How Speech is Recognized to Be Emotional - A Study Based on Information Decomposition

Haoran Sun, Lantian Li*, Thomas Fang Zheng, Dong Wang*
Center for Speech and Language Technologies, BNRist, Tsinghua University
E-mail: lilt@cslt.org; wangdong99@mails.tsinghua.edu.cn

***Abstract*—The way that humans encode their emotion into speech signals is complex. For instance, an angry man may increase his pitch and speaking rate, and use impolite words. In this paper, we present a preliminary study on various emotional factors and investigate how each of them impacts modern emotion recognition systems. The key tool of our study is the SpeechFlow model presented recently, by which we are able to decompose speech signals into separate information factors (content, pitch, rhythm). Based on this decomposition, we carefully studied the performance of each information component and their combinations. We conducted the study on three different speech emotion corpora and chose an attention-based convolutional RNN as the emotion classifier. Our results show that rhythm is the most important component for emotional expression. Moreover, the cross-corpus results are very bad (even worse than guess), demonstrating that the present speech emotion recognition model is rather weak. Interestingly, by removing one or several unimportant components, the cross-corpus results can be improved. This demonstrates the potential of the decomposition approach towards a generalizable emotion recognition.**

I. Introduction

Recognizing emotion from speech signals is highly desirable for designing a comfortable human-machine interface. After three decades of research, speech emotion recognition (SER) has gained significant improvement [1]. Early research mostly focused on extracting emotion-related features, forming some 'standard' feature sets such as GeMAPS [2] and COMPARE [3]. By these emotion features, simple classifiers such as hidden Markov model (HMM) or support vector machines (SVM) were employed to determine the emotion [4], [5] . Recently, deep learning methods gained much popularity, in particular the end-to-end architecture based on deep neural nets (DNN) [6]–[13]. Good performance was reported with various types of DNN models, including convolutional neural network (CNN) [14], deep belief network [15], long-short term memory (LSTM) [8] and variational auto-encoders (VAE) [16]. The main advantage of deep learning models is that they can learn emotional cues automatically, so potentially discover features more powerful than human engineering.

In spite of the notable advance in performance, how a speech is recognized by the machine to be emotional is still far from clear. One reason is that human emotions in audio data are very complex, and the expression and perception of a particular emotion is impacted by various factors such as gender, speakers, age, culture, and languages [17]. From the signal processing perspective, Murray et al [18] identified that the quality of voice, the timing of pronunciation units, and the pitch contour are mostly affected in emotional speech. This study is very inspiring and guided the long-standing research on emotion sensitive features. However, it is still not easy to identify by which information factors in the speech signal that *machines* recognize human's emotion, even if we can test and compare the SER performance with individual features and their combinations. This is because there is no guarantee that these features faithfully reflect the underlying information factors (e.g., prosody patterns), and there is no guarantee that the complex temporal/frequency dependencies within the original speech signal can be recovered by reintegrating the separately extracted features. This is in particular true with the DNN-based end-to-end model, as the decision is made largely in a black box.

In this paper, we try to answer the following question: "How an end-to-end DNN model determines emotions". Our main tool is a speech factorization model called *SpeechFlow* [19]. By this model, speech signals can be decomposed into separate information factors, and these factors can be put together to recover the original speech. This analysis-and-synthesis tool offers us an interesting opportunity to manipulate the information load in speech signals, allowing us testing the impact of each individual information factor and their combinations. In this preliminary study, we decompose speech signal into three components: content, rhythm, and pitch. This decomposition was motivated by the importance of timing (rhythm) and pitch in human's emotion perception, as found by Murray [18], as well as the intrinsic association of emotion status and linguistic content [20]–[22]. Fig. 1 illustrates the full diagram of our approach.

We chose the attention-based convolutional RNN (ACRNN) [10] as the SER backbone, due to its good performance reported in the literature. The ACRNN model was trained with IEMOCAP, a popular emotion speech dataset in English, and was tested on IEMOCAP, SAVEE and CSLT-ESDB, where SAVEE and CSLT-ESDB are two datasets in English and Chinese respectively. The results show that among the three information factors (content, rhythm, pitch), rhythm is the most discriminative and generalizable. When the test data is highly mismatched with the training data, for example in the cross-lingual test, removing some information factors may increase rather than decrease the performance. This suggests that to achieve reasonable generalization, information control deserves deeper investigation.

Our contribution is two-fold: (1) We employed speech fac-

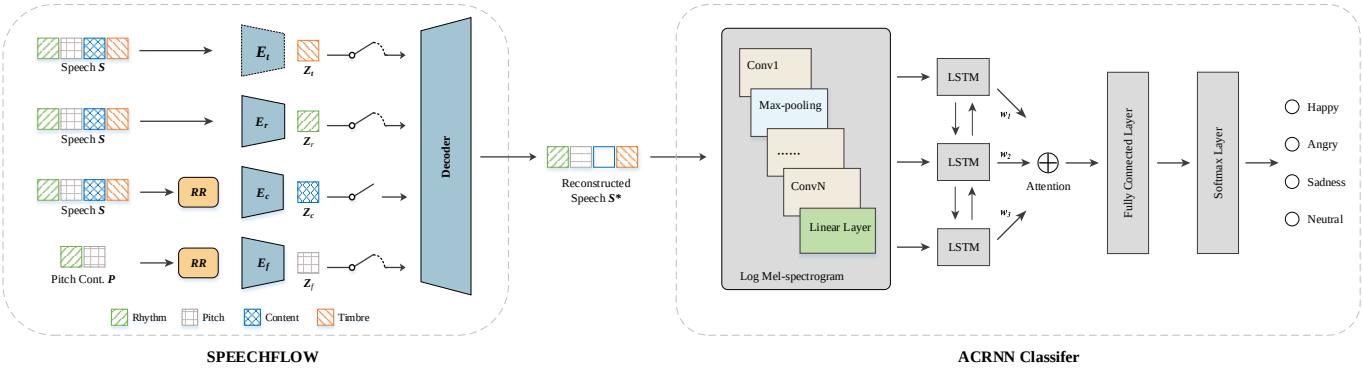


Fig. 1. The full diagram of the work. The SpeechFlow model decomposes speech signals into three information factors: content Z_c, rhythm Z_r, pitch Z_f. These information factors can be manipulated in order to modify information load of the reconstructed speech signal, for instance set the same pitch for all the frames of all the utterances. Note that an additional factor related to speaker trait (timbre) is also represented in the latent space and is denoted by Z_t, but it is only used for speech reconstruction and is not modified in our study. Once the speech signal reconstructed from the modified factors, ACRNN-based SER model is used to predict the emotion of the reconstructed speech.

torization as a novel tool to investigate the decision behavior of DNN, which is new in SER; (2) Using the factorization tool, we studied the salient information factors in SER, and investigated the generalization capacity of different factors in cross-domain and cross-lingual situations. The paper is organized as follows: in Section II, we first briefly introduce the SpeechFlow and ACRNN models, and present our implementation details. Related work is discussed in Section III, and the experiments are presented in Section IV. Finally the paper is concluded by Section V.

II. Methods

In this section, we firstly revisit the SpeechFlow model presented in [19]. This model can decompose speech signals into separate information factors, and then reconstruct the original speech from these factors. Importantly, we can *remove* any individual factor by setting its value to be a constant, which allows us freely manipulating the information load of the speech signal. The second component of our architecture is an attention-based convolutional RNN (ACRNN) [10]. This model can use as the backbone of our SER system, and test the SER performance of speech signals with different information load. The entire architecture of this work is illustrated in Fig.1.

A. SpeechFlow for speech information decomposition

SpeechFlow is a speech factorization model proposed recently [19]. This model can decompose speech signals into separate information factors in an unsupervised way. As shown in Fig.1 (left), speech S is decomposed by SpeechFlow into four information factors: rhythm $\mathbf{Z}_r$, pitch $\mathbf{Z}_f$, content $\mathbf{Z}_c$ and timbre $\mathbf{Z}_t$, by four neural encoders: a *rhythm encoder* $\boldsymbol{E}_r$, a *pitch encoder* $\boldsymbol{E}_f$, a *content encoder* $\boldsymbol{E}_c$ and a *timbre encoder* $\boldsymbol{E}_t$.

The training objective of the model is to reconstruct the original speech S from these information factors, with a *decoder* D. Formally, the reconstructed speech $\hat{S}$ is obtained by:

$$\hat{S} = D(\mathbf{Z}_r, \mathbf{Z}_f, \mathbf{Z}_c, \mathbf{Z}_u), \tag{1}$$

and the objective function $\mathcal{L}$ is formulated as:

$$\mathcal{L} = || \hat{S} - S ||^2. \tag{2}$$

where $||\cdot||$ denotes the ℓ_2-norm. To ensure that the information factors posses their own desired information after model training, the encoders need some special designs, as shown below.

Firstly, the input to the rhythm encoder $\mathbf{E}_r$, content encoder $\mathbf{E}_c$ and timbre encoder $\mathbf{E}_t$ is speech S, whereas the input to pitch encoder $\mathbf{E}_f$ is the normalized pitch contour P. Note that P has been normalized to posses the same mean and variation for all the speakers, so it involves only rhythm and pitch information, without information about speaker trait and linguistic content.

Secondly, a random resampling operation (***RR***) along the temporal axis is conducted before speech $\mathbf{S}$ is fed to the content encoder $\mathbf{E}_c$ and pitch encoder $\mathbf{E}_f$. This operation randomly shrinks or stretches the duration of each speech segment. This operation makes the two encoders lose the true rhythm information, so the complete rhythm information can only pass through the rhythm encoder $\mathbf{E}_r$ and represented by $\mathbf{Z}_r$.

Thirdly, the speaker identity vector is used as the timbre information. We choose a pre-trained speaker recognition model as the timbre encoder $\mathbf{E}_t$. This model is based on deep speaker embedding [23], [24], and the produced speaker vectors are used as $\mathbf{Z}_t$. Note that $\mathbf{Z}_t$ involves only the timber information.

Finally, the dimensions of all the information factors are much lower than the input speech. This limited dimensionality forms the so-called *information bottleneck*, which means that none of the individual factors can represent the whole signal, and all the factors must cooperate together to achieve the learning goal, i.e., reconstruct the original speech. To make this cooperation effective and economic, each factor has to focus

on the information that it can easily supply and others cannot, thus leading to the desired information decomposition [19].

Put them together, the entire encoding process can be formulated as follows:

$$\begin{aligned}\boldsymbol{Z}_r &= \boldsymbol{E}_r(S),\\ \boldsymbol{Z}_f &= \boldsymbol{E}_f(\boldsymbol{RR}(P)),\\ \boldsymbol{Z}_c &= \boldsymbol{E}_c(\boldsymbol{RR}(S)),\\ \boldsymbol{Z}_t &= \boldsymbol{E}_t(S).\end{aligned}$$

where ***RR*** denotes random resampling.

SpeechFlow offers a powerful analysis-and-synthesis tool, by which we can freely manipulate the information factors, hence modifying the information load in the reconstructed speech. For example, we can set any information factor to be a constant, so that remove the corresponding information from the reconstructed speech. In this study, we focus on the impact of three information factors: content ($\mathbf{Z}_c$), rhythm ($\mathbf{Z}_r$), and pitch ($\mathbf{Z}_f$). The impact of the timbre factor $\mathbf{Z}_t$ will be left for future investigation.

B. ACRNN-based emotion recognition

If the SpeechFlow model has been well-trained, we can use it to manipulate the information load in speech signals. This allows us to conduct an ablation study to evaluate which information factor is the most important for emotion recognition. We first generate speech signals with particular information involved, and then fed the speech into an attention-based convolutional RNN (ACRNN) [10] classifier for emotion recognition, as shown in Fig.1 (right).

The ACRNN structure consists of four components, as detailed below. More details about ACRNN can be found in [10].

- **CNN component:** This component involves a 3-D convolutional layer followed by a max-pooling layer, upon which 5 convolutional layers and 1 full-connection layer are stacked. This component aims to learn local patterns of emotion traits.
- **RNN component:** It is a single-layer RNN with LSTM units. This component is designed to model the long-term patterns of emotion traits.
- **Attention component:** This component aggregates the frame-level representations along the temporal axis, to form an utterance-level emotion representation. In particular, the attention mechanism weights each frame according to the information that contains related to emotion prediction.
- **Prediction component:** It involves a full-connection layer and a softmax layer, and the output units of the softmax correspond to the emotion classes.

Note that for each configuration of the information factor selection in SpeechFlow, we need retrain the ACRNN model. This is the case even we select *all* the information factors (i.e., do not intentionally remove any information).

III. Related work

Speech factorization has been extensively used in speech coding, speech synthesis, and voice conversion [25]–[28], but the application in SER is rare. Li et al. [29] proposed a cascade factorization approach, where content and speaker traits are sequentially extracted in prior, and these information factors are used as conditional inputs of an end-to-end SER system. Peri et al. [30] employed an adversarial training to purify emotion embeddings, by using both audio and visual streams, though it is more a feature extraction approach rather than a complete factorization approach.

The research on the decision process of DNN in SER is also not extensive. Jalal et al. [31] investigated how an attention-based DNN model focuses on the important segment of the speech signal when performing SER. This work is different from ours as we focus on important information factors rather than important segments.

IV. Experiments

A. Data

The corpora used in our experiments are summarized in Table I. Considering that the sets of emotion labels are not completely the same for different corpora, we chose to use the four overlapped emotion classes: ***A**ngry*, ***H**appy*, ***S**ad* and ***N**eutral*. All the speech signals were uniformly formatted to 16kHz, 16-bits to ensure data consistency.

Since English and Chinese are different in pronunciation, we trained separate SpeechFlow models for the two languages, using VCTK and AISHELL-3 respectively. Two ACRNN SER models were trained, one with IEMOCAP and the other with CSLT-ESDB[1]. For each training, the entire corpus was split into a training set (80%), a validation set (10%) and a test set (10%). The speakers and utterances in the validation and test sets did not appear in the training set. The SAVEE dataset is too small to be used for model training, we therefore used it as a test set only.

B. Configurations

1) SpeechFlow: The SpeechFlow model was implemented using the source code published online[2]. We mostly followed the settings in the original repository, including network structures, data preprocessing steps, and the training scheme. The only change we made is the timbre encoder, which is one-hot speaker codes. This one-hot code cannot be extended to represent speakers outside of the training dataset, therefore not suited for our task.

To solve this problem, we introduced a speaker encoder to realize the function of the timbre encoder. The speaker encoder is able to generate continuous vectors to represent the trait of speakers. These continuous vectors, often called *speaker vectors*, are generalizable to speakers in any database. In our experiments, we chose the d-vector model to implement the speaker (timbre) encoder [23], [24]. The model was constructed following the Kaldi SITW recipe [37].

[1] http://data.cslt.org

[2] https://github.com/FantSun/Speechflow.

TABLE I
CORPORA DESCRIPTION

Corpus	Language	Content	Emotion Types	# Utters	# Spks	# Hours	Usage
VCTK [32]	English	Newspaper	-	5,376	20	5	SpeechFlow training
AISHELL-3 [33]	Chinese	Natural Text	-	6,162	20	6	SpeechFlow training
IEMOCAP [34]	English	Dialogue	A, H, S, N	2,280	10	2.5	ACRNN training & Emotion evaluation
CSLT-ESDB [35]	Chinese	Emotional Text	A, H, S, N	7,200	30	10	ACRNN training & Emotion evaluation
SAVEE [36]	English	Natural Text	A, H, S, N	300	4	0.5	Emotion evaluation

2) ACRNN: We built the ACRNN SER model using the public source code online[3]. The input feature was 80-dimensional mel-spectrograms, to match the output of the SpeechFlow model. Other configurations were the same as in [10].

C. Qualitative Analysis of SpeechFlow

We firstly verify how SpeechFlow decomposes and recovers speech signals. Specifically, we first decompose the spectrogram of a speech signal into individual information factors, and then remove a particular factor by setting the input of the corresponding encoder to be zero. Finally, the spectrogram can be recovered by the decoder of the SpeechFlow model. The reconstructed spectrograms, when different information factors are removed, are shown in Fig.2.

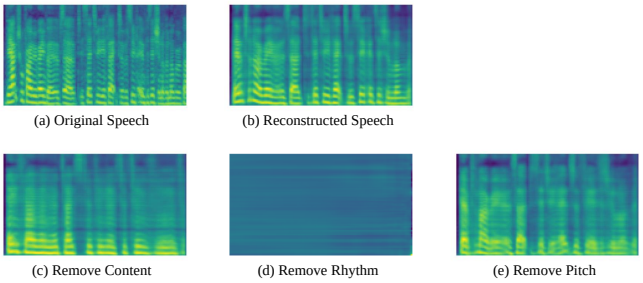


Fig. 2. The spectrogram reconstructed by SpeechFlow. The model was trained with VCTK, and the speech signal was selected from IEMOCAP. (a) spectrogram of the original speech; (b) spectrogram of the reconstructed speech with all the information factors preserved; (c)∼(e) spectrograms of the reconstructed speech with a single information factor removed.

Firstly, by the comparison of Fig.2 (a) and Fig.2 (b), one can observe that the reconstructed spectrogram with all information factors remained is similar to the original spectrogram, and detailed local patterns are retained. It indicates that the SpeechFlow model can successfully reconstruct the speech signal to a large extent. This is the foundation of all the following experiments. Secondly, removing any information factor leads to significant change in the reconstructed spectrogram, and the change is mutually different when different factors are removed.

D. Basic results

In this section, we report the SER performance when different information factors are removed. As the first step, we chose IEMOCAP as the training data to build the ACRNN model, and tested the performance on IEMOCAP and SAVEE. The SpeechFlow model was trained using VCTK. We use the unweighted average recall (UAR) as the metric, and the results are shown in Table II. Note that '✓' denotes that this factor was preserved while "✗" denotes that this factor was removed. For example, in system 4, the ACRNN model was trained with speech reconstructed by preserving the content factor $\mathbf{Z}_c$ only, and the validation and test data were processed in the same way.

TABLE II
UAR(%) RESULTS OF ACRNN MODEL TRAINED ON *IEMOCAP* WITH DIFFERENT INFORMATION FACTOR COMBINATIONS.

No.	Factors			Test Sets	
	Content	Rhythm	Pitch	IEMOCAP	SAVEE
1	-	-	-	59.08	45.00
2	✓	✓	✓	58.46	42.71
3	✗	✗	✗	25.00	25.00
4	✓	✗	✗	43.27	30.21
5	✗	✓	✗	57.85	39.79
6	✗	✗	✓	41.74	24.79
7	✓	✓	✗	57.14	42.50
8	✓	✗	✓	40.20	31.25
9	✗	✓	✓	57.89	37.50

Firstly, we observe that system 2, where all the information factors are preserved, can obtain performance comparable to system 1 on both the two test sets. Besides, if all the information factors are removed (system 3), the decision is completely random.[4] These results double confirmed that SpeechFlow can decompose speech signals into information factors and these factors can be put together to recover the original speech.

Secondly, we find that rhythm obtains better performance than content and pitch (compare system 5 vs. 4 and 6). This observation is consistent in both the within-corpus test (IEMOCAP) and the cross-corpus test (SAVEE). It indicates that rhythm is a salient feature used by the ACRNN model: it is not only discriminative, but also generalizable.

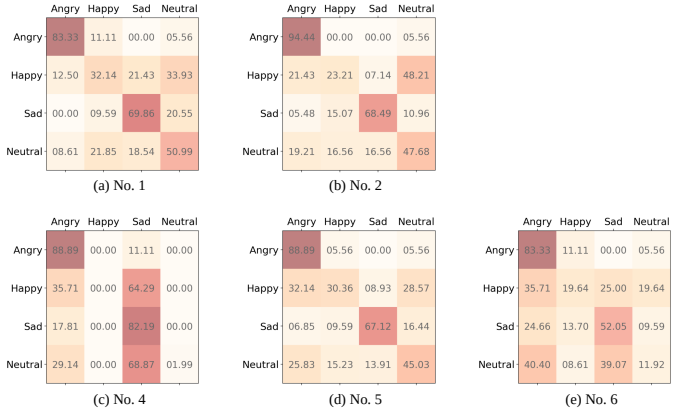
Thirdly, we observe that combining rhythm with other factors did not lead to clear advantage, and sometimes may lead to performance loss. This suggests that DNNs may rely

[3]https://github.com/xuanjihe/speech-emotion-recognition.

[4]Note that there are 4 emotion classes in total so the performance by random guess is 25%.

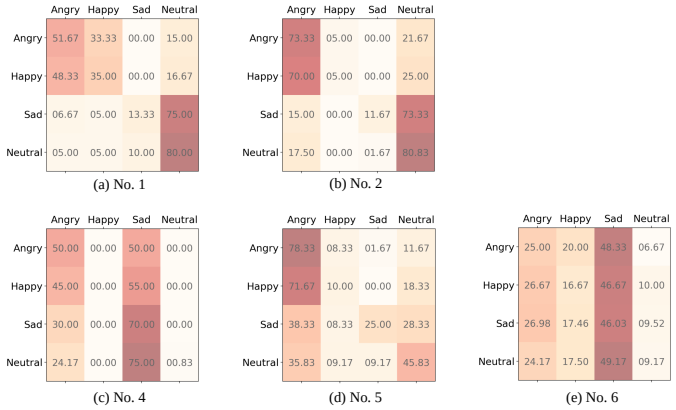
on one or a few information to perform the decision, rather than the full information set. This behavior, however, may be related to the limited training data, and deeper investigation with a larger dataset will give a more convincing conclusion.

To assist the analysis for the SER performances, the confusion matrices for the IEMOCAP test and the SAVEE test are shown in Fig.3 and Fig.4, respectively. It can be found that the performance tendency is quite similar on the two test sets. For example, for system 4 (content only), all the utterances tend to be recognized as *sad*. Moreover, the inter-emotion confusion is more balanced with system 5 (rhythm only) compared to system 4 (content only) and system 6 (pitch only), double confirming the superiority of rhythm information in DNN-based SER.



(a) No. 1 (b) No. 2 (c) No. 4 (d) No. 5 (e) No. 6

Fig. 3. Confusion matrices of systems in Table II when tested on IEMOCAP. The numbers in the cells represent the percentage that a ground-truth emotion (row label) is recognized as a particular emotion (column label).



(a) No. 1 (b) No. 2 (c) No. 4 (d) No. 5 (e) No. 6

Fig. 4. Confusion matrices of systems in Table II when tested on SAVEE. The meaning of the labels and numbers are the same as in Fig. 3.

E. Cross-corpus results

In this section, we design a cross-lingual test using the IEMOCAP and CSLT-ESDB datasets, which are in different languages. VCTK and AISHELL-3 were firstly employed to train an English SpeechFlow and a Chinese SpeechFlow respectively. Then, IEMOCAP was used to train the English ACRNN model, where the VCTK-based SpeechFlow was used to perform information selection, and CSLT-ESDB was used to train the Chinese ACRNN models, where the AISHELL-3-based SpeechFlow was used for information selection. The UAR results are reported in Table III, where *IEMO* and *ESDB* are the abbreviations of IEMOCAP and CSLT-ESDB, respectively.

TABLE III
UAR(%) RESULTS WITH DIFFERENT INFORMATION FACTOR COMBINATIONS ON CROSS-LINGUAL EMOTION RECOGNITION.

No.	Factors			VCTK-IEMO		AISHELL-ESDB	
	Content	Rhythm	Pitch	IEMO	ESDB	IEMO	ESDB
1	-	-	-	59.08	28.08	30.39	80.58
2	✓	✓	✓	58.46	27.37	30.35	75.46
3	✗	✗	✗	25.00	24.83	25.00	25.69
4	✓	✗	✗	43.27	24.96	25.00	43.62
5	✗	✓	✗	57.85	29.88	21.92	72.67
6	✗	✗	✓	41.74	21.62	25.00	36.54
7	✓	✓	✗	57.14	30.33	19.27	73.08
8	✓	✗	✓	40.20	25.00	24.77	44.96
9	✗	✓	✓	57.89	29.33	21.66	71.58

Firstly, it can be observed that although the results of the within-lingual tests are highly promising, the performance of the cross-lingual tests is very bad, sometimes even worse than guess (25%). Besides, by removing one or several unimportant components, the cross-lingual performance can be increased rather than decreased (ref. to System 2 and System 5, 7, 9, with IEMO training and ESDB test). This demonstrates the potential of the decomposition approach towards a generalizable emotion recognition, and also suggests that information selection/control might be important for cross-lingual SER.

Nevertheless, since all the results are so bad in the cross-lingual tests, no conclusions can be said convincing. Perhaps the only thing we can make sure is that the present DNN-based SER system is not well generalized and more research is required.

V. CONCLUSIONS

In this paper, we presented a preliminary study on how DNN-based SER models make decisions on emotions. To answer this question, we employed the SpeechFlow model to decompose speech signals into separate information factors, and then comprehensively studied the impact of each information factor and their combinations to the SER performance. Our results on three emotion corpora showed that rhythm has more discrimination and generalization capability on SER, at least in within-corpus tests. However, with more challenging mismatch such as in the cross-lingual test, all these factors and their combinations failed to deliver reasonable performance. This demonstrated that the current speech emotion model is still unreliable and can not be applied 'in the wild'. As for the future work, we will study more powerful factorization models, in order to pursue more independent factors. Moreover, we need larger emotion datasets to verify the observations in this study. Finally, the impact of cultural discrepancy deserves

deeper investigation, by which we may explain the weak performance in our cross-lingual test.

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[3] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH*, 2013.

[4] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011.

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[6] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Interspeech*, 2017, pp. 1089–1093.

[7] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Deep temporal models using identity skip-connections for speech emotion recognition," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1006–1013.

[8] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[9] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[10] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[11] X. Cheng, X. Zhang, M. Xu, and T. F. Zheng, "Mmann: Multimodal multilevel attention neural network for horror clip detection," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 329–334.

[12] S. Kwon *et al.*, "MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach," *Expert Systems with Applications*, vol. 167, p. 114177, 2021.

[13] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," *arXiv preprint arXiv:1904.03833*, 2019.

[14] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[15] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, "Random deep belief networks for recognizing emotions from speech signals," *Computational intelligence and neuroscience*, vol. 2017, 2017.

[16] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," *arXiv preprint arXiv:1712.08708*, 2017.

[17] S. Kwon *et al.*, "Att-net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, p. 107101, 2021.

[18] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[19] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[20] D. Grandjean, T. Bänziger, and K. R. Scherer, "Intonation as an interface between language and affect," *Progress in brain research*, vol. 156, pp. 235–247, 2006.

[21] M. D. Pell, A. Jaywant, L. Monetta, and S. A. Kotz, "Emotional speech processing: Disentangling the effects of prosody and semantic cues," *Cognition & Emotion*, vol. 25, no. 5, pp. 834–853, 2011.

[22] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[23] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[24] L. Li, Z. Tang, Y. Shi, and D. Wang, "Gaussian-constrained training for speaker verification," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6036–6040.

[25] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.

[26] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *arXiv preprint arXiv:1804.02812*, 2018.

[27] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.

[28] K. Zhou, B. Sisman, and H. Li, "Vaw-gan for disentanglement and recomposition of emotional elements in speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 415–422.

[29] L. Li, D. Wang, Y. Chen, Y. Shi, Z. Tang, and T. F. Zheng, "Deep factorization for speech signal," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5094–5098.

[30] R. Peri, S. Parthasarathy, C. Bradshaw, and S. Sundaram, "Disentanglement for audio-visual emotion recognition using multitask setup," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6344–6348.

[31] M. A. Jalal, R. Milner, and T. Hain, "Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition." in *INTERSPEECH*, 2020, pp. 4113–4117.

[32] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[33] X. X. S. Z. M. L. Yao Shi, Hui Bu, "AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines," 2015. [Online]. Available: https://arxiv.org/abs/2010.11567

[34] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[35] F. Bie, D. Wang, T. F. Zheng, J. Tejedor, and R. Chen, "Emotional adaptive training for speaker verification," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–4.

[36] W. Wang, *Machine Audition: Principles, Algorithms and Systems: Principles, Algorithms and Systems*. IGI Global, 2010.

[37] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.