SIP (2020), vol. 9, e17, page 1 of 12
doi:10.1017/ATSIP.2020.14

ORIGINAL PAPER

Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning

BAGUS TRIS ATMAJA[1,2] AND MASATO AKAGI[1]

The majority of research in speech emotion recognition (SER) is conducted to recognize emotion categories. Recognizing dimensional emotion attributes is also important, however, and it has several advantages over categorical emotion. For this research, we investigate dimensional SER using both speech features and word embeddings. The concatenation network joins acoustic networks and text networks from bimodal features. We demonstrate that those bimodal features, both are extracted from speech, improve the performance of dimensional SER over unimodal SER either using acoustic features or word embeddings. A significant improvement on the valence dimension is contributed by the addition of word embeddings to SER system, while arousal and dominance dimensions are also improved. We proposed a multitask learning (MTL) approach for the prediction of all emotional attributes. This MTL maximizes the concordance correlation between predicted emotion degrees and true emotion labels simultaneously. The findings suggest that the use of MTL with two parameters is better than other evaluated methods in representing the interrelation of emotional attributes. In unimodal results, speech features attain higher performance on arousal and dominance, while word embeddings are better for predicting valence. The overall evaluation uses the concordance correlation coefficient score of the three emotional attributes. We also discuss some differences between categorical and dimensional emotion results from psychological and engineering perspectives.

Keywords: Speech emotion recognition, Multitask learning, Feature fusion, Dimensional emotion, Affective computing

Received 11 December 2019; Revised 27 April 2020

I. INTRODUCTION

Speech emotion recognition (SER) has become more important recently. It is now used in call-center applications for analyzing both customers and call-center staff. In other applications, a voice assistant application uses SER technology to detect the affective state and mood of the user for more natural interaction and wellbeing measurement. In the near future, other applications like driver emotion detection, affective humanoid robots, and voice analysis of gaming users will enter the market.

The research leading to the implementation of the various applications above is the backbone of the acceleration from psychological theory to engineering implementation. Research on automatic speech emotion recognition started at the end of the 1990s, following the success of emotion recognition from facial expressions. This research using only the speech modality is intended to extract human emotion when other modalities, like facial expressions and movements, cannot be recorded. At that time, call-center applications, which are now being implemented, were the target of future implementation. Nevertheless, this research area is still growing and suffering from some difficulties.

The growth of speech/vocal emotion research is mainly conducted in terms of recognizing/predicting categorical emotions. In this approach, the predictor is trained to detect the emotion of a given utterance, whether it is joy, calm, anger, sorrow, or another emotion. The current method applying deep learning techniques for the classifier has achieved high accuracies, as reported in [1–4]. Many researchers in psychology have argued, however, that it is necessary to go beyond categorical emotions [5]. Apart from enabling degrees of emotions in a continuous space, using emotional dimensions offers other advantages such as investigation of cognitive mental states and measurement of productivity and burnout [6].

In dimensional emotions, some dimensions or attributes of emotions are mapped onto a two- (2D), three- (3D), or four-dimensional (4D) space. The 2D space consists of two attributes: valence (positive versus negative) and arousal/activation (high versus low excitation). Most

[1]Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[2]Sepuluh Nopember Institute of Technology, Kampus ITS Sukolilo, Surabaya 60111, Indonesia

Corresponding author:
Bagus Tris Atmaja
Email: bagus@jaist.ac.jp

research on dimensional emotion recognition has used this 2D approach for describing emotional degrees [7–9]. Other researchers have used a 3D approach in which the third dimension is either dominance (weak versus strong) [10, 11] or likability/liking [12, 13]. Finally, a 4D model incorporates expectancy (predictiveness by the subject with respect to conversation) [14, 15].

The research described here uses a 3D emotion model with valence, arousal, and dominance: the VAD model. While a 2D emotion model can be used to characterize categorical emotion, the prediction of dominance can be used to perform other analysis of aspects such as the energy or urgency of speech. In addition to observing blended emotion, the response or performance of emotion recognition for particular emotion attributes can be analyzed on a continuous scale. Furthermore, in automatic emotion recognition development, one advantage of using dimensional emotion is that there is no need to balance the distribution of the dataset.

The benefits of the dimensional approach above require high performance on recognizing emotional dimensions. For instance, if the performance of dimensional emotion recognition is weak, then conversion from dimensional to categorical emotion cannot be performed. Unfortunately, the current results on dimensional SER suffer from this condition [7, 16]. Another issue is whether it is necessary to combine acoustic features from the speech modality with other features like facial expressions and head and hand movements [17].

Our current research is conducted to tackle both issues, by using speech or speech/acoustic features and word embeddings to improve the performance of SER. Because the target implementation is a speech-based application (e.g. a call-center application), the only modality used to extract emotion from human speech is the acoustic information. Text information can now be extracted accurately, however, because of the advancement of automatic speech recognition (ASR) technology. Hence, text transcription can be used to generate word embeddings, providing text features for the emotion classifier. Thus, by using both acoustic and linguistic information, we expect a significant improvement in dimensional SER.

Although the use of both acoustic and linguistic information for SER is not new, most reported works used categorical emotion recognition rather than dimensional emotion recognition. A challenge called the Continuous Audio/Visual Emotion Challenge (AVEC) has been held since 2012 to promote this dimensional affective approach [15]. One result on the use of acoustic and linguistic information, a report by Tian *et al.* [18], showed low performances, i.e. mean accuracy of 69.9% for AVEC2012 dataset and 55.3% for Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. The authors used hierarchical fusion to combine several acoustic and linguistic features.

Using accuracy for dimensional emotion recognition may not be relevant, as the task is regression. Another researcher reported low concordance correlation coefficient (CCC) score on predicting valence, arousal, and dominance simultaneously on two-group datasets using acoustic features [11]. The metric used in the latter report is more relevant and has become the gold standard for measuring dimensional emotion recognition performance. Since the evaluation metric use CCC as the score, the use of CCC loss as the cost function is more reasonable than other cost functions.

Because the goal of dimensional emotion recognition is to predict three outputs, i.e. valence–arousal–dominance (VAD), simultaneously, we propose CCC-based multitask learning (MTL). In this MTL approach, the model is trained to leverage the results from the three emotional dimensions. In the traditional approach, which is called single-task learning (STL), the model is only trained to maximize the performance of a single variable by minimizing a loss function. Our MTL approach uses the opposite of the CCC as the loss function, with the dimensional emotion models trained to minimize the loss function for three variables. Section 4 describes the details of this MTL approach.

The contribution of this paper is the combination of acoustic and text information for dimensional SER via concatenation networks with the CCC-based MTL proposal to train those networks, which is not shown in the previous reported works. We extend our work to evaluate the following sub-issues: (1) which pair of architectures (Acoustic + Text) including their optimal dropout rates performs best; (2) evaluation of different MTL methods with zero parameters, two parameters, and three parameters; and (3) the similarities and differences in conducting categorical and dimensional emotion recognition on the same dataset. First, however, the next three sections describe the related work, feature sets, and the architectures of dimensional SER systems.

II. RELATED WORK

In this section, we reviewed some research done in the past closely related to this research theme. The advantages and disadvantages of its proposed method will be briefly discussed. At the end of this section, we summarized how this research differs from those reviewed research and contributes to filling the existing gaps.

In recent years, there are several attempts to improve the performance of SER. Parthasarathy and Busso proposed a framework to learn shared feature representations that maximize the performance of regression models [11]. Their model used mean squared error (MSE) of valence, arousal, and dominance with two weighting factors. Atmaja and Akagi [19], however, showed that correlation-based loss function is more effective than MSE for MTL dimensional SER.

Sridhar *et al.* [20] used higher regularization (dropout rate) for valence than for arousal and dominance. They reported an improvement from CCC score of 0.29 to 0.31 on MSP-Podcast dataset [21]. Apart from that proposal using higher regularization, an approach using lexical feature was proposed by Aldeneh *et al.* [22] using

pretrained word2vec with Mel Filterbank for the acoustic feature. They converted the regression task into the classification task by dividing valence scores into negative, neutral, and positive categories. The method improved unweighted average recall (UAR) from 0.59 with acoustic modality to 0.694 with acoustic-lexical modalities on the IEMOCAP dataset. Using a similar idea, Zhang *et al.* used acoustic and lexical features to recognize valence from the speech on three valence categories IEMOCAP dataset. Instead of extracting lexical feature of words, they extracted lexical feature of phonemes, i.e. 40-dimensional unique phoneme including an "out of vocabulary" label. The method improved UAR from 0.64 with acoustic modality to 0.74 with acoustic-lexical (phonemes) modality. All of those research aimed to improve valence prediction only on categorical emotion (valence) recognition.

In [7], the authors used semantic features from the affective dictionary along with the MFCC feature to predict valence and arousal. They reduced mean absolute error from 1.98 and 1.29 (acoustic), for valence and arousal, to 1.40 and 1.28 (acoustic + semantic). Although the authors attempted to predict the degree of dimensional emotions, only valence and arousal are counted. The use of the affective word is also limited by the number of vocabularies; words that are not listed in the dictionaries can not be included in the computation of semantic features.

In [1, 23], the authors used different deep learning architectures to predict categorical emotion from both speech and text. Some authors used phonemes instead of text for predicting emotion category, such as in [3, 24] and another author compared text feature from ASR with manual transcription to investigate the effectiveness of its combination with acoustic features for categorical emotion recognition [25]. Those research, although used audio and text features, only predicted categorical emotion.

Huang *et al.* proposed to use audio-word-based embedding with convolutional neural network (CNN) [26]. First, a set of low-level descriptors (LLD) are extracted, then, a set of audio words is generated from an audio codebook. The word2vec-based embedding is applied to obtain word vector for each audio word. While word embeddings are used to quantify the similarities among audio words, CNN is adopted to characterize the temporal variation of LLD. Using this technique, the authors reported an accuracy improvement of 5.6% compared to long short-term memory (LSTM) and 3.4% compared to LLD raw features. Besides predicting emotion categories only, this proposed technique may suffer from the computation time when it is implemented in real-time application. It needs several steps to be processed in a sequence: LLD extraction, audio codebook generation, word2vec extraction, and CNN modeling.

In contrast, our work incorporated MTL with CCC losses not only for speech or text networks but also for a combination of both. We also investigate MTL models with different numbers of parameters and their impact on overall performance. We target three emotion dimensions in floating-point degrees, while others convert those degrees into categories. Additionally, we use input features which can be quickly extracted, e.g. mean and standard deviation (std) of LLDs from speech features and word embeddings that are feasible for future real-time implementation.

III. FEATURE SETS

The two most important aspects of machine learning are feature extraction and classification. For this paper, we evaluated both aspects for SER. While the main idea was the fusion of features from acoustic and text information, the classifiers were also evaluated for optimality. This section describes what the features were and why they were used in this research.

Features as the input of SER system play an important role in its performance. We evaluated two acoustic features, including both LLD and high-level statistical functions, and three different text features. We introduce the mean and std of pyAudioAnalysis [27] to investigate the effectiveness of these functionals, which proved effective for GeMAPS feature set [9]. In total, four acoustic features and three text features are searched to find the best one for each modality. This pair will be evaluated along with other pairs.

Figure 1 shows the block diagram of proposed system. Each acoustic feature set is evaluated in the acoustic network, as well as each text feature set. A concatenation network joins a pair of acoustic-text networks to evaluate those different feature sets. The proposed MTL system minimizes the cumulative errors from three variables: valence, arousal, and dominance, simultaneously based on a defined loss function. For now, we will describe the input features used in that system.

A) Acoustic features

We extend the work of our previous research [1] with the Geneva Minimalistic Acoustic Parameter Set (GeMAPS). GeMAPS represents a proposal to standardize the acoustic features used for voice research and affective computing [28]. The feature set consists of 23 acoustic LLDs such as

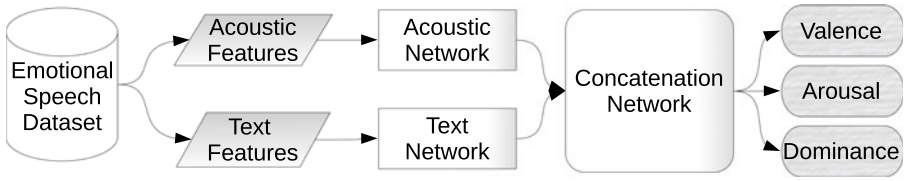

Fig. 1. Block diagram of proposed dimensional emotion recognition system from speech features and word embeddings; the multitask learning will train the weights of the network based on acoustic and text input features to predict degrees of valence, arousal, and dominance simultaneously.

Table 1. Acoustic feature sets: GeMAPS [28] and pyAudioAnalysis [27]. The numbers in parentheses indicate the total numbers of features (LLDs).

	GeMAPs (23)	pyAudioAnalysis (34)
LLDs	Intensity, alpha ratio, Hammarberg index, spectral slope 0–500 Hz, spectral slope 500–1500 Hz, spectral flux, 4 MFCCs, Fo, jitter, shimmer, harmonics-to-noise ratio (HNR), harmonic difference H1-H2, harmonic difference H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.	Zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, 13 MFCCs, 12 chroma vectors, chroma deviation.
HSFs	Mean, std	Mean, std

Fo, jitter, shimmer, and formants, as listed in Table 1. As an extension of GeMAPS, eGeMAPS also includes statistics derived from the LLDs, such as the minimum, maximum, mean, and other statistical values. Including the LLDs, the total number of features in eGeMAPS is 88. The statistical values are often called high-level statistical functions (HSF). The authors of [9] found, however, that using only the mean and standard deviation from the LLDs in GeMAPS gave better results than using eGeMAPS and audiovisual features. We thus incorporated that finding in our research.

In addition, we extended the idea of using the mean and std to another feature set called pyAudioAnalysis (pAA) [27], which was also used in the previous research [1] upon which this research was built. Specifically, pAA is an open-source Python library that provides a wide range of audio analysis procedures, including feature extraction, classification of audio signals, supervised and unsupervised segmentation, and content visualization. Of those capabilities, only feature extraction was used in this research. A total of 34 features were extracted for each frame, including time-domain, spectral-domain, and other features as listed in Table 1. Although pAA was designed for multipurpose audio analysis, some research has reported its effectiveness for affective speech research, including speech emotion recognition and depression classification (e.g. [25]).

For both GeMAPS and pAA feature sets, frame-based processing was applied with a 25 ms window length and a 10 ms hop size. The longest utterance was used to define the frame margins, as utterances shorter than the longest one were padded with zeros to achieve that length. The total feature vectors from all utterances were concatenated to give the resulting feature sizes of (10 039, 3409, 23) for GeMAPS and (10 039, 3411, 34) for pAA. The difference in the number of longest-frame sequences between GeMAPS and pAA was due to the different processing methods. By extracting both the LLDs and HSFs for each feature set, we obtained four acoustic feature set variants, as listed in Table 1. Note that "HSF" in the table refers only to the mean and std of each corresponding LLD. The size of the HSF input was (10 039, 46) for GeMAPS and (10 039, 68) for pAA.

B) Word embeddings

Numeric vectors are needed to feed the input text to a deep learning system. One of the common features used in text processing is word embeddings or lexical features (text features). A word embedding is a vector representation of a word. Numerical values in the form of vectors are used to enable a computer to process text data, as it can only process numerical values. The values are points (numeric data) in a space whose number of dimensions equals the vocabulary size. The word representations embed those points in a feature space of lower dimension [29]. In the original space, every word is represented by a one-hot vector, with a value of 1 for the corresponding word and 0 for other words. The element with a value of 1 is converted into a point in the range of the vocabulary size.

Table 2. Word embedding feature sets.

Feature	Description
WE	Word embedding obtained directly from word sequence in text transcription.
GloVe	WE weighted by pretrained GloVe embeddings (300 dimensions, 2.2M vocabulary size).
FastText	WE weighted by pretrained FastText word vectors (300 dimensions, 2M vocabulary size).

To obtain a vector of each word in an utterance, that utterance in the dataset must be tokenized. Tokenization is the process of dividing an utterance into a number of constituent words. For example, the text "That's out of control" from the IEMOCAP dataset is tokenized as ["That's", "out", "of", "control"]. Thus, if the number of vocabulary items is 3438, then the obtained word vector contains integers less than or equal to 3438, for instance:

```
text_vector = [42, 44, 11, 471].
```

In the above text vector example, each word is coded into a one-dimensional vector, e.g. "out" is coded into 44. In this research, we used a 300-dimensional vector with the 100 longest token sequences, e.g. "out" has a 300-dimensional vector, as well as all 3438 vocabularies. A set of zeros can be padded before or after the obtained vector to obtain a fixed-length word embedding for each utterance. This research used pre-padded zeros.

Instead of directly converting word sequences into vectors, several researchers have obtained effective vectors from certain words [30–32]. The vectors of those words can be used to weight word vectors obtained previously. We thus used two pretrained word embedding models to weight the original word embeddings (WE): the GloVe embedding [31] and FastText [32]. Table 2 describes those word embeddings. Note that we did not synchronize speech and text features

per word (but per sentence) since we assumed that processing of both information is independent of each other. A spoken word may have different time segments, depends on how it is pronounced, but the semantic of that word remains the same. The WE, GloVe, and FastText treat each word to have the same vector regardless of the context.

IV. DIMENSIONAL SPEECH EMOTION RECOGNITION SYSTEM

A) Architecture for unimodal feature

The conventional modern SER process consists of two main blocks: an acoustic feature and an acoustic network/classifier. The acoustic feature is fed into the acoustic network to predict the output of the emotion label, given input data-label pairs. The performance of the acoustic networks is then evaluated to determine the combination of the acoustic network and text network in the early fusion method.

Two main deep learning architectures were evaluated in this research: LSTM and a CNN. These two architectures are the most-used deep learning systems. Apart from that, we also sought to re-observe the finding in [33] that a CNN achieved higher performance in SER. In this case, we used acoustic features similar to those reported there, i.e. the mean and std of GeMAPS LLDs.

Figure 2 shows the architectures of both the LSTM and CNN used for dimensional SER from a unimodal feature. Although this figure represents the acoustic network, the only difference from the text network was the last layer: the text network used a dense layer instead of a flatten layer. This structure was found effective from an empirical study. Both architectures consisted of three main layers. We designed them to have a similar number of trainable parameters for comparison.

For acoustic-based emotion recognition, four different features were evaluated as inputs to the unimodal system. For instance, first consider feeding pAA LLDs into the LSTM network. Batch normalization was performed at the first layer before entering the LSTM stack. We kept the full sequence output of the last LSTM layer (i.e. return_sequence=True) and flattened it with a dropout rate of $p = 0.3$. After that, three dense layers with one unit were added to generate a continuous value of valence, arousal, and dominance. The model was trained with a CCC-based loss function (explained in Section 5). Then, the same workflow was applied for the pAA HSF, GeMAPS LLD, and GeMAPS HSF feature sets. For each architecture (LSTM or CNN), we chose the result with the highest performance, one out of four, to combine with the result from the other modality, i.e. the text network. Hence, we obtained the two best architectures from the acoustic network and the two best architectures from the text network.

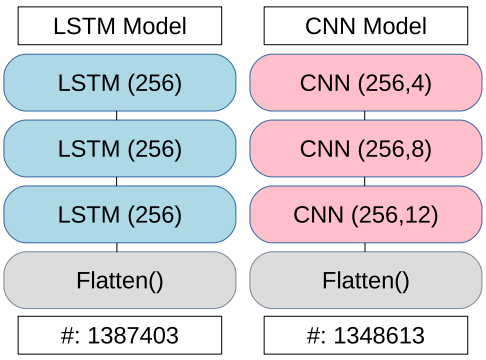

Fig. 2. Overview of the LSTM and CNN models used in the acoustic network. A number in parentheses represents the number of units, with the second number for the convolutional layers representing the kernel size. The numbers after # represent the numbers of trainable parameters. For the text network, the flatten layer was replaced by a dense layer.

For text-based emotion recognition, the three features explained previously were fed separately into the text network, which did not use a batch normalization layer as in the acoustic network. Instead, an embedding layer was used for converting positive integers into dense vectors, and for weighting with a pretrained model, if necessary (for the GloVe and FastText features). Then, the three LSTM or CNN layers were stacked. Another difference from the acoustic network was that the last LSTM layer only outputted the last sequence (i.e.return_sequence=False) and combined it in a dense layer with 64 units and a dropout rate of $p = 0.3$, before the output was split into three one-unit dense layers. As in the acoustic network, the last three dense layers were used to predict a continuous value of valence, arousal, and dominance. All the LSTM layers in this research were built using a GPU-based CuDNN [34] implemented with the Keras toolkit [35].

In total, we obtained eight results from the acoustic network (4 feature sets $\times$ 2 architectures) and six results from the text network (3 feature sets $\times$ 2 architectures). The best architectures from the acoustic and text networks were then paired in a combined system. In addition, the best architectures from each modality were paired with a dense network to investigate our previous findings [1]. To obtain consistent results, a fixed random number was initialized at the beginning. All experiments were also bounded by a callback method using early stopping criteria with a patience value of 10. This means that, if the training process did not improve within 10 epochs, it stopped and used the last highest-score model to predict the test data.

B) Architecture for bimodal feature fusion

A naive method for combining multimodal features in pattern recognition is by using an early fusion method that can be performed, for instance, by feature concatenation or network concatenation. The former approach combines features (e.g. the acoustic and text features) into the same network, while the latter approach combines the networks for different modalities. We used network concatenation by combining the acoustic and text networks with dense layers. This strategy was based on the assumption that human perception processes acoustic and linguistic information differently. Apart from that assumption, the combination of

two networks also had the benefit of not requiring the two feature sets to have the same size or dimension.

Mathematically, the combined bimodal network was formulated as in equation (1). Here, $f(y)$ denotes the output of the corresponding layer; W_1, W_2 denote the weights from previous layers (a: acoustic; t: text) and the current hidden layer, respectively; x_a and x_t are the acoustic feature and word embedding, respectively; b is a bias; and g is a linear activation function. Thus, the output of the first dense layer, after concatenation, was the following:

$$f(y) = W_2 g([W_{1a}^\top xa + b_{1a}; W_{1t}^\top xt + b_{1t}]) + b_2. \quad (1)$$

The above process continued because we used two dense layers with 64 and 32 units before splitting the output of the second dense layer into the last layers. The last layers are three one-unit dense layers with linear activations for regressing the degree of valence, arousal, and dominance.

We constructed the combined bimodal networks by concatenating the previous unimodal networks. The essential modification was the insertion of a concatenation layer with dense layers after the flatten or dense layer following the last LSTM or CNN layer. This concatenation block, with a dropout rate of $p = 0.4$, was followed by three one-unit dense layers. These last layers were used to predict the degree of valence, arousal, and dominance as in the unimodal network. Figure 3 shows a typical implementation of the combined bimodal network fusion using the HSFs of pAA for the acoustic network and GloVe embeddings for the text network.

Eight pairs of acoustic-text networks were evaluated in combined bimodal systems. Four pairs derived from the highest scores for each architecture in each modality, while the other four pairs derived from pairing the best architecture from each modality with a dense network. We evaluated all eight pairs under the same conditions (number of units, dropout rate, batch size). For each pair, we experimented 20 times and observed its deviation with the same callback criteria as used for the unimodal networks. This repetition was conducted to determine the most reliable pair.

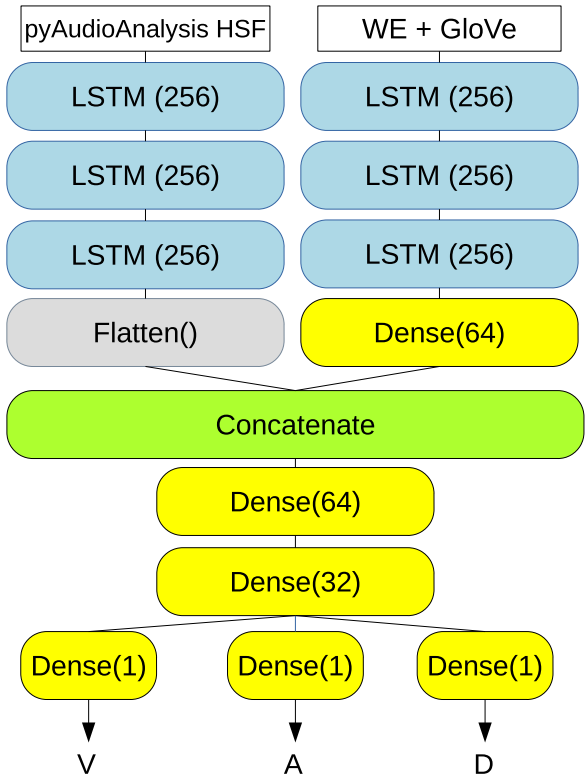

Fig. 3. Architecture of a combined system using LSTMs for both the acoustic and text networks. HSF: high-level statistical functions; WE: word embedding; V: valence; A: arousal; D: dominance.

V. MULTITASK LEARNING

MTL is an approach to jointly learn multidimensional targets in a training process to estimate those targets. For example, if a target consists of variables y_1 and y_2, instead of optimizing either y_1 or y_1, MTL would optimize both variables. In contrast, the traditional approach of optimizing a single variable is called STL.

In this research, an MTL approach is proposed to simultaneously learn three emotional dimensions: valence, arousal, and dominance. The proposed approach begins from a CCC, because the goal is to optimize the CCC score. Here, the CCC is a measure of the relation between predicted and true value, with the score penalized if the prediction shifts the true value. The CCC is formulated as

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \quad (2)$$

where ρ_{xy} is the Pearson coefficient correlation between x and y, σ is the standard deviation, and μ is the mean. This CCC formulation is based on Lin's calculation [36]. The range of CCC is from -1 (perfect disagreement) to 1 (perfect agreement). Therefore, the CCC loss function (CCCL) to maximize the agreement between the true value and the predicted emotion can be defined as

$$CCCL = 1 - CCC. \quad (3)$$

In STL, the loss function would be one for either valence ($CCCL_V$), arousal ($CCCL_A$), or dominance ($CCCL_D$). In MTL, when the total CCC loss is used as a single metric for all arousal, valence, and dominance, $CCCL_{tot}$ is the following combination of those three CCC loss functions:

$$CCCL_{tot} = CCCL_V + CCCL_A + CCCL_D. \quad (4)$$

We refer to this MTL equation as "MTL without parameters", because there is no weighting among valance, arousal, and dominance. In this case, the relation among the three emotional dimensions is determined by joint learning in the training process. As it has been stated that these three emotional dimensions are related in a systemic manner [5], we introduce two parameters to weight the valance and

arousal, with the weight for dominance determined by subtracting those two parameters from 1. This MTL with two parameters is defined as

$$CCCL_{tot} = \alpha\, CCCL_V + \beta\, CCCL_A + (1 - \alpha - \beta)\, CCCL_D, \quad (5)$$

where α and β are the weighting factors for the valence and arousal loss functions, respectively. This proposed MTL is similar to that defined in [11]. While those authors used the MSE as the loss function, we have proposed using this CCC-based loss function. In addition, we can add a parameter γ for dominance to obtain independent scales among valence, arousal, and dominance. The resulting MTL with three parameters is defined as

$$CCCL_{tot} = \alpha\, CCCL_V + \beta\, CCCL_A + \gamma\, CCCL_D. \quad (6)$$

For comparison with the previous MTL without parameters; α, β, and γ were set to 1 in that equation (4), which can be seen as a special case in this MTL with three parameters.

These MTL approaches compare the predicted output from the three one-unit dense layers with the ground truth labels. The training process mechanism relies on the above loss function. Hence, the performance of the produced model is based on this mechanism, too. The choice of loss function is a critical aspect of machine learning, and we thus proposed this MTL based on the CCC loss to learn valence, arousal, and dominance concurrently.

We implemented MTL approaches for both LSTM and CNN and both unimodal and bimodal systems. We split last dense layer for each system into three dense layers with one-unit node and compute the loss functions above (equations (4), (5), and (6)). The results presented in the unimodal feature (Section 6.2) are obtained using MTL with two parameters (equation (5)), while the results presented in bimodal feature fusion (Section 6.3) are obtained by both MTL with two parameters and all three MTL approaches.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The code to reproduce these experimental results is available in the following repository: http://github.com/bagustris/dimensional-ser. The raw dataset used for experiment resource can be obtained from https://sail.usc.edu/iemocap.

A) Dataset

We used the IEMOCAP dataset, which is provided by the University of Southern California (USC). It includes recordings of 10 actors in dyadic sessions with markers on the speech and transcriptions. It also includes information on the actors's facial expressions and head and hand movements during both scripted and spontaneous spoken communication scenarios, but that data were not used in this research. The dataset is freely available upon request, including its labels for both categorical and dimensional emotions. While previous research used categorical emotions [1], this research targeted dimensional emotion scores. These scores in terms of valence, arousal, and dominance ranged from 1 to 5 and were annotated via a Self-Assessment Manikin (SAM) evaluation. Some labels with scores lower than 1 or higher than 5 were normalized to 1 or 5, respectively. All labels were then converted to floating-point values in a scale $[-1, 1]$ when fed to a deep learning system. The conversion is given by the following MinMax scaling [37],

$$x_{std} = \frac{x - x_{min}}{x_{max} - x_{min}}, \quad (7)$$

$$x_{scaled} = x_{std} \times (max - min) + min, \quad (8)$$

where x is the original label score, max is 1, min is -1, and x_{scaled} is the new label score.

The speech data used to extract acoustic features had a 16 kHz single channel per sentence. The manual transcription of speech in the dataset was also used to generate word embeddings from word sequences, instead of using automatic transcription. No further preprocessing was applied to either feature, except as explained in this paper.

We split the dataset into training and test portions with a ratio of 7869:2170. The splitting criterion was a speaker-independent condition in which we excluded the last session (denoted as Session 6) in the training process (i.e. leave one session out, or LOSO). About 20% of the training data were used for validation (1573 utterances). We did not use a speaker-dependent condition, because the speaker-independent condition was more challenging and more suitable for real-life analysis.

B) Results on unimodal feature

A unimodal feature is a feature set from either acoustic or text (e.g. pAA LLD). At first, we trained each feature set on both LSTM and CNN classifier independently. Tables 3 and 4 list our unimodal dimensional emotion results from those acoustic and text networks, respectively. Each table lists the scores for valence, arousal, and dominance in terms of the CCC, along with averaged CCC scores to determine which method performed better. We grouped the results by modality and architecture. They all use the same metric scale and were obtained under the same conditions. Hence, these results can be compared directly with each other.

In the acoustic-based modality, we obtained consistent results among the feature sets on both architectures. From bottom to top, the performance order was pAA LLD, GeMAPS LLD, GeMAPS HSF, and pAA HSF. Thus, although GeMAPS performed better for LLDs, the HSF for pAA performed best on both the LSTM and CNN architectures. This result supports the previous finding that the mean and standard deviation outperform the LLDs defined in GeMAPS. Furthermore, we can generalize this finding to the means and standard deviations from the other feature sets, as well. In our case, the HSF for pAA performed better than the HSF for the affective-designed GeMAPS.

Table 3. CCC score results on the acoustic networks.

Feature set	*V*	*A*	*D*	Mean
		LSTM		
pAA LLD	0.0987	0.5175	0.3536	0.3233
pAA HSF	0.1729	**0.5804**	**0.4476**	**0.4003**
GeMAPS LLD	0.1629	0.5070	0.4433	0.3711
GeMAPS HSF	**0.1818**	0.5306	0.4332	0.3819
		CNN		
pAA LLD	0.0687	0.3665	0.3382	0.2578
pAA HSF	**0.1310**	**0.5553**	**0.4431**	**0.3764**
GeMAPs LLD	0.0581	0.4751	0.4203	0.3178
GeMAPs HSF	0.0975	0.4658	0.4170	0.3268

Table 4. CCC score results on the text networks.

Feature set	*V*	*A*	*D*	Mean
		LSTM		
WE	0.3784	0.3412	0.3638	0.3611
GloVe	**0.4096**	**0.3886**	**0.3790**	**0.3924**
FastText	0.4017	0.3718	0.3771	0.3835
		CNN		
WE	0.3740	0.3285	0.3144	0.3390
GloVe	**0.3843**	0.3646	**0.3911**	**0.3800**
FastText	0.3786	**0.3648**	0.3147	0.3527

Comparing the LSTM and CNN architectures, we found that the LSTM performed better than the CNN did. In terms of all emotional dimensions and the average, the score obtained by the highest-performing LSTM was higher than that obtained by the highest-performing CNN. We thus chose the cases with the best scores from the LSTM and CNN architectures in the acoustic networks to combine with the text networks.

As for the text networks, word embeddings with pre-trained GloVe embeddings performed better than either word embeddings without weighting or word embeddings weighted by the FastText model did. The text networks also showed that the LSTM with GloVe embedding is better than the CNN with the same input feature. In this dimensional emotion recognition, however, the highest performance of a text network was lower than the highest performance of an acoustic network. As with the acoustic networks, we chose two networks, GloVe with LSTM and GloVe with CNN, to combine in the bimodal network fusion.

C) Results on bimodal feature fusion

1) Performance of bimodal networks

According to their unimodal network performance, eight pairs of bimodal acoustic-text networks were evaluated. Table 5 lists their performance results in the same way as for the unimodal results. Among the eight pairs, the combination of the LSTM acoustic network and the LSTM text network achieved the best performance. This result in bimodal feature fusion is linear with respect to the obtained results for the unimodal networks, in which the LSTM performed best on both the acoustic and text networks.

In terms of both emotional dimensions and the average CCC score, the LSTM + LSTM pair outperformed the other bimodal pairs. Moreover, the deviation of the LSTM + LSTM pair was also the lowest. We can also state that, apart from attaining the highest performance, the LSTM + LSTM pair also gave the most stable results. This suggests that the LSTM not only attained comparable results to the CNN with a similar number of trainable parameters but also attained better performance, which differs from what was reported in [33].

To our knowledge, one reasonable explanation for why the LSTM performed better is the use of the full sequence instead of the final sequence in the last LSTM layer. In most applications, the last layer in an LSTM stack only returns the final sequence, so that it can be combined with the outputs of other layers (e.g. a dense layer). In our implementation, however, we returned all sequences outputted from the last LSTM layer and flattened the output before combining with the output of another dense layer. This strategy may keep more relevant information than what is returned by the final sequence of the last LSTM layer. On the other hand, we only observed this phenomenon on the acoustic network. In the case of the text network, the last LSTM layer returning the final sequence performed better than the LSTM returning all sequences did. In that case, we directly coupled the output of the last LSTM layer with that of a dense layer.

If we choose the highest unimodal score as a baseline, i.e. the HSF for pAA, then the relative improvement of the highest bimodal score was 23.97%. We also performed a significance test among the bimodal pair results and observed a significant difference between an LSTM + LSTM pair and other pairs such as a CNN + LSTM pair on a two-tail paired test. The small p-value ($\simeq 10^{-5}$) indicated the strong difference obtained by the LSTM + LSTM and CNN + LSTM pairs. While the CNN + LSTM pair obtained the third-highest score, the second-best performance was by a Dense + CNN pair with CCC = 0.485. The significance test result between the LSTM + LSTM pair and this pair was $p = 0.0006$. The more similar the performance of two acoustic-text network pairs was, the higher the p-value between them was. We thus assert that the LSTM + LSTM pair had a strong difference from the other pairs with $p < 0.005$.

2) Evaluation of MTL with weighting factors

As an extension of the main proposal to jointly learn the valence, arousal, and dominance from acoustic features and word embeddings by using MTL, we also evaluated some weighting factors for the MTL formulation (equations (4), (5), and (6)). In contrast, the above results were obtained using MTL with no parameters (equation 4). Thus, the following results show the effect of the weighting parameters on the MTL method.

MTL with two parameters is an approach to capture the interrelation among valence, arousal, and dominance. In equation 5, the gains of valence and arousal are provided independently, while the gain of dominance depends on the other gains. This simple weighting strategy may represent the relation among the emotional dimensions if the

Table 5. Results of bimodal feature fusion (without parameters) by concatenating the acoustic and text networks; each modality used either an LSTM, CNN, or dense network; batch size = 8.

Acoustic + Text	*V*	*A*	*D*	Mean
LSTM + LSTM	**0.418 ± 0.010**	**0.571 ± 0.017**	**0.500 ± 0.017**	**0.496 ± 0.010**
LSTM + CNN	0.256 ± 0.052	0.531 ± 0.031	0.450 ± 0.036	0.412 ± 0.030
CNN + LSTM	0.401 ± 0.020	0.545 ± 0.016	0.478 ± 0.015	0.476 ± 0.012
CNN + CNN	0.399 ± 0.015	0.541 ± 0.020	0.475 ± 0.014	0.472 ± 0.012
LSTM + Dense	0.274 ± 0.050	0.553 ± 0.019	0.484 ± 0.015	0.437 ± 0.018
CNN + Dense	0.266 ± 0.038	0.497 ± 0.059	0.457 ± 0.047	0.407 ± 0.040
Dense + LSTM	0.368 ± 0.105	0.564 ± 0.015	0.478 ± 0.025	0.470 ± 0.043
Dense + CNN	0.398 ± 0.015	0.570 ± 0.013	0.487 ± 0.015	0.485 ± 0.013

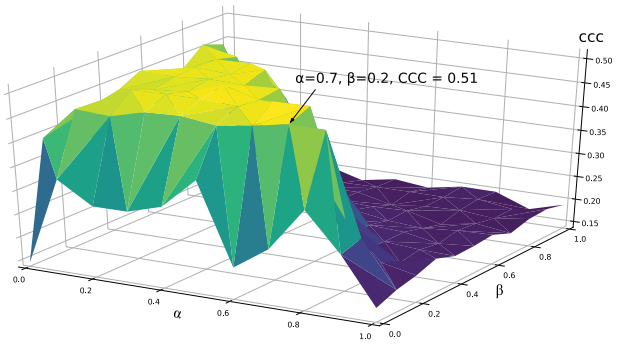

Fig. 4. Surface plot of different α and β factors for MTL with two parameters. The best mean CCC score of 0.51 was obtained using $\alpha = 0.7$ and $\beta = 0.2$. Both factors were searched simultaneously/dependently.

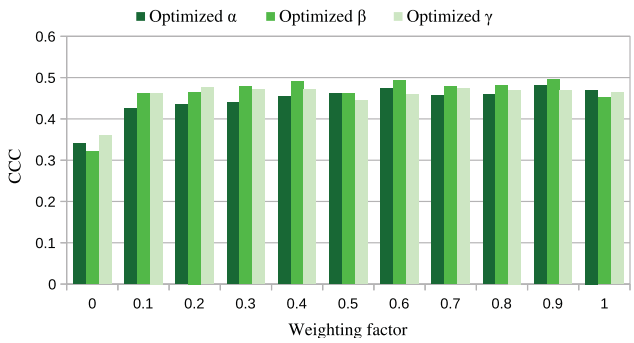

Fig. 5. CCC scores for MTL with three parameters obtained to find the optimal weighting factors. Linear search was performed independently on each parameter. The best weighting factors for the three parameters were $\alpha = 0.9$, $\beta = 0.9$, and $\gamma = 0.2$.

obtained results are better than the results without this weighting strategy.

Figure 4 shows a surface plot of the impact of varying α and β from 0.0 to 1.0 with the corresponding average CCC score. Clearly, performance improvement could be obtained by using proper weighting factors in two-parameter MTL. We found that $\alpha = 0.7$ and $\beta = 0.2$ were the best weighting factors, and they were also used by the text network. In the unimodal network, the best factors for MTL with two parameters were $\alpha = 0.7$ and $\beta = 0.2$ for the text network, and $\alpha = 0.1$ and $\beta = 0.5$ for the acoustic network. These factors were used to obtain the unimodal results above. We cannot be sure whether these same obtained factors for the bimodal network were contributed by the unimodal network or caused by other factors. Investigation on the cause of this finding is a challenging issue for both theoretical and empirical studies.

Next, MTL with three parameters provided all values for three variables, with the scale of each emotional dimension independent of the other emotional dimensions. MTL with no parameters is also a subset of MTL with three parameters, with $\alpha = 1.0$, $\beta = 1.0$, and $\gamma = 1.0$. We optimized the weighting factors with three parameters by using linear search independently on each emotion dimension. Figure 5 shows the impact of the weighting factors on MTL with three parameters. In this scaling strategy, the best weighting factors were $\alpha = 0.9$, $\beta = 0.9$, and $\gamma = 0.2$. The obtained result of $CCC = 0.497$ with these factors was lower, however, than that obtained by MTL with two parameters, i.e. $CCC = 0.508$. While the previous Table 5 presented results with batch size = 8, results in Table 6 are obtained with batch size = 256, to speed up computation process. The results listed in Table 6 show that MTL with two parameters obtained the best performance among the MTL methods. This result suggests that MTL with two parameters may be better at representing the interrelation among the emotional dimensions.

3) Evaluation of dropout for different modalities

To extend our results and discussion, we investigated the impact of the dropout rate for the acoustic and text networks in bimodal feature fusion. In this evaluation, we varied the dropout rate from each modality before concatenating them. The goal of the evaluation, at first, was to investigate the dropout rates for the different modalities.

Figure 6 shows the impact of different dropout rates and the obtained CCC scores. From the figure, using dropout rates of $p = 0.1$ and $p = 0.4$ for the acoustic and text networks, respectively, achieved the best score of $CCC = 0.510$. In fact, these dropout rates were used to obtain the above results on the bimodal network.

From the obtained dropout rates, we believe that this factor depends on the size of the feature/input rather than on modality differences. The acoustic network used the smaller HSF for pAA, a 68-dimensional vector, as compared to the word embedding's size of 100 sequences × 300-dimensional word vectors. Because the goal of using dropout is to avoid overfitting, it is reasonable that, on small data, the dropout rate is small, while on larger data, the

Table 6. Results of MTL with and without parameters for bimodal feature fusion (LSTM + LSTM); batch size = 256.

MTL method	*V*	*A*	*D*	Mean
No parameter	0.409 ± 0.015	0.585 ± 0.011	**0.486 ± 0.016**	0.493 ± 0.010
Two parameters	**0.446 ± 0.002**	**0.594 ± 0.003**	0.485 ± 0.003	**0.508 ± 0.002**
Three parameters	0.419 ± 0.012	0.589 ± 0.012	0.483 ± 0.011	0.497 ± 0.008

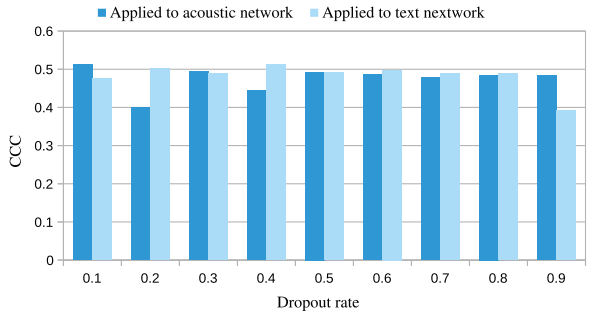

Fig. 6. Analysis of dropout rates applied to the acoustic and text networks before concatenating them. The dropout rates were applied independently on either network while keeping a fixed rate for the other network.

dropout rate increases. Hence, in this research, we believe that dropout rates depend on the size of the input rather than its modality.

D) Discussion in terms of categorical emotions

This paper on dimensional SER using bimodal features is an extension of a similar approach for the categorical method. We found both similarities and differences as compared to the previous categorical research. Here, we limit the discussion to the best bimodal pairs and the impact of feature sets from different modalities.

In dimensional SER, we found more consistent results. We observed low variation among the experiments, while the previous categorical research only used the highest accuracy from many experiments. Both the categorical and dimensional approaches gained performance improvement over unimodal emotion recognition by combining acoustic features and word embeddings. Using pairs of acoustic-text networks, we found that LSTM networks on both modalities performed best in dimensional emotion recognition. This result was also supported by a small standard deviation and significant differences with respect to other results. In the categorical research, a Dense + LSTM pair attained the highest result, followed by a Dense + CNN pair. We observed high performance in some of the 20 experiments with the Dense + LSTM pair. Its average performance ranked fifth, however, among the eight acoustic-text network pairs. The Dense + CNN pair, which was the second best in the categorical emotion research, also ranked second in this dimensional emotion approach. This result from dimensional emotion recognition was supported by the fact that the LSTM also attained the highest performance on unimodal emotion recognition. Similar unimodal results were also observed in the categorical approach, in which the LSTM architecture performed the best among all the architectures.

A second important finding is that we found different results between categorical and dimensional emotion recognition from the feature/modality perspective. Feature set/modality, which attained the highest performance in the categorical approach, is different from the dimensional approach. In the categorical approach with the IEMOCAP dataset, word embeddings gave the highest performance in the unimodal model, as reported in [1, 2, 4, 25]. In contrast, in the dimensional approach, the average performance of acoustic features gave better performance over text features. This phenomenon can be explained by the fact that text features (word embeddings) contribute to valence more than acoustic features do (see Tables 3 and 4). While the authors in [22] found this result, the authors in [7, 8, 12] extended it to find that, for arousal, acoustic features contribute more than text features do. Our results here further extend the evidence that text features contribute more in valence prediction, while acoustic features give more accuracy in arousal and dominance prediction. Given this evidence, it is more likely that acoustic features will obtain higher performance in the unimodal case as compared to text features, because they provide better performance for two of the three emotional dimensions. As suggested by Russell [38], however, a categorical emotion can be characterized by its valence and arousal only. This relation shows why text features achieve better performance than acoustic features do on categorical emotion.

As a final remark, we emphasize some of our findings on combining acoustic and text features for dimensional emotion recognition. Dimensional emotion recognition is scientifically more challenging than categorical emotion recognition. In this work, we achieved more consistent results than what we did in categorical emotion recognition. The combination of LSTM networks for both the acoustic and text networks achieved the highest performance on bimodal feature fusion, as the same architecture did on unimodal emotion recognition. Our proposal on using MTL for simultaneously predicting valence, arousal, and dominance worked as expected, and we found that MTL with two parameters represented the interrelation among the emotional dimensions better than other MTL methods did.

VII. CONCLUSIONS

This paper has reported an investigation of using acoustic features and word embeddings for dimensional SER with MTL. First, we conclude that using acoustic features and word embeddings can improve the prediction of valence,

arousal, and dominance. Word embeddings help improve valence prediction, while acoustic features contribute more for arousal and dominance prediction. All the emotional dimensions gained prediction improvements on bimodal acoustic and text networks, the highest improvement was obtained using LSTM + LSTM architectures pair. Second, our proposed MTL with two parameters could improve the prediction of all emotional dimensions as compared to MTL with no parameters. The weighting factors given to valence and dominance may represent the interrelation among the emotional dimensions. We think, however, that this formulation only partially represents that interrelation, because the obtained improvement was still small. The formulation can be improved for future research by implementing other strategies, particularly those based on psychological theories and experiments. Third, we found a mismatch between categorical and dimensional emotion recognition. For categorical emotion, text features obtained better results than acoustic features did, but for dimensional emotion, the result was the opposite. This can be explained by the argument that categorical emotion only relies on the valence-arousal space, in which the higher valence prediction obtained by word embeddings may result in better categorical emotion prediction than prediction by acoustic features.

In summary, a combination of speech features and word embeddings can solve the drawback of dimensional SER. The low score of the valence dimension in acoustic-based SER is improved by word embeddings. The combination of both features not just improved valence but arousal and dominance dimensions too. MTL also works as expected; it can simultaneously predict the degrees of three emotion dimensions instead of predicting one by one dimension using STL. This human perception-inspired strategy may mimic how our emotion perception from speech works. Based on the obtained performances, however, there is room for next improvements when comparing human emotion perception (i.e. labels) with automatic emotion recognition by computers/robots.

Future research directions can be taken to improve the performance of dimensional SER. While the state-of-the-art categorical method achieves more than 70% accuracy, the current dimensional approach still suffers from low performance in terms of the CCC score. For reporting dimensional emotion recognition performance, we encourage other researchers to use the CCC measure to enable benchmarking with respect to both previous and future research.

ACKNOWLEDGEMENT

None.

FINANCIAL SUPPORT

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

CONFLICT OF INTEREST

None.

REFERENCES

[1] Atmaja B.T.; Shirai K.; Akagi M.: Speech emotion recognition using speech feature and word embedding, in *2019 Asia-Pacific Signal Information Processing Association Annual Summit Conf. (APSIPA ASC)*, Lanzhou, 519–523.

[2] Tripathi S.; Beigi H.: Multi-modal emotion recognition on IEMOCAP dataset using deep learning, CoRR. http://arxiv.org/abs/1804.05788.

[3] Yenigalla P.; Kumar A.; Tripathi S.; Singh C.; Kar S.; Vepa J.: Speech emotion recognition using spectrogram & phoneme embedding, in *Interspeech 2018*, 3688–3692.

[4] Yoon S.; Byun S.; Dey S.; Jung K.: Speech emotion recognition using multi-hop attention mechanism, in *ICASSP 2019 – 2019 IEEE Int. Conf. Acoustic Speech Signal Processing*, Brighton, United Kingdom, 2822–2826.

[5] Gunes H.; Pantic M.: Automatic, dimensional and continuous emotion recognition. *Int. J. Synth. Emot.*, **1** (1) (2010), 68–99.

[6] Mäntylä M.; Adams B.; Destefanis G.; Graziotin D.; Ortu M.: Mining valence, arousal, and dominance, in *Proc. 13th Int. Working Mining Software Repositories – MSR '16*, ACM Press, New York, New York, USA, 247–258.

[7] Karadogan S.G.; Larsen J.: Combining semantic and acoustic features for valence and arousal recognition in speech, in *2012 3rd Int. Working Cognitive Information Processing*, IEEE, Baiona, 2012, 1–6.

[8] Eyben F.; Wöllmer M.; Graves A.; Schuller B.; Douglas-Cowie E.; Cowie R.: On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *J. Multimodal User Interfaces*, **3** (1) (2010), 7–19.

[9] Schmitt M.; Schuller B.: Deep recurrent neural networks for emotion recognition in speech, in *DAGA*, 1537–1540

[10] Grimm M.; Kroschel K.; Mower E.; Narayanan S.: Primitives-based evaluation and estimation of emotions in speech. *Speech Commun.*, **49** (1011) (2007), 787–800.

[11] Parthasarathy S.; Busso C.: Jointly predicting arousal, valence and dominance with multi-task learning, in *Interspeech*, 1103–1107.

[12] Chen S.; Jin Q.; Zhao J.; Wang S.: Multimodal multi-task learning for dimensional and continuous emotion recognition, in *Proc. 7th Annu. Work. Audio/Visual Emot. Chall.*, ACM, Mountain View, California, USA, 2017, 19–26.

[13] Ringeval F. *et al.*: AVEC 2017 – Real-life depression, and affect recognition workshop and challenge, in *AVEC 2017 – Proc. 7th Annual Working Audio/Visual Emot. Challenge, co-located with MM 2017*, 3–9.

[14] Moore J.D.; Tian L.; Lai C.: Word-level emotion recognition using high-level features, in *Int. Conf. Intelligent Text Processing and Computational Linguistics*, Springer, Berlin, Heidelberg, 2014, 17–31.

[15] Schuller B.; Valstar M.; Eyben F.; Cowie R.; Pantic M.: AVEC 2012 – The continuous audio/visual emotion challenge, in *ICMI'12 – Proc. ACM Int. Conf. Multimodal Interaction*, 449–456.

[16] Tits N.; Haddad K.E.; Dutoit T.: ASR-based features for emotion recognition: a transfer learning approach, in *Proc. First Grand Challenge Workshop Human Multimodal Language*, 48–52.

[17] El Ayadi M.; Kamel M.S.; Karray F.: Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognit.*, **44** (3) (2011), 572–587.

[18] Tian L.; Moore J.; Lai C.: Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features, in *2016 IEEE Spoken Language Technology Workshop*, IEEE, 565–572.18

[19] Atmaja B.T.; Akagi M.: Evaluation of error and correlation-based loss functions for multitask learning dimensional speech emotion recognition. http://arxiv.org/abs/2003.10724.

[20] Sridhar K.; Parthasarathy S.; Busso C.: Role of regularization in the prediction of valence from speech, in *Interspeech 2018*, ISCA, ISCA, 941–945.20

[21] Lotfian R.; Busso C.: Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput.*, **10** (4) (2019), 471–483.

[22] Aldeneh Z.; Khorram S.; Dimitriadis D.; Provost E.M.: Pooling acoustic and lexical features for the prediction of valence, in *ICMI 2017 – Proc. 19th ACM Int. Conf. Multimodal Interaction*, ACM, Glasgow, UK, 2017, 68–72.

[23] Yoon S.; Byun S.; Jung K.: Multimodal speech emotion recognition using audio and text. *Spok. Lang. Technol. Work.* (2018), Greece, Athens, 2018, 112–118. http://arxiv.org/abs/1810.04635.

[24] Zhang B.; Khorram S.; Provost E.M.: Exploiting acoustic and lexical properties of phonemes to recognize valence from speech, in *ICASSP, IEEE Int. Conf. Acoustic Speech Signal Processing – Proc.*, Brighton, United Kingdom, 2019, 5871–5875.

[25] Sahu S.; Mitra V.; Seneviratne N.; Espy-Wilson C.: Multi-modal learning for speech emotion recognition: an analysis and comparison of ASR outputs with ground truth transcription, in *Proc. Annual Conf. Int. Speech Communication Association INTERSPEECH*, Graz, Austria, 2019, 3302–3306.

[26] Huang K.Y.; Wu C.H.; Hong Q.B.; Su M.H.; Zeng Y.R.: Speech emotion recognition using convolutional neural network with audio word-based embedding, in *2018 11th Int. Symp. Chinese Spoken Language Processing*, IEEE, Taipei City, Taiwan, 265–269. https://ieeexplore.ieee.org/document/8706610/.

[27] Giannakopoulos T.: PyAudioAnalysis: an open-source python library for audio signal analysis. *PLoS ONE*, **Vol. 10 no 12** (2015), 1-17, https://github.com/tyiannak/pyAudioAnalysis/.

[28] Eyben F. *et al.*: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Trans. Affect. Comput.*, **7** (2) (2016), 190–202.

[29] Goodfellow I.; Bengio Y.; Courville A.: *Deep Learning*, MIT Press, Cambridge, Massachusetts, USA, 2016. https://www.deeplearningbook.org/

[30] Mikolov T.; Chen K.; Corrado G.; Dean J.: Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013), https://arxiv.org/abs/1301.3781.

[31] Pennington J.; Socher R.; Manning C.D.: GloVe: global vectors for word representation, in *Conf. Empirical Methods Natural Language Processing*, 1532–1543.

[32] Mikolov T.; Grave E.; Bojanowski P.; Puhrsch C.; Joulin A.: Advances in pre-training distributed word representations, in *Lr. 2018 – 11th Int. Conf. Language Resources Evaluation*, 52–55.

[33] Schmitt M.; Cummins N.; Schuller B.W.: Continuous emotion recognition in speech – Do we need recurrence?, in *Interspeech 2019*, ISCA, Graz, Austria, 2808–2812.

[34] Chetlur S. *et al.*: cuDNN: Efficient Primitives for Deep Learning, Technical report (2014)

[35] Chollet, F., Others; Keras, https://keras.io (2015).

[36] Lin L.I.K.: A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45** (1) (1989), 255–68.

[37] Pedregosa F. *et al.*: Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12** (2011), 2825–2830.

[38] Russell J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **Vol 39 no 6** (1980), 1161–1178.

Bagus Tris Atmaja received bachelor and master degrees in engineering physics from the Sepuluh Nopember Institute of Technology in 2009 and 2012, respectively, where he is now employed as a researcher in acoustics. Currently, he is also a Ph.D. student at the Japan Advanced Institute of Science and Technology, Nomi, Japan, focusing on speech emotion recognition. His main research interests are speech processing including speech enhancement, source separation, and speech (emotion) recognition.

Masato Akagi received his B.E. degree from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. Eng. degrees from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of JAIST and is now a full professor. His research interests include speech perception, modeling of speech perception mechanisms in humans, and the signal processing of speech.