Active Learning for Speech Emotion Recognition Using Deep Neural Network

Mohammed Abdelwahab and Carlos Busso

Multimodal Signal Processing (MSP) laboratory, Department of Electrical and Computer Engineering

The University of Texas at Dallas, Richardson TX 75080, USA

Mohammed.Abdel-Wahab@utdallas.edu, busso@utdallas.edu

Abstract—**Deep neural networks** (DNNs) have consistently pushed the state-of-the-art performance in many fields, including speech emotion recognition. However, DNN-based solutions require vast amounts of labeled data for training. In speech emotion recognition, the cost and time needed to annotate data with emotional labels can be prohibitive. The available corpora normally have a few thousand recordings collected by a limited number of speakers. As a result, models trained on such corpora fail to generalize to samples from new domains. This study explores practical solutions to train DNNs for speech emotion recognition with limited resources by using *active learning* (AL). We assume that data without emotional labels from a new domain are available and we have resources to select a limited number of recordings to be annotated with emotional labels. We actively select samples using *greedy sampling* (GS) and uncertainty-based methods, evaluating the performance on regression problems where the goal is to predict scores for arousal and valence. We show that the use of active learning leads to competitive performance with limited training data.

Index Terms—**Speech emotion recognition, active learning, multitask autoencoder.**

I. INTRODUCTION

Speech emotion recognition (SER) is an important problem with many applications across fields. The advancements in *deep neural networks* (DNNs) have provided better architectures to build SER systems achieving state-of-the-art performance. A requirement to train the millions of parameters of these DNNs is a large train set with labeled data. This requirement is an important drawback when these models are intended to be used in new domain with limited labeled data. The process of data annotation can be expensive and time-consuming [1]. Several approaches have been proposed to minimize the amount of labeled data needed to train a classifier in a new domain, including semi-supervised learning [2]–[7], domain adaptation [8]–[11] and active learning [12]–[18]. This study focuses on active learning to select the most informative samples to improve DNN models. We consider cases where the resources to annotate samples in the new domain are limited.

Active learning has been widely used to iteratively select training samples that maximizes the model's performance. Several studies have used active learning for SER problems [12]–[15]. Most of these studies have considered algorithms focusing on classical machine learning approaches such as *support vector machines* (SVMs) [12]–[14], [16], [19], [20]. Recently, studies have proposed general active learning methods for DNNs [21], [22], but these approaches have not been

commonly used in SER problems. While there are no universal data acquisition functions that work well in all scenarios, many heuristic approaches have been proposed that were shown to perform well in some domains [23], [24]. The choice of which data acquisition function to use depends on the specific task. Some acquisition functions become computationally expensive as the dimensions in the feature space or the pool of available unlabeled data increase. These two problems are key factors in SER problems. The feature representation in SER often consists of a high dimensional vector [25]. Furthermore, speech can be easily collected with ubiquitous sensors with microphones, so we often have abundance of unlabeled data. It is important to investigate flexible acquisition functions that are appropriate for SER problem implemented with DNNs.

This study considers various uncertainty and greedy data acquisition functions used to select samples for training SER models using DNN. We evaluate the prediction of the emotional attributes arousal (calm versus active) and valence (negative versus positive) using regressors. We show that the use of active learning can lead to improvements in the performance of the system over a classifier trained with randomly selected samples, especially when the train set is limited (i.e., between 200 and 400 short speech recordings). With a few hundred labeled samples, we are able to achieve performances that approach the results obtained by training the DNN models with the labels of all training data. We obtain better performance when the models are pretrained using an autoencoder with unlabeled data. We also show that greedy sampling on the feature space outperforms random sampling at each sample size for both arousal and valence, where most of the differences in performance are statistically significant. The approach not only increases the performance, but also reduces the variability of the results when the models are trained multiple times with different initializations.

This paper is organized as follows. Section II briefly describes related work in active learning for emotion recognition. Section III explains the motivation of our work, describing the different acquisition functions used in the study. Section IV gives the details of our experimental evaluation. Section V presents our findings. Finally, Section VI summarizes the study, discussing potential areas of improvements.

II. RELATED WORK

One of the barriers in training SER systems is the lack of big corpora with emotional labeled data. Abdelwahab and Busso

[26] studied different neural network structures and factors affecting the performance achieved by SER systems built with DNNs, showing consistent performance improvements as more training data is seen by the model. For a new domain, the limited availability of labeled data can severely hinder the model's performance. Various approaches have been suggested to address this issue, including the aggregation of multiple corpora [27], [28], the use of a similar corpus combined with supervised adaption of the models to the new domain [8], [14], semi-supervised learning methods [5]–[7] and self-training approaches [3], [16]. This section reviews studies relying on active learning.

Zhang et al. [29] proposed an algorithm that uses the labelers' agreement to learn uncertainty models for each of the classes. They showed that by selecting samples from the minority class, the system was able to significantly reduces the amount of labeled data needed to maintain the performance for an unbalanced emotion binary classification problem. Zhang et al. [16] combined active learning with uncertainty sampling with multi-view learning to greatly minimize the amount of labeled data needed to reliably train a classifier. They trained binary SVMs for SER that achieved performance comparable to models trained on the whole train set, while only using a quarter of the labeled data. Zhang et al. [12] proposed a dynamic active learning approach. Instead of minimizing the amount of labeled data used by the models, they minimized the amount of annotations per sample. The proposed approach used medium uncertainty sampling to select samples for annotations. Instead of annotating each sample by a fixed number of raters, the selected samples are annotated until an agreement threshold is met. Models trained using this approach had comparable performance while significantly reducing the annotation cost. Zhang et al. [3] proposed an approach to minimize noise accumulation in self-training and co-training approaches. The proposed approach does not remove the selected samples from the pool of unlabeled data once they are added to the train set. Therefore, the assigned label to these samples can fluctuate from iteration to iteration as the models are updated. They kept track of the changes in the assigned labels, minimizing the noise accumulation in self-training and co-training approaches. This method, combined with multi-modal views, led to significant performance improvements.

Active learning has also been successfully used in facial expression analysis. Senechal et al. [17] proposed an active learning method to select positive examples for target *action units* (AUs). The approach increased the diversity of the training samples in terms of the number of individuals and number of expressions, which led to measurable improvements. Muhammad and Alhamid [18] combined active learning with face detection to select and annotate images from a huge pool of images from various social networks. The annotated images were used to build an ensemble of *extreme learning machines* (ELM) for emotion recognition.

A key problem in active learning is to define criteria to identify samples to be annotated. Wu et al. [20] proposed greedy sampling algorithms for regression problems that diversify

samples in the train set. They did not use deep learning, only linear regression models. The algorithms focus on maximizing the feature diversity of samples selected in the feature space, label space, or a combination of both. They show that greedy sampling approaches outperform random sampling. Wu and Huang [19] applied the greedy algorithms in multitask speech emotion regression.

III. METHODOLOGY

Deep neural networks often require big train sets to achieve good performance. Studies have shown the performance improvement in SER problems that can be achieved by just increasing the train set [26]. An important problem is how to achieve good performance even when the size of the train set is small. The process of annotating enough data is expensive and time-consuming. In this context, active learning approaches offer appealing solutions for models to reach their potential performance, while minimizing the amount of labelled data. Some data acquisition functions do not scale well with high dimensional features commonly used in training DNNs, especially in SER problems. While various acquisition functions or their variations have been shown to work well for DNNs in classification problems [23], [30], this is not necessary the same for regression based problems such as the prediction of valence, or arousal scores (e.g., emotional attributes are interval descriptors where regression is the common formulation). This study investigates the effectiveness of several data acquisition functions for regression models in SER problems.

A. Uncertainty (Variance) based Acquisition function

The first class of data acquisition function selects samples using uncertainty. The key idea is to select samples that the existing model is the most uncertain about (some studies have selected samples with medium uncertainty [12], but we did not evaluate this option).

Dropout: Gal et al. [21], [22] showed that dropout can be used as an approximate to Bayesian inference. By sampling the model predictions at various dropout configurations, we can represent the models' uncertainty without sacrificing computational complexity or performance. We referred to this model as *dropout*.

B. Greedy Sampling (GS) Based Acquisition Functions

The second class of data acquisition function is *greedy sampling* (GS), which selects samples by maximizing the diversity in the train set. It chooses samples with the maximum minimum distance to samples previously selected. This approach can be applied in the feature space, label space or a combination of both.

Feature space (GS_x): This approach searches for diversity in the feature space. We use the L_2 norm as the distance metric to estimate the difference between samples (Eq. 1, where x_i and x_j are the feature vector of two samples). By increasing the diversity in the feature space, the train set has representative samples from the target domain. While the original formulation considers the feature vectors [20], our

implementation projects the samples into the embedding of the autoencoder to reduce the search space (see Sec. IV-C).

$$d_x^{i,j} = \|x_i - x_j\|_2 \tag{1}$$

Label space (GS_y): This criterion searches for diversity on the label space (Eq. 2). We estimate the absolute difference between the predicted value of a sample ($\hat{y}_i$) and the true value of the labels of the samples already annotated (y_j). This criterion increases the diversity of the labels in the train set, so the classifier is built with representative samples spanning the possible values of the target attribute.

$$d_y^{i,j} = |\hat{y}_i - y_j| \tag{2}$$

Feature and label space (GS_{xy}): This approach combines the feature and label spaces to increase the diversity on both spaces. As proposed in Wu et al. [20], we use the multiplication of the respective metrics as the combination metric (Eq. 3).

$$d_{xy}^{i,j} = d_x^{i,j} * d_y^{i,j} \tag{3}$$

For the three criteria, we estimate the distances between unlabeled samples in the pool and samples in the train set. We select k samples with the maximum minimum distances, which are included in the train set. We update the models with the new train set, repeating the process until the desired number of samples is reached.

C. Random Sampling (RS) Based Acquisition Functions

As a baseline, we randomly select a given number of samples from the unlabeled set.

IV. EXPERIMENTAL SETTINGS

A. The MSP-Podcast Corpus

The analysis relies on the MSP-Podcast database [31], which is a collection of spontaneous audio recordings downloaded from audio-sharing websites. The dataset follows the ideas presented by Mariooryad et al. [32] to collect naturalistic emotional database. The recordings contain natural speech from different speakers, covering various topics, and under different recording conditions. The audio recordings are split into short turns with duration between 2.75 and 11 seconds using a commercial diariazation algorithm. Shorter segments are discarded and longer segments are split. The segments are processed to ensure good speech quality, removing segments with low *signal-to-noise ratio* (SNR), phone-quality, overlapping speech or music. Emotional models trained on existing emotional corpora are used to retrieve emotional segments from the pool of unlabeled recordings, which are then annotated using a crowdsourcing platform. The protocol corresponds to a modified version of the framework proposed by Burmania et al. [33], which track in real time the performance of the evaluators.

Each speech segment is annotated by at least five annotators into emotional categories, and emotional attributes (valence, arousal and dominance). The survey for emotional attributes was conducted with self-assessment manikins (SAMs) using a seven point Likert scale. The ground truth per speech segment is the average score across annotators. This study uses the version 1.1 of the corpus, which includes 22,630 labeled samples. The test set has 7,181 segments from 50 speakers (25 males, 25 females), the development set has 2,614 segments from 15 speakers (10 males, 5 females) and the train set has the remaining 12,830 segments. This study assumes that the unlabeled pool of speech segments corresponds to the train set (12,830 segments). We assume that the label of a segment becomes available after it is selected by a data acquisition function. More information about this corpus is provided by Lotfian and Busso [31].

B. Acoustic Features

The evaluation uses the standard Interspeech 2013 *computational paralinguistics challenge* (ComParE) [25] feature set, extracted with the OpenSMILE toolkit [34]. The feature set is generated by extracting 65 *low-level descriptors* (LLDs) using a sliding window over the speech segment. The low-level descriptors include various features such as pitch, jitter, energy, and *Mel-frequency cepstral coefficients* (MFCCs). After extracting the LLDs, a set of statistical functionals are calculated over them, resulting in *high-level descriptors* (HLD) features. The statistical functionals include functions such as mean, standard deviation, inter-quartile ranges, rising and falling slopes. This approach creates a 6,373 dimensional feature vector, regardless of the duration of the speech segment.

We separately normalize each feature to have zero mean and a unit standard deviation. The mean and the variance of the data are calculated considering only the values of the features that fall within the 2.5% and 97.5% quantiles to avoid outliers skewing these values. After normalization, we clip any normalized feature with absolute value greater than 3.1 times its standard deviation.

C. Network Structure

This study builds DNN-based regression models with limited data. To improve the learning of the models, this evaluation uses the multitask architecture shown in Figure 1, which includes an autoencoder. The encoder has an input layer with 6,373 nodes, followed by a hidden layer with 1,024 nodes and an embedding layer with 256 nodes. The embedding of the encoder is fed to an output layer that predicts the emotional attribute (primary task), and to the decoder that reconstructs the input. The decoder mirrors the structure of the encoder (auxiliary tasks). The network is trained to minimize both the reconstruction and regression loss:

$$\mathcal{L} = \lambda_1 MSE + \lambda_2 (1 - CCC) \tag{4}$$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \|x - \hat{x}\|^2 \tag{5}$$

$$CCC = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{6}$$
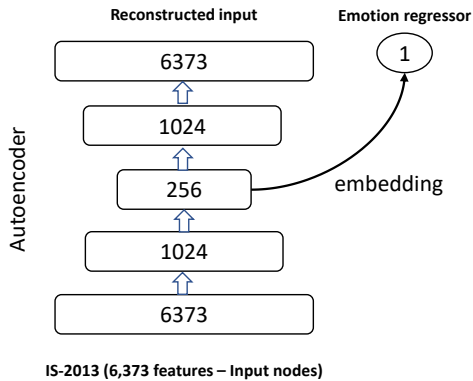
Fig. 1. Multitask autoencoder network used to evaluate active learning methods. The primary task is the emotion regression problem for either arousal or valence. The auxiliary task is an autoencoder which can be trained with unlabeled data.

where MSE is the mean squared error, and CCC is the *concordance correlation coefficient* (CCC). $\mu_{\hat{y}}$ and μ_y are the means of the predicted and actual values, and $\sigma_{\hat{y}}^2$ and σ_y^2 are their corresponding variances. ρ is the Pearson correlation coefficient between the predicted and actual values. The key advantage of this architecture is that data without emotional label can be used to train the autoencoder, pretraining the models.

The analysis in this paper considers cases with a limited train set. We consider 50, 100, 200, 400, 800 and 1200 samples. The active learning methods are implemented with the following steps. First, we consider all the unlabeled samples from the train set to train the autoencoder considering only the reconstruction loss. Second, we select 50 samples using greedy sampling in the feature space. The samples from the unlabeled pool are projected into the embedding of the autoencoder, where the distances between samples are estimated. These 50 samples are used as the starting point for all the models, with the exception of the baseline model (random sampling). We follow this approach since greedy sampling methods on the label space and the dropout method require a pretrained regressor, while the greedy sampling method on the feature space can be implemented with just the embedding of the autoencoder. Then, we incrementally add k new samples using the alternative data acquisition functions. The SER models are updated using the regression and reconstruction losses with all the training samples available at this point, including the k new samples (one epoch, batch size equals to $\min(\#samples, 256)$). Then, we select k more samples until reaching 100 samples. While we implement this approach with $k = 10$, we study the sensitivity of the models to the parameter k in Section V-C. Once we reach 100 samples, we evaluate the performance of the system. We start with the pretrained autoencoder model trained on the unlabeled data using only the reconstruction loss. Then, we train the models with the 100 samples using the regression and reconstruction losses. We train the model for 50 epochs, with a batch size of 64. By
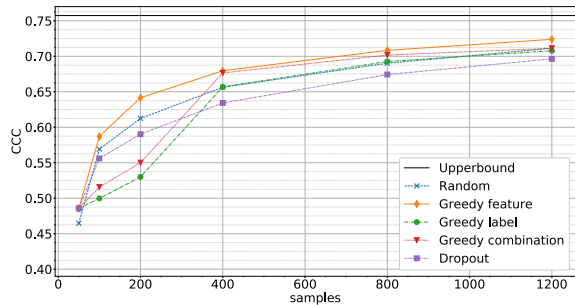
(a) Arousal with multitask autoencoder
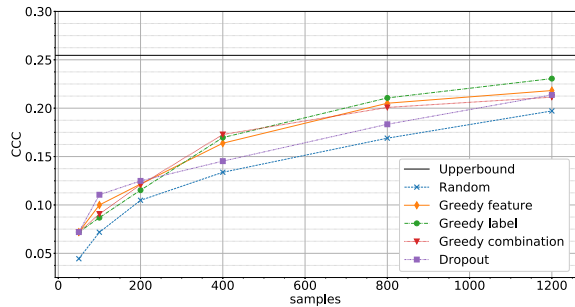
(b) Valence with multitask autoencoder

Fig. 2. Performance of the multitask autoencoder as a function of the number of samples in the train set. The figures compare the data acquisition functions considered in this study. Upperbound is the result using all training data.

retraining the models again instead of adapting the models used to select the last k samples, we avoid improvements solely due to longer training. This process is repeated until reaching 200 samples, starting with the model trained with 100 samples. We continue this process for 400, 800 and 1200 samples. To avoid adjusting hyper parameters, we arbitrarily set $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$ (Eq. 4). We use Adam optimizer with a learning rate of $8e - 5$.

V. RESULTS

We evaluate each setting 20 times with different initialization, reporting the average CCC.

A. Performance as a Function of Number of Selected Samples

Figure 2 shows the average CCC achieved by the active learning models as we increase the size of the train set. Figures 2(a) and 2(b) show the performance for arousal and valence, when the models are initialized using the weights of the pretrained autoencoder model, as explained in Section IV-C. The figures also show the upper bound performance (solid lines), which corresponds to the within corpus performance achieved by the multitask autoencoder architecture when we use the entire train set (12,830 speech segments).

As the number of training samples increases, Figure 2 shows that the performances achieved with active learning methods get closer to the within corpus performance. Notice that even at 1,200 samples, these methods use less than 10% of the train set. In general, we observe that greedy sampling methods

	Arousal [CCC]				Valence [CCC]			
# Samples	100	200	400	800	100	200	400	800
RS	.57	.61	.66	.69	.07	.10	.13	.17
GS_x	.58	.64*	.68*	.71*	.10*	.12	.16*	.21*
GS_y	.50	.53	.66	.69	.09	.12	.17*	.21*
GS_{xy}	.52	.55	.68	.70	.09	.12	.17*	.20*
dropout	.56	.59	.63	.67	.11*	.12	.15	.18

lead to better performance than the dropout method. We also observe that the differences in performances observed across active learning methods reduces as we add more samples. For valence, the performance for random sampling is always lower than the ones obtained with active learning methods. For arousal, the performance of greedy sampling on the feature space is always greater than random sampling. Methods that use greedy sampling on the label space (GS_y and GS_{xy}) perform worse than random sampling when the size of the train set is limited. However, their performances recover as we add more samples (e.g., 400 samples). Greedy sampling on the label space relies on predictions made by the DNN models. Therefore, this method is less reliable when the number of available samples is limited.

Table I shows the performance achieved with active learning methods for 100, 200, 400 and 800 samples. The network corresponds to the multitask autoencoder architecture. The asterisk in the table indicates statistically significant improvements over random sampling (one tailed Fisher Z-transformation test) asserting significance at p-value < 0.05. The improvement in performance for the greedy sampling method on the feature space is always significantly better than the performance obtained with random sampling, with the exception of arousal for 100 samples and valence with 200 samples. One drawback of GS_x is the computational cost of calculating the distances between samples, especially if the dimension of the feature space is high. We mitigate this cost by calculating the distance by projecting the samples into the lower dimension embedding of the proposed architecture. Dropout based sampling statistically outperforms random sampling only for valence when the sample size is small. More work is needed to understand why this method does not perform well in this problem, as expected from previous results in other domains [30], [35].

B. Role of Autoencoder

We also evaluate the DNN models trained without pretraining with the autoencoder. Instead of using the autoencoder, we use random weight initialization training the models using only the regression loss. Figures 3(a) and 3(b) show the performance for arousal and valence for this model. As expected, Figures 2 and 3 show that using pretrained weights for initialization improves the performance of the models. The performance are consistently lower when we use random
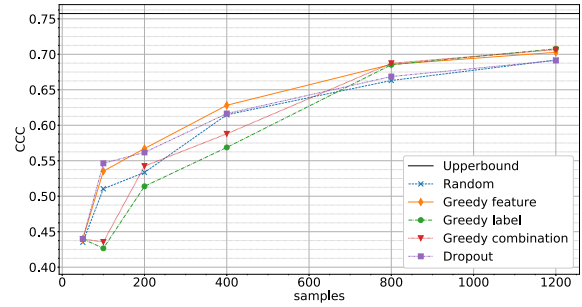
(a) Arousal without autoencoder
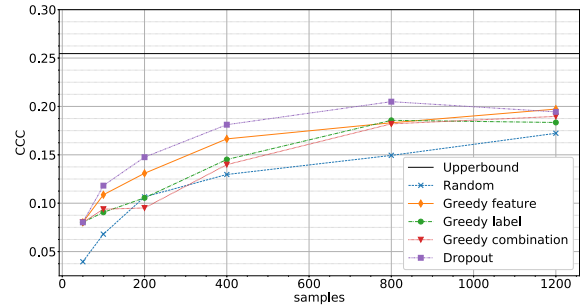
(b) Valence without autoencoder

Fig. 3. Performance of the models trained from scratch, using random weight initialization as a function of the number of samples in the train set. The CCC values are lower than using the multitask autoencoder framework (Fig. 2).

weight initialization, especially when the number of samples is limited.

Figure 3 shows similar trends as the ones observed in Figure 2. Greedy sampling on the feature space consistently outperforms random sampling at each size of the train set for arousal and valence. For arousal, greedy sampling on the label space and greedy sampling combination perform worse than the baseline for small size of the train set. Dropout based uncertainty sampling achieves the best performance for valence. For arousal, it performs well when the sample size is small, but as the sample size increases, it performance becomes similar to the random sampling performance. It is unclear why dropout based uncertainty sampling work well only when the models are not pretrained.

C. Sensitivity to k

We evaluate the sensitivity of the greedy sampling acquisition functions to the number of samples selected before updating the models. We set $k = 1$ in this evaluation, where the model is updated at each iteration. We only consider the greedy sampling approaches in this evaluation for computational reasons. The dropout method requires intensive resources so it is not efficient to implement our approach with $k = 1$.

Figure 4 shows the results for arousal and valence, using the multitask auto encoder framework with $k = 1$ and $k = 10$. We show the results for 100, 200 and 400 samples. We estimate the results for 800 and 1,200 samples but we did not find statistical
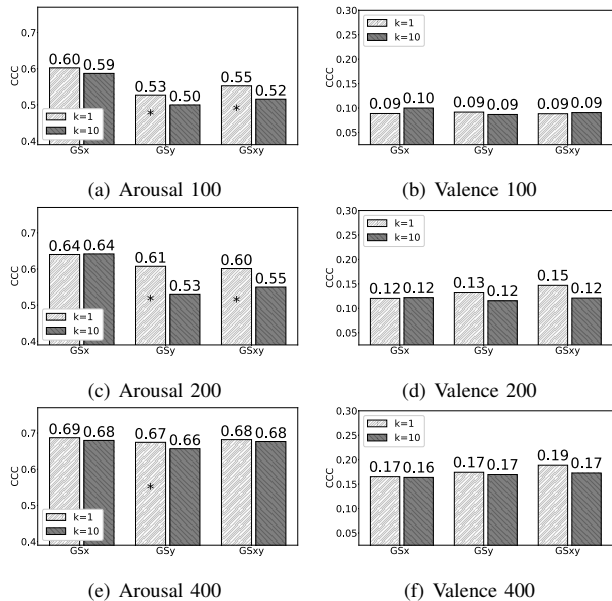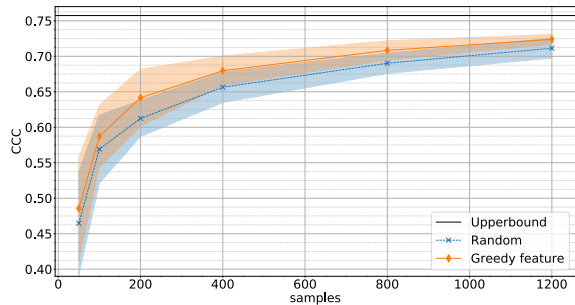
(a) Arousal 100

(b) Valence 100

(c) Arousal 200

(d) Valence 200

(e) Arousal 400

(f) Valence 400

Fig. 4. Average CCC values obtained with the multitask autoencoder framework using $k=1$ and $k=10$. The parameter k corresponds to the number of sentences selected at each iteration before updating the DNN models. The asterisk in the bars indicates that the differences are statistically significant.

differences between models trained with $k = 1$ and $k = 10$ when the size of the train set increases. The asterisks in the bars indicate statistically significant improvements obtained by using $k = 1$ (one tailed Fisher Z-transformation test with p-value<0.05). The results for arousal show that both GS_y and GS_{xy} perform significantly better when samples are selected one at a time. For valence, we observed some improvements, however the differences are not statistically significant. We also observe that greedy sampling on the feature space (GS_x) is insensitive to k. By using a larger value for k, we are able to further reduce the computational cost of GS_x.

D. Consistency of the Results

The last part of the analysis evaluates the consistency of the models' performance. Deep learning models are often sensitive to the initialization of the parameters, so studying the variations in the results is important, especially with a limited train set. For better visualization, we only consider the greedy sampling on the feature space, which is one of the best models considered in this study. As we mentioned, we implement the approach 20 times starting with different initializations. Figure 5 shows the mean and standard deviation of the CCC values achieved by greedy sampling on the feature space and random sampling. The lines correspond to the mean values, and the spreads of the curves correspond to the standard deviation.

The figures show that the standard deviation of the CCC values achieved by the greedy sampling method decreases faster as the sampling size increases, compared to the ones using random sampling. We not only achieve better performance with active learning, but also the results are more consistent across different initializations.

(a) Arousal with multitask autoencoder
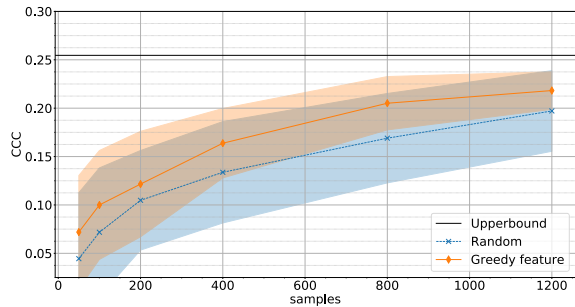
(b) Valence with multitask autoencoder

Fig. 5. Analysis of the variance of the multitask autoencoder framework across different initializations. The figures shows the mean and standard deviation of the SER models using greedy sampling on the feature space (GS_x) and random sampling. The lines correspond to the mean values and the spreads of the curves correspond to the standard deviation.

VI. CONCLUSIONS

This study demonstrated that active learning can significantly reduce the amount of labeled data needed to achieve a reasonable performance in a new domain. Compared to random sampling, we showed that greedy sampling approaches remain a viable option for speech emotion regression problems achieving not only higher performance, but also lower variance. Greedy sampling on the label space are less effective when the train set is limited, since the poorly trained DNNs models are used to select samples. As we introduced more data, the differences in performance across data acquisition functions reduce, indicating that the selection of data acquisition function is a more important decision when we can only annotate few samples to train the models.

As part of our future work, we want to consider different acquisition functions with smaller computational cost. We also want to include functions that consider both aleatoric and epistemic uncertainty. With this approach, we can capture the uncertainty present in difficult samples as well as the uncertainty caused by the lack of data. If we can assess the difficulty of the samples without the need of extra annotations, we can also combine active learning with curriculum learning [36]. This approach will allow the models to select easier samples early in the process, leading to models with better generalization.

REFERENCES

[1] D. Braha, *Data Mining for Design and Manufacturing: Methods and Applications*. Norwell, MA, USA: Kluwer Academic Publishers, October 2001.

[2] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *ACM international conference on Multimedia (MM 2014)*, Orlando, FL, USA, November 2014, pp. 801–804.

[3] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5185–5189.

[4] I. Cohen, N. Sebe, F. Cozman, and T. Huang, "Semi-supervised learning for facial expression recognition," in *ACM SIGMM international workshop on Multimedia information retrieval (MIR 2003)*, Berkeley, CA, USA, November 2003, pp. 17–22.

[5] M. Schels, M. Kächele, M. Glodek, D. Hrabal, S. Walter, and F. Schwenker, "Using unlabeled data to improve classification of emotional states in human computer interaction," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 5–16, March 2014.

[6] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *ArXiv e-prints (arXiv:1905.02921)*, pp. 1–13, May 2019.

[7] ——, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[8] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 5058–5062.

[9] ——, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[10] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2369–2372.

[11] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain adversarial training for accented speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 4854–4858.

[12] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *International conference on Multimodal interaction (ICMI 2015)*, Seattle, WA, USA, November 2015, pp. 275–278.

[13] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5000–5004.

[14] ——, "Incremental adaptation using active learning for acoustic emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5160–5164.

[15] D. Le and E. Mower Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2015)*, Xi'an, China, September 2015, pp. 146–152.

[16] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, January 2015.

[17] T. Senechal, D. McDuff, and R. el Kaliouby, "Facial action unit detection using active learning and an efficient non-linear kernel approximation," in *IEEE International Conference on Computer Vision Workshop (ICCVW 2015)*, Santiago, Chile, December 2015, pp. 10–18.

[18] G. Muhammad and M. Alhamid, "User emotion recognition from a larger pool of social network data using active learning," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10 881–10 892, April 2017.

[19] D. Wu and J. Huang, "Affect estimation in 3D space using multi-task active learning for regression," *IEEE Transactions on Affective Computing*, pp. 1–12, 2019.

[20] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Information Sciences*, vol. 474, pp. 90–105, February 2019.

[21] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, Cambridge, UK, October 2016.

[22] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML 2016)*, New York, NY, USA, June 2016, pp. 1050–1059.

[23] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Long Island, NY, USA: Morgan & Claypool Publishers, July 2012.

[24] S. Dasgupta, "Analysis of a greedy active learning strategy," in *International Conference on Neural Information (NIPS 2004)*, Vancouver, BC, Canada, December 2004, pp. 337–344.

[25] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[26] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088.

[27] Z. Zhang, F. Weninger, M. Wollmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*, Waikoloa, HI, USA, December 2011, pp. 523–528.

[28] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *Interspeech 2011*, Florence, Italy, August 2011, pp. 3285–3288.

[29] Z. Zhang, J. Deng, E. Marchi, and B. Schuller, "Active learning by label uncertainty for acoustic emotion recognition," in *Interspeech 2013*, Lyon, France, August 2013, pp. 2856–2860.

[30] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *International Conference on Machine Learning (ICML 2017)*, vol. 70, Sydney, NSW, Australia, August 2017, pp. 1183–1192.

[31] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. To appear, 2019.

[32] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[33] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[34] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[35] E. Tsymbalov, M. Panov, and A. Shapeev, "Dropout-based active learning for regression," in *Analysis of Images, Social Networks and Texts (AIST 2018)*, ser. Lecture Notes in Computer Science, W. van der Aalst, V. Batagelj, G. Glavaš, D. Ignatov, M. Khachaym, S. Kuznetsov, O. Koltsova, I. Lomazova, N. Loukachevitch, A. Napoli, A. Panchenko, P. Pardalos, M. Pelillo, and A. Savchenko, Eds. Moscow, Russia: Springer, Cham, July 2018, vol. 11179, pp. 247–258.

[36] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 815–826, April 2019.