An efficient framework for constructing speech emotion corpus based on integrated active learning strategies

Fuji Ren, *Senior Member, IEEE,* Zheng Liu, and Xin Kang

Abstract—Speech emotion recognition has been developed rapidly in recent decades because of the appearance of machine learning. Nevertheless, lack of corpus remains a significant issue. For actual speech emotion corpus construction, many professional actors are required to perform voices with various emotions in specific scenes. In the process of data labelling, since the number of samples of different emotion categories is extremely imbalanced, it is difficult to efficiently label the samples. Hence, we proposed an integrated active learning sampling strategy and designed an efficient framework for constructing speech emotion corpora in order to address the problems presented above. Comparing experiments with other active learning algorithms on 13 datasets, our method was shown to improve sampling efficiency. In addition, it is able to select small category samples to be labelled with preference in imbalanced datasets. During the actual corpus construction experiments, our method can prioritize selecting small class emotion samples. As even when the amount of labelled data is less than 50%, the accuracy rate still can reach 90%. This greatly enhances the efficiency of constructing the speech emotion corpus and fills in the gaps.

Index Terms—Affective computing, Speech emotion recognition, Corpus construction, Active learning, Imbalanced dataset

1 Introduction

Affective computing is currently one of the most crucial research topics that aim to enable intelligent systems recognize, feel, interpret and generate human emotions [1]. It has broad prospects in many practical applications. In robotics, affective computing is essential for designing collaborative or service robots which can interact with humans smoothly [2] [3]. In education, it can be used for improving the learning process and knowledge transfer according to the mental state of students [4] [5]. In psychology, potential psychological problems such as depression can be discovered in advance by analyzing the information of the patients [6] [7].

Since the concept of affective computing was first proposed by Picard in the 1990s, research on affective computing has attracted a lot of attention from academia and industry. Fasel [8] made a detailed analysis and summary of the most effective facial expression recognition system in the past ten years. Coan [9] used brain waves captured by sensors to analyze changes in emotions. Ren [10] proposed a mental state transfer network to describe the dynamic process of human emotions. Poria [11] reviewed the unimodal to multimodal affective computing.

In general, affective computing research fields could be classified into the following major categories. Emotion recognition based on facial expressions [12] [13]: Human emotions can often be revealed naturally and directly through facial expressions. Emotion recognition based on language text [14] [15]: The content of human communication and written text contains the authors' opinions, attitudes and emotions at the time. Emotion recognition based on physiological information [16] [17]: It can detect human emotional states, such as heartbeat, blood pressure, etc., by monitoring and analyzing the user's physiological signals. In addition, there are also emotion recognition based on body movements, speech, and so on.

Speech plays a crucial role in human society since it is one of the information conveyed by human communication. With the joint efforts of researchers in recent decades, speech emotion recognition has been rapidly developed [18] [19] [20], especially the emergence of machine learning in recent years has greatly promoted this boom [21] [22] [23]. Machine learning relies on large amounts of data for training, which makes the datasets in the current research on speech emotion recognition become precious and indispensable. The construction of corpus labeled data samples is a huge project with extremely high cost, which often requires countless manpower and material resources. It not only requires professional actors to record, but also the recorded voice fragments have to be manually screened and annotated.

The current methods of constructing datasets usually have the following shortcomings: 1. Most of the datasets are obtained by professional actors performing in a given scene, so the scene is relatively single, the emotional performance is relatively exaggerated, and the naturalness is not high. 2. In the process of labeling speech utterances, due to the uneven distribution of different types of emotions, it is impossible to prioritize the screening of rare samples. The annotator is required to listen to all the speech utterances one by one, and the labelling cost is high. 3. Datasets are mostly collected in specific scenarios. In actual application scenarios, the accuracy of speech emotion recognition will

- *Fuji Ren is with University of Electronic Science and Technology of China, E-mail: renfuji@uestc.edu.cn*
- *Zheng Liu and Xin Kang are with the Institute of Technology and Science, Tokushima University, Tokushima, Japan. E-mail: c502047006@tokushima-u.ac.jp; kang-xin@is.tokushima-u.ac.jp*

Manuscript received XX, 2021; revised XX, 2021.

be low due to different languages, different speaking styles, and different application scenarios.

Lack of emotion corpus restricts the development of deep learning in the field of speech emotion recognition. To cope with the problem of insufficient datasets, researchers often use data enhancement in speech emotion experiments to increase the number of corpus samples [24] [25] [26], which expands the amount of data. Since the target voice is generated by the algorithm, it is impossible to ensure the quality of the voice. There are also some studies [27] [28] [29] [30] that make up for the shortcomings of the corpus through the form of cross-corpus, but it can't solve the problems fundamentally because of insufficient datasets. In the process of constructing the corpus, some tools for labelling emotion types were invented [31] [32] [33], which improved the efficiency of constructing the corpus to some extent but required the annotator to listen to the voice from the beginning to the end. Therefore, how to construct a corpus suitable for specific application scenarios quickly, efficiently and at low cost is the focus of this paper.

According to our knowledge, there is not much research and application for the improvement of the efficiency of labeling samples in the construction of speech emotion datasets. The purpose of active learning [34] [35] is to select the most valuable samples for the current model through certain screening strategies. It is mainly composed of five core parts, including: unlabeled sample pool, screening strategy, labeling experts in related fields, labeled sample pool, and target classification model. Active learning combines the above five parts into the same process, and updates the performance of the classification model, unlabeled sample pool, and labeled data set in an iterative training method until the target model reaches the preset performance or no longer provides labeled data.

In this paper, we present an active learning method to improve the construction efficiency of speech emotion data sets in the labelling process of constructing the emotion corpus. Our method is divided into the following steps. First, use a small number of labeled samples to train a logistic regression classifier, predict all unlabeled samples, and use the predicted probability to find the samples with the most unclear category, which we call uncertainty sampling. Then, use the feature relationship of all unlabeled samples to find the sample at the feature center in the feature space, and filter out the most a representative sample, which is representative sampling. Next calculate the feature distance between the unlabeled sample and the labeled sample one by one and select the samples that are different from the labelled samples, which is diversity sampling. Finally, by using the probability of the unlabeled data and labeled data, select the unlabeled samples which make all categories have the same distribution. Therefore, the samples selected by our method have the characteristics of strong uncertainty, strong class representativeness, strong feature diversity, and strong complementarity between classes. Compared with other active learning algorithms, the experimental results show that our algorithm is significantly better than other algorithms in screening small class samples.

To solve the problem of the lack of datasets in the field of SER, we designed a framework for speech emotion corpus construction. The framework integrates the newly proposed active learning strategy, and at the same time uses voice activity detection, feature extraction and other technologies. The raw voice of our framework is not limited to the performance of the actors, so it can be adapted to voice data collection in different scenarios. The framework first splits long raw data into segments by using endpoint detection. After simple filtering, uses opensmile [36] feature extraction tool to extract the features of all filtered segments as unlabeled datasets. Finally, by using the proposed active learning method, screen out the samples for labelling. The framework can be applied to data collection in different voice scenarios, and has the characteristics of high naturalness, flexibility, and high efficiency. In this work, we take TV drama videos as the data source. After subjective listening and discriminating experiments, our scheme has achieved excellent results.

The main contributions of our work are summarized as follows:

1) We proposed a new active learning strategy using cross entropy distance for the sampling of imbalanced datasets.
2) Based on the new strategy, we proposed an integrated active learning method, which can not only improve the efficiency of overall sampling, but also give priority to the small class samples of imbalanced dataset.
3) We designed a framework for constructing emotional speech emotional corpus. The proposed framework can be applied to the audio collection in different scenarios.
4) Regarding the lack of corpus in the field of speech emotion recognition, our work provides an efficient method for constructing a speech emotion corpus.

The rest of the paper is arranged as follows. The second section introduces the research work related to our research. The third section introduces the newly proposed active learning strategy and the framework for constructing the speech emotion corpus. The fourth section shows the comparative experiments between our active learning strategy and other methods, as well as the subjective evaluation results in the actual construction of the emotional corpus. We also have a discussion on the efficiency of our proposed method strategy combination, calculation efficiency, parameter selection, and speech emotion corpus construction. In the fifth part, we make a summary of our work and a description of the future work.

2 Related Work

2.1 Speech Emotion Recognition

Speech emotion recognition is a kind of classification supervised learning. There are two main types of emotion description, discrete emotion description model and dimensional emotion description model. The discrete emotion description model is popular in the psychology field. Paul Ekman [37] proposed the six basic emotions of Anger, Disgust, Fear, Happiness, Sadness, and Surprise, which are widely used in computer science, especially affective computing. Plutchik, R. [38]. proposed the Plutchik mood wheel with 8 emotions., and other complex emotions are composed

of a combination of these eight emotions. Regarding the classification of emotion types, there are many discussions in the academic circle such as [39] [40]. The frequency of expression of emotions in daily life is different. For example, neutral emotions are often more than other emotions. Therefore, the process of emotion corpus construction faces the problem of imbalanced emotion data sampling.

Regardless of research or practical application, the construction of the speech emotion corpus in a specific scenario is a very important link. IEMOCAP corpus is widely used in the field of speech emotion recognition, which was collected through a form of impromptu performance performed by 10 actors in a specific environment [41] . The German emotional speech corpus Emodb was recorded by the Technical University of Berlin, which collected speech data of a total of 10 Germans showing different emotions [42]. And the Belfast sentiment corpus [43] was obtained by the interpretation of 5 paragraphs by a total of 40 recorders from Queen University. For the construction of the speech emotion corpus, Cowie et al. [31] developed the FEELTRACE emotion tagging system, which is based on the establishment of emotion valence and activation models and provides standardized tools for tagging speech emotion data. Morris et al. [32] constructed a labelling tool with cartoon characters as prototypes based on the three dimensions of the emotional PAD model. The demeanor and image size of the cartoon characters respectively represent the intensity of emotion. There are also some tools that use web development technology to build multi-modal emotion markings [33]. These tools have improved the efficiency of speech emotion corpus construction.

The extraction of speech features is an indispensable link in the research of speech emotion recognition. Speech features are mainly divided into three categories, acoustic features, deep features, and hybrid features. Acoustic features are mainly used to describe the tone, amplitude, timbre and other information of the sound, such as formants, Mel cepstrum coefficients, and so on. The acoustic features can be extracted by the feature extraction tools, such as Librosa [44] and Opensmile [36], which can not only extract the inherent acoustic features of speech, but also low-level descriptions of speech in the form of statistics. Some classic machine learning models such as SVM, KNN, GMM, etc. often use acoustic features to classify emotions [45] [46]. Due to the limited number of acoustic features, deep features can be further extracted from speech spectrogram by using neural networks [47]. The papers [23] [48] [49] [50] [51] use high-level features extracted from acoustic features, [50] directly extract original audio, and [51] [52] [53] reconstruct the features by using representation learning. Hybrid features use a variety of information in the context at the same time, not only at the speech level, but also incorporate other modal features to increase the recognition rate of speech emotion recognition [54] [55] [56] [57].

2.2 Active Learning

In classification tasks, supervised learning usually relies on many manually labeled training samples, and the process of labeling samples is very expensive [58]. The emergence of active learning has played a significant role in reducing the cost of marking. Regarding active learning research, the most important thing is how to choose the most valuable samples for labeling [34] [35] [59]. There are some classic sampling strategies developed so far, such as the random sampling strategy [60], which mainly randomly selects a certain proportion of samples from unlabeled samples and provides them to the model. Uncertain sampling strategy [61], mainly by combining the characteristics of the sample itself, calculating and selecting the least easily distinguishable sample, the sample with the best value. QBC query strategy [62], this algorithm will train multiple classifiers from different perspectives, and jointly screen samples for labeling experts. Active learning has attracted the attention of many researchers, and many methods have been expanded on this basis. For example, Density [63] uses a density map to find the most representative sample for labeling from all unlabeled samples. LAL [64] uses the pre-trained regression model to calculate the sample prediction error to determine the sample to be labelled. Query [65] finds the most informative and representative samples, and provides them to the annotators for labelling.

Over the years, active learning has been widely used in different fields. Goudjjl et al. [66] combined active learning with the SVM model, and selectively selected informative samples to train the SVM model, which achieved excellent results in text classification. Yan et al. [67] addressed the problem of difficult sample labeling in long text classification, combined with active learning to screen long text samples with high labeling value, and condenses the long text into words to facilitate labeling by the annotator. The papers [68] [69] combined active learning and deep learning to improve the efficiency of image segmentation and labeling in the biological field. Cao et al. [70] combined active learning and the CNN model to label complex spectral pictures. In the field of speech emotion recognition, Mohammed et al. [71] proposed an iterative fast converging incremental adaptation algorithm that combines active learning and supervised domain adaptation to address the lack of generalization of speech emotion classifiers in real applications. Mohammed et al. [72] also used greedy sampling and DNN model to conduct speech emotion recognition experiments, and the results show that active learning can improve the performance when the training set is limited. Vaaras et al. [73] combined CPC and various dimensionality reduction methods to explore the performance of clustering-based active learning under different feature conditions.

Regarding the construction of speech emotion corpus, although the emergence of labeling tools has improved the accuracy of the annotator's judgment and the convenience of labeling, it often needs the annotator to annotate all samples from beginning to end. Therefore, to further reduce the cost of labeling, we propose a novel active learning method based on class imbalanced data sampling and a framework for speech emotion corpus construction. Experiments show that our method has achieved excellent results.

3 METHOD

3.1 Active Learning Strategies

In our work, we propose an integrated active learning method. The proposed method includes a commonly used

logistic regression pre-classifier and four sampling strategies, uncertain sampling, representative sampling, diverse sampling and complementary sampling. Especially for complementary sampling, it is able to preferentially select rare samples and provide them to the annotator for labeling according to the distribution of the current samples.

For the logical classifier, suppose we have a total data set C with K categories. The dataset C contains the labeled data set L and the unlabeled data set U , namely $C = L+U$, where $L = \{(X_i, Y_i), ..., (X_m, Y_m)\}$. $X_i \in \mathbb{R}^N$ represents the features of the sample with N dimensions; $Y_i \in \mathbb{B}^k$represents the category of sample with K dimensions, and each dimension represents a different category value.

We train K binary logistic regression models for K different categories:

$$F_k : \mathbb{R}^N \to \mathbb{B} \tag{1}$$

Therefore, for each binary logistic regression, we can get:

$$Y_k = f_k(X) = \frac{1}{1 + exp(-W_k * X)} \tag{2}$$

Among them, W_k is the weight of each logistic regression classification model, Y_k is the probability of the logistic regression of the k-th category of each binary classification, and its value is between 0 and 1, and then P_k is used to represent Y_k in the following paper.

Uncertain sampling is mainly used for sampling ambiguous sample points at the decision boundary, which helps to clarify the decision boundary. Therefore, combined with the predicted probability of the sample, we use cross entropy to find the largest uncertain sample point. The following is the calculation formula for uncertain sampling. For sample X, the cross entropy H of each category is calculated separately, and the largest cross entropy is selected as the uncertainty value of the sample. We will sort all the samples and select the samples with the greatest uncertainty for subsequent sampling.

$$U(X) = max\left\{H\left(P_k\right) | k = 1, ..., K\right\} \tag{3}$$

$$H(P_k) = -P_k \log P_k - (1 - P_k) \log(1 - P_k) \tag{4}$$

Representative sampling is a kind of sampling that can best represent all unlabeled sample points. The sample point closest to the center of the unlabeled samples will be selected through the distance calculation of all sample features as the most representative sample point for labeling. In the following formula, for each unlabeled sample X, the feature distances of N dimensions between X and all other unmarked sample points will be calculated, and the average value is calculated as the feature similarity of the sample points.

$$R(X) = \frac{1}{|U| - 1} \sum_{X' \in U - X} -Dis_{eu}(X, X') \tag{5}$$

$$Dis_{eu}(X, X') = \sqrt{\sum_{i=1}^{N} (X_i - X'_i)^2} \tag{6}$$

Diversity sampling is mainly to find the unlabeled sample point closest to the feature center of all labeled samples. We assume that the feature center of the labeled sample is the boundary center of feature diversity, so that by looking for the closest to the feature center can effectively reduce the overfitting of the trained model. We use the following formula to calculate and find the point with the smallest feature distance between the unlabled sample X and all the labeled points X′.

$$D(X) = \min_{X' \in L} Dis_{eu}(X, X') \tag{7}$$

Complementary sampling can give preference to samples of small categories as much as possible, so that the overall category distribution of labeled data will become similar. Cross entropy is used to measure the similarity of the distribution using the category probability, the smaller the value, the more similar the distribution. In the following formula, P represents the sum of the category distribution with the probability of unlabeled points and the category of all labelled points, and Q is a reference, which represents the distribution of K same proportions, $(1/K, ..., 1/K)$. When the sum of the distribution of the unlabeled point we selected and all labelled points is close to the Q of the equal distribution, the cross-entropy value is the smallest, then the unlabeled points will be selected for labelling.

$$D(X) = \min_{X' \in L} Dis_{ce}(P(L(X)), Q) \tag{8}$$

$$L(X) = L \cup (X, Y) \tag{9}$$

$$Dis_{ce}(P, Q) = -\sum_{i=1}^{K} Q_i \log P_i \tag{10}$$

Algrithm 1 is our proposed integrated active learning strategy. Each strategy retains a certain percentage of samples. Through the integration of the above strategies, the samples selected by our strategy have the characteristics of strong uncertainty, strong class representativeness, strong feature diversity, and strong complementarity of imbalanced classes.

Algorithm 1: The Integrated Active Learning Strategy

Input: input L U λ ;
L: Labeled Data ;
U: Unlabeled Data ;
λ: Query Ratio
Output: ouput selection result L;
L: Updated Labeled Data;
A total of m times selection;
for $i = 1; i \leq m$ **do**
  Use feature and label of L to train K classifiers LR;
  Use feature of U and LR, get U that contains prediction possibilitis;
  Select samples by using the four strategies;
  U_u=select_samples(λ, U(U));
  U_r=select_samples(λ, R(U_u));
  U_d=select_samples(λ, D(U_r));
  U_c=select_samples(λ, C(U_c));
  update L and U ;
  L=L + U_c ;
  U=U - U_c ;

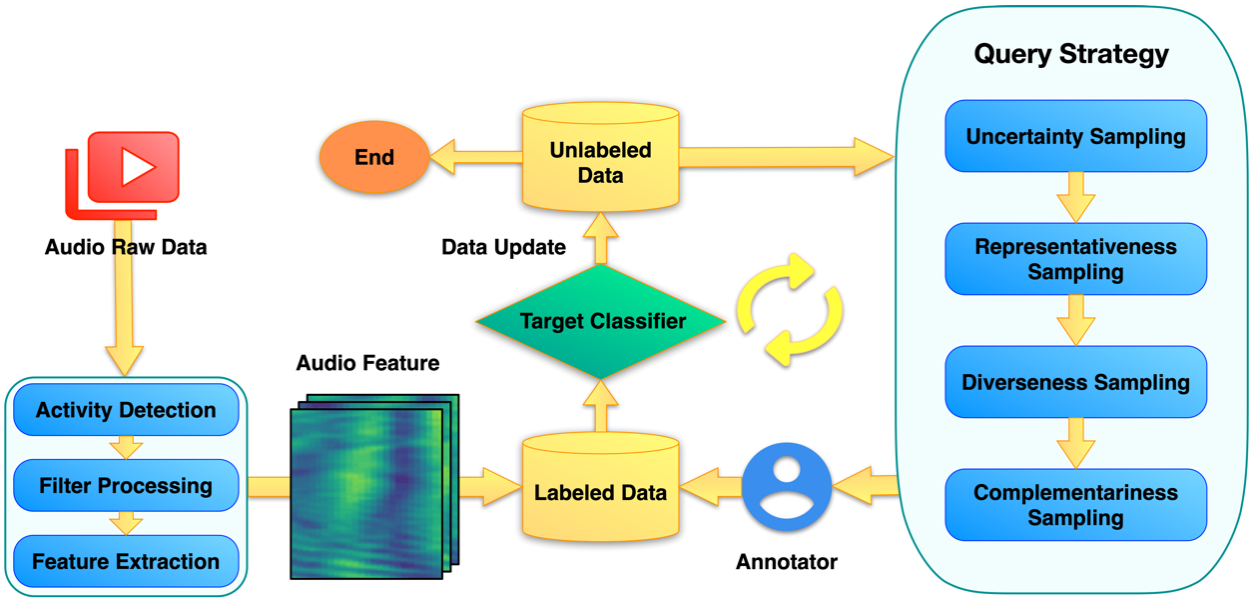


Fig. 1: Flow chart of the speech emotion corpus construction

3.2 Speech Emotion Corpus Construction Framework

When speech emotion recognition is applied in a specific scene, it is usually necessary to construct an emotion corpus corresponding to the specific scene, which often consumes large costs. To solve this problem, we propose a framework for constructing speech emotion corpus based on our active learning strategies.

The framework is shown in Figure 1. The audio processing part is mainly divided into three steps. First, the voice is split into segments through voice activity detection. Since the segments are split from the original data, data cleaning and manual filtering are required. After simple filtering, all the segments will be extracted for features.

For the labelling process, sample selection is an important part of active learning. It consists of five components, labeled data set, unlabeled data set, classifier, annotator, and selection strategy. All the features we get from the audio processing part will be divided into unlabeled data and labeled data. First, manually annotate a part of the samples. The classifier trained with this data will do a preliminary pre-classification of other unlabeled data and send all the feature values and probability values into the core selection strategy. Our method will repeatedly filter out a batch of high-quality data, let the annotator label, and the annotator will update the labelled data and unlabeled data respectively, then carry out the next round of screening. After several times of manual labeling, the finally trained target classifier replaces the manual labeling, so as to achieve the purpose of saving cost.

In the entire speech emotion corpus construction framework, the original audio source can be combined with the current application scenario, not only can be obtained from the network video, but also the audio from a specific scene can be recorded, which greatly improves the flexibility of the data source. For the voice activity detection part, we use the auditok tool to split the utterance into segments. In the feature extraction part, we use the openSmile speech feature extraction tool to extract low-level descriptors of acoustic features as language features.

4 EXPERIMENTS AND DISCUSSION

4.1 Active Learning Comparative Experiment

4.1.1 Datasets

To verify the effectiveness of our proposed active learning strategy in imbalanced class datasets, we selected 13 imbalanced class datasets which are commonly used in machine learning experiments and speech emotion recognition filed. Table 1 shows the description of the used dataset in detail, which contains the name of the dataset, the number of instance, class and feature, the detail in different classes. Aggregation, Blood, Diabetes, qsar-biodeg, Vote, Vowel WBC and Thyroid are UCI data sets [74]. SAVEE [75], EMO-DB, CASIA [76],eNTERFACE05 [77] and IEMOCAP [41] are speech emotional datasets. Since CASIA, eNTERFACE05 and IEMOCAP datasets are not so imbalanced datasets, we manually selected a part and processed them into CASIA_im, eNTERFACE05_im and IEMOCAP_im as imbalanced datasets.

4.1.2 Baseline

In addition, we also compared our proposed method with five baseline active learning strategies as follows:

1) QBC [62]: Active learning that trains multiple classifiers from different perspectives, and jointly select the samples for labeling.
2) Random: Random sampling, randomly selecting samples that need to be labeled from the sample pool to be labeled
3) Unc [61]: Uncertainty sampling, select the most uncertain sample considered by the model.
4) Density [63]: Using a density map to find the most representative sample for labeling from all unlabeled samples

TABLE 1: Detail description of the 8 datasets used in the experiment

Dataset	Instance	Class	Feature	Detail
SAVEE	480	7	62	60;60;60;60;120;60;60
EMO-DB	535	7	62	127;81;46;69;71;79;62
CASIA	6000	5	62	1200;1200;1200;1200;1200
CASIA_im	1400	5	62	100;100;1000;100;100
eNTERFACE05	1257	6	62	210;210;210;209;209;209
eNTERFACE05_im	610	6	62	40;60;80;100;150;180
IEMOCAP	4983	4	62	1051;1707;1161;1064
IEMOCAP_im	1600	4	62	300;1000;200;100
Aggregation	788	7	3	45;170;102;273;34;130;34
Blood	748	2	5	570;178
Diabetes	768	2	9	500;268
qsar-biodeg	1055	2	42	356;699
Vote	435	2	17	267;168
Vowel	871	6	4	72;89;172;151;207;180
WBC	683	2	9	444;239
Thyroid	215	3	6	150;35;30

5) LAL [64]: LAL uses the pre-trained regression model to calculate the sample prediction error to determine the samples for labeling.

4.1.3 Evaluation Metrics

For the evaluation index of the experiment, we used the macro-average F1 value to measure the prediction effect of different methods. The F1 value is the harmonic average of the model's accuracy and recall, which indicates the two prediction performances of the model.

$$F_1 = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{11}$$

4.1.4 Experimental Results

In the experiment, we used the Alipy [78] active learning tool developed by Huang Shengjun's team. A total of 5 active learning algorithms were used in our experiments on 8 classes of imbalanced datasets. Figure 2 is the result of our experiment, where the horizontal axis represents the number of labelled samples, and the vertical axis represents the results of the macro average F1 value in the test set using the labelled samples, of which the test set accounts for 30%. The different colored lines in the figure represent the results of different algorithms. The red line is our proposed method and our method achieved excellent results, whether it is a two-class or multi-class imbalanced dataset.

To further illustrate the effectiveness of our method in imbalanced sample selection, we selected the results on the dataset qsar-biodeg and the multi-class Vowel dataset. Figure 3 shows the results on the qsar-biodeg dataset. The horizontal axis represents the total number of samples selected, and the vertical axis represents the number of samples in different classes. The labels 0 and 1 represent two different types of samples. The dataset qsar-biodeg contains two classes with a ratio of 1:2. In the selection of random sampling (d), the selected sample has a uniform proportion, which best shows the proportion of the data set itself. (a)(b)(c)(e) methods do not calculate the class ratio of labelled samples, so they cannot achieve the balance of overall samples. Our algorithm (f) can give priority to the balance between different classes and select rare samples so that the proportions of different samples are close to each other.

Also in the multi-class emotional data set EMO-DB, as shown in Figure 4. Due to the limited number of labeled samples, the trained classification decision boundary cannot accurately determine the feature space of the category. Our method fully considers the category distribution of the labeled samples, so it can prioritize the selection of samples with rare classes to ensure that the selected sample classes are balanced.

4.2 Speech Emotion Corpus Construction Experiment

4.2.1 Datasets

To evaluate the performance of our proposed method in the construction of speech emotion dataset, we perform the selection experiment on several speech emotional datasets. This experiment uses the following datasets commonly used in the field of SER. Each dataset represents a fixed speech annotation scene, so we can better test the annotation effect of our method on the speech emotion datasets.

1) CASIA [76]: A dataset constructed by the China Institute of Automation Science in 2005, which was recorded in a pure environment by 4 professional sound recorders, two men and two women. There are 5 emotions including happiness, sadness, anger, surprise, and neutrality.
2) EMO-DB [79]: A German emotional speech corpus recorded by the Technical University of Berlin, with 10 actors (5 males and 5 females) performing 7 emotions on 10 sentences (5 long and 5 short). The selection of corpus text follows the principle of semantic neutrality and no emotional tendency. Voice recording is done in a professional recording studio, requiring actors to reminisce their own real experience or experience to brew emotions before interpreting a specific emotion to enhance the realism of emotions.
3) eNTERFACE05 [77]: It is an audition emotion dataset that contains six emotions such as anger, disgust, fear, happiness, sadness, and surprise. The dataset contains a total of 1166 video sequences. Of these 1166 video sequences, 264 female recordings (23%) and 902 male recordings (77%).
4) IEMOCAP [41]: Collected by the Sail Laboratory of the University of Southern California, it is a database of actions, multiple modes and multiple peaks. Completed by 10 actors and actresses, the dataset is about 12 hours of audiovisual data, which contains 10 emotions such as anger, neutrality, and excitement.
5) SAVEE [75]: The database contains a total of 480 British English audios from 4 male actors. These recordings have 7 different emotions: angry, disgusted, scared, happy, sad, surprised, neutral.

The detail description of emotion datasets used are shown in Table 1.

4.2.2 Experiments and Results

We conducted two sets of experiments respectively. In the first group of experiments, the number of samples to be

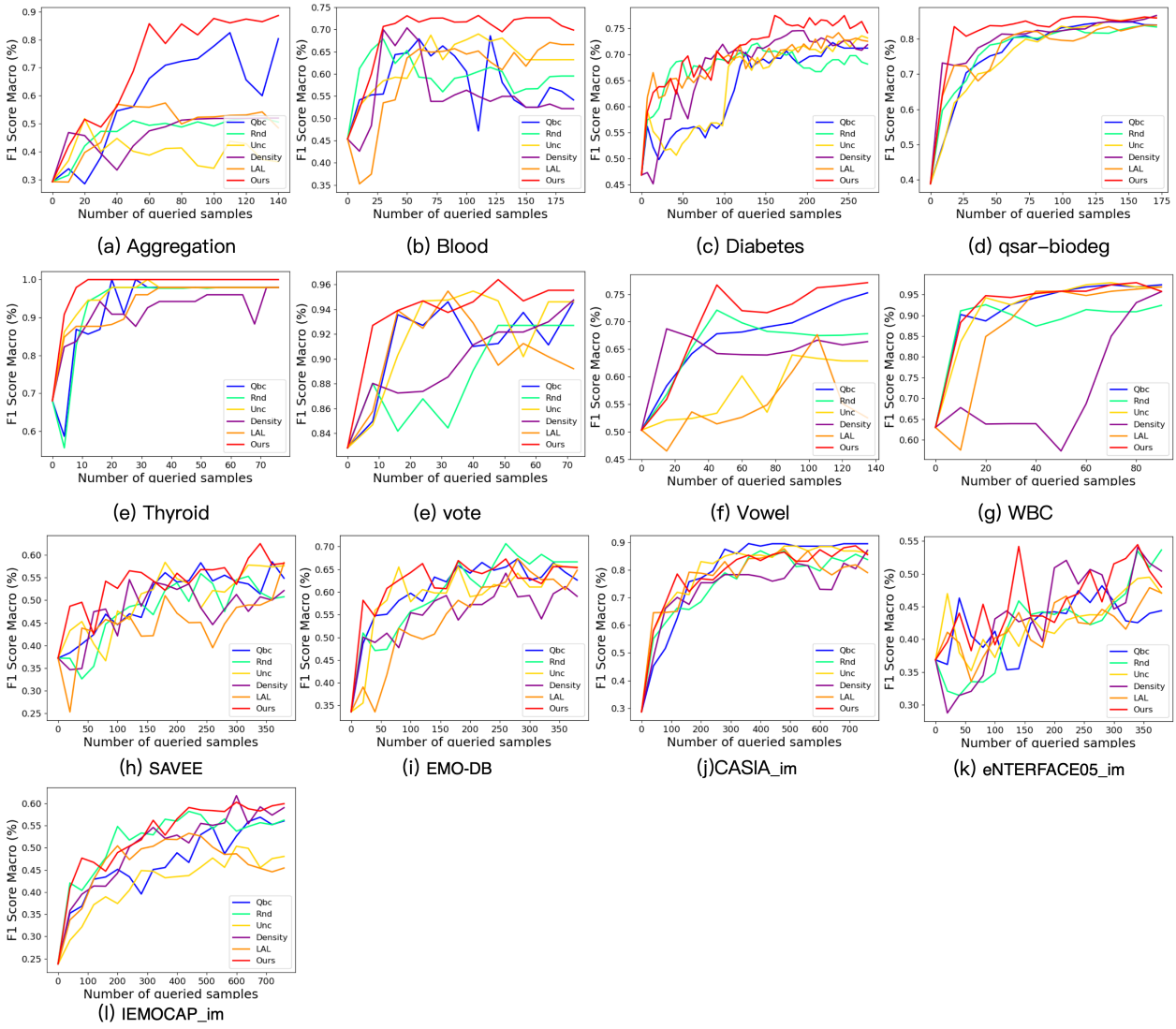


Fig. 2: Comparison of different active learning methods on 15 datasets using the macro-average F1 value

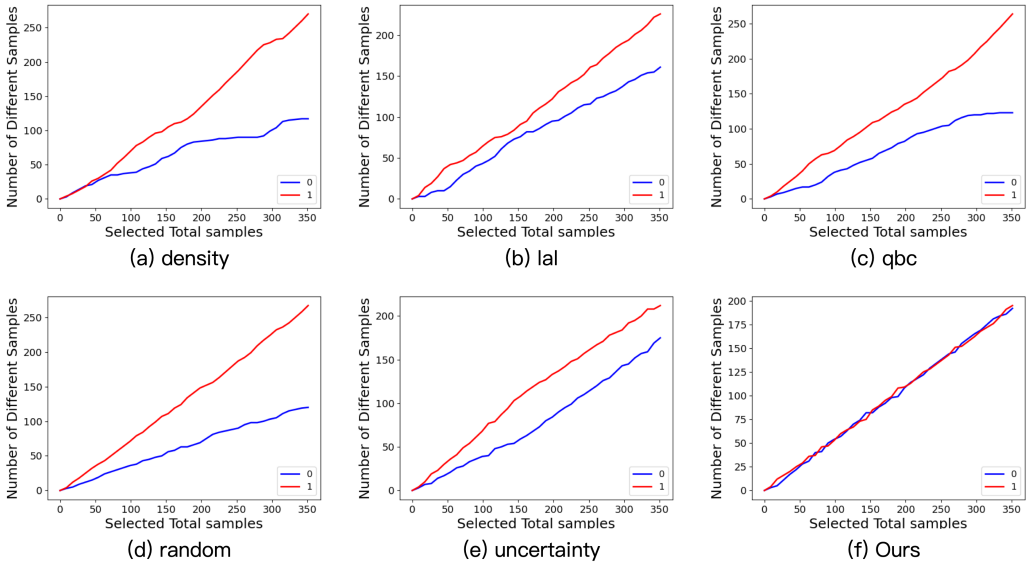


Fig. 3: Selection of samples of different categories in the binary class dataset qsar-biodeg

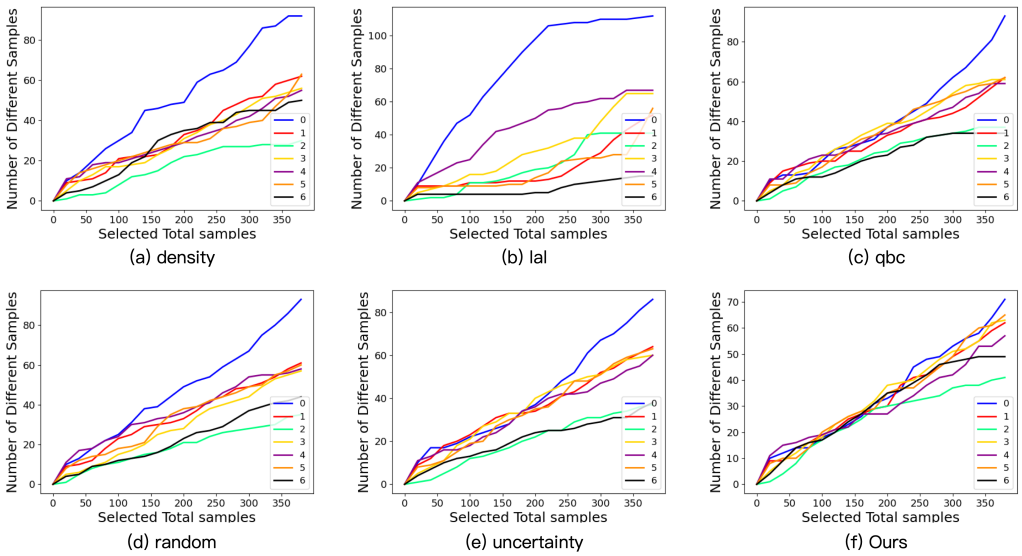


Fig. 4: Selection of samples of different categories in the multi-classification data set EMO-DB

queried is fixed to 20 each query. In the second group of experiments, the samples to be selected are fixed to 5 percentage each query. We use macro average F1 valuation to measure our method. The train set and test set of each experiment were randomly divided five times, so the experimental results were averaged.

Table 2 and Table 3 show the selection efficiency results of our experiments respectively, the number 1 to 10 represent annotation times. The number 1 represents the initial labeling round. We observe that the trained classifier is positively correlated with the number of labeled samples, but the efficiency of proposed method is affected by the sample size of datasets. For example, the total size of CASIA and IEMOCAP data sets are 6000 and 4983, when the number of labelled samples is about 20%, the classifier will have a good performance. For the datasets eNTERFACE05, EMO-DB and SAVEE, the total sample size is small, and the classifier can chive better results when the number of labelled samples is about 40

Based on the above experiments, we used our proposed framework to construct a lightweight speech emotional dataset. The speech audio was derived from Chinese TV dramas. After the processing of video-to-audio, voice activity detection, and simple filtering, 1151 speech segments were filtered out as the unlabeled pool. We divide emotion into four classes: angry, happy, neutral, and hurt. Using the proposed active learning method, we labelled a total of 404 samples on 20 rounds, and finally, get the dataset shown in Table 4.

TABLE 4: The results of actual speech emotion dataset construction

	Anger	Happy	Neutral	Sad	Total
Labeled Sample	50	32	298	24	404
Prediction Sample	16	16	712	3	747
Total Sample	66	48	1010	27	1151

Due to the selection of our proposed active learning strategies, many small-class samples are preferentially selected. The subjective audiometric test shows that when the amount of labelled data is less than 50%, the accuracy rate can reach 90%.

4.3 DISSCUSSION

4.3.1 *Query Strategy Combination*

To verify the efficiency of different strategy combinations, we randomly combined strategies into the following models, where C is a single complementary sampling, U+C is a combination of uncertainty sampling and complementary sampling, and R+C is a combination of representative sampling + Complementary sampling, D+C is combination of diversity sampling and complementary sampling, and U+R+D+C is a fusion of four sampling strategies.

We counted the average of the first 5 querying accuracy values. Table 5 shows the comparison results of different strategy combinations. U+R+D+C model achieves best performance, where U+R+D is used to improve the accuracy of overall dataset, and C is used to give the priority of selecting the small-class samples.

4.3.2 *Ratio Value*

The Ratio value represents the proportion of samples to be retained after querying by the four strategies. To further explore the efficiency of different ratio values, we perform the experiments with parameters 0.5, 0.6, 0.7, 0.8, and 0.9. As the Table 6 shown, we observed that the ratio value around 0.8 can achieve better results.

4.3.3 *Algorithm Running Speed*

We reported the average CPU time of each query time for the datasets with different sample sizes. Figure 5 shows the running speed of our proposed method under the CentOS system with the machine configured as Intel(R) Core(TM) i7-6800K CPU @ 3.40GHz. The horizontal axis represents the number of queries, and the vertical axis represents the

TABLE 2: The active strategy selects 20 samples in each round and the performance of the target classifier accuracy

Dataset	Test	Init	1	2	3	4	5	6	7	8	9	10	All
SAVEE	96	20	0.19±0.04	0.27±0.06	0.38±0.08	0.39±0.06	0.43±0.08	0.46±0.1	0.45±0.06	0.5±0.05	0.5±0.07	0.5±0.06	0.53±0.05
EMO-DB	107	20	0.34±0.06	0.38±0.05	0.43±0.04	0.49±0.03	0.52±0.02	0.55±0.03	0.55±0.05	0.57±0.05	0.61±0.06	0.61±0.05	0.69±0.05
CASIA	1200	20	0.24±0.02	0.31±0.03	0.39±0.04	0.42±0.02	0.43±0.02	0.46±0.02	0.48±0.02	0.5±0.02	0.52±0.02	0.53±0.01	0.62±0.01
CASIA_im	280	20	0.31±0.02	0.37±0.07	0.49±0.04	0.56±0.07	0.61±0.07	0.68±0.04	0.7±0.03	0.72±0.05	0.72±0.04	0.73±0.04	0.83±0.02
eNTERFACE05	251	20	0.16±0.03	0.19±0.02	0.23±0.04	0.24±0.04	0.25±0.04	0.26±0.03	0.28±0.03	0.3±0.03	0.32±0.04	0.33±0.06	0.42±0.04
eNTERFACE05_im	122	20	0.21±0.06	0.25±0.02	0.3±0.04	0.38±0.03	0.38±0.03	0.38±0.04	0.4±0.05	0.4±0.03	0.39±0.04	0.42±0.05	0.44±0.03
IEMOCAP	997	20	0.32±0.06	0.36±0.02	0.38±0.03	0.4±0.04	0.42±0.04	0.42±0.03	0.43±0.04	0.45±0.05	0.46±0.05	0.46±0.04	0.57±0.02
IEMOCAP_im	320	20	0.3±0.03	0.35±0.02	0.37±0.04	0.4±0.04	0.41±0.03	0.42±0.02	0.43±0.02	0.45±0.02	0.47±0.03	0.48±0.02	0.5±0.03

TABLE 3: The active learning strategy selects 5% in each round and the performance of the target classifier accuracy

Dataset	Test	Init	1	2	3	4	5	6	7	8	9	10	ALL
SAVEE	96	19	0.19±0.04	0.29±0.03	0.38±0.06	0.39±0.05	0.44±0.04	0.47±0.08	0.47±0.09	0.49±0.07	0.49±0.05	0.49±0.05	0.54±0.06
EMO-DB	107	21	0.34±0.02	0.47±0.07	0.49±0.08	0.56±0.11	0.59±0.06	0.6±0.04	0.63±0.05	0.66±0.04	0.68±0.04	0.7±0.03	0.72±0.03
CASIA	1200	240	0.54±0.03	0.59±0.02	0.61±0.01	0.63±0.01	0.63±0.01	0.63±0.01	0.63±0.01	0.63±0.01	0.62±0.02	0.64±0.01	0.62±0.01
CASIA_im	280	56	0.41±0.03	0.64±0.07	0.72±0.05	0.73±0.02	0.75±0.04	0.76±0.02	0.78±0.03	0.8±0.03	0.79±0.03	0.78±0.05	0.83±0.03
eNTERFACE05	251	50	0.23±0.02	0.29±0.02	0.31±0.03	0.34±0.02	0.38±0.03	0.38±0.02	0.4±0.04	0.41±0.02	0.42±0.01	0.42±0.01	0.47±0.04
eNTERFACE05_im	122	24	0.21±0.02	0.25±0.03	0.36±0.01	0.31±0.02	0.36±0.08	0.38±0.03	0.4±0.05	0.4±0.06	0.39±0.06	0.41±0.08	0.43±0.06
IEMOCAP	997	199	0.48±0.02	0.54±0.02	0.55±0.01	0.56±0.01	0.56±0.01	0.57±0.02	0.57±0.02	0.57±0.02	0.57±0.01	0.58±0.02	0.57±0.02
IEMOCAP_im	320	64	0.39±0.04	0.44±0.02	0.48±0.03	0.48±0.02	0.51±0.03	0.52±0.02	0.52±0.04	0.51±0.03	0.52±0.04	0.52±0.04	0.5±0.04

TABLE 5: The average accuracy performance of different strategy combinations

	C	U+C	R+C	D+C	U+R+D+C
SAVEE	0.305±0.03	**0.326±0.04**	0.295±0.05	0.296±0.03	0.291±0.03
EMO-DB	0.451±0.05	**0.476±0.05**	0.439±0.03	0.472±0.03	0.455±0.03
CASIA_im	**0.691±0.04**	0.684±0.02	0.68±0.03	0.68±0.03	0.666±0.04
eNTERFACE05_im	0.265±0.03	0.271±0.02	0.267±0.02	0.271±0.04	**0.286±0.03**
IEMOCAP_im	0.45±0.03	**0.465±0.05**	0.458±0.04	0.448±0.05	0.455±0.03
Aggregation	0.548±0.06	0.542±0.01	0.555±0.03	0.536±0.05	**0.564±0.02**
Blood	0.527±0.06	0.532±0.07	0.574±0.07	0.517±0.06	**0.584±0.06**
Diabetes	0.691±0.03	0.688±0.02	0.659±0.04	**0.694±0.02**	0.672±0.03
qsar-biodeg	0.79±0.03	0.796±0.02	0.791±0.03	0.802±0.03	**0.805±0.03**
Vote	0.923±0.02	0.923±0.01	**0.933±0.02**	0.924±0.02	0.932±0.02
Vowel	0.681±0.03	0.656±0.04	**0.687±0.03**	0.684±0.05	0.679±0.04
WBC	**0.941±0.02**	0.934±0.02	0.936±0.01	**0.941±0.02**	0.916±0.02
Thyroid	0.919±0.05	0.919±0.05	0.924±0.06	0.921±0.04	**0.928±0.04**

TABLE 6: The accuracy performance of different Ratio values in the average of first 5 queries

	R_0.5	R_0.6	R_0.7	R_0.8	R_0.9
SAVEE	0.317±0.06	0.313±0.04	0.299±0.06	0.311±0.04	**0.319±0.03**
EMO-DB	0.439±0.01	0.448±0.03	0.434±0.01	**0.46±0.05**	0.437±0.05
CASIA_im	0.589±0.03	0.653±0.02	0.68±0.03	**0.679±0.03**	0.679±0.04
eNTERFACE05_im	0.305±0.03	**0.306±0.06**	0.305±0.03	0.29±0.03	0.291±0.03
IEMOCAP_im	0.476±0.02	**0.479±0.02**	0.468±0.02	0.478±0.02	0.457±0.02
Aggregation	0.504±0.03	0.523±0.04	0.563±0.03	**0.571±0.03**	0.559±0.04
Blood	0.591±0.02	**0.599±0.03**	0.588±0.03	0.594±0.04	0.571±0.05
Diabetes	0.597±0.05	0.605±0.04	0.626±0.05	0.635±0.04	**0.641±0.05**
qsar-biodeg	0.808±0.02	0.807±0.02	**0.813±0.02**	0.799±0.02	0.8±0.01
Vote	0.902±0.04	0.907±0.04	0.91±0.03	0.916±0.02	**0.922±0.02**
Vowel	0.657±0.05	0.686±0.03	0.692±0.04	**0.706±0.06**	0.696±0.04
WBC	0.894±0.03	0.893±0.04	0.907±0.06	0.925±0.02	**0.929±0.02**
Thyroid	0.841±0.07	0.864±0.07	0.86±0.08	**0.884±0.04**	0.857±0.03

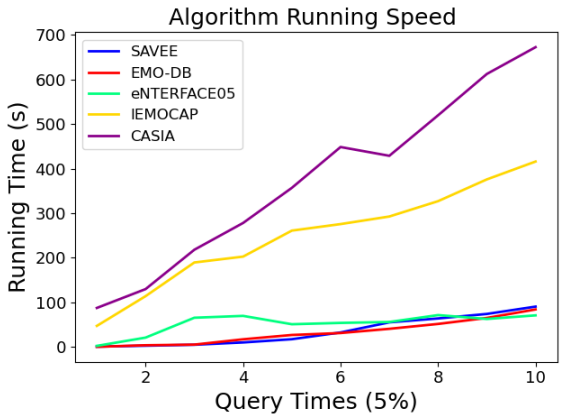


Fig. 5: The running speed of 10 queries on 5 speech emotional datasets

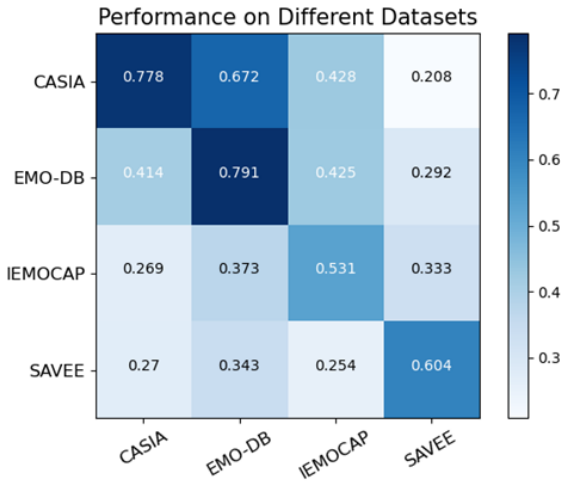


Fig. 6: The cross-prediction results on four datasets

running time of each query. The CASIA and IEMOCAP datasets are too large and consume more time, and other data sets are kept within 100s.

4.3.4 *Necessity of constructing speech emotion dataset*

The extracted features of voice collected in different scenarios have different feature spaces. In the application of the specific scene, it is very important and necessary to construct a new dataset. To verify our thinking, we selected 4 datasets, each of which has 4 common emotions, Anger, Happy, Neutral, and Sad. We divide the training set and the test set with the ratio of 8:2 and train the classifier on each dataset. Then we used the four classifiers to view the performance of the four datasets.

Figure 6 shows the cross-prediction results for four datasets. We observe that the classifier trained on its own dataset has the best performance, but performs poorly on other datasets, which means that the contextual speaking styles in different scenarios will affect the recognition accuracy. Therefore, in the field of SER, it is very necessary to construct new corpus under different application scenarios.

5 CONCLUSION AND FUTURE WORK

To solve the high cost of the speech emotion corpus construction, especially for the problem of difficulty in sampling small-class samples, we propose an integrated active learning strategy and designed a framework for the construction of speech emotion dataset. In comparison experiments with other active learning strategies, our proposed active learning method achieved the best performance for sampling on different datasets, especially for the selection of small-class samples, which is significantly better than other methods. In another actual dataset construction experiment, our method in the process of constructing the dataset, when the total number of labeled samples is less than 50%, the recognition accuracy of emotion classes reaches more than 90%.

The architecture of the active learning method we proposed is serial, which cannot achieve parallel operations. In the future, we will try a parallel sampling architecture to further improve the overall running speed of the method. At present, the sampling algorithm we propose combines the logistic regression classifier for sample selection. In the future, we will explore other classification models to improve the sampling efficiency. In addition, in the actual application of active learning, we will further explore the application of active learning methods in the construction of multimodal emotional corpus to adapt to the increasingly complex mass data.

ACKNOWLEDGMENTS

This research has been supported by JSPS KAKENHI Grant Number 19H04215.

REFERENCES

[1] F. Ren and Y. Bao, "A review on human-computer interaction and intelligent robots," *International Journal of Information Technology & Decision Making*, vol. 19, no. 01, pp. 5–47, 2020.

[2] C. Filippini, D. Perpetuini, D. Cardone, A. M. Chiarelli, and A. Merla, "Thermal infrared imaging-based affective computing and its application to facilitate human robot interaction: A review," *Applied Sciences*, vol. 10, no. 8, p. 2924, 2020.

[3] Z. Liu, X. Kang, S. Nishide, and F. Ren, "Vowel priority lip matching scheme and similarity evaluation model based on humanoid robot ren-xin," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2020.

[4] C.-H. Wu, Y.-M. Huang, and J.-P. Hwang, "Review of affective computing in education/learning: Trends and challenges," *British Journal of Educational Technology*, vol. 47, no. 6, pp. 1304–1323, 2016.

[5] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, and N. B. Hussin, "Affective computing in education: A systematic review and future research," *Computers & Education*, vol. 142, p. 103649, 2019.

[6] C. Zucco, B. Calabrese, and M. Cannataro, "Sentiment analysis and affective computing for depression monitoring," in *2017 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2017, pp. 1988–1995.

[7] S. Jayawardena, J. Epps, and E. Ambikairajah, "Evaluation measures for depression prediction and affective computing," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6610–6614.

[8] C. Sumathi, T. Santhanam, and M. Mahadevi, "Automatic facial expression analysis a survey," *International Journal of Computer Science and Engineering Survey*, vol. 3, no. 6, p. 47, 2012.

[9] J. A. Coan and J. J. Allen, "Frontal eeg asymmetry as a moderator and mediator of emotion," *Biological psychology*, vol. 67, no. 1-2, pp. 7–50, 2004.

[10] F. Ren, "Affective information processing and recognizing human emotion," *Electronic notes in theoretical computer science*, vol. 225, pp. 39–50, 2009.

[11] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[12] X. Wang, J. Gong, M. Hu, Y. Gu, and F. Ren, "Laun improved stargan for facial emotion recognition," *IEEE Access*, vol. 8, pp. 161 509–161 518, 2020.

[13] Z. Huang, F. Ren, M. Hu, and S. Chen, "Facial expression imitation method for humanoid robot based on smooth-constraint reversed mechanical model (srmm)," *IEEE Transactions on Human-Machine Systems*, vol. 50, no. 6, pp. 538–549, 2020.

[14] X. Wang, L. Kou, V. Sugumaran, X. Luo, and H. Zhang, "Emotion correlation mining through deep learning models on natural language text," *IEEE transactions on cybernetics*, 2020.

[15] J. Deng and F. Ren, "Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning," *IEEE Transactions on Affective Computing*, 2020.

[16] Y. Gu, H. Yan, M. Dong, M. Wang, X. Zhang, Z. Liu, and F. Ren, "Wione: One-shot learning for environment-robust device-free user authentication via commodity wi-fi in man–machine system," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 3, pp. 630–642, 2021.

[17] F. Ren, Y. Dong, and W. Wang, "Emotion recognition based on physiological signals using brain asymmetry index and echo state network," *Neural Computing and Applications*, vol. 31, no. 9, pp. 4491–4501, 2019.

[18] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[19] M. B. Mustafa, M. A. Yusoof, Z. M. Don, and M. Malekzadeh, "Speech emotion recognition research: an analysis of research focus," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 137–156, 2018.

[20] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[21] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 373–382, 2018.

[22] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[23] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[24] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition." in *INTERSPEECH*, 2019, pp. 2828–2832.

[25] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, and S. Narayanan, "Data augmentation using gans for speech emotion recognition." in *Interspeech*, 2019, pp. 171–175.

[26] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Cnn+ lstm architecture for speech emotion recognition with data augmentation," *arXiv preprint arXiv:1802.05630*, 2018.

[27] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition." *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, 2019.

[28] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 307–318, 2019.

[29] H. Luo and J. Han, "Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2047–2060, 2020.

[30] W. Zhang, P. Song, D. Chen, C. Sheng, and W. Zhang, "Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression," *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[31] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[32] J. D. Morris, "Observations: Sam: the self-assessment manikin; an efficient cross-cultural measurement of emotional response," *Journal of advertising research*, vol. 35, no. 6, pp. 63–68, 1995.

[33] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–8.

[34] B. Settles, "Active learning literature survey," 2009.

[35] R. Gilyazev and D. Y. Turdakov, "Active learning and crowdsourcing: A survey of optimization methods for data labeling," *Programming and Computer Software*, vol. 44, no. 6, pp. 476–491, 2018.

[36] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[37] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[38] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[39] W. Gerrod Parrott, "Ur-emotions and your emotions: Reconceptualizing basic emotion," *Emotion review*, vol. 2, no. 1, pp. 14–21, 2010.

[40] N. H. Frijda, S. Markam, K. Sato, and R. Wiers, "Emotions and emotion words," in *Everyday conceptions of emotion*. Springer, 1995, pp. 121–143.

[41] 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.

[42] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[43] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[44] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[45] L. Sun, S. Fu, and F. Wang, "Decision tree svm model with fisher feature selection for speech emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 1–14, 2019.

[46] M. J. Al Dujaili, A. Ebrahimi-Moghadam, and A. Fatlawi, "Speech emotion recognition based on svm and knn classifications fusion," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, p. 1259, 2021.

[47] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE access*, vol. 7, pp. 19 143–19 165, 2019.

[48] 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.

[49] M. Sajjad, S. Kwon *et al.*, "Clustering-based speech emotion recognition by incorporating learned features and deep bilstm," *IEEE Access*, vol. 8, pp. 79 861–79 875, 2020.

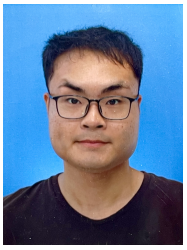
[50] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," *arXiv preprint arXiv:1904.03833*, 2019.

[51] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

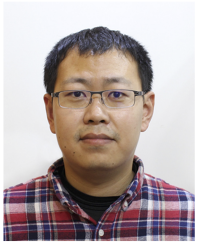
[52] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-y. Lee, "Audio albert: A lite bert for self-supervised

learning of audio representation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 344–350.
[53] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.
[54] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61 672–61 686, 2020.
[55] W. Wang, P. A. Watters, X. Cao, L. Shen, and B. Li, "Significance of phonological features in speech emotion recognition," *International Journal of Speech Technology*, vol. 23, no. 3, pp. 633–642, 2020.
[56] C. Zheng, C. Wang, and N. Jia, "An ensemble model for multi-level speech emotion recognition," *Applied Sciences*, vol. 10, no. 1, p. 205, 2019.
[57] W. Jiang, Z. Wang, J. S. Jin, X. Han, and C. Li, "Speech emotion recognition with heterogeneous feature unification of deep neural network," *Sensors*, vol. 19, no. 12, p. 2730, 2019.
[58] S. B. Kotsiantis, I. Zaharakis, P. Pintelas *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.
[59] C. Schröder and A. Niekler, "A survey of active learning for text classification using deep neural networks," *arXiv preprint arXiv:2008.07267*, 2020.
[60] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *ICML*, vol. 2, no. 4. Citeseer, 2000, p. 6.
[61] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 148–156.
[62] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 287–294.
[63] S. Ebert, M. Fritz, and B. Schiele, "Ralf: A reinforced active learning formulation for object class recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3626–3633.
[64] K. Konyushkova, R. Sznitman, and P. Fua, "Learning active learning from data," *Advances in neural information processing systems*, vol. 30, 2017.
[65] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," *Advances in neural information processing systems*, vol. 23, 2010.
[66] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, "A novel active learning method using svm for text classification," *International Journal of Automation and Computing*, vol. 15, no. 3, pp. 290–298, 2018.
[67] Y.-F. Yan, S.-J. Huang, S. Chen, M. Liao, and J. Xu, "Active learning with query generation for cost-effective text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 6583–6590.
[68] S. Xie, Z. Feng, Y. Chen, S. Sun, C. Ma, and M. Song, "Deal: Difficulty-aware active learning for semantic segmentation," in *Proceedings of the Asian Conference on Computer Vision*, 2020.
[69] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 399–407.
[70] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4604–4616, 2020.
[71] M. Abdelwahab and C. Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5160–5164.
[72] ——, "Active learning for speech emotion recognition using deep neural network," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.
[73] E. Vaaras, M. Airaksinen, and O. Räsänen, "Analysis of self-supervised learning and dimensionality reduction methods in clustering-based active learning for speech emotion recognition," *arXiv e-prints*, pp. arXiv–2206, 2022.
[74] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml
[75] S. Haq, P. J. Jackson, and J. Edge, "Speaker-dependent audio-visual emotion recognition." in *AVSP*, vol. 2009, 2009, pp. 53–58.
[76] J. T. F. L. M. Zhang and H. Jia, "Design of speech corpus for mandarin text to speech," in *The Blizzard Challenge 2008 workshop*, 2008.
[77] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.
[78] Y.-P. Tang, G.-X. Li, and S.-J. Huang, "ALiPy: Active learning in python," Nanjing University of Aeronautics and Astronautics, Tech. Rep., Jan. 2019, available as arXiv preprint https://arxiv.org/abs/1901.03802. [Online]. Available: https://github.com/NUAA-AL/ALiPy
[79] K. Mannepalli, P. N. Sastry, and M. Suman, "A novel adaptive fractional deep belief networks for speaker emotion recognition," *Alexandria Engineering Journal*, vol. 56, no. 4, pp. 485–497, 2017.

Fuji Ren received his Ph. D. degree in 1991 from the Faculty of Engineering, Hokkaido University, Japan. From 1991 to1994, he worked at CSK as a chief researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor of the Faculty of Engineering, Tokushima University. He is a Chair Professor of University of Electronic Science and Technology of China from 2022. His current research interests include Natural Language Processing, Artificial Intelligence, Affective Computing, Emotional Robot. He is the Academician of The Engineering Academy of Japan and EU Academy of Sciences. He is a senior member of IEEE, Editor-in-Chief of International Journal of Advanced Intelligence, a vice president of CAAI, and a Fellow of The Japan Federation of Engineering Societiesa Fellow of IEICE, a Fellow of CAAI. He is the President of International Advanced Information Institute, Japan.



Zheng Liu received the B.S. degree from the department of computer science in Donghua University, Shanghai, China, and M.S. degree from Tokushima University, Tokushima, Japan 770-8506. He is currently pursuing Ph.D. degree in Tokushima University. His research insterests include speech emotion recognition,machine learning, and humanoid robot.



Xin Kang received his Ph.D degree from Tokushima University, Tokushima, Japan, in 2013, his M.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2009, and his B.E. degree from Northeastern University, Shenyang, China, in 2006. He is currently an assistant professor in Tokushima University. His research interests include machine learning, text emotion prediction, and natural language generation. Faculty of Engineering, Tokushima University, 2-1, Minamijyousanjima-cho, Tokushima 770-8506 Japan