Speech Communication 140 (2022) 11–28

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion

Bagus Tris Atmaja [a,*,1], Akira Sasou [a], Masato Akagi [b]

[a] National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki 305-8560, Japan
[b] Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan

ARTICLE INFO

Keywords:
Speech emotion recognition
Affective computing
Audiotextual information
Bimodal fusion
Information fusion

ABSTRACT

Speech emotion recognition (SER) is traditionally performed using merely acoustic information. Acoustic features, commonly are extracted per frame, are mapped into emotion labels using classifiers such as support vector machines for machine learning or multi-layer perceptron for deep learning. Previous research has shown that acoustic-only SER suffers from many issues, mostly on low performances. On the other hand, not only acoustic information can be extracted from speech but also linguistic information. The linguistic features can be extracted from the transcribed text by an automatic speech recognition system. The fusion of acoustic and linguistic information could improve the SER performance. This paper presents a survey of the works on bimodal emotion recognition fusing acoustic and linguistic information. Five components of bimodal SER are reviewed: emotion models, datasets, features, classifiers, and fusion methods. Some major findings, including state-of-the-art results and their methods from the commonly used datasets, are also presented to give insights for the current research and to surpass these results. Finally, this survey proposes the remaining issues in the bimodal SER research for future research directions.

1. Introduction

Speech is a sensory modality to express and communicate emotions. In a speech chain, humans convey their messages from the speakers' brain to the listeners' brain via speech. By speaking, speakers not only express their thoughts into speech information but also communicates their speech information. The information in speech includes emotion. The speaker wants the listener to be able to perceive their emotions, for instance, by modulating the intonation into their voice. This speech chain shows how humans and humans communicate their emotions through speech.

Speech emotion recognition (SER) is a part of affective computing – computing that relates to, arises from, or influences emotions (Picard, 1995) – that focuses on recognizing emotion from humans' voices. SER is an attempt to make a computer to be able to recognize the expressed emotions in a given utterance. The earliest reported research on SER, perhaps, was the work of Dellaert et al. (1996). The study explored prosodic features with several statistical pattern recognition techniques to classify the emotional content of utterances. Under a limited number of data (1000 utterances), the system achieved a comparable performance close to humans.

Instead of using acoustic information only, one of the key insights in the two decades of SER development is the fusion of acoustic and linguistic information (Schuller, 2018). As predicted in human emotion perception, the addition of linguistic data enriches the SER system and helps machines recognize human emotion better. The addition of linguistic information also means doubling the data (input features), and more data tends to be more effective. Given the evidence, there is a shift in recent SER research from unimodal acoustic analysis to bimodal acoustic–linguistic information fusion. Hence, this introduction will briefly describe unimodal acoustic analysis for SER and the shift from unimodal to bimodal information fusion.

1.1. Unimodal acoustic analysis

Utilizing speech to identify humans' emotions roots from the correlation between voice and emotion. There is strong evidence that humans can recognize other's emotions from their voices. For instance, Mozziconacci (2002) stated "*speech variability [prosody] corresponding to the expressiveness in the speech is not random and that a better understanding of this variability would be praiseworthy*". For dimensional emotion (recognizing degree of valence, arousal, and dominance), it is observed that fundamental frequency correlates with valence while rhythmic and spectral characteristics of voice correlate with arousal

* Corresponding author.
E-mail address: b-atmaja@aist.go.jp (B.T. Atmaja).
[1] Currently on leave from Department of Engineering Physics, Institut Teknologi Sepuluh Nopember, Indonesia.

https://doi.org/10.1016/j.specom.2022.03.002
Received 21 April 2021; Received in revised form 23 November 2021; Accepted 4 March 2022
Available online 26 March 2022

(Mairano et al., 2019). A review by Schuller (2018) showed the root correlation between acoustics and human emotion.

Among multimodal data, speech has been widely chosen for recognizing emotion since it is less private than other data, e.g., image and video. We argue that speech is less private than image and video data since it contains less information than both image and video. For instance, one can attribute information from images and video (e.g., the physical appearance of a person in a photo) and search/find related information about a person using these data easier than using audio information. Using speech to recognize emotion benefits future implementation of speech-related technologies, e.g., voice assistants and telephone conversations. A milestone conducted by Petrushin (1999) proves a pilot study to implement SER for call center applications. This laboratory-scale study, at that time, showed a potential application for SER technologies. Nowadays, this SER technology that could recognize speaker emotions in telephone conversations is available limited in the commercial market, while its research is still ongoing.

Research on SER has been focused on two main areas: finding biomarkers that are highly correlated to emotion and building models based on these biomarkers. The first focus resulted in two main types of acoustic features commonly adopted in SER community: low-level descriptor (LLD) and high-level statistical function (HSF). LLDs were used to evaluate acoustic characteristics related to emotion, mostly by modeling temporal and spectral information. First, an utterance of the speech signal is divided into several frames. Second, a window function is applied on each frame, and the specific acoustic features are extracted on this frame. The LLDs from all frames are concatenated to obtain feature representation for a single utterance. If needed, these LLDs are padded with zeros to produce the same vector size as other utterances.

Instead of concatenating acoustic features from all frames in an utterance, a global value per acoustic feature can be calculated by aggregating these acoustic features. The role of this HSF is to model temporal variations and contour of different LLDs from all frames in an utterance (Mirsamadi et al., 2017). Among many statistical functions, mean values and standard deviations were found usefully for (dimensional) speech emotion recognition (Schmitt et al., 2019; Atmaja and Akagi, 2021). Although El Ayadi et al. (2011) argued that this HSF suffers from the loss of temporal information and the small size of features, Atmaja and Akagi (2021) showed that HSF obtained better results than LLD in the same dataset and model.

Combining local and global features is a way to compromise the advantages and disadvantages of each feature extraction method. eGeMAPS (Eyben et al., 2016) is a fusion of GeMAPS (containing 23 LLD) and their functionals resulting in 88 parameters. Vlasenko et al. (2007) combined frame-level and turn-level information (LLD and HSF) for robust speech emotion recognition. The results emphasized feature integration on different levels of feature extraction. However, there is no study found investigating the trade-off among the use of LLD, HSF, and hybrid features. Neither a way to accelerate the extraction of functionals, which requires the calculation of LLD, was proposed.

The progressive research on SER led to practical implementation in the commercial industry. Nowadays, SER has been implemented in various applications, both web/cloud-based applications and standalone applications. Although it is useful to analyze the subject's affective states, these emerging affective recognizer technologies have been criticized by others. Researchers in psychology argued that due to individuals' high variability, the emotional categories do not have an essence; the correlation between particular facial expressions and the corresponding basic emotions was not strongly supported (Barrett et al., 2019). While this argument was attributed to categorical emotion, it may be better to model the emotion in other than categorical form.

To this end, acoustic information is required to extract emotional knowledge from speech data. However, using acoustic features only may be insufficient. Since linguistic features also can be derived from speech, it is reasonable to fuse acoustic with linguistic information to observe such improvements.

1.2. From unimodal to bimodal information fusion

Among many other issues, multimodal information fusion is a challenging task in pattern recognition. Recent studies (Poria et al., 2017; Atmaja and Akagi, 2020; Atmaja et al., 2019) confirm that multimodal classifiers outperform unimodal classifiers. The multimodal information employed speech, text, and video data. The unimodal information utilized only speech data. In SER itself, one of the main issues in searching for a more predictive feature is whether it suffices to explore acoustic features only, or it is necessary to combine acoustic features with other modalities (El Ayadi et al., 2011). For speech, both acoustic and linguistic features can be extracted. Thus, two pieces of information can be fused to evaluate the effectiveness of information fusion from a single speech modality without the need for additional modalities.

Aside from unnecessary additional measurements, the correlation among acoustic, linguistic, and emotion in human communication is also predicted in both experimental psychology and neuroscience (Nygaard and Queen, 2008; Liebenthal et al., 2016). The result of experiments suggested that humans used emotional words to strengthen emotional tone. Neuroscience experiments show the neural dynamics and interaction between verbal (linguistic) and non-verbal (vocalization) channels. As in this human ability to perceive emotion from acoustic and linguistic information, machines and computers should be able to recognize humans' emotions from bimodal acoustic and linguistic information fusion more precisely than from unimodal information.

The use of linguistic information for SER is also reasonable from an affective computing point of view. In natural language processing, linguistic information is extracted from the text for sentiment analysis. This textual information was also used to detect emotion in text (Alm et al., 2005; Mulcrone, 2012; Calvo and Kim, 2013). In these works, textual information shows encouraging results on both categorical and dimensional emotion recognition from text. Fusing acoustic and linguistic information may improve the performance of SER more significantly than other strategies.

Indeed, bimodal emotion recognition by fusing acoustic and linguistic information shows significant performance improvement. Refs. Schuller et al. (2004, 2005), Eyben et al. (2010), Ye and Fan (2014) and Tian et al. (2016) show the usefulness of fusing acoustic and linguistic in different strategies to improve SER performance. Different acoustic and linguistic features were fused using different classifiers. Different fusion strategies were evaluated to investigate the effectiveness of the fusion methods.

Speech delivers messages that go beyond words. In this understanding, word meaning is not enough to convey a message; acoustic information is needed. Acoustic information alone is also not enough to deliver a message. It is not only how it is said (acoustic) but also what is said (linguistic). The fusion of acoustic and linguistic information makes clear the messages from speaker to listener. This concept is illustrated in Fig. 1.

This survey paper differs from the previous ones (e.g., El Ayadi et al. (2011) and Akçay and Oguz (2020)) in several ways. In these previous surveys, the belief for emotional speech is primarily about how it is said rather than what is said. This paper reviews the studies on combining both pieces of "how" and "what" information. Acoustic features contain information on how it is being said. Linguistic features contain information on what is being said. Fusing both pieces of information, which are extracted from a speech, will improve the clarity of the message, including the expressed emotion. The perception of emotion will also be improved by the fusion of this bimodal information. This process is in line with the data-information-knowledge (DIK) concept in information science.

The process of recognizing emotion from speech consists of two main steps. The first is extracting acoustic representation and linguistic information from speech data; the second is extracting emotional knowledge from acoustic and linguistic information. The output of the

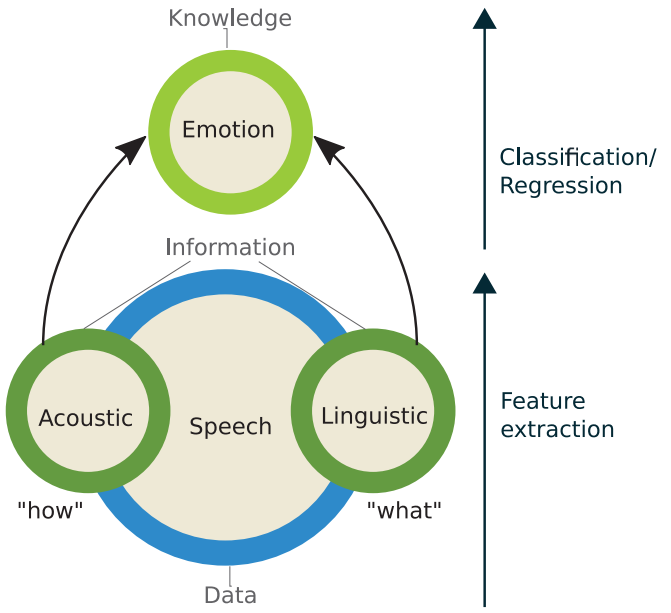

Fig. 1. A concept of bimodal emotion recognition by combining acoustic and linguistic information.

first step becomes the input of the second step. Feature extraction extracts two pieces of information from speech — acoustic and linguistic. For categorical emotion, the process from information to knowledge is classification. For dimensional emotion, the process is regression. These processes are commonly performed within machine learning or deep learning. The acoustic and linguistic information are fused in this step, which can be implemented in various ways.

This paper aims to review the current studies of bimodal emotion recognition by utilizing acoustic and linguistic information. The main scope of this study includes the datasets, emotion models, features, classifiers, and fusion methods used in bimodal SER. The main scope is extended to explore the major findings and highlight the remaining challenges for bimodal SER.

The rest of this paper can then be organized as follows. Section 2 describes related work and the difference of this work from the previous surveys. Sections 3–6, and 7 present datasets, emotion models, features, classifiers, and fusion methods. Section 8 discusses the major findings and highlights the remaining issues. Finally, Section 9 concludes this survey paper.

2. Related work

Surveys or reviews on speech emotion recognition have been presented in many forms (El Ayadi et al., 2011; Anagnostopoulos et al., 2012; Sailunaz et al., 2018; Akçay and Oguz, 2020; Wu et al., 2014). However, no survey has been found focusing on the fusion of acoustic and linguistic information for SER. Most SER surveys only focus on acoustic information while others focus on the multimodal fusion of acoustic information with other modalities, including biological signals, texts (linguistic), and videos (visual).

A survey paper by El Ayadi et al. (2011) thoroughly reviewed features, classification schemes, and datasets for SER. That survey paper by El Ayadi et al. has influenced the research of SER in many areas: the importance of local feature vs. global feature, the effect of such preprocessing methods, and the necessity to combine speech with other modalities. That paper by El Ayadi et al. also motivates this survey paper in which bimodal acoustic–linguistic emotion recognition has recently become a trend in SER research.

Anagnostopoulos et al. (2012) performed a survey on SER research conducted between 2000–2011. The survey paper focused on features and classification schemes used in the SER paper during that period. The strong point of this survey paper is the discussion about the different classification schemes used for SER, including their advantages and disadvantages. The authors argued the need for hybrid and ensemble classifiers since no single classifier has performed consistently for various tasks and datasets. In addition, the author highlighted the potency of using linguistic information along with acoustic information. At that time, the author proposed to utilize salient words to assess emotion. Nowadays, deep learning-based linguistic information has been adopted widely in bimodal SER instead of dictionary-based lexicon or keyword spotting mechanism.

Sailunaz et al. (2018) attempted to survey emotion detection from text and speech. However, instead of fusing acoustic (speech) and linguistic (text), the authors completed a survey within each modality, either speech emotion recognition or text emotion recognition. Both modalities are reviewed independently. The survey paper focused on the features used in speech and text emotion recognitions. The authors concluded the necessity for investigating acoustic features related to specific emotions and the necessity to improve text emotion recognition for emotionally implicit text.

Recently, Akçay and Oguz (2020) reviewed the advancements in the main building blocks of SER research: emotion models, datasets, feature, preprocessing methods, supporting modalities, and classifiers. The author distinguished between their paper and the previous survey paper in the availability of these building blocks. Compared to the previous studies, this survey paper is the complete one covering these SER building blocks. Hence, a comparison can be made between this paper and that paper (Akçay and Oguz, 2020).

In addition to the recent survey paper by Akçay and Oguz (2020), our paper focuses on the addition of linguistic information for improving SER performance. A new topic by the addition of linguistic information appears: fusion methods to combine acoustic information and linguistic information. Although the linguistic features and the fusion block are the only novelties of this survey paper from the SER building blocks point of view, we expand our coverage to review the emotion models, datasets, acoustic features, and classifiers employed by acoustic–linguistic emotion recognition research. This paper surveys more than 100 papers related to acoustic–linguistic emotion recognition from the first appearance (Lee et al., 2002) until the recent findings in 2021 (Santoso et al., 2021; Atmaja and Akagi, 2021). Similar to a survey paper conducted by Wu et al. (2014) that focused on audiovisual emotion recognition, we, to the best of our knowledge, do not see the same topic has been discussed for acoustic–linguistic (audiotextual) emotion recognition.

3. Datasets

The first main block of the bimodal acoustic–linguistic emotion recognition system is the dataset. Dataset is the raw (input) for the emotion recognition process. We introduce the following three datasets since these are the common ones used in bimodal acoustic–linguistic emotion recognition. In addition to serving as a starter kit on acoustic–linguistic SER research, the datasets can be used for benchmarking the performance of the various proposed SER methods.

1. IEMOCAP

IEMOCAP, which stands for interactive emotional dyadic motion capture database, contains dyadic conversations with markers on the face, head, and hands. The recordings thus provide detailed information about the actors' facial expressions and hand movements during both scripted and spontaneous spoken communication scenarios (Busso et al., 2008). Both speech data per dialogue and per sentence are available. The IEMOCAP dataset is freely available upon request, including its labels for categorical and dimensional emotion. The dataset provides both categorical and dimensional emotions. For categorical emotion, there are ten unique categories: neutral, happy, anger, surprise, fear,

disgust, frustration, excitement, and others. For dimensional emotion, there are three labels: labels: valence, arousal, dominance. The scores of the dimensional labels are the average scores of two evaluators. The dimensional emotion scores for valence, arousal, and dominance (VAD) are meant to range from 1 to 5 as a result of self-assessment manikin (SAM) evaluation. It should be noticed for bimodal emotion recognition researchers to thoroughly inspect the labels since the performance will depend on the reliability of these labels. For instance, the author stated that the range of dimensional labels is from 1 to 5, but it has been found that some labels are below 1 and above 5. These outliers must be treated before processing these labels in the SER system. It is also common for dimensional emotion to convert the 5-point scale to a floating-point value range [−1, 1] when they are fed to a classifier. The classifier is usually a deep neural network (DNN) system.

The total length of the IEMOCAP dataset is about 12 h, or 10 039 turns/utterances, from ten actors in five dyadic sessions (two actors each). The speech modality used to extract acoustic features usually is a set of files in the dataset with a single channel per sentence. The sampling rate of the speech data was 16 kHz. The manual transcription in the dataset without additional preprocessing is commonly used for the text data except for comparing it with automatic speech recognition (ASR) outputs.

2. MSP-IMPROV

MSP-IMPROV (Busso et al., 2017), developed by the Multimodal Signal Processing (MSP) Lab at the University of Texas, Dallas, is a multimodal emotional database obtained by applying lexical and emotion control in the recording process while also promoting naturalness. The dataset provides audio and visual recordings, while text transcriptions are obtained via ASR provided by the authors upon request. The annotation method for the recordings was the same as for IEMOCAP, i.e., SAM evaluation, with ratings by at least five evaluators.

The MSP-IMPROV dataset contains 8438 turns/utterances, including four scenarios: "Target-improvised", "Target-read", "Other-improvised", and "Natural-interaction". For bimodal acoustic–linguistic fusion, it may be necessary to remove lexical-controlled target sentences. These sentences are inside "Target-improvised" and "Target-read" scenarios. There are five categorical labels (anger, happy, neutral, sad, and others) and three-dimensional labels (valence, arousal, dominance) available in the dataset.

The speech data in the MSP-IMPROV dataset are available per-sentence audio file. The original sampling rate was 44.1 kHz with a single channel recording (mono). The audio bit rate was 705 kbps.

3. USOMS-e

Ulm State of Mind in Speech-elderly (USOMS-e) dataset is the corpus used in the elderly emotion sub-challenge in the INTERSPEECH 2020 computational paralinguistic challenge. The whole dataset subset is recorded with 87 subjects aged 60–95 years; 55 of the subjects were male, and the rest 32 were female. The dimensional emotion labels were given in valence and arousal divided into three categories: low, medium, and high.

Table 1 shows different properties of the common datasets used for acoustic–linguistic emotion recognition, including USOMS-e dataset. For the USOMS-e dataset, labels are given per each story (long utterance). The label on the dataset is given for valence and arousal (VA) on both alphabetic and numeric symbols, i.e., low ('L' or '0'), medium ('M' or '1'), and high ('H' or '2'). The original baseline paper (Schuller et al., 2020) chose alphabetic labels between both. Since the duration for each story is long, the authors provided chunks as smaller segments of utterances of five seconds. Note that the number of chunks is different for each story. For instance, there are 34 chunks in the first story and 46 chunks in the second story.

4. Other datasets

Although we only describe the three most common datasets used in bimodal acoustic–linguistic emotion, there are other datasets that have been explored for bimodal SER research. Table 2 lists those known datasets used in acoustic–linguistic emotion recognition. Note that if no information available in the reference paper, we set it as "No", e.g., the transcription and the availability of the dataset. The "Used in" column refers to examples of research papers in which the dataset was used. The minimum criterion for a dataset to be included in the list is used in a reference considering linguistic information for SER.

The table shows the domination of certain characteristics of the dataset over others. First, English language datasets are more available than any other language due to their nature. This low availability of non-English datasets raises a necessity to build SER datasets in other languages, particularly for acoustic–linguistic fusion purposes. Second, categorical emotion dominated the model of emotion, among other emotion descriptions. This emotion model's availability leads to more research in categorical than other models (which will be shown later). Finally, most of the dataset is still closed for the public. It means no access to the dataset or no information is provided to access the dataset. It is essential to open the dataset for accelerating SER research while keeping confidential issues (e.g., private data of respondents).

4. Emotion models

The second building block of SER is the emotion model. The choice of the emotion model is mostly based on the availability of emotion labels in the dataset. However, it is important to choose an emotion model beyond the availability of the labels. According to Grandjean et al. (2008), there are at least three views to model humans' emotions: categorical emotion, dimensional emotion, and componential appraisal emotion. The following description describes these models used actively in affective computing research. The purpose of introducing these three models is to provide a unified view of emotion from the psychological side. Knowing the three models, researchers in speech communication could understand the current models' limitations, improve the current models, or incorporate the models into their SER research.

4.1. Categorical model

Categorical emotions, also known as basic emotions, are discrete emotions that are independent of each other in their manifestations. Although the original idea is to organize affective states into their emotion families (rather than discrete emotions); however, most researchers agree that there are six basic emotions: anger, fear, enjoyment, sadness, disgust, and surprise. The first five emotions are backed by robust and consistent evidence, while the evidence for the surprise is not as firm (Ekman, 1992). Nevertheless, these six basic emotions have been standard in categorical emotions.

Before Ekman coined the terms of basic emotions, Plutchik and Kellerman (1980) have defined basic eights bipolar emotions: joy (reproduction), sorrow/sadness (deprivation), acceptance/trust (incorporation), disgust (rejection), surprise (orientation), anticipation (exploration), anger (destruction), fear (protection). These eight emotions can be illustrated as a wheel of emotion. Each emotion can mix with other emotions to make up another emotion, as mixing colors.

Instead of six, recent research suggests that four latent expressive patterns were commonly observed in facial expressions (Jack et al., 2016). However, instead of mentioning the name of basic emotions, the research utilized the term basic "action unit pattern" (AU Pattern), from one to four. Although backed by scientific evidence, this finding did not have any practical implementation yet.

Ekman revised the characteristics which distinguish basic emotion from 9 criteria (Ekman, 1992) to 11 criteria (Ekman, 2005). The new criteria resulted in 15 emotions: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame. Du et al. (2014) shows 21 categories of facial expressions by a

Table 1
Different properties of the common datasets used in acoustic–linguistic emotion recognition.

	IEMOCAP	MSP-IMPROV	USOMS-e
# samples	10039	8438	261 stories (7778 chunks)
# speakers	10	12	87
# hours	≈12	>9	≈10.8
Emotion model	Categorical + VAD	Categorical + VAD	VA [L, M, H]
Transcription	Manual	ASR	Manual + ASR
Characteristic	Acted	Acted, naturalness	Naturalness, elderly people

Table 2
Speech datasets for acoustic–linguistic emotion recognition.

Dataset	Language	Emotion description	Transcription	Reference (Year)	Used in	Availability
Fermus III	English	Categorical	No	Schuller (2002)	Schuller et al. (2004) and Rigoll et al. (2005)	No
ITSPOKE	English	Negative, Neutral, Positive	No	Litman et al. (2004)	Litman and Forbes-Riley (2004, 2006)	No
Emo-DB	German	Categorical	No	Burkhardt et al. (2005)	Schuller et al. (2005, 2008)	Publicly available
IEMOCAP	English	Categorical, Dimensional	Yes (Manual)	Busso et al. (2008)	Atmaja and Akagi (2020)	By request
VAM	German	Categorical, Dimensional	No	Grimm et al. (2008)	Schuller (2011) and Grimm et al. (2007)	By request
FAU AIBO	German	Categorical	Yes	Steidl (2009)	Metze et al. (2009) and Polzehl et al. (2011)	No
UAH Corpus	Spanish	Categorical	No	Callejas et al. (2011)	Griol et al. (2019)	No
SEMAINE	English	Categorical, Dimensional	Yes	McKeown et al. (2012)	Schuller et al. (2012)	By registration
EMOV	English	Valence, Arousal	No	Karadogan and Larsen (2012)	Karadogan and Larsen (2012)	No
Image Description	Spanish	Categorical	No	Griol and Molina (2015)	Griol et al. (2019)	No
Let us Go	English	Categorical	Yes	Griol et al. (2016)	Griol et al. (2019)	No
MSP-IMPROV	English	Categorical, Dimensional	Yes (ASR)	Busso et al. (2017)	Atmaja and Akagi (2021)	By request
IDEC	Indonesia	Categorical	No	Kurniawati et al. (2017)	Kurniawati et al. (2017)	No
Call center data	English	Negative, Neutral, Positive	Yes (ASR)	Cho et al. (2018)	Cho et al. (2018)	No
CMU-MOSEI	English	Categorical	Yes (ASR)	Zadeh et al. (2018)	Khare et al. (2020)	Publicly available
MELD	English	Categorical	Yes	Poria et al. (2019)	Ho et al. (2020)	Publicly available
MSP-Podcast	English	Categorical, Dimensional	No	Lotfian and Busso (2019)	Pepino et al. (2020)	By request
SEWA DB	Six languages	Valence, Arousal, Liking	Yes	Kossaifi et al. (2019)	Tzirakis et al. (2021)	By registration
USOMS-e	German	Valence, Arousal	Yes (ASR+Manual)	Schuller et al. (2020)	Soğancıoğlu et al. (2020)	By request

facial action coding system analysis. Furthermore, Cowen and Keltner (2017) found 27 emotional experiences from facial expression across self-report methods. The growth of the number of categorical emotions based on facial expression measurements confirms the high variability of humans' expressed emotions. Darwin argued that the biological category, including the emotion category, does not have an essence; it is hard to map one-to-one facial expressions to emotional states.

4.2. Dimensional model

Instead of dividing emotion into several categories, a dimensional emotion views emotion as continuous values/degrees of attributes in valence-arousal space (2D) or valence-arousal-dominance (3D) space. Valence is the degree of positive or negative emotion, arousal refers to the level of activation from sleepiness (low) to awakeness (high), and dominance is the degree of control over the emotion (Gunes and Pantic, 2010). In this theory, an emotion or affective state is not independent of one another. Rather, they are related one to another in a systemic manner (in 2D or 3D space). Russell (1980) argued that the previous emotion categories could be mapped within 2D valence-arousal space. An illustration of VA space with several emotion categories is shown in Fig. 2.

The search for higher dimensions for dimensional emotion is a worthwhile study. Mehrabian and Russell (1974) developed three numerical dimensions, VAD, to assess environmental perception, experience, and psychological responses. However, Fontaine et al. (2017) have found that the world of dimensional emotion is not two or three dimensions, but four dimensions. The fourth dimension is unpredictability. In order of importance, the order of dimensional emotions is valence, dominance, arousal, and predictability. Fortunately, five years before that study was published, an emotion recognition challenge that involved four-dimensional emotions was held in 2012 (Schuller et al., 2012). In this challenge, four-dimensional emotions are arousal, expectancy, power/dominance, and valence. Expectancy, which represents the predictiveness of the subject's feeling, is very similar to predictability/unpredictability in the latter report.

4.3. Appraisal model

The third emotion model, the hybrid model or appraisal model, can be viewed as an extension of the dimensional model. In this model, emotion categories are spanned between bipolar dimensions. For instance, "impatience" is located in the upper part of the arousal axis (Scherer, 2005). This study of appraisal-based emotion theory leads to the development of the Geneva emotion wheel (GEW) rating study. This hybrid model has two similarities with the previous 4D dimensional emotion model. First, the hybrid model also uses four attributes/dimensions: valence, dominance/power, arousal, and conducive/obstructive (instead of predictiveness). Second, in version 2.0 of GEW, two axes used to draw emotion terms are valence and dominance/power, which are the two most important emotional attributes according to Fontaine et al. (2017). In version 3.0, the model used the simpler words with more degrees for each emotional word (six degrees instead of five degrees in version 2.0). Nevertheless, the use of the hybrid model in SER currently is not familiar in the SER research community, perhaps due to these labels' availability in the dataset. On another side, all SER research employed either first, second, or combination of both views as target emotion.

Aside from three models of emotion, there is a recent study that suggests that emotion annotation – which represents the emotion model – is ordinal by nature (Yannakakis et al., 2021). We saw that this approach is close to the appraisal model, which extends the dimensional model to a categorical model with different orders. While this approach is not intended to model the emotion, it is a good way to annotate emotions for both categorical and dimensional models. We encourage the future speech emotion dataset builder to follow this approach, i.e., to use ordinal annotation by asking annotators to rank a preference of emotion among options (e.g., via SAM evaluation).

5. Features

The input feature to the SER system is the most important issue for developing bimodal information SER. If the input is not informative for

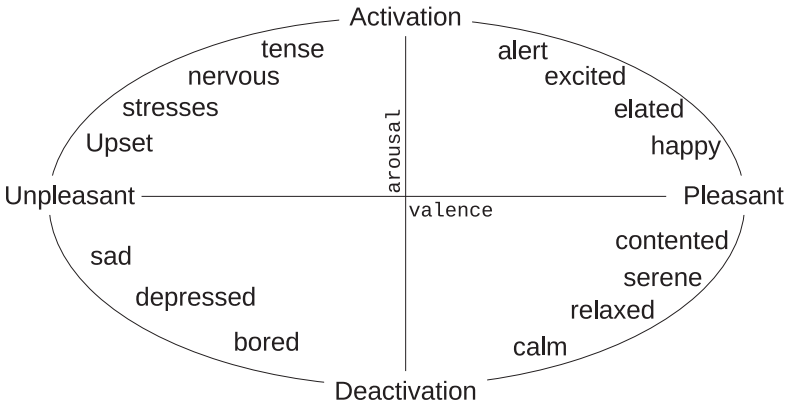

Fig. 2. Graphical representation of circumplex model (2D valence-arousal space).

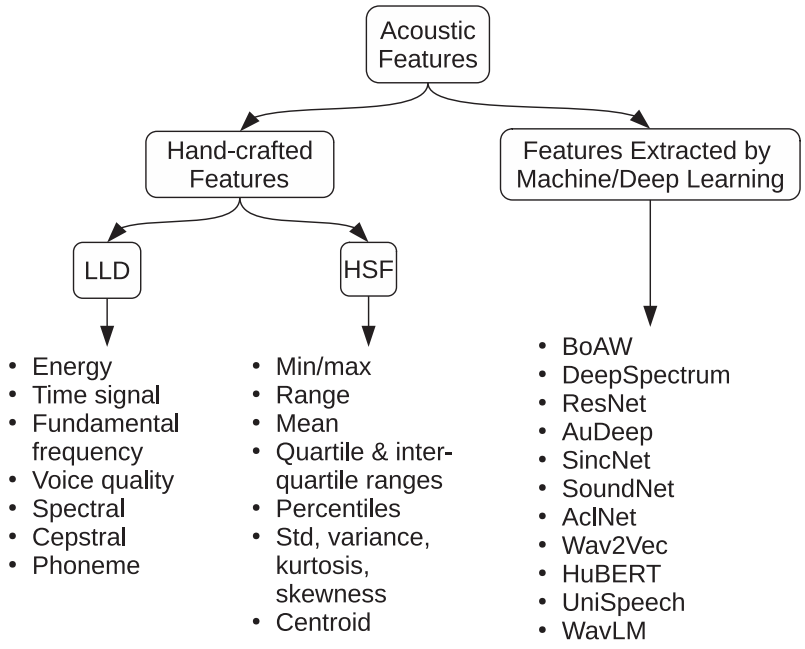

Fig. 3. Classification of acoustic features for SER.

predicting emotion or does not correlate with the predicted emotion, the prediction results will suffer from the low performance. In principle: garbage in, garbage out. The following classifications are useful features for SER from acoustics and linguistics. The classification is based on the criteria of the extraction process, whether it is conducted manually via formulation or physical modeling or generated by model or data-driven learning.

5.1. Acoustic features

The correlation of acoustic features with emotion has been studied for many years (Scherer, 2005; Mairano et al., 2019). The main classification of acoustic features for SER is the classical and modern approaches, i.e., handcrafted features vs. deep learning-based features. Classical handcrafted features employed acoustic features extracted per frame. These features are often called local features or low-level descriptors (LLDs). On the other hand, statistical features computed from LLDs are new ways to capture the dynamics among frames. The features generated by this latter feature extraction method are called global features, suprasegmental features, high-level features, or high-level statistical functions (HSFs). Fig. 3 shows the classification of acoustic features for SER.

Eyben et al. (2010) divided LLD and HSF into five groups: signal energy, fundamental frequency (perception: pitch), voice quality, cepstral, time signal, and spectral. Prosodic features (f_o, duration, intensity, voice quality) have been known to have a strong correlation with emotion from a psychology point of view (Frick, 1985; Mozziconacci, 2002; Liebenthal et al., 2016). In acoustics, prosody is implemented into several acoustic features, including LLD and HSF. Väyrynen (2014) made a distinction between prosodic and acoustic (non-prosodic) features. His study reported that a combination of prosodic and acoustic features achieved performance comparable to human reference on basic emotion recognition.

Both Lee et al. (2002) and Schuller et al. (2004) employed f_o and energy-based acoustic features for the SER task. The former applied both LLD and HSF of f_o and energy features, while the latter only applied HSF of f_o and energy features. The latter reference found that pitch-based features correlated with the performance of SER more than energy-based features.

As a 'default' feature on most ASR systems, MFCC has been explored for the SER task. Metze et al. (2009) has found that MFCC is the most informative acoustic feature compared to other evaluated acoustic features. Tripathi et al. (2019) found that MFCC performed better than spectrogram features on unimodal acoustic SER.

The shift from MFCC to mel filterbank (MFB) features in ASR motivates SER researchers to adopt a similar direction. Aldeneh et al. (2017) extracted 40 MFB features for dimensional SER tasks on the IEMOCAP dataset. Zhang et al. (2019) employed a similar MFB with 40-dimensional with z-normalization on categorical IEMOCAP and MSP-IMPROV datasets. Both kinds of research showed fair performances (50%–65% accuracy) of MFB features for the SER task.

Phoneme, the smallest unit of speech, has been investigated to be useful for the SER task. Zhang et al. (2019) furthermore combined MFB with phoneme for the same SER task. A combination of phoneme with MFB outperforms MFB-only of phoneme-only input features. Yenigalla et al. (2018) combined phoneme embedding with a spectrogram. The phoneme embedding is generated from the word2vec model (Mikolov et al., 2013) and IEMOCAP speech data. The combination of phonemes with spectrogram achieves the highest accuracy among individual features.

Since most classifiers in modern SER systems have used deep learning methods, it is reasonable to extract an acoustic representation of speech in an end-to-end manner via deep learning methods. In INTERSPEECH 2020 ComParE challenge, two deep learning-based features were given in the baseline system, DeepSpectrum and AuDeep. The provided DeepSpectrum features with ResNet50 network achieved the highest unweighted average recall (UAR) on the elderly emotion sub-challenge test set. Although there is a movement to use DNN-based feature extraction, the majority of SER research still relies heavily on handcrafted acoustic features.

5.2. Linguistic features

Since this paper surveys fusion of acoustic and linguistic information for SER, it is necessary to introduce the common linguistic features used in text processing. Linguistic features are the realization of linguistic information. It is also called text features, textual features, lexical features, language features, or semantic features. Aside from different meanings of linguistic and lexical terms in information processing, i.e., language vs. word meaning, these terms also have a different meaning in book/article writing, particularly the term "text features". In book/article writing, the text features include writing components such as a glossary, bold typeface, title, headings, captions, and labels. In information science, text or linguistic features are features extracted from written or spoken text. Thus, the term linguistic features is a preferable term to text features to avoid confusion among readers.

Linguistics features used in emotion recognition represent numerical values related to the emotional states in a word. The simplest way to build linguistic features for emotion detection is emotional keyword spotter (Chuang and Wu, 2004). In this framework, every word is assumed to have a correlation with emotion categories. For instance, the word "disappointed" can be represented as [(2, 0.2), (3, 0.6)] where 2 represents "angry" emotion and 3 represents "sadness" emotion. Both 0.2 and 0.6 represent degrees of emotion's intensity. This emotional keyword spotter can be expanded into an emotional phrase spotter (Schuller et al., 2004).

The first systematic linguistic representation of a document, perhaps, is TF–IDF (term-frequency inverse document frequency). TF is defined as the frequency of a word in a particular document/utterance. IDF is defined as a logarithm of the total number of documents' ratio to the total number containing that word. TF–IDF is the multiplication of TF with IDF.

Bag-of-Words (BoW) is a numerical feature vector to represent "words in a bag". First, a fixed integer is assigned to each word occurring in any document, i.e., building a dictionary from a corpus by assigning a word to integer indices. Second, count the number of occurrences of each word and store it as the value of feature j where j is the index of word w in the dictionary (Pedregosa et al., 2011). These BoW features can be expanded for acoustic and visual modalities (BoAW and BoVW).

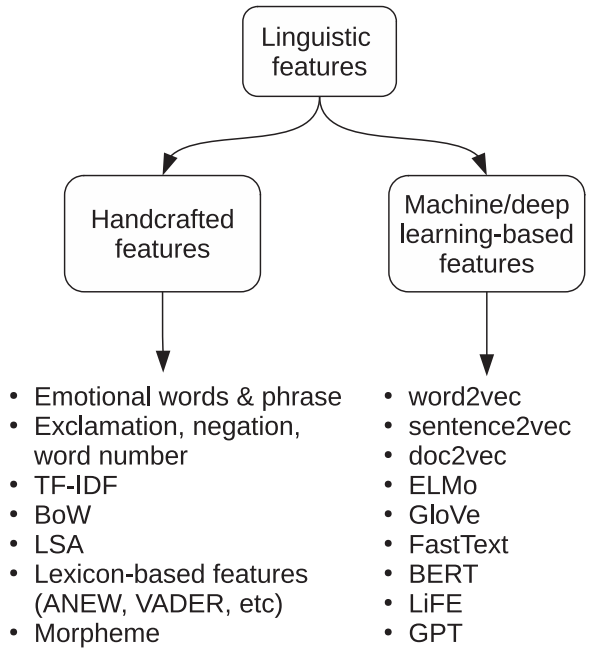

Fig. 4. Classification of linguistic features used in SER.

Several lexicon dictionaries have been developed to inform the 'emotion score' of emotional words. These dictionaries include DAL (Whissell, 2009), ANEW (Warriner et al., 2013), VADER (Hutto and Gilbert, 2014), and NRC (Mohammad, 2018). Using these dictionaries allows direct measurement of emotional words in the given utterances. For instance, the word "arose" has values of 2.11, 2.00, and 1.40 for pleasantness, activation, and imagery. These values are on a 3-point scale; different dictionaries have different scales.

The search for vector representation from a word led to the research of word embedding or word vector. In this approach, a deep neural network is used to train a large corpus (i.e., a Wikipedia corpus) to generate word vectors based on an algorithm. This approach has resulted in a new paradigm in the vector representation of linguistic information of a word. Several models exist, including word2vec, GloVe, FastText, and BERT.

Fig. 4 shows the classification of linguistic features used in the SER task. In contrast to acoustic features, there is a tendency to move to deep learning-based features from handcrafted features. The chosen linguistic feature is usually based on the complexity of the task and the size of the data.

5.2.1. Word embedding

A classifier needs a set of input features to model input–output relation. One of the common features used in text processing is word embedding (WE). A word embedding is a vector representation of a word. A numerical value in the form of a vector is used to make the computer to be able to process text data since it only processes numerical values. This value is the points (numeric data) in the space of a dimension, in which the size of the dimension is equal to the vocabulary size. The word representations embed these points in a feature space of lower dimension (Goodfellow et al., 2015). A one-hot vector represents every word; a value of 1 corresponds to this word and 0 for others. This element with a value of 1 will be converted into a point in the range of vocabulary size.

To obtain a vector of each word in an utterance, first, this utterance in the dataset must be tokenized. Tokenization is the process of dividing an utterance by the number of constituent words. For example, the text "That's out of control" from the IEMOCAP dataset will be tokenized as ["That's", "out", "of", "control"]. Suppose the number of vocabulary

is 2182 (number of words in IEMOCAP dataset with six emotion categories), then the obtained word vector is something similar to

$$text\ vector = [42, 44, 11, 471].$$

An embedding layer will convert those positive fixed integers into dense vectors of fixed size. For instance, each 1-dimensional word vector in the utterance will be converted into 2-dimensional dense vector,

$$[42, 44, 11, 471] \rightarrow$$
$$[[0.12, 0.3], [0.12, 0.29], [-0.54, 0.2], [0.71, 0.23]].$$

The higher dimensions are used to obtain a better representation of a word vector. A number of 50-, 100-, and 300-dimensional vectors are commonly employed to build pre-trained word vectors from a large corpus.

A set of zeros can be padded in front of or behind the obtained vector to obtain the fixed-length vector for all utterances. The size of this zero-sequence can be obtained from the longest sequence, i.e., an utterance within the dataset that has the longest words, subtracted by the length of a vector in the current utterance.

5.2.2. Pre-trained word embeddings

A study to vectorize certain words has been performed by several researchers (Mikolov et al., 2013; Pennington et al., 2014; Mikolov et al., 2019). The vector of those words can be used to weigh the word vector obtained previously. The following word embedding techniques are commonly used in research on speech emotion recognition involving linguistic information.

word2vec

The classical word embedding paradigm used unsupervised (handcrafted) learning algorithms such as LSA, n-gram, and similar methods. Due to advancements in neural network theory supported by computer hardware's speedup, word vector search shifted to deep learning-based algorithms. Mikolov et al. (2013) developed word representation using the so-called word2vec (word to vector) using a neural network language model trained in two steps. First, continuous word vectors are learned by using a simple model, and then the n-gram neural net language Model (NNLM) is trained on top of these distributed representations of words (Mikolov et al., 2013). Two new model architectures are proposed to obtain a word vector: the Continuous-Bag-of-Word (CBoW) architecture to predict the current word based on the context and the skip-gram architecture to predict surrounding words given the current word.

From those two approaches, skip-gram was founded as an efficient method for learning high-quality distributed vector representations that capture precise syntactic and semantic word relationships (Mikolov et al., 2013). The objective of the skip-gram model is to maximize the average log probability,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<j<c, c\neq 0} \log p(w_{t+j}|w_t), \tag{1}$$

where T is the number of words in a sequence and c is the size of the training context (which can be a function of the center word w_t). Larger c results in more training examples and can lead to higher accuracy at the expense of the training time. The basic skip-gram formulation of $p(w_{t+j}|w_t)$ can be defined using the softmax function, and computational efficiency can be approached by a hierarchical softmax (Mikolov et al., 2013).

GloVe

Pennington et al. (2014) combined global matrix factorization and local context window methods for learning the space representation of a word. In the GloVe (Global Vectors) model, the statistics of word occurrences in a corpus is the primary source of information available to all unsupervised methods for learning the word representations. Although several methods exist, the question remains as to how meaning is generated from these statistics and how the resulting word vectors might represent that meaning. Glove captured the global statistics from the corpus, for example, a Wikipedia document or a common crawl document.

In the GloVe model, the cost function is given by

$$\sum_{i,j=1}^{V} f(X_{i,j})(u_{i,j}^T v_j + b_i + c_j - \log X_{i,j})^2, \tag{2}$$

where:

- V is the size of the vocabulary,
- X denotes the word co-occurrence matrix (so $X_{i,j}$ is the number of times that word j occurs in the context of word i)
- the weighting f is given by $f(x) = (x/x_{\max})^\alpha$ if $x < x_{\max}$ and 1 otherwise,
- $x_{\max} = 100$ and $\alpha = 0.75$ (determined empirically),
- u_i, v_j are the two layers of word vectors,
- b_i, c_j are bias terms.

In a simple way, GloVe is a weighted matrix factorization with the bias terms.

FastText

Mikolov et al. (2019) improved word2vec CBoW model by using some strategies, including subsample frequent words technique. This modification of word2vec is trained on large text corpora such as news collection, Wikipedia, and web crawl. They named the pre-trained model with that modification as FastText. The following probability p_{disc} of discarding a word is used by FastText to subsample the frequent words:

$$P_{disc}(w) = 1 - \sqrt{t/f_w}, \tag{3}$$

where f_w is the frequency of the word w, and t is a parameter > 0.

FastText also counts the classical n-gram word representation by enriching the word vector with a bag of character n-gram vectors learned from a large corpus. In this computation, each word is decomposed into its character n-grams N, and each n-gram n is represented by a vector x_n. The new word vector is then simply the sum of both representations,

$$v_w + \frac{1}{|N|}\sum_{n\in N} x_n, \tag{4}$$

where v_w is the old word vector. The set of n-grams N is limited to 3 to 6 characters in practical implementation.

BERT

The previous aforementioned word embeddings – word2vec, GloVe, FastText – generates word representations in a context-free model. It means the same word that appears in a different phrase has the same word representation, e.g., word "book" in "mathematics book" and "book a hotel". Instead of using a context-free model, BERT (Bidirectional Encoder Representations from Transformers) was built upon pre-training contextual representation (Devlin et al., 2018).

BERT is different in many ways from its predecessors. Apart from contextual representation, the main contribution of BERT is to employ bidirectional pre-training for language representation. Unlike its predecessors, which model languages in a unidirectional way, i.e., from left to right as a writing/reading system, BERT used two unsupervised tasks for pre-training models. The first task is the masked language model; the second task is the next sentence prediction (NSP). The BERT model's dimension for each word depends on the number of hidden layers used in the architecture. This number is either 768-dimensions for the base model or 1024-dimensions for the large model.

Apart from the pre-trained model, BERT provides a fine-tuning model. Fine-tuning allows BERT to model several tasks, single or text

pairs, by swapping out the corresponding inputs and outputs. Fine-tuning can be seen as adjusting the pre-trained model according to the context, i.e., the dataset. Hence, fine-tuning can only be done after obtaining the pre-trained model and is relatively expensive. Fine-tuning is suitable for a specific task, like SER, rather than general linguistic tasks.

6. Classifiers

This section reviews the four most used classifiers in speech emotion recognition. One is a machine learning classifier, i.e., a support vector machine (SVM). Others are three deep learning classifiers, i.e., multilayer perceptron (MLP), convolutional neural networks (CNN), and long short-term memory (LSTM) neural networks. The brief descriptions of these classifiers are given below.

6.1. SVM

SVM is a useful machine learning classifier for, generally, small datasets. For categorical emotion recognition, SVM applies acoustic or linguistic features for the given labels. This SVM applied to the classification task is called support vector classification (SVC). For dimensional emotion recognition, SVM applies regression analysis to map them to the given scores. This SVM for regression task is called support vector regression (SVR).

SVM can accept unimodal or multimodal inputs. In bimodal emotion recognition from acoustic and linguistic information, SVM can be utilized in two-stage scheme for evaluation of the emotion recognition system from DNNs outputs. In bimodal information fusion, each prediction from the acoustic and text networks is fed into the SVM. From two values (e.g., valence predictions from the acoustic and text networks), the SVM learns to generate a final predicted degree (e.g., for valence).

6.2. MLP

MLP is a classical feedforward neural network that projects input data into linearly separable space using non-linear transformation. A hidden layer is an intermediate layer between inputs and outputs, containing many perceptrons (also called units or nodes). An MLP commonly refers to more than one hidden layer. The MLP used in most SER tasks is similar to the definition of connectionist learning proposed by Hinton (1989). A deeper layer MLP usually consists of many layers to enable deep learning hierarchically. This neural network architecture is also known as dense networks or fully connected (FC) networks.

MLP is powerful for combining acoustic and linguistic network for network concatenation. Mathematically, the fusion of acoustic and linguistic information using MLP could be formulated as in Eq. (5),

$$f(y) = W_2\ g([W_{1a}^{\top} xa + b_{1a}; W_{1l}^{\top} xl + b_{1l}]) + b_2. \tag{5}$$

Here, $f(y)$ denotes the output of the corresponding layer; W_1, W_2 denote the weights from previous layers (a: acoustic; l: text), i.e., dense layer after LSTM for each network, and the current hidden layer, respectively; x_a and x_l are the acoustic features and word embeddings, respectively; b is a bias; and g is an activation function.

Schuller et al. (2004) utilized MLP for combining acoustic and linguistic information. Their evaluation using MLP showed lower errors than a fusion method by means of logical "OR". Griol et al. (2019) compared baseline majority-class method to MLP for evaluating the effect of context on categorical SER task. The result shows that MLP outperforms the baseline method in six out of eight scenarios. Zhang et al. (2019) used MLP in all experiments involving acoustic features, phoneme, and combination of both; MLP showed its effectiveness on both single-stage and multi-stage SER tasks.

6.3. CNN

CNN is a class of neural networks that contains convolutional layers. Convolution is a mathematical operation between two functions by measuring the overlap of both when one function ("input") is flipped and shifted by another function ("kernel"). The resulting output, which is the goal of a convolution layer, is a feature map. This convolution operation is similar to cross-correlation; cross-correlation does not flip the second function. Convolution is also can be seen as cross-correlation with a scalar bias. In deep learning literature, the convolution terminology views cross-correlation as convolution since many deep learning frameworks did not take bias into account by default.

The convolutional network is often applied to image-like data. Time-series data, including acoustic features in vectors, can be fed into convolutional networks using 1-dimensional (1D) CNN. To take the most benefit of CNN, spectrogram and MFB features are frequently used as input to the SER system. For text processing, the main idea for CNN is to compute vectors for n-grams (e.g., 2-, 3-, and 4-g) and group them afterward. CNN is commonly used for both speech and language processing.

Apart from convolutional layers, CNN typically still needs a fully-connected layer (FC or MLP). The feature map as the output of the convolution layer is fed into MLP to obtain desired outputs. Although recently it has been found unnecessary (Springenberg et al., 2014), a CNN commonly uses pooling layers after convolutional layers for mitigating and reducing spatial representation (Zhang et al., 2020).

Yenigalla et al. (2018) experimented with CNN for categorical SER by inputting phoneme, spectrogram, and combination of both. The combination of both phoneme embeddings and spectrogram achieved the highest performance. The architecture of each phoneme and spectrogram network was convolution layer, max pooling, and FC layer. Both networks are concatenated with an FC layer to obtain the outputs.

Instead of phoneme and spectrogram, Huang et al. (2018) proposed to use bag-of-audio-words for the input of the CNN-based SER system. The architecture was similar to Yenigalla et al. (2018), i.e., convolution, pooling, and FC layer. The result shows that the use of BoAW outperforms raw acoustic features.

Cho et al. (2018) combined acoustic and linguistic information for categorical SER; the acoustic inputs used an LSTM network, while the linguistic inputs used a multi-resolution CNN. A multi-resolution CNN is utilized to predict categorical emotion given the utterance by employing word embedding, convolution layer, and global mean pooling. The combination of acoustic network with LSTM, linguistic network with CNN, and emotion vector (e-vector) is fed into SVM and achieved the highest performance compared to unimodal results.

While most bimodal SER research used CNN for linguistic (due to image-like data) and LSTM for acoustic (due to time series data), Sebastian and Pierucci (2019) proposed the opposite, LSTM for text and CNN for speech. The CNN architecture contains two convolution layers and two FC layers. The acoustic features are 6373 features extracted using INTERSPEECH 2013 ComParE feature set (ComParE2013). The linguistic features are word embeddings pre-trained with FastText embeddings. In this case, the performance of CNN-based text emotion recognition is the lowest among other models, while the combination of early and late fusions topped the performances.

Cai et al. (2019) combined a CNN architecture with bidirectional LSTM and attention layer for acoustic emotion recognition. The improved dual-channel architecture was called CNN-Bi-LSTM-Attention (CBLA) model. On both unimodal and multimodal (with BLSTM for linguistic emotion recognition), CBLA outperforms an MLP model by considering both global and temporal information in the data.

6.4. LSTM

Long Short-Term Memory (LSTM) neural networks is an extension of a recurrent neural network. The idea of using LSTM networks comes from an approach that human has the persistence to keep memory long in a short-term period. Humans do not begin their thought from scratch every second. When reading a paper, a reader understands each word based on the understanding of the previous words. Humans do not throw everything away and begin thinking again from scratch. The humans' thoughts have persistence. LSTM has networks in the loop to allow information to persist as in humans' thoughts (Olah, 2015).

Three gates are introduced in LSTMs: the input gate ($\mathbf{I}_t$), the forget gate ($\mathbf{F}_t$), and the output gate ($\mathbf{G}_t$). In addition to that, there are memory cells that take the same shape as the hidden state. A memory cell is just a fancy version of a hidden state, custom-engineered to record additional information.

LSTM has dominated the classifier used in both ASR and SER. Although the use of LSTM over CNN for emotion recognition task has been challenged (Schmitt et al., 2019), the opposite also applies (Macary et al., 2020). Since the data (i.e., input features) are sequence, a recurrent neural network is a straightforward way to process these data. In addition, LSTM is able to model long-range context in emotional features to map it with the emotional labels. Tian et al. (2015) used the LSTM classifier to build hierarchical neural networks.

Instead of unidirectional LSTM, bidirectional LSTM (BLSTM) was utilized to learn information both from the past and the future inside the network (unidirectional LSTM only learns from the past). In Cai et al. (2019), BLSTM is used for the textual network rather than the acoustic network. This bidirectional LSTM is often combined with an attention model to boost the performance of the SER task (Atmaja and Akagi, 2019). However, using BLSTM doubles the model's complexity, making the model may not be suitable for real-time applications.

6.5. Other classifiers

While research on speech processing area focuses on the acoustic features correlated with emotion in speech, the research in machine learning and artificial intelligence (AI) focuses on the AI architecture that suits emotion recognition task. As stated previously, there are debates over the most suitable architecture for the emotion recognition task. Current results show that recurrent neural network (RNN)-based architectures, like LSTM and GRU, has dominated and has been used in production for acoustic–linguistic emotion tasks. However, there is a trend to use the more advanced architectures like domain adversarial neural network (DANN) (Lian et al., 2020), generative adversarial network (GAN) (Chang and Scherer, 2017), conditional adversarial auto-encoder (CAAE) (Kim et al., 2020) and attention-based neural network.

Among them, the bimodal SER trend shows more adaptation of attentive neural networks among other architectures. In the following sections, the benefit of employing the attention-based neural networks is highlighted to fuse acoustic and linguistic information at the feature level (extracted using DNNs). Nevertheless, also in the next section and in the later section, DANN architecture achieved the state-of-the-art (SOTA) for the IEMOCAP dataset.

7. Fusion methods

Multimodal fusion in technology is the combination of information that comes from different sources. This terminology is similar but different to human multimodal perception. In human multimodal perception, the information comes from different sensors (sensory organs); this requirement is not necessary for technology. Multimodal fusion can be viewed as multisensor data fusion. In this terminology, the 'sensor' is the soft sensor. Acoustic and linguistic feature extractors can be regarded as soft sensors in bimodal acoustic–linguistic fusion. While Ref. Poria et al. (2017) defines multimodality as the presence of more than one modality or channel, we encourage to call the fusion of two modalities bimodal for the sake of clarity.

Fusing acoustic and linguistic features has been attempted at the early stage of speech emotion recognition research. The first work on fusing acoustic with linguistic information has been performed by Lee et al. (2002) by combining acoustic and language features at the decision level using logical "OR". If at least one decision corresponds to a specific emotion, then the result is this specific emotion. This earliest work only involved negative and non-negative emotion categories.

The linguistic information used for fusion is extracted from the text. Hence, the feature is also called text/textual features. In the early text processing research, the linguistic feature is extracted by hand, like term frequency–inverse document frequency (TF–IDF) or bag-of-word (BoW). Nowadays, the extraction of linguistic features is done automatically by training large datasets, e.g., Wikipedia. This process resulted in pre-trained word vectors like word2vec, GloVe, FastText, ELMo, and BERT, as explained previously. Emotional words or word values from affective lexicon dictionaries are also often used to represent linguistic information. These lexicon-based features are commonly utilized when the fusion of acoustic and linguistic information is aimed at recognizing dimensional emotion.

Fusing acoustic and linguistic information for SER can be accomplished in several ways. Fig. 5 shows the classification. Early fusion combines acoustic and linguistic information at the feature level; late fusion combines results from acoustic and linguistic information at the decision level. Early fusion, furthermore, can be split into three main categories: feature concatenation, networks/model concatenation, and hierarchical model. Hierarchical model, as proposed in Majumder et al. (2018) and Tian et al. (2016), can be regarded as early fusion since the method fuses features at a different level of layers, not at the decision level.

7.1. Early fusion approach

7.1.1. Feature concatenation

The simplest fusion of acoustic and linguistic information is by feature concatenation [Fig. 5(a)]. In this scheme, both acoustic and linguistic features are concatenated and are fed into the same networks; a single model received two input features (acoustic features and linguistic features). Hazarika et al. (2018) combined acoustic and linguistic information at feature level with a self-attention mechanism. They obtained a significant improvement of accuracy from 62.5% to 72.2% on the IEMOCAP dataset. Similarly, Atmaja et al. (2020) improved the accuracy of valence prediction from 49% to 56.3% on USOMS-e dataset using an acoustic–linguistic feature concatenation method. Recent papers show that instead of concatenating handcrafted features, an attention-based network is adopted to concatenate hidden representations (output of hidden layer) extracted from DNNs of acoustic and linguistic data (Hazarika et al., 2018; Priyasad et al., 2020).

7.1.2. Model/network concatenation

A step further in the early fusion of acoustic and linguistic information is by concatenating models or networks [Fig. 5(b)]. In this scheme, each modality has different models: acoustic network and linguistic network. For example, LSTM is used for linguistics, while CNN is used for acoustic. A concatenation layer is then added to the top of these networks. Atmaja and Akagi (2020) fused acoustic and linguistic at model level and improved dimensional SER performance from the highest performance of any single modality. This model concatenation is also the most used fusion type for fusing bimodal acoustic and linguistic information found in INTERSPEECH 2020 papers.

A typical model of network or model concatenation is shown in Fig. 6. LSTM is often used for both acoustic and linguistic classifiers due to its nature; an utterance is a sequence of (spoken) words. CNN is used for the speech classifier if the input is spectrogram-like data, including

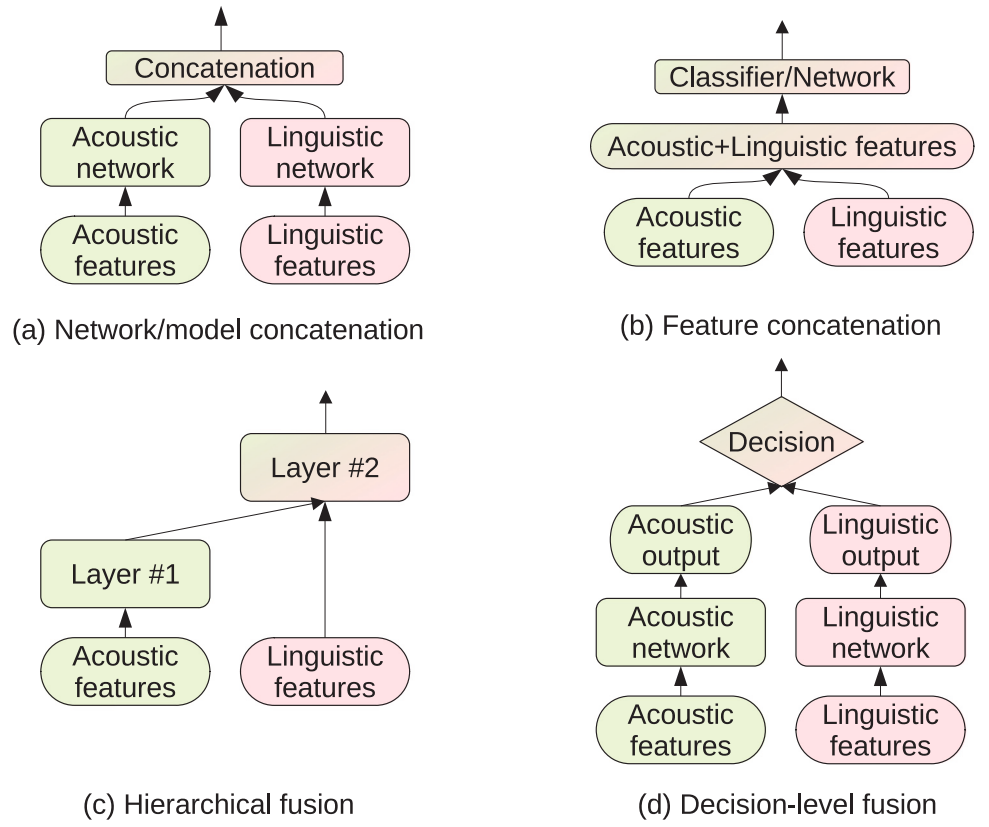

Fig. 5. Different schemes of fusing acoustic with linguistic information; (a), (b), (c): early fusion approach; (d): late fusion approach.

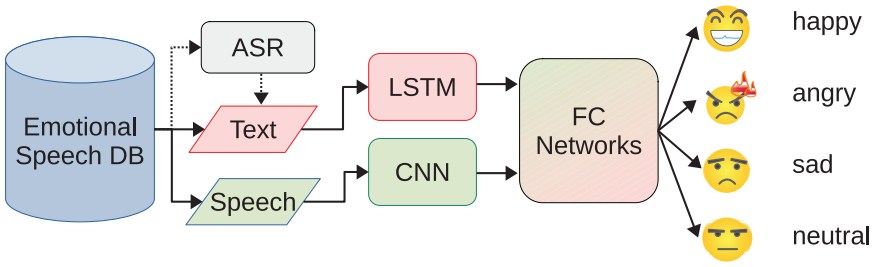

Fig. 6. A typical model/network fusion for categorical acoustic–linguistic emotion recognition; the dashed lines shows the flow to obtain automatic transcription.

mel-spectrogram, filterbank, and their variants. FC networks behave as the concatenation layer, combining outputs of LSTMs and CNNs. The output layer's size depends on the number of labels to be predicted. The loss function is either a categorical cross-entropy for categorical emotion or a mean squared error (MSE) for dimensional emotion. The latter is often replaced by concordance correlation coefficient (CCC) loss since the goal is to maximize CCC.

7.1.3. Hierarchical fusion

Inspired by the human cognitive model that processes multimodal information at different levels, a hierarchical model [Fig. 5(c)] is proposed to go beyond the same-level information fusion. Tian et al. (2016) proposed hierarchical fusion (HL) model for bimodal acoustic–linguistic information fusion. They argued that HL performs better than early and late fusions. Majumder et al. (2018) proposed HL with context modeling to combine multimodal information for sentiment analysis. They found that a combination of acoustic and linguistic information achieved the second-highest performance score after a combination of acoustic, linguistic, and visual information.

7.2. Late fusion approach

Instead of fusing acoustic and linguistic information at the feature level, both information can be fused at the decision level [Fig. 5 (d)]. In this scheme, each modality is processed independently until its results are generated. The results are then processed by the decision function to obtain the final results (final prediction). A typical decision function utilizes majority voting or ensemble methods. Chen and Zhao (2020) fused acoustic and linguistic information at feature level for categorical SER and used three classifiers to obtain three emotion predictions separately: acoustic, linguistic, and acoustic+linguistic. The final decision fusion averages the predictions of three classifiers. Using this technique, they improved accuracy from 70.83% with bimodal feature concatenation to 71.06% with decision fusion from three classifiers. In Cho et al. (2018), SVM was employed to combine outputs of acoustic network with LSTM, text network with CNN, and e-vector. The fusion of three outperformed any unimodal result. Atmaja et al. (2020) used two-stage processing by deep learning and SVM for bimodal emotion recognition; predictions from each modality using deep learning are fed into SVM to obtain the final degrees of dimensional emotion. They

improved the CCC from 0.508 with an early fusion to 0.536 with a late fusion.

In addition to these four types of fusion methods, one can mix two fusion types at the same time. We called this fusion type hybrid fusion. As shown in Table 3, there is only two references proposed this idea (Sebastian and Pierucci, 2019). Using a combination of early fusion (feature concatenation) and late fusion, they achieved moderate results on the IEMOCAP dataset. Although this hybrid fusion is more complex than others, the obtained result is not better than a single early fusion or a late fusion conducted during the same period of research (~2019).

It is not clear which fusion scheme performs best. Planet and Iriondo (2012) found that "feature-level fusion revealed as the best scheme to merge the acoustic and the linguistic information", while Atmaja and Akagi (2021) found the opposite. Pepino et al. (2020) confirmed the previous finding that no significant difference among fusion types.

8. Discussion

8.1. Major findings

In INTERSPEECH 2020, which is one of the most prestigious speech conferences, at least 49 papers were found accepted in the theme of speech emotion recognition. Among these papers, 11 papers presented bimodal emotion recognition from acoustic and linguistic information fusion. Model (network) fusion at the feature level dominates the fusion type of bimodal acoustic–linguistic fusion due to its simplicity. This model fusion is incorporated by six out of these 11 papers. The research used one of USOMS-e, IEMOCAP, or CMU-MOSEI datasets.

As one of the earliest available speech emotion datasets with its transcription, IEMOCAP was the most used dataset in SER considering linguistic information. Table 3 summarizes some bimodal emotion recognition results on the IEMOCAP dataset by utilizing acoustics and linguistic information. The results shown in the table show significant improvements over unimodal acoustic emotion recognition. For instance, the previous review on SER (Akçay and Oguz, 2020) highlighted accuracies of 54% (2014) and 63.5% (2017). In a recent study, Yeh et al. (2020) achieved 66% of accuracy using listen, attend, and spell (LAS) model for multitasking ASR and SER. In contrast, the fusion of acoustic and linguistic information performed by Lian et al. (2020) topped 82% of weighted accuracy (WA).

It is interesting to see that all methods achieved SOTA for weighted accuracy and unweighted accuracy (UA) both employed an attention module on the top of their classifiers. For UA, the attention model on the top of multi-scale convolutional layers leads to 81.4% of accuracy (Peng et al., 2021) on cross-validation evaluation. The DANN method evaluated by Lian et al. (2020), utilized GRU classifiers with multi-head self-attention layer achieved a high UA with 82.68 for non-cross-validation evaluation. Similarly, the self-attentive layer added on both time-synchronous and time-asynchronous models achieved the highest accuracy for both UA (83.08%) and WA (83.22%) for non-cross-validation evaluation. Priyasad et al. (2020) also employed an attention mechanism to fuse acoustic and linguistic data and is achieving SOTA for WA with cross-validation. Since the data is near balance for each emotion class, reporting both WA and UA is necessary (WA for balanced data, UA for unbalanced data). It is not easy to judge which one is better between the two. Nevertheless, since both methods provided WA, the method proposed by Lian et al. (2020) is 1.3% more accurate than that of Peng et al. (2021) in terms of weighted accuracy.

Almost all research reported in Table 3 used similar experimental settings. All references except (Zhang et al., 2019; Lee et al., 2020; Ho et al., 2020) used IEMOCAP data from four categories: anger, happiness/excitement, neutral, and sadness. This data is also known as 'IEM4', containing either 5531, 5530, or 4936 utterances based on the processing method used by the authors. Both Zhang et al. (2019) and Ho et al. (2020) focused on improvised speech among scripted (the result shown in the Table 3 is for mixed/all portion). Lee et al. (2020) used seven categories with 7486 utterances. Either five or ten folds cross-validation was evaluated to judge the performance of different speakers (LOSO, leave one speaker/session out). The last session from two speakers is left for the test partition in almost all reported papers. A high score obtained by Lian et al. (2020) employed five folds from IEM4 dataset (5531 utterances). The transcription used in this DANN-based fusion, as well as other data in Table 3, is manual transcription instead of ASR outputs.

At least seven authors in Table 3 reported the performance of bimodal SER using ASR outputs combined with acoustic information in addition of manual transcription. Kim and Shin (2019) reported degradation of WA and UA to 66.6% and 68.7% while Xu et al. (2019) obtained 70.4% and 69.5% by utilizing the same Google speech recognition system as the previous report. Yoon et al. (2019) proposed an attention mechanism for the fusion method and revealed scores of 73.0% (WA) and 73.9% (UA). Peng et al. (2021) evaluated their multi-scale CNN and attention method on bimodal SER using Google ASR. They topped with 78.0% (WA) and 79.1% (UA), which is the SOTA for IEM4 dataset from ASR-processed transcript fused with acoustic information. In line with these findings, Sahu et al. (2019) reported a drop of 4% accuracies from manual to automatic transcription for acoustic–linguistic emotion recognition. Heusser et al. (2019) and Peng et al. (2021) reported a smaller degradation of 2% for both WA and UA while Wu et al. (2021) reported 3.7%. These findings show that the current ASR systems are sufficient for extracting linguistic information for acoustic–linguistic emotion recognition. There is still room for improvement in the performance of linguistic-only emotion recognition since the achieved word error rate (WER) is only about 40% for emotional speech.

Accelerating bimodal SER research by fusing acoustic and linguistic information can be triggered by providing both dataset and challenge. USOMS-e dataset and the elderly emotion sub-challenge (ESC) in INTERSPEECH 2020 contribute to research bimodal acoustic–linguistic emotion recognition in several ways, mainly on feature extraction methods, classification models, and fusion types. Table 4 shows research results on USOMS-e dataset with various methodologies. The distribution of the fusion method is almost equal; early fusion with feature concatenation is more adopted than late fusion (decision-level fusion). In this dataset, a fusion type that achieved the best performance for valence prediction is different from that of arousal prediction. The requirement of two models (two fusion methods) may not be effective for future practical implementation on predicting valence and arousal simultaneously.

In contrast to IEMOCAP data, experiment settings in research papers reported on the USOMS-e dataset are almost the same. The organizer of the ESC — computational paralinguistic challenge already provided the structured dataset (by splitting the data into training, evaluation, and test sets) along with its baseline code; the authors proposed their methods in feature extraction, classification, or fusion methods. The SOTA for this dataset was achieved using an ensemble method (Soğancıoğlu et al., 2020). For valence prediction, five different features were chosen, including TF–IDF, FastText word embeddings, high-level polarity features, FastText+Polarity features, and dictionary-based linguistic features in German and English. The fusion of the classifiers from these five feature sets utilized an ensemble method with label fusion strategies. For arousal prediction, the ensemble method based on the Fisher vector (FV) combines acoustic-based arousal prediction with the baseline systems (Schuller et al., 2020).

From the dataset point of view, IEMOCAP (particularly IEM4) was overused. It is necessary to explore other datasets as well as other emotion models. The cross-corpus evaluation is now a major challenge in acoustic-only SER; however, there is no work found in cross-corpus bimodal acoustic–linguistic emotion recognition. The necessity of using datasets besides English is also important to judge the generalization of the proposed model.

Table 3
Weighted and unweighted accuracies (WA & UA) of acoustic–linguistic emotion recognition on IEMOCAP dataset. Linguistic feature is extracted from ground-truth text. *Italics* indicate results without cross-validation (CV). **Bolds** indicate the best results in CV.

Reference	Acoustic feature	Linguistic feature	Fusion type	WA %	UA %
Jin et al. (2015)	ACO (+) Cepstrum (+) Cepstral-BoW	Lex-BoW (+) Lex-eVector	Decision	69.2	68.4
Cho et al. (2018)	eGeMAPS	E-vector, WE	Decision	64.97	65.9
Hazarika et al. (2018)	ComParE2013	WE	Feature	72.1	71.9
Sebastian and Pierucci (2019)	ComParE2013	FastText	Hybrid	61.2	60.2
Sahu et al. (2019)	pyAudioAnalysis (34)	GloVe	Model	–	68.18
Cai et al. (2019)	pyAudioAnalysis (34)	GloVe	Model	70.1	71.25
Li and Lee (2019)	Emobase2010	GloVe	Model	–	70.3
Xu et al. (2019)	pyAudioAnalysis (34)	GloVe	Model	72.5	70.9
Zhang et al. (2019)	mel filertbank (MFB)	phonemes (40)	Feature	–	73.79
Kim and Shin (2019)	BN	word2vec + ANEW	Model	73.7	75.5
Atmaja et al. (2019)	pyAudioAnalysis (34)	WE	Model	75.49	–
Tripathi et al. (2019)	MFCC	word2vec	Model	76.1	69.5
Heusser et al. (2019)	IS09	BERT	Model	73.5	71.0
Yoon et al. (2019)	MFCC	WE	Feature	76.5	77.6
Lee et al. (2020)	MFCC	GloVe	Feature	57.9	48.7
Bhosale et al. (2020)	mel spectrogram + Δ + $\Delta\Delta$	DeepSpeech-1	Model	68.11	63.15
Pepino et al. (2020)	GeMAPS (36)	BERT	Decision	–	65.1
Chen and Zhao (2020)	Log-mel filterbank	ALBERT	Decision	71.06	72.05
Shen et al. (2020)	MFCC + Δ + $\Delta\Delta$	GloVe	Feature	75.9	76.4
Feng et al. (2020)	Log-mel filterbank	WE	Model	68.6	69.7
Liu et al. (2020)	MFCC	GloVe embedding	Model	72.39	70.08
Krishna and Patil (2020)	Raw waveform	GloVe	Feature	–	72.82
Ho et al. (2020)	MFCC	BERT	Model	73.23	–
Siriwardhana et al. (2020)	VQ-word2vec + Speech-BERT	GPT-2 tokenizer + RoBERTa	Model	–	75.45
Kim et al. (2020)	LLD+BN	word2vec+BoW+ANEW	Model	74.37	76.91
Priyasad et al. (2020)	SincNet	Bi-RNN+CNN	Feature	**80.51**	79.22
Lian et al. (2020)	ComParE2013	ELMo	Feature	82.68	–
Santoso et al. (2021)	MFCC + CQT + f_o	BERT	Model	76.1	75.9
Wang et al. (2021)	MFCC + Δ + $\Delta\Delta$ + Transformer	WE (Transformer)	Hybrid	76.8	77.1
Wu et al. (2021)	Filterbank + f_o + Δ	GLoVe + BERT	Model	77.57	78.41
Wu et al. (2021)	Filterbank + f_o + Δ	GLoVe + BERT	Model	83.08	83.22
Peng et al. (2021)	MFCC + Δ + $\Delta\Delta$ + X-vector	GloVe	Model	80.3	**81.4**

Table 4
Unweighted accuracy results of valence (V) and arousal (A) predictions on the USOMS-e dataset.

Reference	Acoustic feature	Linguistic feature	Fusion type	Val	Aro
Schuller et al. (2020)	ResNet50	BLatt	Unimodal	50.4%	49.0%
Juli (2020)	ComParE2013 + oXv	bert-as-a-service	Feature	61.0%	48.8%
Yang et al. (2020)	ComParE (+) BoAW (+) ResNet50 (+) AuDeep (+) FV-MFCC	BoW (+) TFIDF (+) Sparse PMI (+) Sparse NGD (+) PMI-BoW (+) NGD-BoW (+) PMI-TFIDF (+) NGD-TFIDF	Decision	59.0%	54.3%
Soğancıoğlu et al. (2020)	Fisher Vector (FV)	TFIDF (+) FastText (+) Polarity (+) Fastext+Polarity (+) Dictionary	Decision	**63.7%**	**57.5%**
Atmaja et al. (2020)	ResNet50	BLAtt (Val), Gmax (Aro)	Feature	50.4%	56.3%
Viraraghavan et al. (2020)	ComParE2013	WE	Feature	36.3%	–
Boateng and Kowatsch (2020)	ResNet50	SBERT	Unimodal	57.8	50.4

Although the recent advancements in signal processing enable more advanced feature extraction strategies, the current trends show handcrafted acoustic feature is more meaningful than deep learning-based features for the SER task. Four of the five highest weighted accuracies (WA) in the IEMOCAP dataset (Table 3) obtained their results using handcrafted features ((Lian et al., 2020; Wu et al., 2021; Peng et al., 2021)); only (Priyasad et al., 2020) employed SincNet layers to extract acoustic features. In contrast, all five highest WA used deep learning-based linguistic features, either WE, GloVe, or ELMo.

One possible reason for that different finding in acoustic and linguistic feature extraction data is the nature of the data itself. Linguistic data, i.e., text, is available abundantly on the internet. Using deep learning with more data tends to be more effective, it is reasonable to model connections among samples in the text data and extract their representations. For speech, which is less than text in the number of data, handcrafted features deriving information related to emotion mathematically or physically are still superior to deep learning-based feature extraction. Handcrafted acoustic features are also easy to interpret and usually cost lower in computation than DNN-based features.

The recurrent-based neural network, including LSTM, BLSTM, and GRU, still dominated classifiers used in the SER task. Besides RNN, CNN and attention-based networks are the common classifiers in the SER task. The current trends on these classifiers are to evaluate the consistency among various datasets. Either RNN, CNN, or attention-based mechanism arguably performs well. Currently, combining different classifiers also performed best on this SER task (as Lian et al. (2020) did with GRU and attention layer).

It was found that model concatenation dominates fusion type in acoustic–linguistic emotion recognition. A number of 12 research reported model fusion in Table 3 for combining acoustic and linguistic information in the IEMOCAP dataset. Others are feature and hybrid fusions. FC layer is usually employed to concatenate two or more models. Half of the model fusion employed FC layers while the rest used attention models or rule-based system. Interestingly, the SOTA result achieved by Lian et al. (2020) employed feature fusion with a multi-head self-attention mechanism. The authors used a single GRU classifier to receive acoustic and linguistic features with a DANN framework.

Finally, to enable future benchmarks, it is important to stick to the metrics presented in current papers: overall accuracy (weighted/unbalanced accuracy, WA) and class accuracy (unweighted/balanced accuracy, UA). Additional scores can be added, such as the F1 score and recall. The confusion matrix is also important for categorical emotion analysis. It enables in which emotions the proposed method performed

Table 5
Comparison of SER performance on difference multimodal fusions; A: acoustic, L: linguistic, V: visual.

Reference	Dataset	Metric	A+L %	A+V %	L+V %	A+L+V %
Majumder et al. (2018)	IEMOCAP	WA	76.0	69.6	75.6	76.8
Poria et al. (2018)	IEMOCAP	WA	70.8	52.2	68.6	71.6
Sebastian and Pierucci (2019)	IEMOCAP	WA	61.2	–	–	60.1
		UA	60.2	–	–	58.3
		F1	61.2	–	–	59.9
Delbrouck et al. (2020)	IEMOCAP	F1	74	–	–	71.5
Khare et al. (2020)	CMU-MOSEI	WA	66.2	60.9	–	67.1
		F1	78.4	76.4	–	78.8

good and bad. For dimensional emotion, CCC is the gold-standard metric, among others.

8.2. Comparison with other multimodal fusions

The fusion methods explained above have been adopted not only for acoustic–linguistic emotion recognition but also for other multimodal fusions. In Table 5, a brief performance comparison of acoustic–linguistic (A+L) fusion with acoustic–visual (A+V) and acoustic–linguistic–visual (A+L+V) fusions are presented on different datasets and metrics. It can be concluded that A+L only suffers from A+L+V. In other words, A+L fusion outperformed other bimodal fusions in most cases reported in that table.

Except for a report by Sebastian and Pierucci (2019), all multimodal evaluations are performed in the same early fusion method. Sebastian and Pierucci (2019) showed that their late fusions of A+L surpassed recently proposed methods with A+L+V. This finding is similar to emotion recognition (IEMOCAP) and binary sentiment classification (CMU-MOSEI) results reported by Delbrouck et al. (2020). In both cases, A+L also surpassed A+L+V. Note that although some authors also reported sentiment classification results [e.g., Poria et al. (2018)], we do not include it. Sentiment only predicts polarity of expression, whether it is positive, neutral, or negative. This term is very close, if not the same, to valence in dimensional emotion. It also has been found in Poria et al. (2018) that the order of contribution of modalities for the IEMOCAP and CMU-MOSI datasets is acoustic, visual, and linguistic. In the MOUD dataset, the order of contribution for predicting sentiment is linguistic, acoustic, and visual.

Given the competitive results of A+L fusion among other fusions, it is worth continuing research on acoustic–linguistic emotion recognition. The problems found in the previous research could be investigated along with confirming the previous findings. Some related issues below are suggested for future research on bimodal acoustic–linguistic emotion recognition.

8.3. Future directions

Although it has been researched for almost twenty years, the research of bimodal acoustic–linguistic fusion for SER is not ready well for implementation. Several issues below are the most fundamental ones among many.

As shown in Fig. 7, there is a bottleneck between the processing of acoustic and linguistic data. The acoustic side extracts acoustic features directly from speech while the linguistic side waits for the output of ASR. Although the current ASR technology could transcribe speech into text in a small latency, still, there is a time gap between linguistic and acoustic processings. Kim and Shin (2019) proposed a bottleneck acoustic feature for early acoustic–linguistic fusion; however, this feature is not intended to tackle the bottleneck issues between both processings.

Since there is a bottleneck between acoustic and linguistic processing, there is a chance to accelerate the computation process by obtaining the linguistic information directly from acoustic information. This linguistic information may also be embedded in prosodic features (Fujisaki, 2003). Feng et al. (2020) used acoustic-to-word representation trained on ASR to fuse with linguistic information. However, their method still requires transcription for generating linguistic features. The challenge here is to obtain linguistic information directly without a need for text or transcription.

The fusion of acoustic and linguistic information has been found to be effective for both dimensional and categorical speech emotion recognition. However, no detailed study was found on investigating when linguistic necessary is needed. In several cases, such as in short utterances, linguistic information may not be necessary. The investigation to find a threshold for the necessity of adding linguistic information is worth of study for future research. This research direction can also be expanded to investigate each modality's contribution weight, confirming the previous results.

Fusion of acoustic and linguistic for SER also challenges the necessity for aligning acoustic features with respected words. Tzirakis et al. (2021) argued that alignment between acoustic and linguistic embedding spaces enriches the speech representations. Xu et al. (2019) found that alignment between original speech and recognized text helps to improve the performance when both acoustic and linguistic modalities are fused. Both Lee et al. (2020) and Liu et al. (2020) aligned acoustic and text features using attention-based BSLSTM networks and obtained significant improvements over their baselines. These reported research using aligned techniques are still less superior than context-based DAN (Lian et al., 2020) on the same IEMOCAP dataset. It is of interest to see the performance of that technique with alignment (the reported result is without alignment).

Text-independent is still a difficult task for bimodal SER. Pepino et al. (2020) showed that the performance of bimodal SER with text-independent features is not as good as the performance without text-independent features (split by script). Atmaja and Sasou (2021) reported a similar finding that text-independent is more difficult for Japanese acoustic-based SER. Their works highlight the necessity to tackle the limitation of bimodal SER under different scripts for training and test partitions. One may enlarge the training set to include more linguistic information in the training phase. One may also focus on the smaller dataset to train the model in which linguistic information for the test phase is not available in the training phase.

It is not the only emotion that can be obtained from speech but also others, such as gender, age, and words. The feature representation used for emotion recognition may overlap with these other tasks. Instead of predicting emotion only, predicting multitask output is more beneficial for future applications. This problem can be approached by multitask or transfer learning. In this muti-output prediction, one may incorporate information from other tasks into the SER model. An example is adding gender and age features to improve SER (e.g., based on Zhao et al. (2018)) after obtaining these pieces of information. One can also incorporate a language model from ASR to improve SER since a similar method was reported to work for sentiment analysis (Shon et al., 2021).

Finally, one model that worked better on one dataset shall work better on other datasets. In fact, this model generalization is still a problem in acoustic–linguistic emotion recognition. The contribution of a language to emotion information may differ from other languages. An investigation of the effect of linguistic information in different languages is also necessary to accelerate the implementation of bimodal acoustic–linguistic emotion recognition for multilingual speech. Also, training a model with different (cross languages) datasets is a merit study for the future to find some adjustments for language-specific emotion recognition.

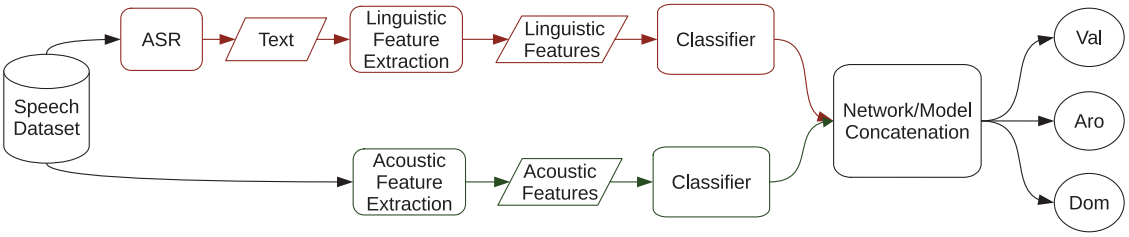

Fig. 7. Utilizing ASR outputs for dimensional SER (predicting degree of valence[Val], arousal[Aro], and dominance [Dom]) with model concatenation.

9. Conclusions

In this paper, a comprehensive review of bimodal speech emotion recognition from acoustic and linguistic information fusion is presented. The study focuses on five main components of SER from bimodal information: datasets, emotion models, features, classifiers, and fusion methods. These main blocks must exist in the speech emotion recognition method by fusing acoustic and linguistic information. There are three emotion models developed in psychological research; however, most SER research focused on the categorical model. There is a move to extract acoustic features in the feature extraction step by using deep learning methods, while deep learning-based linguistic features already dominated text processing research, including SER from linguistic information. However, the majority of SER methods, including those currently achieving state-of-the-art results, still rely on handcrafted acoustic features. Then, the common classifiers for bimodal SER are briefly described. Although more advanced DNN architectures have been developed, bimodal SER still relies heavily on SVM, MLP, CNN, and LSTM architectures, with recurrent-based neural networks dominating these classifiers. Some major findings besides these in the five SER building blocks have also been discussed. Finally, some raised issues in SER research are highlighted for future research directions in speech emotion recognition by fusing acoustic and linguistic information.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan. Parts of this study were also conducted where the first author is still affiliated with the Japan Advanced Institute of Science and Technology (JAIST) and Institut Teknologi Sepuluh Nopember (ITS).

References

Akçay, M.B., Oguz, K., 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 116, 56–76. http://dx.doi.org/10.1016/j.specom.2019.12.001.

Aldeneh, Z., Khorram, S., Dimitriadis, D., Provost, E.M., 2017. Pooling acoustic and lexical features for the prediction of valence. In: ICMI 2017 - Proceedings of the 19th ACM International Conference on Multimodal Interaction, Vol. 2017-Janua. ACM, pp. 68–72. http://dx.doi.org/10.1145/3136755.3136760.

Alm, C.O., Roth, D., Sproat, R., 2005. Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05. pp. 579–586. http://dx.doi.org/10.3115/1220575.1220648, arXiv:arXiv:1011.1669v3 URL: http://l2r.cs.uiuc.edu/.

Anagnostopoulos, C.N., Iliou, T., Giannoukos, I., 2012. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artif. Intell. Rev. 43 (2), 155–177. http://dx.doi.org/10.1007/s10462-012-9368-5.

Atmaja, B.T., Akagi, M., 2019. Speech emotion recognition based on speech segment using LSTM with attention model. In: 2019 IEEE International Conference on Signals and Systems (ICSigSys). IEEE, pp. 40–44. http://dx.doi.org/10.1109/ICSIGSYS.2019.8811080.

Atmaja, B.T., Akagi, M., 2020. Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. APSIPA Trans. Signal Inf. Process. 9 (May), e17. http://dx.doi.org/10.1017/ATSIP.2020.14.

Atmaja, B.T., Akagi, M., 2021. Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM. Speech Commun. 126, 9–21. http://dx.doi.org/10.1016/j.specom.2020.11.003.

Atmaja, B.T., Hamada, Y., Akagi, M., 2020. Predicting valence and arousal by aggregating acoustic features for acoustic-linguistic information fusion. In: 2020 IEEE Region 10 Conference (TENCON). IEEE, pp. 1081–1085. http://dx.doi.org/10.1109/TENCON50793.2020.9293899, URL: https://ieeexplore.ieee.org/document/9293899/.

Atmaja, B.T., Sasou, A., 2021. Effect of different splitting criteria on the performance of speech emotion recognition. In: 2021 IEEE Region 10 Conference (TENCON). IEEE, pp. 760–764. http://dx.doi.org/10.1109/TENCON54134.2021.9707265, URL: https://ieeexplore.ieee.org/document/9707265/.

Atmaja, B.T., Shirai, K., Akagi, M., 2019. Speech emotion recognition using speech feature and word embedding. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, Lanzhou, pp. 519–523. http://dx.doi.org/10.1109/APSIPAASC47483.2019.9023098, URL: https://ieeexplore.ieee.org/document/9023098/.

Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D., 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. Psychol. Sci. Public Interest 20 (1), 1–68. http://dx.doi.org/10.1177/1529100619832930.

Bhosale, S., Chakraborty, R., Kopparapu, S.K., 2020. Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7189–7193. http://dx.doi.org/10.1109/ICASSP40776.2020.9054621.

Boateng, G., Kowatsch, T., 2020. Speech emotion recognition among elderly individuals using multimodal fusion and transfer learning. In: Companion Publication of the 2020 International Conference on Multimodal Interaction. ACM, New York, NY, USA, pp. 12–16. http://dx.doi.org/10.1145/3395035.3425255, URL: https://dl.acm.org/doi/10.1145/3395035.3425255.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A database of german emotional speech. In: Proc. Interspeech 2005. pp. 1517–1520, URL: http://www.expressive-speech.net/emodb/.

Busso, C., Bulut, M., Lee, C.-C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. IEMOCAP: INteractive emotional dyadic motion capture database. Lang. Resour. Eval. 42 (4), 335–359. http://dx.doi.org/10.1007/s10579-008-9076-6.

Busso, C., Parthasarathy, S., Burmania, A., Abdelwahab, M., Sadoughi, N., Provost, E.M., Member, S.S., Parthasarathy, S., Member, S.S., Burmania, A., Abdelwahab, M., Sadoughi, N., Mower Provost Member, E., 2017. MSP-IMPROV: AN acted corpus of dyadic interactions to study emotion perception. Trans. Affect. Comput. 8 (1), 67–80. http://dx.doi.org/10.1109/TAFFC.2016.2515617.

Cai, L., Hu, Y., Dong, J., Zhou, S., 2019. Audio-textual emotion recognition based on improved neural networks. Math. Probl. Eng. 2019, http://dx.doi.org/10.1155/2019/2593036.

Callejas, Z., Griol, D., López-Cózar, R., 2011. Predicting user mental states in spoken dialogue systems. EURASIP J. Adv. Signal Process. 2011 (1), 1–21. http://dx.doi.org/10.1186/1687-6180-2011-6.

Calvo, R.A., Kim, S.M., 2013. Emotions in Text: Dimensional and Categorical Models. Technical Report 3, URL: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8640.2012.00456.x.

Chang, J., Scherer, S., 2017. Learning representations of emotional speech with deep convolutional generative adversarial networks. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2746–2750. http://dx.doi.org/10.1109/ICASSP.2017.7952656.

Chen, M., Zhao, X., 2020. A multi-scale fusion framework for bimodal speech emotion recognition. In: Proc. Interspeech 2020. pp. 374–378. http://dx.doi.org/10.21437/Interspeech.2020-3156.

Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y., Dehak, N., 2018. Deep neural networks for emotion recognition combining audio and transcripts. In: Proc. Interspeech 2018, Vol. 2018-Septe. ISCA, pp. 247–251. http://dx.doi.org/10.21437/Interspeech.2018-2466.

Chuang, Z.J., Wu, C.-h., 2004. Multi-modal emotion recognition from speech and text. J. Comput. Linguist. Chin. Lang. Process. 9 (2), 45–62, URL: http://www.aclweb.org/anthology/O/O04/O04-3004.pdf.

Cowen, A.S., Keltner, D., 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc. Natl. Acad. Sci. USA 114 (38), E7900–E7909. http://dx.doi.org/10.1073/pnas.1702247114.

Delbrouck, J.-B., Tits, N., Dupont, S., 2020. Modulated fusion using transformer for linguistic-acoustic emotion recognition. In: International Workshop on Natural Language Processing beyond Text. pp. 1–10. http://dx.doi.org/10.18653/v1/2020.nlpbt-1.1, arXiv:2010.02057.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: International Conference on Spoken Language Processing, ICSLP, Proceedings. pp. 1970–1973. http://dx.doi.org/10.1109/icslp.1996.608022.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: PRe-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 URL: http://arxiv.org/abs/1810.04805.

Du, S., Tao, Y., Martinez, A.M., 2014. Compound facial expressions of emotion. Proc. Natl. Acad. Sci. USA 111 (15), http://dx.doi.org/10.1073/pnas.1322355111.

Ekman, P., 1992. An argument for basic emotions. Cogn. Emot. 6 (3–4), 169–200. http://dx.doi.org/10.1080/02699939208411068.

Ekman, P., 2005. Basic emotions. In: Handbook of Cognition and Emotion. http://dx.doi.org/10.1002/0470013494.ch3.

El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit. 44 (3), 572–587. http://dx.doi.org/10.1016/j.patcog.2010.09.020.

Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., Andre, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P., 2016. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7 (2), 190–202. http://dx.doi.org/10.1109/TAFFC.2015.2457417.

Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R., 2010. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. J. Multimodal User Interfaces 3 (1), 7–19. http://dx.doi.org/10.1007/s12193-009-0032-6.

Feng, H., Ueno, S., Kawahara, T., 2020. End-to-end speech emotion recognition combined with acoustic-to-word ASR model. In: Interspeech 2020. ISCA, ISCA, pp. 501–505. http://dx.doi.org/10.21437/Interspeech.2020-1180.

Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., Phoebe, C., Fontaine, J.R.J., Scherer, K.R., Roesch, E.B., Ellsworth, P.C., 2017. The world of emotions is not two-dimensional. Psychol. Sci. 18 (12), 1050–1057.

Frick, R.W., 1985. Communicating emotion: The role of prosodic features. Psychol. Bull. 97 (3), 412–429. http://dx.doi.org/10.1037/0033-2909.97.3.412.

Fujisaki, H., 2003. Prosody, information, and modeling with emphasis on tonal features of speech. In: Workshop on Spoken Language Processing.

Goodfellow, I., Bengio, Y., Courville, A., 2015. Deep Learning Book. MIT Press.

Grandjean, D., Sander, D., Scherer, K.R., 2008. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. Conscious. Cogn. 17 (2), 484–495. http://dx.doi.org/10.1016/j.concog.2008.03.019.

Grimm, M., Kroschel, K., Mower, E., Narayanan, S., 2007. Primitives-based evaluation and estimation of emotions in speech. Speech Commun. 49 (10–11), 787–800. http://dx.doi.org/10.1016/j.specom.2007.01.010, URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639307000040.

Grimm, M., Kroschel, K., Narayanan, S., 2008. The vera am Mittag german audio-visual emotional speech database. In: 2008 IEEE International Conference on Multimedia and Expo. pp. 865–868. http://dx.doi.org/10.1109/ICME.2008.4607572.

Griol, D., Iglesias, J.A., Ledezma, A., Sanchis, A., 2016. A two-stage combining classifier model for the development of adaptive dialog systems.. Int. J. Neural Syst. 26 (1), 1650002. http://dx.doi.org/10.1142/S0129065716500027.

Griol, D., Molina, J.M., 2015. A sentiment analysis classification approach to assess the emotional content of photographs. In: Mohamed, A., Novais, P., Pereira, A., Villarrubia González, G., Fernández-Caballero, A. (Eds.), Ambient Intelligence - Software and Applications. Springer International Publishing, Cham, pp. 105–113.

Griol, D., Molina, J.M., Callejas, Z., 2019. Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances. Neurocomputing 326–327, 132–140. http://dx.doi.org/10.1016/j.neucom.2017.01.120.

Gunes, H., Pantic, M., 2010. Automatic, dimensional and continuous emotion recognition. Int. J. Synth. Emot. 1 (1), 68–99. http://dx.doi.org/10.4018/jse.2010101605.

Hazarika, D., Gorantla, S., Poria, S., Zimmermann, R., 2018. Self-attentive feature-level fusion for multimodal emotion detection. In: Proceedings - IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018. IEEE, pp. 196–201. http://dx.doi.org/10.1109/MIPR.2018.00043.

Heusser, V., Freymuth, N., Constantin, S., Waibel, A., 2019. Bimodal speech emotion recognition using pre-trained language models. In: ASRU. Singapore, arXiv:1912.02610 URL: https://gluebenchmark.com/leaderboardhttp://arxiv.org/abs/1912.02610.

Hinton, G.E., 1989. Connectionist learning procedures. Artificial Intelligence 40 (1–3), 185–234. http://dx.doi.org/10.1016/0004-3702(89)90049-0.

Ho, N.-h., Yang, H.-j., Kim, S.-h., Lee, G., 2020. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access 8, 61672–61686. http://dx.doi.org/10.1109/ACCESS.2020.2984368, URL: https://ieeexplore.ieee.org/document/9050806/.

Huang, K.-Y., Wu, C.-H., Hong, Q.-B., Su, M.-H., Zeng, Y.-R., 2018. Speech emotion recognition using convolutional neural network with audio word-based embedding. In: 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, pp. 265–269. http://dx.doi.org/10.1109/ISCSLP.2018.8706610, URL: https://ieeexplore.ieee.org/document/8706610/.

Hutto, C.J., Gilbert, E.E., 2014. VADER: A Parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International Conference on Weblogs and Social Media (ICWSM-14).

Jack, R.E., Sun, W., Delis, I., Garrod, O.G.B., Schyns, P.G., 2016. Four not six: Revealing culturally common facial expressions of emotion. J. Exp. Psychol. [Gen.] 145 (6), 708–730. http://dx.doi.org/10.1037/xge0000162.

Jin, Q., Li, C., Chen, S., Wu, H., 2015. Speech emotion recognition with acoustic and lexical features. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Vol. 2015-Augus. IEEE, pp. 4749–4753. http://dx.doi.org/10.1109/ICASSP.2015.7178872, arXiv:arXiv:1011.1669v3.

Juli, M., 2020. Exploring text and audio embeddings for multi-dimension elderly emotion recognition. In: Proc. Interspeech 2020. pp. 2067–2071.

Karadogan, S.G., Larsen, J., 2012. Combining semantic and acoustic features for valence and arousal recognition in speech. In: 2012 3rd International Workshop on Cognitive Information Processing (CIP). IEEE, pp. 1–6. http://dx.doi.org/10.1109/CIP.2012.6232924, URL: http://ieeexplore.ieee.org/document/6232924/.

Khare, A., Parthasarathy, S., Sundaram, S., 2020. Multi-modal embeddings using multi-task learning for emotion recognition. In: Proc. Interspeech 2020. ISCA, ISCA, pp. 384–388. http://dx.doi.org/10.21437/Interspeech.2020-1827, arXiv:2009.05019.

Kim, E., Shin, J.W., 2019. DNN-Based emotion recognition based on bottleneck acoustic features and lexical features. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. pp. 6720–6724.

Kim, E., Song, H., Shin, J.W., 2020. Affective latent representation of acoustic and lexical features for emotion recognition. Sensors 20 (9), 1–12. http://dx.doi.org/10.3390/s20092614.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Schuller, B., Star, K., et al., 2019. SEWA DB: A Rich database for audio-visual emotion and sentiment research in the wild. arXiv preprint arXiv:1901.02839.

Krishna, D.N., Patil, A., 2020. Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In: Proc. Interspeech 2020, pp. 4243–4247.

Kurniawati, P., Lestari, D.P., Khodra, M.L., Kurniawati, Pipin; Lestari, Dessi Puji; Leylia Khodra, M., 2017. Speech emotion recognition from Indonesian spoken language using acoustic and lexical features. In: Oriental COCOSDA. IEEE, pp. 189–195. http://dx.doi.org/10.1109/ICSDA.2017.8384467, URL: https://ieeexplore.ieee.org/document/8384467/.

Lee, C.M., Narayanan, S.S., Angeles, L., Pieraccini, R., 2002. Combining acoustic and language information for emotion recognition. In: ICLSP-2002. pp. 873–876.

Lee, Y., Yoon, S., Jung, K., 2020. Multimodal speech emotion recognition using cross attention with aligned audio and text. In: Proc. Interspeech 2020. pp. 2717–2721. http://dx.doi.org/10.1109/SLT.2018.8639583, arXiv:1810.04635.

Li, J.-L., Lee, C.-C., 2019. Attentive to individual: A multimodal emotion recognition network with personalized attention profile. In: Interspeech 2019. ISCA, ISCA, pp. 211–215. http://dx.doi.org/10.21437/Interspeech.2019-2044.

Lian, Z., Tao, J., Liu, B., Huang, J., Yang, Z., Li, R., 2020. Context-dependent domain adversarial neural network for multimodal emotion recognition. In: Proc. Interspeech 2020, pp. 394–398.

Liebenthal, E., Silbersweig, D.A., Stern, E., 2016. The language, tone and prosody of emotions: Neural substrates and dynamics of spoken-word emotion perception. Front. Neurosci. 10 (NOV), 506. http://dx.doi.org/10.3389/fnins.2016.00506.

Litman, D.J., Forbes-Riley, K., 2004. Predicting student emotions in computer-human tutoring dialogues. In: 42nd Annual Meeting on Association for Computational Linguistics. pp. 351–es. http://dx.doi.org/10.3115/1218955.1219000.

Litman, D.J., Forbes-Riley, K., 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Commun. 48 (5), 559–590. http://dx.doi.org/10.1016/j.specom.2005.09.008.

Litman, D.J., Rosé, C.P., Forbes-Riley, K., VanLehn, K., Bhembe, D., Silliman, S., 2004. Spoken versus typed human and computer dialogue tutoring. In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (Eds.), Intelligent Tutoring Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 368–379.

Liu, P., Li, K., Meng, H., 2020. Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition. In: Proc. Interspeech 2020, pp. 379–383.

Lotfian, R., Busso, C., 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEE Trans. Affect. Comput. 10 (4), 471–483. http://dx.doi.org/10.1109/TAFFC.2017.2736999.

Macary, M., Lebourdais, M., Tahon, M., Estève, Y., Rousseau, A., 2020. Multi-corpus experiment on continuous speech emotion recognition: Convolution or recurrence? In: International Conference on Speecn and Computer. pp. 304–314. http://dx.doi.org/10.1007/978-3-030-60276-5_30.

Mairano, P., Zovato, E., Quinci, V., 2019. Do sentiment analysis scores correlate with acoustic features of emotional speech? In: AISV Conference.

Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E., Poria, S., 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowl.-Based Syst. 161, 124–133. http://dx.doi.org/10.1016/j.knosys.2018.07.041, https://linkinghub.elsevier.com/retrieve/pii/S0950705118303897.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M., 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. 3 (1), 5–17. http://dx.doi.org/10.1109/T-AFFC.2011.20.

Mehrabian, A., Russell, J.A., 1974. An Approach to Environmental Psychology. the MIT Press.

Metze, F., Polzehl, T., Wagner, M., 2009. Fusion of acoustic and linguistic speech features for emotion detection. In: Proc. International Conference on Semantic Computing (ICSC). Berkeley, CA, pp. 153–160. http://dx.doi.org/10.1109/ICSC.2009.32.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. In: International Conference on Learning Representations. arXiv:1301.3781.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A., 2019. Advances in pre-training distributed word representations. In: LREC 2018 - 11th International Conference on Language Resources and Evaluation. pp. 52–55, arXiv:1712.09405.

Mirsamadi, S., Barsoum, E., Zhang, C., S., M., 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2017. pp. 2227–2231. http://dx.doi.org/10.1109/ICASSP.2017.7952552.

Mohammad, S., 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 174–184. http://dx.doi.org/10.18653/v1/P18-1017.

Mozziconacci, S., 2002. Prosody and emotions. In: Speech Prosody 2002. pp. 1–9.

Mulcrone, K., 2012. Detecting emotion in text. In: UMM CSci Senior Seminar Conference.

Nygaard, L.C., Queen, J.S., 2008. Communicating emotion: Linking affective prosody and word meaning. J. Exp. Psychol.: Hum. Percept. Perform. 34 (4), 1017–1030. http://dx.doi.org/10.1037/0096-1523.34.4.1017.

Olah, C., 2015. Understanding LSTM networks – colah's blog. URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825—-2830, arXiv:1201.0490.

Peng, Z., Lu, Y., Pan, S., Liu, Y., 2021. Efficient speech emotion recognition using multi-scale CNN and attention. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3020–3024. http://dx.doi.org/10.1109/ICASSP39728.2021.9414286, arXiv:2106.04133.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543.

Pepino, L., Riera, P., Ferrer, L., Gravano, A., 2020. Fusion approaches for emotion recognition from speech using acoustic and text-based features. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6484–6488. http://dx.doi.org/10.1109/ICASSP40776.2020.9054709, URL: https://ieeexplore.ieee.org/document/9054709/.

Petrushin, V.A., 1999. Emotion in speech: Recognition and application to call centers. Proc. Artif. Neural Netw. Eng. 710, 22–30.

Picard, R.W., 1995. Affective Computing. Technical Report.

Planet, S., Iriondo, I., 2012. Comparative study on feature selection and fusion schemes for emotion recognition from speech. Int. J. Interact. Multimedia Artif. Intell. 1 (6), 44. http://dx.doi.org/10.9781/ijimai.2012.166.

Plutchik, R., Kellerman, H., 1980. Emotion, Theory, Research, and Experience. Academic Press.

Polzehl, T., Schmitt, A., Metze, F., Wagner, M., 2011. Anger recognition in speech using acoustic and linguistic cues. Speech Commun. 53 (9–10), 1198–1209. http://dx.doi.org/10.1016/j.specom.2011.05.002.

Poria, S., Cambria, E., Bajpai, R., Hussain, A., 2017. A review of affective computing: From unimodal analysis to multimodal fusion. Inf. Fusion 37, 98–125. http://dx.doi.org/10.1016/j.inffus.2017.02.003, URL: https://linkinghub.elsevier.com/retrieve/pii/S1566253517300738.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R., 2019. MELD: A Multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 527–536. http://dx.doi.org/10.18653/v1/p19-1050.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., Hussain, A., Cambria, E., 2018. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. IEEE Intell. Syst. 33 (6), 17–25. http://dx.doi.org/10.1109/MIS.2018.2882362, URL: https://ieeexplore.ieee.org/document/8636432/.

Priyasad, D., Fernando, T., Denman, S., Sridharan, S., Fookes, C., 2020. Attention driven fusion for multi-modal emotion recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3227–3231. http://dx.doi.org/10.1109/ICASSP40776.2020.9054441.

Rigoll, G., Müller, R., Schuller, B., 2005. Speech emotion recognition exploiting acoustic and linguistic information sources. Specom 61–67.

Russell, J.A., 1980. A circumplex model of affect. J. Personal. Soc. Psychol. 39 (6), 1161–1178. http://dx.doi.org/10.1037/h0077714.

Sahu, S., Mitra, V., Seneviratne, N., Espy-Wilson, C., 2019. Multi-modal learning for speech emotion recognition: An analysis and comparison of ASR outputs with ground truth transcription. In: Interspeech 2019. ISCA, ISCA, pp. 3302–3306. http://dx.doi.org/10.21437/Interspeech.2019-1149.

Sailunaz, K., Dhaliwal, M., Rokne, J., Alhajj, R., 2018. Emotion detection from text and speech: a survey. Soc. Netw. Anal. Min. 8 (1), 1–26. http://dx.doi.org/10.1007/s13278-018-0505-2.

Santoso, J., Yamada, T., Makino, S., Ishizuka, K., Hiramura, T., 2021. Speech emotion recognition based on attention weight correction using word-level confidence measure. In: Interspeech 2021. ISCA, ISCA, pp. 1947–1951. http://dx.doi.org/10.21437/Interspeech.2021-411.

Scherer, K.R., 2005. What are emotions? And how can they be measured? Soc. Sci. Inf. 44 (4), 695–729. http://dx.doi.org/10.1177/0539018405058216, URL: http://journals.sagepub.com/doi/10.1177/0539018405058216.

Schmitt, M., Cummins, N., Schuller, B.W., 2019. Continuous emotion recognition in speech — Do we need recurrence? In: Proc. Interspeech 2019. ISCA, ISCA, pp. 2808–2812. http://dx.doi.org/10.21437/Interspeech.2019-2710.

Schuller, B., 2002. Towards intuitive speech interaction by the integration of emotional aspects. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Vol. 6. pp. 481–486. http://dx.doi.org/10.1109/icsmc.2002.1175635.

Schuller, B., 2011. Recognizing affect from linguistic information in 3D continuous space. IEEE Trans. Affect. Comput. 2 (4), 192–205. http://dx.doi.org/10.1109/T-AFFC.2011.17.

Schuller, B.W., 2018. Speech emotion recognition two decades in a nutshell. Commun. Acm 61 (5), http://dx.doi.org/10.1145/3129340.

Schuller, B.W., Batliner, A., Bergler, C., Messner, E.-M., Hamilton, A., Amiriparian, S., Baird, A., Rizos, G., Schmitt, M., Stappen, L., Baumeister, H., Macintyre, A.D., Hantke, S., 2020. The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. In: Proc. Interspeech 2020. ISCA, pp. 2042–2046. http://dx.doi.org/10.21437/Interspeech.2020-0032.

Schuller, B., Müller, R., Lang, M., Rigoll, G., 2005. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: Proc. Interspeech 2005, pp. 805–808.

Schuller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1. IEEE, pp. 577–580. http://dx.doi.org/10.1109/icassp.2004.1326051.

Schuller, B., Valstar, M., Eyben, F., Cowie, R., Pantic, M., 2012. Avec 2012 - the continuous audio/visual emotion challenge. In: ICMI'12 - Proceedings of the ACM International Conference on Multimodal Interaction. pp. 449–456. http://dx.doi.org/10.1145/2388676.2388776.

Schuller, B., Vlasenko, B., Arsic, D., Rigoll, G., Wendemuth, A., 2008. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. In: 2008 IEEE International Conference on Multimedia and Expo, ICME 2008 - Proceedings. pp. 1333–1336. http://dx.doi.org/10.1109/ICME.2008.4607689.

Sebastian, J., Pierucci, P., 2019. Fusion techniques for utterance-level emotion recognition combining speech and transcripts. In: Proc. Interspeech 2019. ISCA, ISCA, pp. 51–55. http://dx.doi.org/10.21437/Interspeech.2019-3201.

Shen, G., Lai, R., Chen, R., Zhang, Y., Zhang, K., Han, Q., Song, H., 2020. WISE: Word-level interaction-based multimodal fusion for speech emotion recognition. In: Proc. Interspeech 2020, pp. 369–373.

Shon, S., Brusco, P., Pan, J., Han, K.J., Watanabe, S., 2021. Leveraging pre-trained language model for speech sentiment analysis. In: Interspeech 2021. ISCA, ISCA, pp. 3420–3424. http://dx.doi.org/10.21437/Interspeech.2021-1723, arXiv:2106.06598.

Siriwardhana, S., Reis, A., Weerasekera, R., Nanayakkara, S., 2020. Jointly fine-tuning "BERT-like" self supervised models to improve multimodal speech emotion recognition. In: Proc. Interspeech 2020. pp. 3755–3759, arXiv:2008.06682 URL: http://arxiv.org/abs/2008.06682.

Soğancıoğlu, G., Verkholyak, O., Kaya, H., Fedotov, D., Cadée, T., Salah, A.A., Karpov, A., 2020. Is everything fine, grandma? Acoustic and linguistic modeling for robust elderly speech emotion recognition. In: Proc. Interspeech 2020. ISCA, ISCA, pp. 2097–2101. http://dx.doi.org/10.21437/Interspeech.2020-3160.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. In: 3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings. pp. 1–14, arXiv:1412.6806 URL: http://arxiv.org/abs/1412.6806.

Steidl, S., 2009. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech (Ph.D. thesis). p. 250.

Tian, L., Moore, J.D., Lai, C., 2015. Recognizing emotions in dialogues with acoustic and lexical features. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 737–742. http://dx.doi.org/10.1109/ACII.2015.7344651, URL: http://ieeexplore.ieee.org/document/7344651/.

Tian, L., Moore, J., Lai, C., 2016. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In: 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 565–572. http://dx.doi.org/10.1109/SLT.2016.7846319, URL: http://ieeexplore.ieee.org/document/7846319/.

Tripathi, S., Kumar, A., Ramesh, A., Singh, C., Yenigalla, P., 2019. Deep learning based emotion recognition system using speech features and transcriptions. In: International Conference on Computational Linguistics and Intelligent Text Processing. arXiv:1906.05681.

Tzirakis, P., Nguyen, A., Zafeiriou, S., Schuller, B.W., 2021. Speech emotion recognition using semantic information. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, ISBN: 978-1-7281-7605-5, pp. 6279–6283. http://dx.doi.org/10.1109/ICASSP39728.2021.9414866, URL: https://ieeexplore.ieee.org/document/9414866/.

Väyrynen, E., 2014. Emotion Recognition from Speech Using Prosodic Features (Ph.D. thesis). University of Oulu.

Viraraghavan, V.S., Gavas, R.D., Ramakrishnan, R.K., 2020. Role of emotion words in detecting emotional valence from speech. In: SMM20, Workshop on Speech, Music and Mind 2020. ISCA, ISCA, pp. 26–30. http://dx.doi.org/10.21437/SMM.2020-6.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007. Combining frame and turn-level information for robust recognition ofemotions within speech. In: Proc. Interspeech 2007, pp. 2712–2715.

Wang, Y., Shen, G., Xu, Y., Li, J., Zhao, Z., 2021. Learning mutual correlation in multimodal transformer for speech emotion recognition. In: Interspeech 2021. ISCA, ISCA, pp. 4518–4522. http://dx.doi.org/10.21437/Interspeech.2021-2004.

Warriner, A.B., Kuperman, V., Brysbaert, M., 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. Behav. Res. Methods 45 (4), 1191–1207. http://dx.doi.org/10.3758/s13428-012-0314-x, URL: http://link.springer.com/10.3758/s13428-012-0314-x.

Whissell, C., 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. Psychol. Rep. 105 (2), 509–521. http://dx.doi.org/10.2466/PR0.105.2.509-521, URL: http://journals.sagepub.com/doi/10.2466/PR0.105.2.509-521.

Wu, C.H., Lin, J.C., Wei, W.L., 2014. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. In: APSIPA Trans. Signal Inf. Process.. 3, pp. 1–18. http://dx.doi.org/10.1017/ATSIP.2014.11.

Wu, W., Zhang, C., Woodland, P.C., 2021. Emotion recognition by fusing time synchronous and time asynchronous representations. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 2021-June. IEEE, pp. 6269–6273. http://dx.doi.org/10.1109/ICASSP39728.2021.9414880, arXiv:2010.14102 URL: https://ieeexplore.ieee.org/document/9414880/.

Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., Li, X., 2019. Learning alignment for multimodal emotion recognition from speech. In: Proc. Interspeech 2019. ISCA, ISCA, pp. 3569–3573. http://dx.doi.org/10.21437/Interspeech.2019-3247, arXiv:Interspeech.2019-3247.

Yang, Z., An, Z., Fan, Z., Jing, C., Cao, H., 2020. Exploration of acoustic and lexical cues for the interspeech 2020 computational paralinguistic challenge. In: Proc. Interspeech 2020, pp. 2092–2096.

Yannakakis, G.N., Cowie, R., Busso, C., 2021. The ordinal nature of emotions: An emerging approach. IEEE Trans. Affect. Comput. 12 (1), 16–35. http://dx.doi.org/10.1109/TAFFC.2018.2879512.

Ye, W., Fan, X., 2014. Bimodal emotion recognition from speech and text. Int. J. Adv. Comput. Sci. Appl. 5 (2), http://dx.doi.org/10.14569/ijacsa.2014.050204.

Yeh, S.-l., Lin, Y.-s., Lee, C.-c., 2020. Speech representation learning for emotion recognition using end-to-end ASR with factorized adaptation. In: Proc. Interspeech 2020, 2020, pp. 536–540.

Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., Vepa, J., 2018. Speech emotion recognition using spectrogram & phoneme embedding. In: Proc. Interspeech 2018. pp. 3688–3692. http://dx.doi.org/10.21437/Interspeech.2018-1811.

Yoon, S., Byun, S., Dey, S., Jung, K., 2019. Speech emotion recognition using multi-hop attention mechanism. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2822–2826. http://dx.doi.org/10.1109/ICASSP.2019.8683483.

Zadeh, A., Liang, P.P., Vanbriesen, J., Poria, S., Tong, E., Cambria, E., Chen, M., Morency, L.P., 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proc. ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Vol. 1. Association for Computational Linguistics, pp. 2236–2246. http://dx.doi.org/10.18653/v1/p18-1208.

Zhang, B., Khorram, S., Provost, E.M., 2019. Exploiting acoustic and lexical properties of phonemes to recognize valence from speech. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 2019-May. IEEE, pp. 5871–5875. http://dx.doi.org/10.1109/ICASSP.2019.8683190, URL: https://ieeexplore.ieee.org/document/8683190/.

Zhang, A., Lipton, Z.C., Li, M., Smola, A.J., 2020. Dive into deep learning.

Zhao, H., Ye, N., Wang, R., 2018. Transferring age and gender attributes for dimensional emotion prediction from big speech data using hierarchical deep learning. In: 2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS). IEEE, pp. 20–24. http://dx.doi.org/10.1109/BDS/HPSC/IDS18.2018.00018, URL: https://ieeexplore.ieee.org/document/8552276/.