Using Prosodic Phrase-Based VQVAE on Audio ALBERT for Speech Emotion Recognition

Jia-Hao Hsu[1], Chung-Hsien Wu[2] and Tsung-Hsien Yang[3]
[1,2]Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, TAIWAN
E-mail: [1]jiahaoxuu@gmail.com, [2]chunghsienwu@gmail.com Tel: +886-6-2757575-62500-2801
[3]Telecommunication Laboratories Chunghwa Telecom Co., Ltd., Taoyuan, Taiwan
E-mail: [3]yasamyang@cht.com.tw Tel: +886-3-4244169

***Abstract*— Speech emotion recognition has been an important field in the research of human-computer interaction. Understanding the user's emotions from speech help the system to grasp the user's underlying information, such as user satisfaction with the service. This research attempts to detect the emotion of the user's speech recorded by the customer service dialogue systems for telecommunication applications. This study proposes the prosodic phrase-based Vector Quantized Variational AutoEncoder (VQVAE) as the feature extraction module in the pre-trained model, Audio ALBERT (AALBERT). Two steps are added before fine-tuning the pre-trained AALBERT model, including prosodic phrase segmentation and prosodic phrase-based VQVAE model. The speech segments are extracted using the prosodic phrase segmentation algorithm, in which each segment is supposed to contain only a single emotion. The VQVAE model is trained to obtain quantized important prosodic phrase vectors. In the experiment, the speech corpus collected by the telecom customer service system was used for evaluation, and the ablation study shows that the method proposed can effectively improve the performance of the pre-trained model, and the accuracy reached 91.41%. It can be seen that feature extraction using prosodic segmentation and prosodic phrase quantization has a certain potential in the field of speech emotion recognition.**

I. INTRODUCTION

Speech emotion recognition (SER) is an important function in human-computer interaction interfaces. Understanding human emotions can bring convenience to many applications and improve the quality of human-machine interaction [1]. Among them, user satisfaction is valuable but difficult to obtain in various services. The previous methods of obtaining customer satisfaction, such as market satisfaction surveys. It usually takes a lot of time to obtain results, and it is quite difficult to deal with customer dissatisfaction immediately. Therefore, this study proposes an approach to recognizing the customer's emotional state of speech signals in telecommunications customer service.

The two most commonly used emotion definitions are dimensional emotion theory and discrete emotion theory [2]. The former defines the dimensional space of emotional state [3]. Common dimensional spaces include arousal (emotional intensity), valence (the pleasantness of a stimulus), and dominance (degree of control). The latter clusters on these dimensional spaces and defines a list of discrete categories [4], such as "angry" and "happy." This study is applied to the customer service corpus, so it mainly explores the emotional response of each user's speech valence in the dimension space.

As the research on signal processing becomes more and more mature, considerable progress has been made in the extraction of speech features. Among emotion-related speech features, prosodic features [5], low-level descriptions (LLDs), and high-level descriptions (HLDs) [6] are the basic features for emotion recognition. Compared with the raw waveform as input, past research has pointed out that these speech features can be used as model input to obtain better recognition accuracy [7]. Such traditional emotional features have high interpretability and relatively low computational cost, which have shown their importance in speech emotion recognition [8, 9]. However, it is always a bottleneck in the traditional feature extraction method for SER. This is because the expression of emotions may vary from person to person, and traditional methods are difficult to effectively understand the contextual information in the signal [10, 11].

The feature extraction method of deep learning can better solve these problems in recent years [12, 13]. Early deep learning models used a large number of convolution and linear layers to learn locally important features from the global signal such as VGG [14] and ResNet [15]. And the sequence models emerged to solve the contextual relationship in time series, and deep models were able to derive more implicit features from a large amount of data. Sequence models such as GRU [16] and LSTM have been used to learn the context of signals and more accurately recognize emotions [17, 18]. Recently, architectures considering vector quantization (VQ) have gradually been used, such as VQVAE [19], wav2vec 2.0 [20], etc. In the VQ architecture, the model learns a codebook of discrete features from the training data, and uses quantization as a feature extraction method. The discrete features of VQ are more similar to some natural modal phenomena [19]. For example, the objects in pictures can be disassembled into discrete pieces of information, and a sentence can be segmented into discrete pieces of word information. Using this discrete information allows the model to learn to find obvious and important information. The VQ-related method is also often regarded as a feature extraction process.

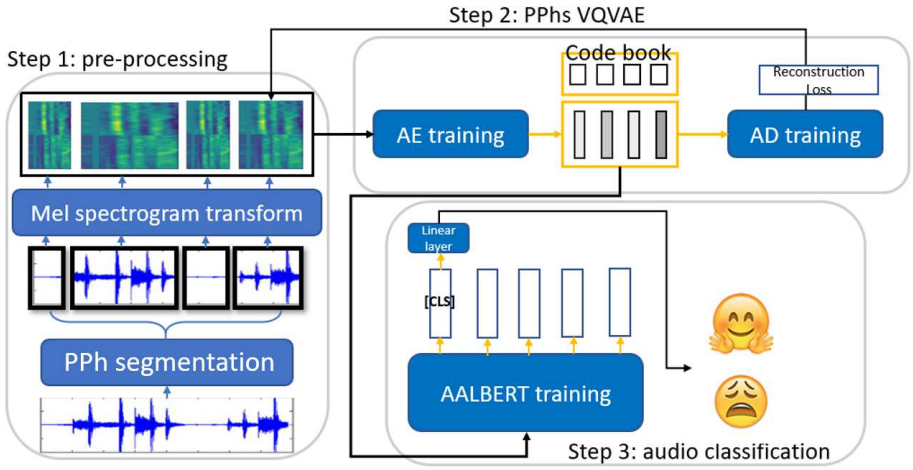


Fig. 1 *Proposed System Architecture*

In recent years, larger self-supervised models have emerged, using a huge corpus for pre-training on prediction and restoration tasks. This type of model is based on a Transformer and has many variants according to different downstream tasks, such as GPT2 [21], ALBERT [22], etc. They are pre-trained and provide other studies to fine-tune the model with a small amount of data to achieve good results in various fields. However, compared with traditional models, it is less interpretable and is difficult to adjust the model or extract features according to human understanding methods [23]. Often, better breakthroughs can only be achieved by deepening the model or increasing the amount of data. This study improves these giant pre-trained models by changing ALBERT and adding feature extraction steps related to prosodic features. Instead of just changing the model parameters, emotional-related processing is added to improve the system's ability to discriminate emotions.

The main contribution of this study is to improve the performance of the large pre-trained model for emotion recognition. We use a huge Chinese speech corpus to pre-train the audio ALBERT model [24] for speech feature restoration, and propose a feature extraction step related to emotion. Prosodic phrase segmentation of audio signals is applied so that each segment represents only a single emotional expression. Then we use the Vector Quantized Variational Autoencoder model (VQVAE) [19] to train the prosodic phrase feature codebook, which is then used to quantize the input signal into prosodic phrase feature vectors. We use the telecom-related Chinese voice customer service corpus to fine-tune and evaluate the model. We further analyze whether the added feature extraction step can improve the accuracy of emotion recognition.

II. METHODS

The system architecture proposed in this study is shown in Fig 1. First, we perform prosodic phrase segmentation of the customer's voice signal. The segmented prosodic phrase-based signals are used to extract Mel spectral features, and these spectral features are fed into subsequent models to extract deeper emotional features. We use the Vector Quantized Variational Autoencoder to train the prosodic phrase feature codebook and the prosodic phrase vector encoder. Each segmented prosodic phrase vector encoded by the autoencoder is then input into the Audio ALBERT (AALBERT) to extract feature embedding considering contextual information. Finally, the embedding vector of the start token in AALBERT is classified into the final emotion result through the linear layer. The process of each step will be described in detail below.

A. Prosodic phrase segmentation (PPS)

This study employs the speech prosodic phrase segmentation method proposed in [7] to segment the speech signals. The segmentation process includes silence segment detection, verbal/non-verbal segment identification, and prosodic turning point detection. The purpose of the prosodic phrase segmentation is to segment the original signal into segments, in which each segment is supposed to contain only a single emotion or prosodic expression. Past studies have shown that this method can enable the model to learn better about emotional information in segments, and avoid confusion caused by multiple emotional expressions in the same recognition interval [25].

B. Vector Quantized Variational Autoencoder

We perform short-time Fourier transform of the aforementioned segmented audio to obtain the spectral features of each segment, and VQVAE is applied to perform feature encoding. VQVAE is adopted to perform dimensionality reduction on the input spectral features. It reserves commonly used vectors and quantizes other less common ones. The quantization method is to train a codebook (or embedding space) during the training process, and the vector encoded by the auto-encoder selects the vector closest to the input vector from this codebook. The training loss is shown in Equation 1, and the codebook (prosodic feature space) is also fine-tuned so that the vectors in the codebook can match all the vectors appearing in the corpus as much as possible. In Equation 1, $\boldsymbol{z_e}$ is the vector encoded by the auto-

encoder, and $\boldsymbol{z_q}$ is the vector obtained after quantization. In the formula, e represents the vector to be used for quantization in the codebook, and x is the spectral feature of the current input. Each vector $\boldsymbol{z_e}$ is quantized by matching to the nearest $\boldsymbol{e_j}$ in the codebook, as shown in Equation 2. If the minimum distance between $\boldsymbol{z_e}$ and $\boldsymbol{e_j}$ in the codebook is greater than the threshold, $\boldsymbol{z_e}$ is added into the codebook as $\boldsymbol{e_{k+1}}$. Because we use prosodic phrase segments when training VQVAE, and each segment represents only one emotion, the vectors after the encoder have different representations from each other due to different prosody or emotion. Therefore, the coding differences of this encoder are mainly based on emotion, voice performance and rhythm. The final codebook is also affected by the differences in prosody and emotion, and the subsequent audio can be quantized into vectors according to the differences in emotion or prosody, so as to strengthen the model's understanding of prosodic and emotional expression.

$$\mathrm{L} = logp\left(x\middle|z_q(x)\right) + \|sg[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - sg(e)\|_2^2 \quad (1)$$

$$z_q(x) = e_j, j = arg\min_i \|z_e(x) - e_i\|_2 \quad (2)$$

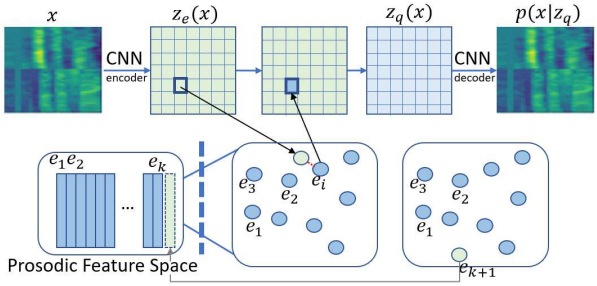


Fig. 2 *VQVAE structure*

C. *Audio ALBERT*

The Audio ALBERT model, a self-supervised speech representation model, applies Transformer-based models to audio tasks. It follows the masking training method and self-attention mechanism in pre-training, so that the model can achieve reasoning of the contextual information of the audio signals. For model pre-training, a few frames are randomly masked. After passing through multiple converter layers and prediction layers, the training target should accurately predict the masked frames back to the original frames, as shown in Fig 3. In this study, the spectral features used by the original AALBERT are replaced by the prosodic phrase features obtained by the aforementioned VQVAE, so it was also necessary to re-train the AALBERT. And this study is test on Chinese speech corpus, which is better to be pre-trained with Chinese corpus. We first perform prosodic phrase segmentation and VQVAE model quantization to obtain prosodic phrase features of the audio signals and then pre-train the AALBERT. Finally, the Chinese speech emotion corpus is used to train the AALBERT model to obtain audio embedding features. By adding the linear layer to the AALBERT model as the classification layer, the loss obtained after classifying the audio embedding features is used to fine-tune the AALBERT model so that it can embed the audio features into the emotional space.

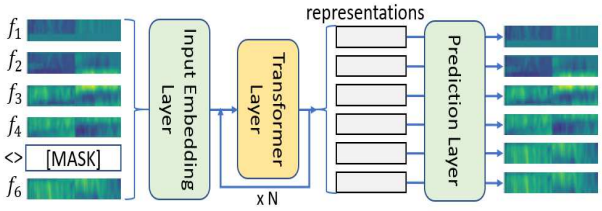


Fig. 3 *AALBERT pre-training step.*

III. EXPERIMENTS

This study experimentally evaluated whether the proposed method can improve the performance of the SER system. The input was the user's voice collected from the Chinese Telecom Customer Service Dialogue System. The outputs were the user's sentiment toward the telecom's services, categorized into neutral and negative emotional states.

A. *Dataset*

Librispeech is a common and widely used large-scale English speech corpus, mainly used in the field of ASR. This study used three datasets: train-clean100, train-clean-360, and train-other-500 in the Librispeech corpus for pre-training the AALBERT.

King-ASR is a large amount of Chinese speech corpus often used by Chinese ASR systems. This study used King-ASR 044 for the pre-training of AALBERT. The corpus is 2184 hours long, and the language is the same as the telecom speech corpus used in this study.

Telecommunications Speech Corpus (TSC) is the original corpus recorded in the customer service system. The speech corpus was manually labeled into two emotional categories, and it is ensured that each label of the data is certified by at least three annotators. For the actual application environment, users who used the customer service system mainly asked questions or expressed dissatisfaction with the service. There were almost no happy or obvious positive emotions in the corpus. Therefore, this study took the neutral and negative categories as the classification targets.

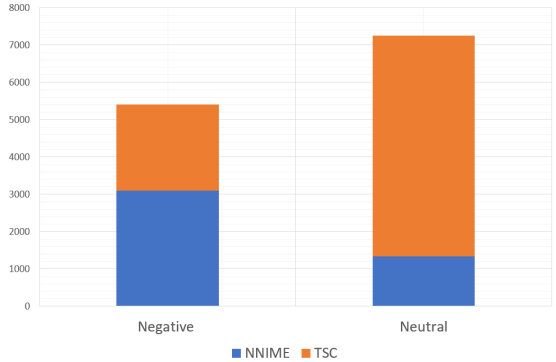
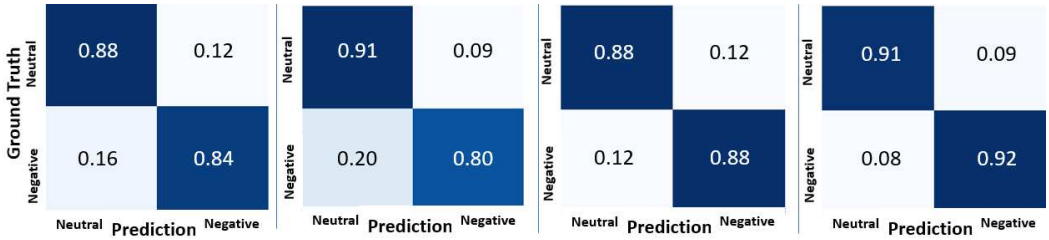


Fig. 4 *Data distributions for NNIME and TSC*



(B1) (B2) (B3) (B4)

Fig. 5 *Confusion matrices of the methods in ablation study.*

NNIME is a Chinese multimodal corpus containing 6701 sentences from 44 participants. Many scenes with limited emotions were performed for the performers according to their real-life feelings. The corpus had dimensional labels (valence and arousal) and discrete labels (angry, happy, sad, neutral, frustrated, surprised). The NNIME corpus was selected in this study to balance the unevenness of the telecommunication corpus. Therefore, this Chinese emotional corpus was selected. The data labeled with valences 4 and 5 were removed from the corpus, and the remaining corpus labeled with valences 1 and 2 were defined as negative class data, and 3 as neutral data.

B. Experimental Results

To evaluate the reliability of the method proposed in this study, we evaluated the original AALBERT and the ablation study by adding VQVAE permutation vector and prosodic phrase segmentation respectively. We first determined the optimal parameter configuration of each model in the system by experiments, and the parameter configuration is shown in the following table. In each experiment, the telecom corpus was divided into 5 folds for 5-fold cross-validation evaluation, and all the selected NNIME corpora were used as the training corpus. The model evaluation indicators were the correct rate of binary classification and F1 score.

In the evaluation experiments of the AALBERT model, different pre-training corpora were used for pre-training, and the results are shown in Tab. 1. The best recognition results were obtained by combining the two corpora. However, using a large amount of Chinese corpus King-ASR-044 provided a better emotion recognition rate than the Librispeech. But Librispeech also provided AALBERT information on the expression of emotions in the different languages.

Tab. 1 *Experimental results of pre-training AALBERT.*

Data	Accuracy	F1 score
Libri-speech	82.56±0.12	83.06±0.09
King-ASR 044	86.42±0.15	86.85±0.11
Both	**86.73±0.14**	87.11±0.09

Tab. 2 shows the evaluation results of the effectiveness of the ablation study of the method proposed in this study. From the experimental results, it can be seen that adding VQVAE and prosodic phrase segmentation to AALBERT respectively improved the performance, but the improvement of the model was not obvious. The improvement of adding only VQVAE was slightly higher than that of the other, which showed that VQVAE can extract important features and is an effective feature extraction method. Finally, simultaneously applying two mechanisms can get a more obvious performance improvement. It was shown that training the VQVAE with prosodic phrase segments yielded a codebook related to emotion. Without the addition of prosodic phrase segmentation, the trained codebook was not highly correlated with emotion. The final best performance of this system achieved 91.41%. Fig 5 shows the confusion matrix. The method of adding VQVAE can improve the recognition of neutral emotions, while adding prosodic phrase segmentation can improve the recognition of negative emotions. The prosodic phrase features enabled the model to better understand the representation of negative emotions.

Tab. 2 *Ablation study of the proposed method*

No.	Methods	Accuracy	F1 score
B1	**AALBERT**	86.73±0.14	87.11±0.09
B2	**B1 + VQVAE**	88.16±0.16	88.19±0.10
B3	**B1 + PPS**	87.73±0.12	88.29±0.08
B4	**B3 + VQVAE**	**91.41±0.19**	91.75±0.12

IV. CONCLUSIONS

This study proposes a method with prosodic-related processing to speech emotion recognition, which is applied to the telecommunication customer service corpus. A feature extraction method using prosodic phrase segmentation with VQVAE model is proposed to train a prosody phrase-related codebook to quantize the original audio with prosody- and emotion-related feature vectors. It can be seen from the experiments that the method proposed in this study can improve the performance of the model for emotion recognition when used on a large-scale pre-training model. And it achieved the best performance of 91.41% on the binary classification of telecom corpus. The reliability of the proposed method is verified, and it can cooperate with the popular large-scale models. It can be applied in a wide range in the future and can be flexibly applied to other models.

REFERENCES

[1] B.A. Erol, A. Majumdar, P. Benavidez, P. Rad, K.-K.R. Choo, and M. Jamshidi, "Toward artificial emotional intelligence for cooperative social human–machine interaction," *IEEE Transactions on Computational Social Systems,* vol. 7, no. 1, pp. 234-246, 2019.

[2] K.R. Scherer, "Towards a prediction and data driven computational process model of emotion," *IEEE Transactions on Affective Computing,* vol. 12, no. 2, pp. 279-292, 2019.

[3] J.A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality,* vol. 11, no. 3, pp. 273-294, 1977.

[4] P. Ekman, "An argument for basic emotions," *Cognition & emotion,* vol. 6, no. 3-4, pp. 169-200, 1992.

[5] R.W. Frick, "Communicating emotion: The role of prosodic features," *Psychological bulletin,* vol. 97, no. 3, p. 412, 1985.

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, and E. Marchi, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[7] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, "Speech emotion recognition considering nonverbal vocalization in affective conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 29, pp. 1675-1686, 2021.

[8] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *2009 international conference on information engineering and computer science*, 2009: IEEE, pp. 1-4.

[9] L. Chen, X. Mao, Y. Xue, and L.L. Cheng, "Speech emotion recognition: Features and classification models," *Digital signal processing,* vol. 22, no. 6, pp. 1154-1160, 2012.

[10] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, "Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN," *Sensors,* vol. 17, no. 7, p. 1694, 2017.

[11] K.-Y. Huang, C.-H. Wu, and M.-H. Su, "Attention-based convolutional neural network and long short-term memory for short-term detection of mood disorders based on elicited speech responses," *Pattern Recognition,* vol. 88, pp. 668-678, 2019.

[12] 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)*, 2016: IEEE, pp. 5200-5204.

[13] Q.-B. Hong, C.-H. Wu, M.-H. Su, and C.-C. Chang, "Exploring Macroscopic and Microscopic Fluctuations of Elicited Facial Expressions for Mood Disorder Classification," *IEEE Transactions on Affective Computing,* vol. 12, no. 4, pp. 989-1001, 2019.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556,* 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[16] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078,* 2014.

[17] K.-Y. Huang, C.-H. Wu, M.-H. Su, and Y.-T. Kuo, "Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model," *IEEE Transactions on Affective Computing,* vol. 11, no. 3, pp. 393-404, 2018.

[18] J.-H. Hsu and C.-H. Wu, "Attentively-Coupled Long Short-Term Memory for Audio-Visual Emotion Recognition," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020: IEEE, pp. 1048-1053.

[19] A. Van Den Oord and O. Vinyals, "Neural discrete representation learning," *Advances in neural information processing systems,* vol. 30, 2017.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems,* vol. 33, pp. 12449-12460, 2020.

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog,* vol. 1, no. 8, p. 9, 2019.

[22] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942,* 2019.

[23] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, and T. Rocktäschel, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems,* vol. 33, pp. 9459-9474, 2020.

[24] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-y. Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021: IEEE, pp. 344-350.

[25] S. Sahoo, P. Kumar, B. Raman, and P.P. Roy, "A segment level approach to speech emotion recognition using transfer learning," in *Asian Conference on Pattern Recognition*, 2019: Springer, pp. 435-448.