INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Active Learning Methods for Low Resource End-To-End Speech Recognition

Karan Malhotra, Shubham Bansal, Sriram Ganapathy

Learning and Extraction of Acoustic Patterns (LEAP) lab, Indian Institute of Science, Bangalore.
{karanm,shubhamb,sriramg}@iisc.ac.in

Abstract

Recently developed end-to-end (E2E) automatic speech recognition (ASR) systems demand abundance of transcribed speech data, there are several scenarios where the labeling of speech data is cumbersome and expensive. For a fixed annotation cost, active learning for speech recognition allows to efficiently train the ASR model. In this work, we advance the most common approach for active learning methods which relies on uncertainty sampling technique. In particular, we explore the use of path probability of the decoded sequence as a confidence measure and select the samples with the least confidence for active learning. In order to reduce the sampling bias in active learning, we propose a regularized uncertainty sampling approach that incorporates an i-vector diversity measure. Thus, the active learning in the proposed framework uses a joint score of uncertainty and i-vector diversity. The benefits of the proposed approach are illustrated for an E2E ASR task performed on CSJ and Librispeech datasets. In these experiments, we show that the proposed approach yields considerable improvements over the baseline model using random sampling.

Index Terms: speech recognition, end-to-end models, active learning, uncertainty sampling, diversity measures.

1. Introduction

The speech recognition paradigm has witnessed tremendous advancements in the recent years owing to the success of deep neural network (DNN) models and sequence models like recurrent neural networks. The initial approaches using hidden Markov modeling (HMM) have been modified with the use of hybrid DNN-HMM models [1]. In the last few years, the E2E neural models which are devoid of HMM based modeling, have been explored for speech recognition [2, 3]. In E2E models, an encoder-decoder neural network is typically trained which converts the sequence of acoustic features directly to the character/word outputs. The recent experiments using end-to-end attention based models [4, 5] have shown comparable performance with the hybrid systems on a range of datasets [6]. Further, the E2E systems are significantly lower in computational complexity compared to traditional hybrid pipeline. The E2E models are attractive for quickly developing speech recognition systems in new languages owing to the sequence-to-sequence modeling which is devoid of any subword or phoneme level transcriptions/alignments. However, the E2E models for speech recognition systems require a large amount of transcribed speech data [7].

Many E2E models developed for speech recognition have conveniently focused on English language where there is an easy access to large amounts of speech data and the availability of orthographic transcriptions. However, there are a number of diverse languages which are low-resource where the collection of speech data is cumbersome and/or the labeling resources are expensive and time consuming. While there is some leverage to record large amounts of speech data with the aid of mobile technology, the labeling difficulties continue to hinder the development of speech technologies for many under-resourced languages. It was reported that labeling of audio can take about 10 times longer than the actual audio duration [8]. In this paper, we address the problem of efficiently utilizing the labeling resources for improving the performance of E2E ASR systems.

Active learning is a branch of machine learning that allows the model itself to choose the data it wants to learn from [9]. Thus, active learning aims at reducing the number of labeled samples required to achieve the desired performance for a specific task. The most common method of query for active learning is the one based on uncertainty sampling [10]. In this technique, the model queries new data labels for which the current model is least certain about its prediction.

In this paper, we explore the application of active learning for E2E ASR models. Unlike the conventional ASR systems, where the acoustic and language models can be separately improved using active learning, the implementation of active learning is more challenging with E2E models. We explore the use of uncertainty sampling [10] which is adapted to the E2E model by computing the uncertainty over sequences in the decoding graph. Also, the generalization performance of ASR systems is influenced by the diversity of the training speakers. Specifically, sampling too many training examples from the same speaker will create a speaker bias in the training data that may degrade the ASR performance on unknown test speakers. As the uncertainty sampling is prone to generate biased samples (potentially sampling too many examples from the same speaker) for ASR training, we propose a regularized uncertainty sampling which incorporates i-vector [11] features based diversity in the confidence based measures of uncertainty sampling. With ASR experiments on Librispeech and CSJ datasets, we show that the proposed approach is more effective than the baseline model using random sampling.

The rest of paper is organized as follows. Sec. 2 highlights the related prior work. In Sec. 3, we discuss the application of uncertainty sampling for E2E ASR systems. Sec. 4 describes the proposed approach to active learning which incorporates diversity in confidence measures of uncertainty sampling. We report the experiments and results on ASR tasks in Sec. 5. In Sec. 6, we conclude with the summary of the work.

2. Related Prior Work

The earliest work in the direction of E2E system for speech recognition used the connectionist temporal classification (CTC) loss function [12]. The E2E systems based on encoder decoder models with attention were inspired by similar approaches in machine translation [5]. Recently, Watanabe et al. [7] proposed a hybrid CTC/Attention structure where the E2E system is trained jointly with the CTC loss and attention based encoder-decoder loss. In this paper, we use the hybrid approach to E2E system as this was found to provide the best

http://dx.doi.org/10.21437/Interspeech.2019-2316

ASR performance in the tasks considered.

The active learning principle has been successfully applied for many natural language processing tasks like information extraction and part-of-speech tagging (a survey of algorithms can be found in [13]). For traditional speech recognition models using HMM based framework, informative data selection is explored in [14]. Efficient acoustic modeling using active learning has also been explored in the past [15, 16, 17]. For conventional ASRs, the confidence based active learning technique was previously proposed by Tur et al. [18], where the confidence measure was calculated using the utterance's confidence scores through exploitation of the lattice output of speech recognizer. In another approach [19], the authors proposed an entropy based technique which aims to maximize the expected value of global entropy reduction over unlabeled dataset and Kuo et al. [20] explored the selection of the samples based on the criterion of minimizing the expected error rate. Recently, [21] gave an approach to select samples based on expected gradient length for CTC based neural network models.

It had also been observed previously that the uncertainty sampling suffers from the problem of sampling bias [22] in which querying just based on the confidence scores leads to the sampling of instances which may not generalize well to the overall data distribution. Klaus [23] gave an approach to incorporate diversity in active learning algorithms while sampling of data for SVMs and in [24] both classifier uncertainty and sample diversity has been incorporated in a single convex optimization problem for the case of batch mode active learning.

In this paper, we observe that the application of uncertainty sampling to ASR leads to generation of samples that have significant speaker bias in the labeled set, i.e. some of the speaker groups are sampled considerably more than the others. Thus, we propose the use of unsupervised i-vector feature [11] based diversity measures which are incorporated in the sampling strategy.

3. Active learning for End to End ASR

In active learning paradigm, the informativeness of new samples can be computed by measuring several types of confidence scores. The most popular one in ASR is the least confidence sampling technique [25] in which the sample with the least certainty (given by an underlying model) is considered as the most informative sample. In conventional ASR, the decoded sequence probability can be obtained using the lattice based decoding.

The E2E ASR converts the feature sequence $\mathbf{X}$ of a given speech sample to the corresponding character sequence. For E2E ASR, the joint decoding phase provides the path probabilities for top N paths (where N is the beam width) for a given feature sequence. Here, path refers to a character sequence for a given speech utterance. These path probabilities are first length normalized to convert all of them into the same scale.

$$P^{LN}(\mathbf{C}_i|\mathbf{X}) = P(\mathbf{C}_i|\mathbf{X})^{1/L_i} \quad \forall i = 1, 2.....N \tag{1}$$

$\mathbf{C}_i$ is i^{th} decoded path and L_i is the length of i^{th} path and $\mathbf{X}$ is the input speech feature sequence. Note that the logarithm of $P(\mathbf{C}_i|\mathbf{X})$ is extracted during E2E decoding and when divided by the length of the path (L_i) followed by application of exponential function, it results into length normalized probability (Eq 1).

Thus, the goal of uncertainty sampling for E2E ASR can be stated as the selection of the unlabeled samples whose most likely path has the lowest probability. The least confidence score $\alpha(\mathbf{X})$ is given as

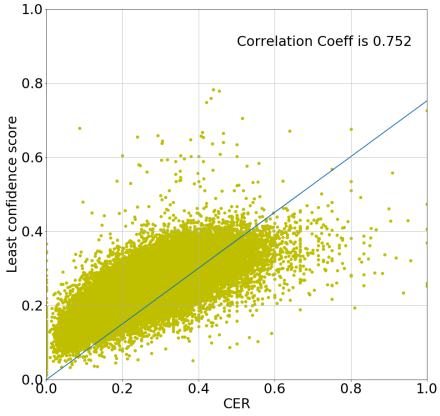

Figure 1: *Scatter Plot showing relationship between Least confidence score and character error rate (CER) for unlabeled data in case of Librispeech. The straight line fit with a slope equal to the correlation coefficient is also highlighted.*

$$\alpha(\mathbf{X}) = 1 - P^{LN}(\mathbf{C}^*|\mathbf{X}) \tag{2}$$

where $\mathbf{C}^*$ is the most likely decoded path for input $\mathbf{X}$. Given an unsupervised dataset, one can estimate the least confident samples by ranking the utterances based on the confidence measure (Eq. 2). Fig. 1 depicts the relationship between the least confidence score (Eq. 2) calculated using the model trained with some initial labeled data from Librispeech dataset (details of the dataset are provided in Sec. 5) and the character error rate (CER) of the samples. The CER is calculated using oracle ground truth transcription for same data samples. As seen in the fig. 1, the correlation coefficient is quite high (0.75) which highlights that the data samples on which the ASR system has low confidence (high value for least confidence score) also have high CER values. Hence, using the labels of these samples in the training data would make the E2E ASR system more robust.

The least confidence based sampling measures may suffer from the problem of sampling bias [22, 23]. Moreover, in our framework as the dataset consist of speech utterances from multiple speakers, the confidence based sampling method leads to selection of samples which is biased towards certain speakers. We hypothesize that speaker diversity is key to ASR training and this measure has to be incorporated in the active learning. A naive random sampling has the advantage of reduced speaker bias, provided underlying data is not biased.

4. Regularized Active Learning

We propose a novel approach of incorporating speaker diversity in confidence based measures for E2E ASRs and it is based on clustering of i-vectors [11]. I-vectors, as introduced in [11], consist of speaker statistical information and can be extracted in an unsupervised manner i.e, without using any speaker label information. The i-vectors are low dimensional feature representation for speech utterances which models both speaker and channel variabilities, and are defined using a factor analysis framework. Previously, i-vectors have been explored for speaker and channel adaptation in conventional HMM based speech recognition systems [26, 27]. In this work, we use the

Algorithm 1 Regularized Active Learning

Input: Unlabeled samples D^U, initially labeled samples D^L, total active learning dataset duration F in hours.
Parameters: Diversity controlling parameter λ, update step T, number of clusters K.

```
// Train a model using D^L
i = 0, D^A = {}
while F > 0 do
  if i%T == 0 then
    // Updating values of P̂_j
    N = Σ_{j=1}^{K} N_j
    P̂_j = N_j/N,  ∀j = 1, 2....K
    where N_j is the number of samples of current labeled
    data (D^L ∪ D^A) present in j^th cluster.
    β(X) = 1 − P̂_{φ(X)},  ∀ X ϵ D^U
  end if
  // Sampling
  X_i* = argmax_{X ϵ D^U} {λα(X) + (1 − λ)β(X)}
  D^A = D^A ∪ X_i*
  D^U = D^U \ X_i*
  F = F − dur(X_i*)
  i++
end while
return D^A
```

i-vector representations as a "speaker" embedding vector.

4.1. Mathematical Formulation

The proposed approach utilizes clustering of the i-vectors corresponding to full data (combined labeled and unlabeled data) and the idea behind clustering of i-vectors is to have utterances with similar statistical characteristics to be clustered together.

Let D^L denote a pre-labeled dataset (on which the seed ASR model has been pre-trained) and let D^U denote the pool of unlabeled dataset. Let D^A denote the subset of unlabeled dataset obtained through the active learning algorithm. Then, the final ASR model will be trained using $D^Q = D^L \cup D^A$.

The i-vector features from $D^L \cup D^U$ are clustered using K-means clustering into K clusters. Let $\phi(\mathbf{X})$ denote the random variable that indicates the cluster identity for an utterance $\mathbf{X}$ obtained using the corresponding i-vector. Note, that $\phi(\mathbf{X})$ can take values from 1 to K, where K is number of clusters. For the set of labeled data finally used for ASR (D^Q), the maximum likelihood estimates corresponding to parameters of probability mass function for $\phi(\mathbf{X})$ can be computed as,

$$P(\phi(\mathbf{X}) = j) = \hat{P}_j = N_j/N, \quad \forall j = 1, 2....K \tag{3}$$

where $N = \sum_{j=1}^{K} N_j$ and N_j is the number of samples of D^Q present in j^{th} cluster.

We propose a regularized version of active sampling which encourages sampling from clusters that have low probability $\hat{P}_j$. This will increase the "speaker" diversity in the final labeled data used for ASR training and the makes the data set D^Q more uniform in terms of sampling from the i-vector clusters. Let,

$$\beta(\mathbf{X}) = 1 - \hat{P}_{\phi(\mathbf{X})} \tag{4}$$

The proposed confidence measure for active learning is based on the convex combination of diversity score and the least confidence score, and is formulated as,

$$\gamma(\mathbf{X}) = \lambda\alpha(\mathbf{X}) + (1 - \lambda)\beta(\mathbf{X}), \tag{5}$$

Table 1: *Entropy (in bits) for the i-vector cluster distribution ($K = 64$) on the final ASR training dataset D^Q as a function of λ. Here, the size of D^L is 20 hours and F is 20 hours.*

Lambda(λ)	Entropy(CSJ)	Entropy(Librispeech)
1	5.963	5.954
0.3	5.975	5.984
0.2	5.980	5.990
0.1	5.988	5.996
0.01	5.999	5.998

where λ is a hyper-parameter and $\lambda \in [0, 1]$. The estimates of $\hat{P}_j$ depend on subset of data D^Q which is the output of the active learning while the active learning algorithm requires the estimates $\hat{P}_j$. Thus, we perform the querying of a single batch through active learning in sub-batches of size T and update the estimates of $\hat{P}_j$ at the end of every sub-batch sampled. The final algorithm is outlined below.

4.2. Algorithm

Algorithm 1 shows the pseudo code for the proposed regularized active learning. The input to the algorithm is the pool of pre-labeled and unlabeled datasets D^L and D^U respectively. Let F denote the desired final sampling batch size (i.e., F is the duration of the data in subset D^A in hours). An initial seed E2E ASR is trained using the pre-labeled dataset D^L. After the ASR training, the unlabeled dataset D^U is decoded using the joint decoding. Then, the least confidence score (Eq. 2) and the diversity score (Eq. 4) are computed for each utterance in the unlabeled dataset.

Initially, D^A is an empty set and the data samples are added iteratively during the active sampling. The parameter T denotes the window of samples after which the values of parameter $\hat{P}_j$, $\forall j = 1, 2....K$ are updated. At each iteration of active learning, a single sample is added to the set D^A which has the maximum value of regularized least confidence score (Eq (5)). After the sampling process is completed the set D^A is sent for labeling. The final E2E ASR is trained with using the $D^Q = D^L \cup D^A$ as the training data.

Table 1 reports the entropy values for the i-vector (speaker) cluster distribution ($K = 64$) $P(\phi(\mathbf{X}))$ obtained using the $D^L \cup D^A$ samples, where D^A is the output of the active learning. Here, we use $K = 64$ and the maximum value of entropy is 6.0 bits. Note that, the set D^A changes for different choice of regularization parameter λ. A decrease in the value of λ increases the weight given to diversity term in Eq. (5) and thereby improves the entropy of distribution $P(\phi(\mathbf{X}))$. In this work, we choose a threshold value on the entropy and select the highest λ value which gives the desired entropy. For the CSJ and Librispeech dataset, the chosen value of λ is 0.01 and 0.2 respectively.

5. Experiments and Results

Fot this work, we use the ESPNET (End to End Speech Processing) toolkit [6] which is a speech recognition toolkit, and it is built on deep learning frameworks like Chainer [28] and Pytorch [29]. This toolkit is a hybrid structure of the two main deep learning ASR frameworks, connectionist temporal classification (CTC) [2] and attention-based encoder-decoder [30]. Moreover, the data pre-processing and feature extraction steps follow the Kaldi ASR [31] toolkit. We have used 80 mel bins to

Table 2: *Performance (WER/CER in %) of different sampling techniques in case of Librispeech test-other for different amounts of active sampling sizes F.*

F **(hours)**	**WER**			**CER**		
	Random	***LC***	***RLC***	***Random***	***LC***	***RLC***
20	46.9	43.3	**42.3**	24.0	22.4	**21.6**
40	38.5	37.4	**37.1**	19.5	18.9	**18.7**
80	35.3	33.3	**32.7**	17.5	16.8	**16.1**
140	32.6	29.9	**29.8**	15.9	**14.4**	14.5

derive filter-bank features which are combined with pitch features to give 83 dimensional features.

For regularized least confidence sampling, 64 dimensional i-vectors are extracted and clustered into $K = 64$ clusters using K-means clustering algorithm with L2 normalized euclidean distance as the distance metric. The initial labeled data is D^L on which a seed model is trained and then the unlabeled data D^U is decoded with the help of the trained seed model. After the decoding phase of D^U, the least confidence method (LC) and the regularized least confidence method (RLC) is employed for active learning from the unlabeled data D^U. A baseline ASR system with the same amount of labeled data is also trained using a naive random sampling ($Random$) of the unlabeled set D^U. The naive random sampling provides the baseline performance for comparison with the proposed active learning methods. The total budget allocation for speech annotation is expressed in form of the number of hours of sampling (denoted as F). We experiment with different values of F for the Librispeech and CSJ datasets.

5.1. Librispeech

LibriSpeech [32] is a dataset of approximately 1000 hours of 16 kHz read English speech. In this paper, we use the 'other' pool of this dataset (subset of 1000 hours) which consists of 500 hours of training data termed as train-other, 5 hours of test and development data termed as test-other and dev-other respectively. The train-other set is split into two parts: 20 hours considered as initial labeled data D^L and 480 hours considered as the unlabeled dataset D^U.

The network architecture consists 8 layers Bi-LSTM encoder of 320 hidden units and a single layer LSTM decoder of 300 hidden units . Adadelta optimizer is chosen for optimizing the multi-objective loss function with 0.5 weight given to both CTC and attention losses. The learning rate is initially chosen as 1.0 . The learning rate is reset to 3.0 after 40 epochs of learning. We observed the resetting of the learning rate improves the convergence of training loss for the E2E ASR. During decoding, the CTC weight is 0.5 and beam width parameter is 20. For decoding of test-other, a character based RNN language model (trained on Librispeech external LM corpus) is used with 0.4 selected as LM weight for all the three sampling techniques.

Table 2 reports the word error rate (WER) and the character error rate (CER) results for the Librispeech experiments for different active sampling sizes (F). The random results also use the same size of training data as the active sampling methods (LC and RLC). It is seen that the active learning methods provide significant improvements over a random sampling approach. The confidence score used in this paper is effective even for large active sampling sizes of 140 hours which corresponds to about 30 % of the unlabeled pool set of 480 hours. The regularization of the active learning using the i-vector diversity further improves the ASR results. On the 20 hour sampling condition, the proposed active sampling approach achieves a relative improvement of 9.8% in WER and about 10 % in CER over the random sampling approach.

Table 3: *Performance (CER in %) of different sampling techniques in case of CSJ eval sets for different choice of active sampling sizes F.*

	F **(hours)**	**CER**		
		Random	***LC***	***RLC***
eval 1	20hrs	23.2	20.5	**20.3**
	40hrs	18.5	17.9	**17.3**
eval 2	20hrs	20.8	18.4	**18.2**
	40hrs	15.8	15.6	**15.0**
eval 3	20hrs	21.5	19.6	**19.1**
	40hrs	17.3	16.9	**16.2**

5.2. Corpus of Spontaneous Japanese (CSJ)

The CSJ[33] is a collection of monologue Japanese speech data including academic lectures and simulated presentations. The full data consists of 581 hours of training set and three types of evaluation sets of 5 hours in total. For the development data, 6 hours of data is sampled form the 581 hours of training data randomly. Separately, 250 hours of data is sampled randomly and split into two parts: 20 hours considered as labeled data i.e D^L and 230 hours considered as unlabeled data i.e. D^U. The network architecture consists 6 layers Bi-LSTM encoder of 320 hidden units and a single layer LSTM decoder of 320 hidden units . Adadelta optimizer is chosen for optimizing the multi-objective loss function with 0.5 weight given to both CTC and attention losses. The learning rate is chosen as 1.0. During decoding the CTC weight selected is 0.3 and beam width parameter is 20.

In Table 3, we report the CER results for the random sampling method and the active learning methods. Similar to the trends seen in the Librispeech corpus, we find that the active learning model (LC) improves significantly over the random sampling method. The regularization in active learning further improves the active learning performance. For example, the regularized least confidence provides around 12.5% relative improvement in CER over the random for the case of sampling 20 hours.

6. Conclusions

In this paper, we advance the active learning methods for application in end-to-end speech recognition. The active learning approach is based on uncertainty sampling using a confidence score. We use the path probability of the top decoded path as the confidence measure and sample the least confident data for active learning. We also propose a novel approach to regularize the confidence measure which takes the diversity of the i-vector features into account. The i-vector features broadly capture the speaker identity of the recording and the regularized active sampling encourages speaker diversity. The ASR experiments on Librispeech and CSJ datasets confirm the benefits of the active learning methods over a random sampling approach.

7. Acknowledgements

This work was partly funded by project grants from the Department of Science and Technology (DST) (ECR01341) and the Department of Atomic Energy (DAE) (DAE0205).

8. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[2] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[4] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[5] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[6] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[7] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[8] X. Zhu, J. Lafferty, and R. Rosenfeld, "Semi-supervised learning with graphs," Ph.D. dissertation, Carnegie Mellon University, language technologies institute, school of , 2005.

[9] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[10] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine learning proceedings 1994*. Elsevier, 1994, pp. 148–156.

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[13] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," *Swedish Inst. of Comput. Sci., SICS Tech. Rep. T2009:06,*, 2009.

[14] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *2007 IEEE workshop on automatic speech recognition & understanding (ASRU)*. IEEE, 2007, pp. 562–565.

[15] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE transactions on speech and audio processing*, vol. 13, no. 4, pp. 504–511, 2005.

[16] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka, "Speech modeling based on committee-based active learning," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4350–4353.

[17] M. Chellapriyadharshini, A. Toffy, V. Ramasubramanian *et al.*, "Semi-supervised and active-learning scenarios: Efficient acoustic model refinement for a low resource indian language," *arXiv preprint arXiv:1810.06635*, 2018.

[18] D. Hakkani-Tür, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2002, pp. IV–3904.

[19] B. Varadarajan, D. Yu, L. Deng, and A. Acero, "Maximizing global entropy reduction for active learning in speech recognition," 2009.

[20] H.-K. J. Kuo and V. Goel, "Active learning with minimum expected error for spoken language understanding," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[21] J. Huang, R. Child, V. Rao, H. Liu, S. Satheesh, and A. Coates, "Active learning for speech recognition: the power of gradients," *arXiv preprint arXiv:1612.03226*, 2016.

[22] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 208–215.

[23] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 59–66.

[24] E. Elhamifar, G. Sapiro, A. Yang, and S. Shankar Sasrty, "A convex optimization framework for active learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 209–216.

[25] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 1070–1079.

[26] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 55–59.

[27] S. Ganapathy, S. Thomas, D. Dimitriadis, and S. Rennie, "Investigating factor analysis features for deep neural networks in noisy speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[28] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, vol. 5, 2015, pp. 1–6.

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[30] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[31] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[33] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese." in *LREC*. Citeseer, 2000.