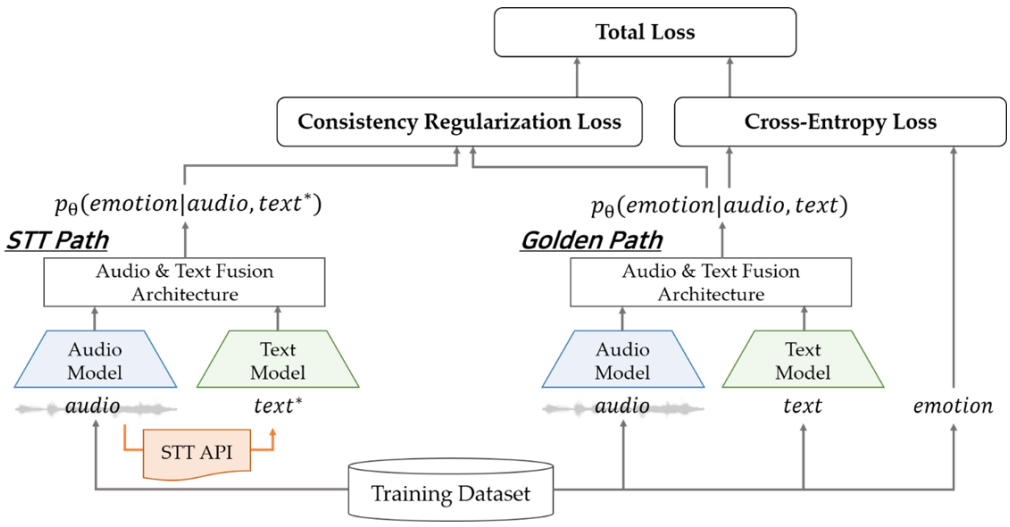

Figure 4. Training Process with Consistency Regularization

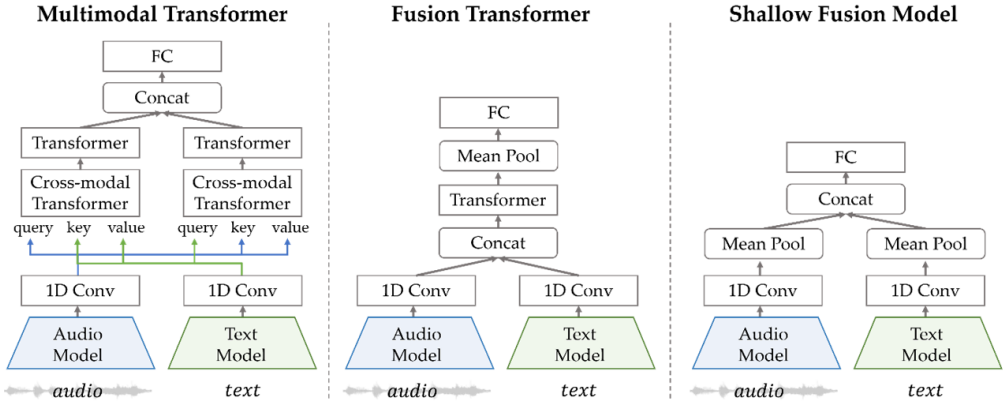

Figure 5. Multi-modal Fusion Architecture

3.3 멀티모달 융합 아키텍처

제안 프레임워크의 융합 아키텍처는 Wav2vec 2.0에서 추출한 음성벡터와 HanBERT에서 추출한 텍스트벡터의 정보를 융합하여 감정을 분류하는 멀티모달 모델이다. 본 연구는 다양한 선행연구의 멀티모달 모델을 변형하여 융합 아키텍처에 적용한다. 본 연구에서 활용한 3가지 융합 아키텍처의 구조는 <Figure 5>와 같다.

Multimodal Transformer(Tsai *et al.*, 2019)는 서로 길이가 다른 변수의 순차적 정보를 반영하고 각 변수에서 공통적으로 나타난 감정을 추출할 수 있는 융합 아키텍처이다. 제안 프레임워크는 Multimodal Transformer 내부의 Cross-modal Transformer를 변형하여 키(key), 쿼리(query), 값(value)에 각각 자기지도학습 모델로부터 추출한 음성벡터와 텍스트벡터를 투입하여 두 정보를 융합한다. Fusion Transformer(Khan *et al.*, 2020)는 순차적인 변수의 정보를 취합할 수 있는 Transformer의 셀프 어텐션(self-attention) 구조를 고려한 융합 아키텍처이다. 본 연구에서는 자기지도학습 모델의 음성벡터와 텍스트벡터의 차원 크기가 다른 것을 고려하여 합성곱 연산을 통해 음성벡터와 텍스트벡터의 차원을 일치시킨 후 두 벡터를 결합(concatenation)하여 Transformer의 입력으로 활용한다. Shallow Fusion Model(Siriwardhana *et al.*, 2020b)은 BERT와 유사한 형태의 음성 및 텍스트 자기지도학습 모델에서 클래스 토큰에 해당하는 첫 번째 벡터를 각각 추출한 후 결합하여 감정을 분류하는 융합 아키텍처이다. 본 연구에서 활용한 음성 자기지도학습 모델은 음성과 텍스트의 크기 따라 음성벡터와 텍스트벡터의 길이가 가변적이기 때문에 평균을 활용하여 크기가 고정적인 벡터를 도출한 후 Shallow Fusion Model에 적용한다.

4. 실험 및 결과

4.1 데이터 설명

본 연구는 AIhub에서 감정 인식 연구용으로 공개된 멀티모달 영상 데이터셋을 활용하여 제안 방법론을 실험하였다. 이 데이터셋은 총 300명의 연기자가 2인 1조를 이루어 다양한 상황을 연기한 영상으로 구성되며, 영상 파일과 함께 영상의 대화내용을 기록한 텍스트, 감정정보 등의 메타파일을 포함하고 있다. 감정 라벨은 기쁨(happy), 놀람(surprise), 분노(angry), 슬픔(sad), 혐오(dislike), 공포(fear), 경멸(contempt), 중립(neutral)으로 총 8가지이며, 자세한 라벨 분포는 <Table 1>과 같다. 데이터셋의 감정라벨에는 데이터 불균형이 존재한다. 데이터 불균형은 모델의 안정적인 학습을 저해하므로 데이터의 수가 가장 적은 경멸(contempt)과 공포(fear)를 제외하고 총 6가지 라벨 데이터셋을 활용하여 제안 방법론을 실험하였다.

4.2 실험 설계

4.2.1 학습 및 평가 방법

본 연구는 멀티모달 영상 데이터를 8:1:1로 나누어 각각 학습 데이터, 검증 데이터, 평가 데이터로 활용하였다. 모든 모델은 학습률(learning rate)이 5e-5로 적용된 Adam(Kingma *et al.*, 2015) Optimizer를 활용하여 학습데이터로 3 에폭(epoch) 학습되었다. 멀티모달 모델 내부 모든 융합 아키텍처의 히든 벡터(hidden vector) 크기는 768로 고정하고, 합성곱 연산의 슬라이딩 윈도우(sliding window) 크기는 3을 적용하였다. 자기지

Table 1. Details of Label Configuration

Label Configuration (Amount, %)								
Happy	Surprise	Angry	Sad	Dislike	Fear	Contempt	Neutral	Total
15706 (19.6%)	11540 (14.4%)	7584 (9.4%)	7808 (9.7%)	19012 (23.7%)	5571 (6.9%)	2944 (3.6%)	9956 (12.4%)	80121 (100%)

도학습 모델을 멀티모달 모델에 적용할 때 간단한 형태의 융합이 더 효과적이라는 선행연구(Siriwardhana *et al.*, 2020a)를 고려하여, Multimodal Transformer와 Fusion Transformer 융합 아키텍처에는 층(layer)의 개수가 1인 Transformer를 적용하고 실험하였다. 자기지도학습 모델이 적용된 멀티모달 모델을 학습할 때, 융합 아키텍처 외에도 자기지도학습 모델을 함께 학습하였다. 학습 시 배치(batch) 크기를 32 또는 64로 설정하고 검증 데이터에서 가장 좋은 성능을 보인 모델을 선택하여 평가 데이터에서 결과를 도출하였다. 실험은 1개의 RTX 2080ti GPU를 활용하여 수행되었으며 배치 크기를 조정하기 위하여 gradient accumulation 기법을 활용하였다. 최종결과는 데이터를 나누는 과정을 포함하여 시드를 변경하면서 총 5회 반복실험을 수행한 뒤 평균값과 분산을 기재하였다.

4.2.2 성능 평가 지표

멀티모달 영상 데이터셋은 <Table 1>에서 확인할 수 있듯이 라벨 불균형이 존재한다. 따라서 평가 데이터의 라벨 별 데이터의수를 반영하여 가중치를 적용한 가중 F1 점수(weighted f1-socre)와 가중 평균 정확도(weighted accuracy)를 평가지표로 활용하였다. <Table 2>는 모델이 예측한 라벨과 실제 라벨을 기반으로 구성한 라벨 e의 정오행렬(confusion matrix)이다. 해당 정오행렬을 바탕으로 라벨 e에 대한 재현율(recall), 정밀도(precision), F1 점수(f1-score), 정확도(accuracy)를 아래와 같이 계산할 수 있다.

$$Recall_e = \frac{TP_e}{TP_e + FN_e} \tag{4}$$

$$Precision_e = \frac{TP_e}{TP_e + FN_e} \tag{5}$$

$$F1_e = \frac{2 \times (Recall_e \times Precision_e)}{Recall_e + Precision_e} \tag{6}$$

$$Acc_e = \frac{TP_e + TN_e}{TP_e + FP_e + TN_e + FN_e} \tag{7}$$

평가 데이터에서 라벨 e의 구성 비율을 $w_e : 0 \leq w_e \leq 1$라고 할 때, 식 (6)과 (7)을 바탕으로 가중 F1 점수와 가중 평균 정확도를 계산하는 수식은 아래와 같다.

$$weighted\ F1 = \sum_e F1_e \times w_e \tag{8}$$

$$weighted\ \ Acc = \sum_e Acc_e \times w_e \tag{9}$$

Table 2. Confusion Matrix of Emotion Label e

Emotion e		Model Prediction	
		Positive	Negative
True Label	Positive	TP_e	FN_e
	Negative	FP_e	TN_e

4.2.3 단일 모델

비교실험을 위해 음성과 텍스트 중 한 가지 데이터만을 활용하여 감정을 인식하는 단일 모델(Wu *et al.*, 2021)을 구축하였다. 단일 모델은 합성곱 신경망(CNN)과 Transformer로 구성되어있으며 다양한 특징정보를 입력으로 받아 감정을 분류한다. 본 실험은 음성 단일 모델의 입력변수로 음성의 스펙트럼 특징인 MFCC와 음성 자기지도학습 모델로부터 추출한 음성 벡터를 활용한다. 그리고 텍스트 단일 모델의 입력변수로 FastText 임베딩 또는 텍스트 자기지도학습 모델인 HanBERT로부터 추출한 텍스트벡터를 활용한다. 이때, FastText 임베딩은 FastText 알고리즘(Bojanowski *et al.*, 2017)을 활용하여 위키피디아 한글 데이터에서 등장한 단어들의 유사정보를 반영한 벡터를 의미한다. 본 연구는 웹 상에 공개된 FastText 임베딩을 활용하였다(https://github.com/ratsgo/embedding).

4.2.4 외부 음성인식 모듈

본 연구는 음성인식 모듈로 구글 클라우드의 STT API와 카카오 Auto Speech Recognition API를 활용하였다. 음성인식 모듈의 성능 평가 지표는 문자 오류율(CER), 단어 오류율(WER), 보정 단어 오류율(sWER)이 있다. 단어 오류율은 음성인식 모듈을 활용하여 추출한 텍스트와 원본 텍스트의 편집 거리(Levenshtein distance)를 단어 단위로 측정한 평가지표이다. 문자 오류율은 문자 단위로 편집 거리를 측정한 평가지표이다. 마지막으로 보정 단어 오류율(sWER)은 한국어의 유연한 띄어쓰기를 반영하여 띄어쓰기를 보정한 후 단어 오류율을 측정한 평가지표(Bang *et al.*, 2020)이다. 한국어 멀티모달 데이터셋에서 세 가지 평가지표(CER/WER/sWER)를 활용하여 음성인식 모듈의 성능을 평가한 결과 구글 음성인식 모듈의 성능은 35.1/63.8/56.7을 기록하였고, 카카오 음성인식 모듈은 56.1/78.3/73.8을 기록하였다.

4.3 실험 결과

4.3.1 원본 텍스트를 활용한 감정인식 성능비교

본 실험은 멀티모달 데이터셋에서 제공하는 음성과 대화내용이 기록된 텍스트를 활용하여 모델을 학습하고 평가한 실험이다. 이는 음성인식 모듈의 음성 인식률이 100%인 환경에서 감정인식 프레임워크의 성능을 평가하는 것과 동일하다. 비교실험을 위해 단일 모델과 함께 MFCC와 FastText를 적용한 Shallow Fusion 멀티모달 모델을 기준모델(baseline model)로 구축하였다. 단일 모델과 감정인식 프레임워크가 적용된 멀티모달 모델의 실험결과는 <Table 3>과 같다.

실험 결과 감정인식 프레임워크를 적용한 멀티모달 모델은 단일 모델보다 높은 성능을 기록하였다. MFCC, FastText 모델을 제외한 대부분의 멀티모달 모델은 단일 모델 중 최고 성능을 기록한 HanBERT보다 가중 F1 점수 기준 최소 3%p 이상 높은 성능을 기록하였다. 멀티모달 모델은 음성 정보와 텍스트 정보를 함께 활용

Table 3. Emotion Recognition Results with Original Text

Model	F1 mean/F1 std	Acc mean/Acc std
Only Text		
Single Model - FastText	0.4116/0.0085	0.4247/0.0073
Single Model - HanBERT	**0.4463/0.0078**	**0.4517/0.0089**
Only Audio		
Single Model - MFCC	0.3539/0.0058	0.3637/0.0073
Single Model - Wav2vec 2.0	**0.4326/0.0058**	**0.4409/0.0043**
Multi-Modal		
Shallow Fusion - MFCC, FastText	0.4364/0.0069	0.4474/0.0057
Shallow Fusion - Wav2vec 2.0, HanBERT	**0.4952/0.0062**	**0.4989/0.0062**
Fusion Transf. - Wav2vec 2.0, HanBERT	0.4906/0.0075	0.4963/0.0067
Multimodal Transf. - Wav2vec 2.0, HanBERT	0.4893/0.0092	0.4958/0.0090

하여 감정을 인식하므로 실험 결과에서 단일 모델보다 더 우수한 것으로 나타났다. 또한, 본 실험에서 자기지도학습 모델을 적용한 멀티모달 모델의 성능이 크게 향상되는 것을 확인할 수 있다. Wav2vec 2.0과 HanBERT를 활용한 Shallow Fusion 멀티모달 모델이 가중 F1 점수 기준 0.4952로 가장 높은 성능을 기록했다. 대량의 데이터로 사전 학습된 자기지도학습 모델은 음성 또는 텍스트에서 중요한 패턴 및 특징을 추출할 수 있으므로 비교적 데이터의 양이 적은 감정인식 분야에서 모델의 성능을 효과적으로 향상시키는 것을 확인할 수 있다.

4.3.2 음성인식 모듈을 활용한 감정인식 성능비교

본 실험은 음성만 활용 가능한 실제 서비스 환경에서 감정인식 프레임워크의 성능을 평가하기 위해 음성인식 모듈에서 추출한 텍스트를 활용하여 멀티모달 모델을 학습하고 평가한다. 다만, 제안 학습 방법론은 모델의 훈련 단계에서 원본 텍스트와 음성인식 모듈에서 추출한 텍스트를 함께 활용하여 모델을 학습한다. 그리고 평가 단계에서는 추출한 텍스트와 음성만을 활용하여 제안 방법론으로 학습된 모델을 평가한다. 본 실험의 상세한 실험 결과는 <Table 4>에서 확인할 수 있다. 해당 표에서 음성인식 모듈을 적용한 모델과 원본 텍스트를 활용한 모델의 차이를 괄호 안에 표기하여 음성인식 모듈 적용한 실제 서비스 환경에서 감정인식 프레임워크의 성능하락을 표기하였다.

실험 결과 음성인식 모듈을 적용한 실제 상황에서는 이상적인 상황(완벽한 음성인식이 가능한 상황)에 비해 감정인식 성능이 하락하는 것을 확인할 수 있다. 특히, 성능이 저조한 카카오 음성인식 모듈을 감정인식 프레임워크에 적용한 실험에서 감정인식 성능 하락이 두드러진다. Shallow Fusion 모델의 감정인식 하락률이 음성인식 모듈을 적용하기 전보다 가중 F1 점수 기준 약 8.3%p 하락하여 HanBERT를 적용한 단일모델보다 낮은 성능을 보였다. 이는 음성인식 모듈에서 추출한 텍스트가 실제 대화내용과 다르기 때문에 노이즈로 작용하여 멀티모달 모델의 감정인식에 부정적인 영향을 미친 것으로 해석할 수 있다.

반면, 제안 학습 방법론을 적용하면 멀티모달 모델의 감정인식 성능이 일정 수준 회복되는 것으로 나타났다. 카카오 음성인식 모듈을 적용하여 성능이 하락한 Shallow Fusion 멀티모달 모델에 제안 학습 방법론을 적용하면 4.3%p 감정 인식 성능이 회복되는 것을 확인할 수 있다. 게다가 구글 음성인식 모듈을 적용한 프레임워크에서 제안 학습 방법론으로 인한 감정인식 성능 회복이 더 크다는 것을 확인할 수 있다. 이는 제안 학습 방법론을 적용하면 모듈에서 추출한 텍스트와 실제 대화

Table 4. Emotion Recognition Results with Original Text

Model	F1 mean/F1 std (%p change)		Recovery
	STT Text	Proposed Method	
Kakao API			
Shallow Fusion - Wav2vec 2.0, HanBERT	**0.4121/0.0440(↓8.3)**	**0.4550/0.0027(↓4.0)**	**4.3%p**
Fusion Transf. - Wav2vec 2.0, HanBERT	0.4281/0.0061(↓6.2)	0.4521/0.0043(↓3.8)	2.4%p
Multimodal Transf. - Wav2vec 2.0, HanBERT	0.4213/0.0066(↓6.8)	0.4572/0.0047(↓3.2)	3.5%p
Google API			
Shallow Fusion - Wav2vec 2.0, HanBERT	**0.4220/0.0463(↓7.3)**	**0.4674/0.0041(↓2.7)**	**4.6%p**
Fusion Transf. - Wav2vec 2.0, HanBERT	0.4313/0.0051(↓5.9)	0.4601/0.0085(↓3.0)	2.9%p
Multimodal Transf. - Wav2vec 2.0, HanBERT	0.4240/0.0332(↓6.5)	0.4670/0.0059(↓2.2)	4.3%p

내용의 차이를 학습에 반영할 수 있으므로 저조한 음성 인식률 때문에 하락한 멀티모달 모델의 성능을 회복시킬 수 있다는 것을 의미한다. 다만, 모듈에서 추출한 텍스트와 실제 대화내용의 차이가 클수록 두 문장 사이의 관계를 모델이 학습하기 위하여 더 많은 데이터가 필요하다. 한국어 감정인식 데이터는 소량이므로 추출한 텍스트와 실제 대화내용의 차이가 비교적 작은 구글 음성인식 모듈을 적용했을 때 멀티모달 모델의 감정인식 성능 회복이 더 크게 나타났다.

4.4 제안 학습 방법론 효과 분석

본 연구는 제안 학습 방법론의 효과를 정밀 분석하기 위하여 멀티모달 모델에서 음성과 텍스트 자기지도학습 모델의 기여도를 Gradient-weighted Class Activation Mapping(Grad-CAM)을 활용하여 측정했다. Grad-CAM은 이미지 분야에서 모델이 특정 라벨을 예측하는데 기여한 객체를 추출하기 위하여 이미지 픽셀의 중요도를 그래디언트로 추정한 분석 방법론이다 (Selvaraju *et al.*, 2017). 감정분석 멀티모달 모델에 Grad-CAM을 적용하면 융합 아키텍처의 입력으로 활용된 각 자기지도학습 모델의 정보벡터가 감정을 분류하는데 기여한 정도를 측정할 수 있다. 본 연구는 테스트 데이터에서 Grad-CAM을 활용하여 자기지도학습 모델의 기여도를 측정한 후, 라벨을 기준으로 평균을 적용하여 자기지도학습 모델의 라벨별 평균 기여도를 계산하였다. <Figure 6>는 Shallow Fusion 융합 아키텍처가 적용된 멀티모달 모델에서 자기지도학습 모델의 라벨별 평균 기여도를 측정한 그림이다.

기여도 측정 결과 혐오(dislike)와 중립(neutral) 감정을 예측할 때, 음성 자기지도학습 모델의 정보 의존도가 각각 53.2%, 51.9%로 비교적 크다는 것을 확인할 수 있다. 멀티모달 데이터셋에서 혐오 감정은 텍스트에는 감정이 드러나지 않는 비꼬는 형태의 대화에서 나타난다. 비꼬는 형태의 대화는 다양한 감정에서 활용될 수 있는 일반적인 단어로 구성된 텍스트와 독특한 느낌의 억양이 포함된 음성의 조합으로 표현된다. 반면, 중립 감정이 나타난 대화는 음성 및 텍스트에 감정을 나타낼 수 있는 특징을 포함하고 있지 않지만 음성의 변화에 따라 다양한 형태의 감정으로 변형될 수 있다. 따라서 멀티모달 모델이 혐오 및 중립 감정을 다른 감정과 구분하기 위하여 텍스트 정보보다 음성 정보를 더 활용한 것으로 추측할 수 있다.

한편, 슬픔(sad) 감정 예측에는 주로 텍스트 자기지도학습 모델의 정보를 활용하는 것으로 나타났다. 데이터셋에서 슬픔을 표현하는 대화에는 다른 감정과는 다르게 슬픔을 나타낼 수 있는 특정 단어가 빈번히 등장한다. 따라서 멀티모달 모델이 텍스트 정보에 대한 의존도가 높은 것으로 추측할 수 있다. 그 이외의 감정 예측에는 두 가지 정보를 공평하게 활용하는 것으로 나타났다. 측정한 기여도를 기반으로 음성인식 모듈 및 제안 학습 방법론을 적용한 멀티모달 모델의 성능 변화를 분석하기 위하여, 테스트 데이터에서 측정한 각 모델의 성능을 정오행렬로 표현한 그림은 <Figure 7>과 같다.

분석 결과 음성인식 모듈을 적용한 멀티모달 모델의 감정인식 성능이 원본 텍스트를 사용한 멀티모달 모델보다 낮은 것을 확인할 수 있다. 특히, 텍스트 자기지도학습 모델의 정보 기여도가 큰 라벨인 슬픔(sad)의 예측성능은 48.8%에서 34.1%로 크게 하락하는 것으로 나타났다. 음성인식 모듈로부터 추출한 텍스트가 실제 대화내용과 다르기 때문에 텍스트에 의존적인 라벨의 감정인식 성능이 하락하는 것으로 추측할 수 있다. 반면, 음성 자기지도학습 모델의 기여도가 큰 라벨인 혐오(dislike)와 중립(neutral)의 감정인식 성능은 조금 상승한 것을 확인할 수 있다. 마지막으로 제안 학습 방법론을 적용한 결과 음성인식 모듈의 저조한 인식률로 인해 하락한 놀람(surprise), 슬픔(sad), 분노(angry) 라벨의 감정인식 성능이 상승한 것을 확인할 수 있다. 이 결과는 제안 학습 방법론을 적용하면 음성인식 모듈 때문에 발생한 텍스트 정보의 오류를 멀티모달 모

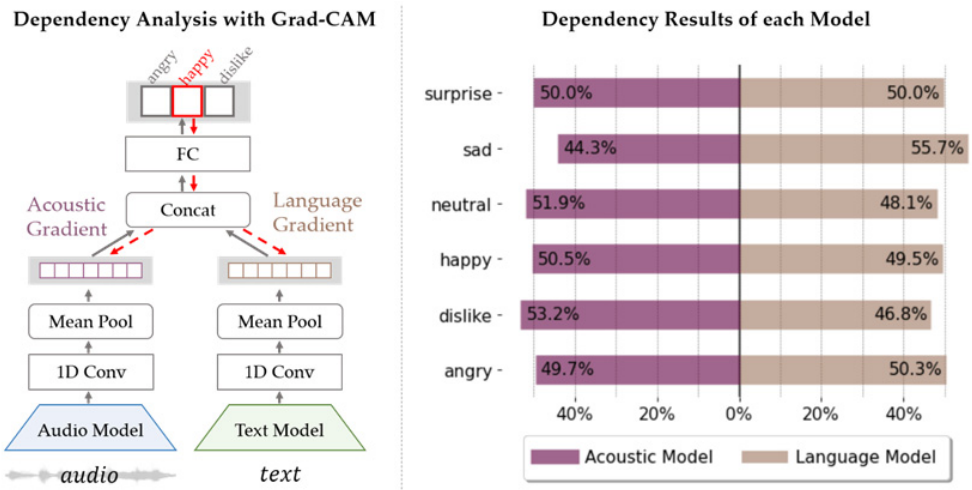

Figure 6. Dependency Results of each Self-supervised Model by Grad-CAM

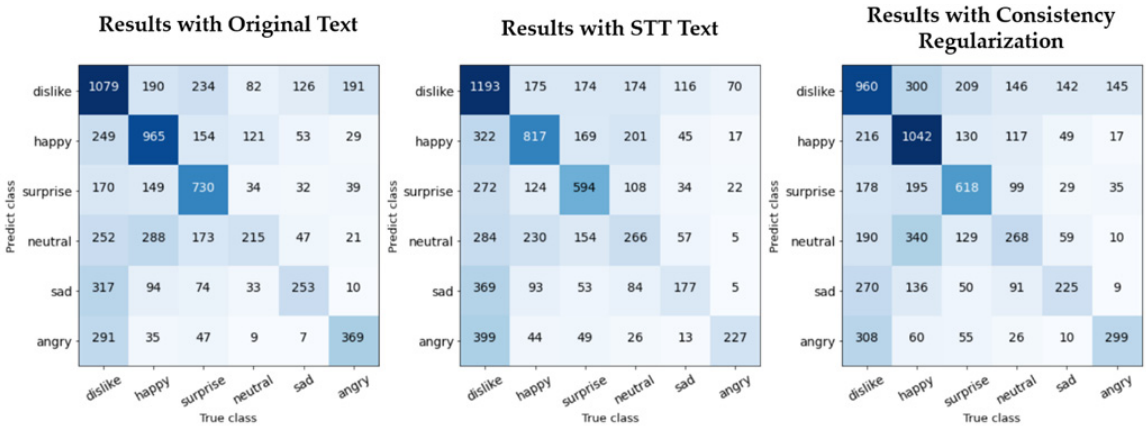

Figure 7. Confusion Matrix Results of Multi-modal Model with Various Settings

델이 보정하여 프레임워크의 감정 인식 성능을 향상시킬 수 있음을 입증하는 것이다.

5. 결 론

본 연구는 실제 서비스 환경에서 주변소음이 포함되어 있는 경우 음성인식 모듈을 적용한 감정인식 프레임워크의 성능이 하락하는 것을 인지하고 이를 극복할 수 있는 학습 방법론을 제안하였다. 제안 학습 방법론은 음성인식 모듈로부터 추출한 텍스트와 실제 대화내용의 차이를 학습에 반영하여 실제 서비스 환경에서 프레임워크의 감정인식 성능 하락을 최소화하였다. 또한, 한국어 감정인식 데이터가 부족한 상황에서 모델의 성능을 향상시키기 위하여 자기지도학습 모델인 Wav2vec 2.0과 HanBERT를 적용한 감정인식 프레임워크를 제안하였다. 사전 학습된 자기지도학습 모델은 음성과 텍스트에서 감정분석에 필요한 특징벡터를 추출할 수 있으므로 소량의 데이터를 활용하여 감정인식 프레임워크의 성능을 향상시켰다. 추가적으로 다양한 퓨전 아키텍처를 적용하여 음성과 텍스트의 특징벡터를 융합하고 감정을 분석함으로써 음성 또는 텍스트만 활용한 단일 모델 대비 우수한 성능을 기록하였다.

본 연구 결과를 바탕으로 다음과 같은 후속 연구 방향을 생각해볼 수 있다. 첫째, 한국어와 영어는 발음 및 억양이 다르기 때문에 음성에서 감정을 나타내는 특징이 다를 수 있다. 본 연구의 제안 프레임워크는 영어 음성데이터로 사전 학습된 Wav2vec 2.0을 음성 자기지도학습 모델로 활용하고 있으므로 감정인식 데이터로 학습하기 전에는 한국어 특화된 감정특징을 추출할 수 없다는 단점을 갖고 있다. 따라서 한국어 감정인식 데이터에서 자기지도학습 모델의 효과를 극대화하기 위해 한국어로 사전 학습된 자기지도학습 모델을 감정인식 프레임워크에 적용하는 연구가 필요할 것이다. 또한, 제안 프레임워크는 텔레뱅킹, AI스피커 등 다양한 환경에서 서비스를 제공해야 하므로 실시간으로 감정인식을 가능해야 한다. 하지만 제안 프레임워크는 감정인식 성능을 향상시키기 위하여 대량의 파라미터를 보유한 자기지도학습 모델을 활용하고 있으므로 연산속도가 느리다는 단점을 갖고 있다. 해당 문제에 대해서는 저용량의 자기지도학습 모델(Jiao *et al.*, 2020; Peng *et al.*, 2021)과 학습된 모델을 경량화하여 연산속도를 향상시키는 기법들(Sanh *et al.*, 2019)이 대안이 될 수 있을 것이다.

참고문헌

Baevski, A., Schneider, S., and Auli, M. (2019), vq-wav2vec: Self-supervised learning of discrete speech representations, arXiv.org > cs > arXiv:1910.05453.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020), wav2vec 2.0: A framework for self-supervised learning of speech representations, arXiv:2006.11477 (cs).

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

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017), Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, **5**, 135-146.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020), A simple framework for contrastive learning of visual representations, *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119: 1597-160.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018), Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805 (cs).

Fayek, H. M., Lech, M., and Cavedon, L. (2017), Evaluating deep learning architectures for Speech Emotion Recognition, *Neural Networks*, **92**, 60-68.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., and Marsic, I. (2018), Multimodal affective analysis using hierarchical attention strategy with word-level alignment, *Proceedings of the Conference*, Association for Computational Linguistics. Meeting, 2225-2235.

Han, K., Yu, D., and Tashev, I. (2014), Speech emotion recognition using

deep neural network and extreme learning machine, *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., and Liu, Q. (2020), TinyBERT: Distilling BERT for Natural Language Understanding, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Findings.

Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F., and Hamfors, O. (2012), Usefulness of sentiment analysis, *Paper presented at the European Conference on Information Retrieval*.

Khan, A. U., Mazaheri, A., Lobo, N. D. V., and Shah, M. (2020), Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering, arXiv:2010.14095 (cs).

Kim, K.-H., Lee, C.-S., Jo, S.-M., and Cho, S.-B. (2015), Predicting the success of bank telemarketing using deep convolutional neural network, *Proceedings of the 7th International Conference of Soft Computing and Pattern Recognition* (SoCPaR).

Kingma, D. P. and Ba, J. (2015), Adam: A Method for Stochastic Optimization, Paper presented at the ICLR (Poster).

Kolesnikov, A., Zhai, X., and Beyer, L. (2019), Revisiting self-supervised visual representation learning, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (2003), Emotion recognition by speech signals, *Paper presented at the Eighth European conference on speech communication and technology*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, Luke, and Stoyanov, V. (2019), Roberta: A robustly optimized bert pretraining approach, arXiv:1907.11692 (cs)..

Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., and Poria, S. J. K.-B. S. (2018), Multimodal sentiment analysis using hierarchical fusion with context modeling, *Knowledge-Based Systems*, **161**, 124-133.

McDuff, D., Rowan, K., Choudhury, P., Wolk, J., Pham, T., and Czerwinski, M. (2019), A multimodal emotion sensing platform for building emotion-aware applications, arXiv:1903.12133 (cs).

Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003), Speech emotion recognition using hidden Markov models, *Speech Communication,* **41**(4), 603-623.

Paleari, M., Huet, B., and Chellali, R. (2010), Towards multimodal emotion recognition: A new approach, *Proceedings of the ACM International Conference on Image and Video Retrieval*, 174-181.

Pantic, M., Sebe, N., Cohn, J. F., and Huang, T. (2005), Affective multimodal human-computer interaction, *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 669-676.

Park, J. (2018), HanBERT: Pretrained BERT Model for Korean.

Peng, Z., Budhkar, A., Tuil, I., Levy, J., Sobhani, P., Cohen, R., and Nassour, J. (2021), Shrinking Bigfoot: Reducing wav2vec 2.0 footprint, arXiv:2103.15760 (cs).

Pepino, L., Riera, P., and Ferrer, L. (2021), Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings, arXiv:2104.03502 (cs).

Rawat, S., Gupta, P., and Kumar, P. (2014), Digital life assistant using automated speech recognition, *Proceedings of the 2014 Innovative Applications of Computational Intelligence on Power, Energy and Controls with their impact on Humanity* (CIPECH).

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019), DistilBERT, a distilled version of BERT: Smaller, faster, Cheaper and Lighter, arXiv:1910.01108 (cs).

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019), wav2vec: Unsupervised pre-training for speech recognition, arXiv:1904.05862 (cs).

Seehapoch, T. and Wongthanavasu, S. (2013), Speech emotion recognition using support vector machines, *Proceedings of 5th International Conference on Knowledge and Smart Technology* (KST).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017), Grad-cam: Visual explanations from deep networks via gradient-based localization, *Proceedings of the IEEE International conference on Computer Vision*.

Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., and Nanayakkara, S. (2020a), Multimodal Emotion Recognition With Transformer-Based Self Supervised Feature Fusion, *IEEE Access*, **8**, 176274-176285.

Siriwardhana, S., Reis, A., Weerasekera, R., and Nanayakkara, S. (2020b), Jointly Fine-Tuning "BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition, arXiv:2008.06682 (eess).

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019), Multimodal transformer for unaligned multimodal language sequences, *Proceedings of the Conference*, *Association for Computational Linguistics*, Meeting.

Tsai, Y. H. H., Liang, P. P., Zadeh, A., Morency, L. P., and Salakhutdinov, R. (2018), Learning factorized multimodal representations, arXiv:1806.06176 (cs).

Wu, M., Li, K., Leung, W. K., and Meng, H. J. P. I. (2021), Transformer Based End-to-End Mispronunciation Detection and Diagnosis, *INTERSPEECH 2021,* 3954-3958.

Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., and Li, X. (2019), Learning Alignment for Multimodal Emotion Recognition from Speech, arXiv:1909.05645 (cs).

Yoon, S., Byun, S., and Jung, K. (2018), Multimodal speech emotion recognition using audio and text, *Proceedings of the 2018 IEEE Spoken Language Technology Workshop* (SLT).

저자소개

김정희: 고려대학교 산업경영공학과에서 2015년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학부에서 석사과정으로 재학 중이다. 연구 분야는 비정형 데이터를 활용한 데이터 마이닝이다.

강필성: 서울대학교 산업공학과에서 2003년 학사, 2010년 박사학위를 취득하였다. 이후 현대카드 과장, 서울과학기술대학교 조교수로 근무하였으며, 현재는 고려대학교 산업경영공학부 부교수로 재직 중이다. 연구 분야는 정형 및 비정형 데이터를 활용한 데이터마이닝 및 기계학습 알고리즘 개발 및 제조/IT/공공분야 응용이다.