
Proceedings Paper:
Wu, C., Ng, R.W.M., Torralba, O.S. et al. (1 more author) (2017) Analysing acoustic model changes for active learning in automatic speech recognition. In: International Conference on Systems, Signals and Image Processing (IWSSIP). International Conference on Systems, Signals and Image Processing (IWSSIP), 22-24 May, 2017, Poznań, Poland. IEEE . ISBN 978-1-5090-6344-4

https://doi.org/10.1109/IWSSIP.2017.7965609

Analysing Acoustic Model Changes for Active Learning in Automatic Speech Recognition

Chenhao Wu*, Raymond W. M. Ng*, Oscar Saz Torralba*, Thomas Hain*
* Speech and Hearing Research Group, University of Sheffield, UK
cwu11@sheffield.ac.uk, wm.ng@sheffield.ac.uk, Oscar.SazTorralba@cirrus.com, t.hain@sheffield.ac.uk

***Abstract*—In active learning for Automatic Speech Recognition (ASR), a portion of data is automatically selected for manual transcription. The objective is to improve ASR performance with retrained acoustic models. The standard approaches are based on confidence of individual sentences. In this study, we look into an alternative view on transcript label quality, in which Gaussian Supervector Distance (GSD) is used as a criterion for data selection. GSD is a metric which quantifies how the model was changed during its adaptation. By using an automatic speech recognition transcript derived from an out-of-domain acoustic model, unsupervised adaptation was conducted and GSD was computed. The adapted model is then applied to an audio book transcription task. It is found that GSD provide hints for predicting data transcription quality. A preliminary attempt in active learning proves the effectiveness of GSD selection criterion over random selection, shedding light on its prospective use.**

***Index Terms*—Active learning; data selection; confidence measures; speaker adaptation**

I. Introduction

Using large amounts of acoustic data can help to adapt an Automatic Speech Recognition (ASR) system very precisely to a speaker, but in many cases ground truth transcriptions are not available for such data and one has to resort to using the ASR hypotheses as pseudo labels. In this case, the presence of errors in the transcripts degrades the performance of adaptation, resulting in non-optimal results overall [1]. One way to deal with errorful labels and transcripts is to perform data selection. If you only select data that is most likely correct, based on assessments by some prior model, you focus on less challenging data that reaffirms existing models, but limits the learning of new attributes. Active learning's objective is to select samples for manual transcription for active development of a model [2]. Active learning approaches have already been studied for the training of acoustic models for ASR. Confidence measuring and uncertainty sampling are the standard data selection methods in speech recognition scenarios [3], [4], further work also explored query-by-committee techniques [5]. In recent years there has been interest in selection strategies based on predicting how models change under new data (expected model change) [6].

In this work a need for speaker and domain adaptation is demonstrated through an ASR task with audio book data. An out-of-domain acoustic model generates errorful transcripts for adaptation. We then select a portion of the data using the new proposed method for manual labelling. The models can then be re-adapted, aiming for better performance [7]. Several techniques can be used to perform the data selection, including random selection and confidence score based selection. This paper proposes to use a new data selection method based on Gaussian Supervector Distance (GSD) between original and adapted models to perform active learning. Gaussian Mixture Model (GMM) parameters of a Hidden Markov Model (HMM) are adapted using Maximum A Posteriori (MAP) adaptation [8], based on Perceptual Linear Perceptron (PLP) features. GSD of MAP adapted models is usually exploited to express difference in speaker or language space – this work proposes new applications in active learning. The proposed GSD method for active learning improved performance over the use of ASR transcripts' baselines and random selection methods.

The rest of the paper is organised as follows: Section II reviews active learning approaches for ASR. Section III describes the GSD method based on expected model change. Sections IV and V present the experimental setup and list the baseline results. Section VI analyses model change and its relation to recognition errors. Section VII presents the results achieved using the GSD method. Finally, Section VIII gives the conclusions for this work.

II. Active learning in ASR

In machine learning literature, active learning is referred to learning through trial and error and query learning [9], [10]. It is an iterative approach combining special query data selection techniques with data correction [11], shown to improve classification performance [12], [13].

Active learning in acoustic modelling for ASR has been used to select the most representative and informative subset from a large set of untranscribed data in order to selectively transcribe that subset [10]. Active learning helps minimise manual transcription costs, by building acoustic models with significantly better accuracy than out of domain models [14].

Active learning has been studied in various areas related to speech technology including spoken language understanding [15] and speaker recognition [16]. Many active learning methods are partly or fully uncertainty based, mainly relying on confidence scores to select data. Applications include reducing need for human transcription [3], [4], [17], accent adaptation in ASR [18] and building emotional acoustic models [19].

Uncertainty-based methods of active learning are example-based, where confidence metrics are computed for each utterance example independently. Given the goal in selecting the most informative data subset for further training, we took a

different approach. Take the typical tied-state triphone model in ASR as an example, a model comprises of thousands of states, each of which can behave differently. In theory one must look for "poorly behaved" model components, determine model behaviours, anchor said behaviours to a data subset and select them for further training. In this paper, we show a first attempt to analyse and interpret model behaviour using model distance between original (out-of-domain) and (unsupervised) adapted models. We used this knowledge to run a trial experiment of speech recognition with audio book data and showed the potential usage of model distance in active learning.

III. Model change as a predictor to transcript quality

Model change is a metric quantifying how the model was changed during adaptation. Speaker recognition is an example of the use of model change. A Gaussian mixture model is trained on the background population. This model, commonly referred to as the Universal Background Model (UBM), is then adapted to individual speakers. Model change between the UBM and the adapted model is analysed, the underlying assumption being that different "principal components" of model change can be correlated with speaker identity, by techniques such as factor analysis and i–vector [20].

In ASR, an acoustic model comprises of a set of units which normally correspond to phonemes or some sensible partitions in the acoustic space. A baseline model is analogous to the UBM. It is trained on abundant out-of-domain data but with mismatched speakers, channels and domains. Model adaptation takes the in-domain data and tunes the model parameters. In this study, we attempt to correlate model adaptation direction, in terms of the model distance between the baseline and adapted model, with transcript quality. As a preliminary study, we take Hidden Markov Models (HMM) for context-dependent phonemes as the exemplar model. The model is adapted by updating the mean statistics using Maximum A Posteriori (MAP) criterion. In theory, this method should be applicable to other acoustic model architectures.

Let s_i denote a clustered GMM–HMM state, then let $\Lambda(s_i)$ denote the supervector formed by concatenation of N Gaussian mean vectors associated with the state, $\Lambda(s_i) = [\mu_1 ... \mu_n ... \mu_N]$. We then define the state-level GSD between the two states in the baseline and adapted models as the Euclidean distance between supervectors of the models. Let Λ_X denote the original model's supervector. Without assuming any knowledge on oracle data, adaptation can use an ASR transcription obtained by decoding with the baseline model. A supervector Λ_X^{ASR} is composed. GSD for a state s_i is then expressed as,

$$GSD(\Lambda(s_i), \Lambda^{ASR}(s_i)) = \sqrt{(\Lambda(s_i) - \Lambda^{ASR}(s_i))^2} \quad (1)$$

GSD is a state-level metric on model distance. A segment-GSD (sGSD) is computed to summarise the model distance in one segment. This is done by iterating through M states, weighting each state-level GSD with its duration in a segment, then taking the average:

$$sGSD(\Lambda, \Lambda^{ASR}) = \frac{1}{T} \sum_{i=1}^{M} l_i \times GSD(\Lambda(s_i), \Lambda^{ASR}(s_i)) \quad (2)$$

l_i is the duration of state i in a segment derived from the forced alignment of the ASR transcript. $T = \sum_{i=1}^{M} l_i$ is the segment duration.

The sGSD indicates an expected model distance. There are several observations and assumptions on the relationship among GSD, sGSD and ASR transcript quality. First, the sGSD value of a segment is correlated with the GSD values of its constitute phonemes (according to Eq. (2)). Second, by using sGSD as a grouping criterion to partition the training set, we can then derive different training subsets where particular types of GSD states are relatively prominent. Meanwhile, it is expected that phoneme (thus state) distribution within each segment will be comparable. Because linguistic constraints bind phonemes together in different segments in a similar way (i.e. phonotactic constraints), a segment with low sGSD will also contain states with high GSD values, but the ratio of high GSD states will be relatively lower. As such, a training subset with low sGSD is still representative for the full phonetic space and error comparison across different sGSD sets are still valid.

If a certain relationship between sGSD and segment error can be established, we can then locate the corresponding training subset to render better labelling, following conventional active learning approaches.

IV. Experimental setup

For evaluation using GSD, experiments were set up using audio book recordings (hereinafter abbreviated as ABA) from the public domain, as large amounts of data from individual speakers are available [21]. Six audio books were used: *A Tramp Abroad, Oliver, Typee, His Grace of Osmonde, Wuthering Heights* and *Emma*, with audio retrieved from the Librivox archives and text from Project Gutenberg[1].

A baseline GMM-HMM model on Perceptual Linear Perceptron (PLP) features was used for the ASR experiments. The baseline model is trained on out-of-domain data (170 hours) from close-talking microphone recordings as used in the AMIDA RT'09 transcription system [22]. The acoustic features comprise 13 PLP features and first and second derivatives. The baseline language model was the one used for the AMIDA RT'09 system [22], based on a 50,000–word vocabulary.

For ABA data, 60 hours of speech were available, containing 811k words from 3 male speakers ($m1$, $m2$, $m3$) and 3 female speakers ($f4$, $f5$, $f6$). The text data was pre-processed to remove metadata inserted in the transcripts, including chapter headings or other elements not spoken in the audio files. Further processing normalised the use of abbreviations and other written forms. Finally, this transcript was aligned to the audio and used as the reference for scoring.

[1] http://www.gutenberg.org/

TABLE I
WER(%) ON ABA-TEST FOR BASELINE AND ADPATED MODELS

	Speaker						
Model	m1	m2	m3	f4	f5	f6	Avg
Baseline	30.4	57.7	28.6	82.8	82.3	46.1	**52.9**
Adapt (ASR)	19.2	33.7	14.6	45.2	61.4	28.6	**33.7**
Adapt (GT)	16.5	16.0	11.2	16.4	21.1	13.7	**15.8**

For experimentation purposes, the data was split into 54 hours for model adaptation (ABA-DEV), and 6 hours for testing (ABA-TEST).

MAP adaptation of the GMM–HMM models was performed on ABA-DEV data independently for each speaker. Adaptation with ASR one-best hypotheses (ASR) and ground truth transcripts (GT) were performed respectively. For ASR adaptation, initial decoding with the baseline model derived the first-best hypotheses, based on which the models were adapted. For GT adaptation, the ground truth transcription was used for MAP adaptation. Baseline ASR results show the model performance without any adaptation. The GT condition mimics the scenario where high quality transcripts are available for model adaptation. The ASR and GT adaptation conditions represent the lower and upper bound performance of adaptation, controlled by the availability of high-quality labelled data. τ was set to 10 for all MAP adaptations for fair comparison.

V. ASR RESULTS WITH BASELINE AND ADAPTED MODELS

In this section, we report the capabilities of baseline and adapted ASR models on ABA-TEST. Results on ABA-TEST obtained with the baseline model and the two adapted models are shown in Table I in terms of Word Error Rate (WER). That the baseline model shows WER, suggests a mismatch between the model training data (meeting speech) and ABA-TEST (audio book data). Four speakers, $f4$, $f5$, $f6$ and $m2$, show especially poor performance. Results with ASR-adapted and GT-adapted models show significant adaptation improvements over the baseline model, which confirms the mismatch between baseline model and test data. With the Ground Truth (GT) transcript adaptation yielding better results by a wide margin, the gap between ASR-adapted and GT-adapted models show the potential improvement of audio book speech recognition with the availability of ground truth transcription from ABA-DEV. As expected, ASR transcripts for speakers with high WER ($f4$, $f5$, $f6$, $m2$) are low quality. Thus the GT-adapted model provides relative WER improvement above 50%, compared with the other speakers $m1$ and $m3$, where the improvement is 14% and 23% respectively.

VI. MODEL DISTANCE AND PHONEME ERROR RATE

A. Error analysis with respect to sGSD

In this section, an analysis is performed on ABA-DEV in an attempt to establish a relationship between sGSD and the transcription quality. Based on each speaker-dependent ASR-adapted model, GSD was computed for every state according to Eq.(1). segment-GSD (sGSD) was then computed for every segment in ABA-DEV using Eq.(2). A histogram of sGSD was computed and five subsets with equal numbers of segments were derived such that the segments they contain have sGSD falling in the 0-20%, 21-40%, 41-60%, 61-80% and 81-100% percentiles respectively.

TABLE II
PHONEME ERROR RATE (%) FOR DURATION OF DATA SORTED BY SGSD

Speaker.	m1	m2	m3	f4	f5	f6
0-20%	19.53	51.40	23.77	76.30	58.58	35.13
21-40%	18.96	42.80	21.56	59.34	54.69	30.47
41-60%	18.31	41.60	19.80	55.10	54.96	29.54
61-80%	18.38	40.31	18.49	49.89	61.48	31.34
81-100%	20.50	41.49	18.36	46.54	76.13	37.77

GSD is a state-level metric. Ideally, comparison with a state-level error metric makes sense as the two metrics are on the same linguistic level. However the confusion patterns among states were obscure and it may have a loose connection to the transcript quality compared with WER. Therefore, we chose to use Phoneme Error Rate (PER) as an evaluation metric for transcript quality to correlate with sGSD.

The PER of the five subsets, derived by their sGSD values, are represented in Table II. For each of the speakers, the correlation between sGSD and PER can be observed from figure 1. The relationship between sGSD and PER is either linear, or U shaped. When the data has low sGSDs, i.e. the 0-20% sGSD percentile data set, the PERs are high. When the data has high sGSDs, i.e. the 80-100% sGSD percentile subset, the PERs showed a mixed trend, where half of the speakers have high PERs and the other half give low PERs.

B. Low-GSD states: Discussion and further validation

The error analysis results above indicate two characteristics in sGSD which reflect transcript quality. First, segments with low sGSD values render high phoneme error rates. It is envisaged that among other complex factors, low GSD states reflect a "failure of adaptation" with low quality transcript. When the ASR transcript for a state is fundamentally wrong, it would have been mapped to different first-best hypotheses phonemes in different examples. As a result, adaptation would be conducted in a noisy condition, resulting in little change on average. Second, segments with high sGSD values also render high phoneme error rates, but this is only true for half the speakers. High GSD value implies a significant shift of model parameters after adaptation. Given the adaptation target is ASR transcripts, it is not clear whether the parameter shifts are correct or wrong. This may be a factor leading to the contradictory PER values among different speakers.

To further validate the effectiveness of sGSD and GSD in the role of active learning, an experiment was carried out to implement a full pipeline of active learning. We focus on the low-GSD regions as it gave a consistent trend for different speakers, suggesting their poor label quality. As explained in Section III, low-sGSD segments contain a fair amount of high-GSD states due to the phonotactic constraints. Therefore a count-based method was used to select segments for relabelling. This count-based method tried to enhance the homogeneity of low-GSD states in the selected segments.

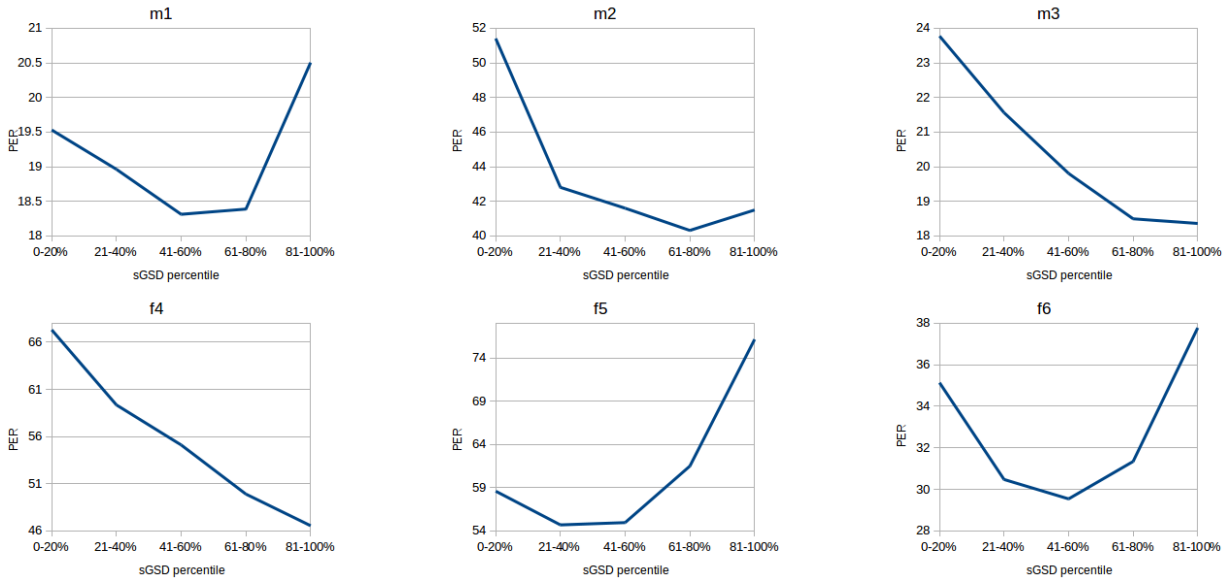

Fig. 1. Relationship between phoneme error rate (PER) using baseline model and sGSD values of segments (ABA-DEV)

In this count-based method, the 40% of states with the lowest GSD values were selected. The occurrences of these states in every segment were counted. The counts were normalised by segment length and the segments were ranked. This ranking method is equivalent to substituting the five-class sGSD quantisation (Section VI-A) with a binary GSD classification scheme (low or high GSD). Meanwhile, the count-based method tried to enhance the homogeneity of the low-GSD states in the selected segments.

To complete the active learning system pipeline, selected segments under the count-based methods had their labels replaced by ground-truth transcriptions. These new transcripts were then combined with the ASR transcripts from the unselected segments, on which model adaptation was performed. Therefore, despite different selection conditions the total duration of the adaptation data set remained the same.

Finally, the adapted acoustic models with different percentages of relabelled data were tested on ABA-TEST and the WERs were compared with the amount of data relabelled. This count-based selection method was also compared with random selection.

VII. Active Learning Experiments

As described in Section VI-B, various amounts of ASR transcripts in ABA-DEV were replaced by ground truth transcripts, mimicking an active learning scenario in which important data is chosen for relabelling. To evaluate the performance, the adapted models were applied to ABA-TEST. A control experiment was run where equivalent amounts of data were selected randomly to contrast with each set derived from the GSD selection method.

Fourteen data sets (7 GSD selection + 7 random selection) were used. For each selection method, the amount of data selected for relabelling varied between 5%, 10%, 20%, 30%, 40%, 50% and 70% (by duration).

Results on ABA-TEST with the GMM-HMM models adapted with different data are shown in Figure 2. All curves started on the left with a lower-bound performance (0% relabelled data, highest WER). For both selection methods, WER decreases with the percentage of relabelled data. Nevertheless, the decrease of WER with the random method is slower and more unpredictable. This is particularly true when we constrain the percentage of relabelled data to 10% and 20%.

Table III compares the WER with the GSD and random selection method particularly at 10% and 20% selection ratios. Apart from speaker $f6$, using the GSD selection method with 10% relabelled data gave almost as good performance as using the random selection method with double (20%) relabelled data.

TABLE III
Comparison between GSD and random data selection for active learning

Data	m1	m2	m3	f4	f5	f6
Random selection						
10%	19.3	31.6	14.1	44.5	51.5	26.5
20%	18.9	29.1	13.5	34.5	48.7	23.8
GSD selection: $Dist(\Lambda(s_i), \Lambda_X^{ASR}(s_i))$						
10%	18.7	30.4	13.7	34.3	49.4	25.9
20%	18.4	27.4	13.2	29.0	42.5	23.4

VIII. Discussion

This paper explored the relationship of model change after adaptation with correct or errorful transcripts. It was observed that errorful transcripts yield different distributions of GSD compared to correct transcripts. Based on that observation, a method was devised to first identify problematic states and then use them to identify problematic sentences as candidates for manual correction.

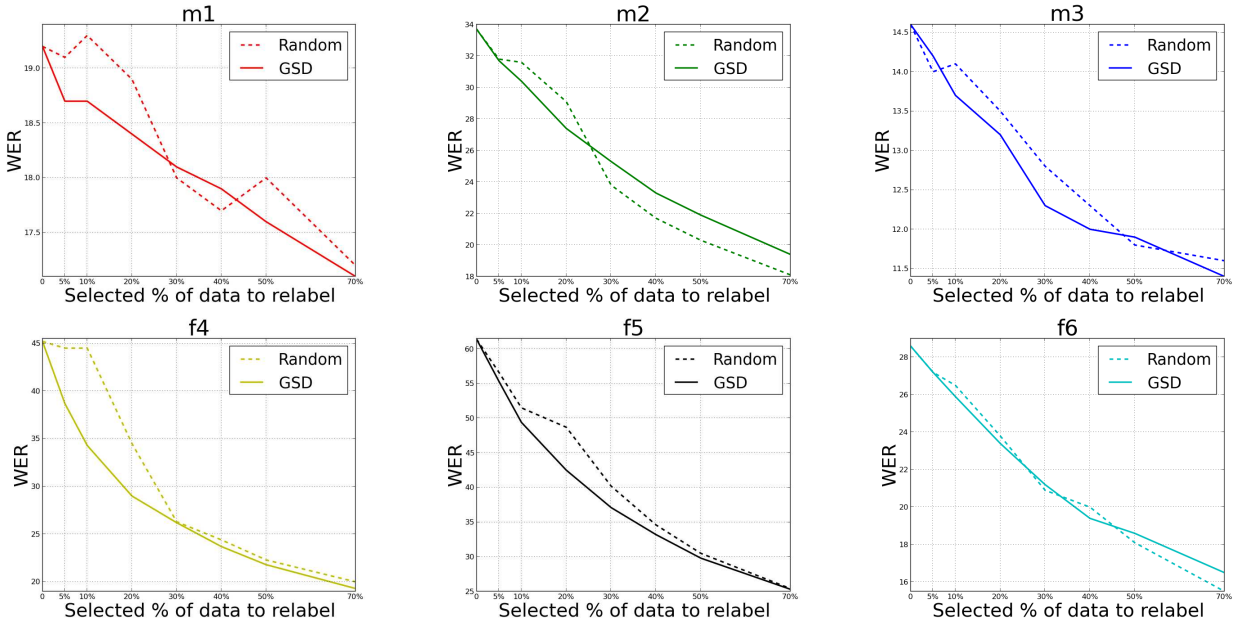

Fig. 2. Per-speaker comparison on word error rate (WER) when different percentages of the data has been relabelled using Ground-Truth. Two data selection methods (random and GSD) were used.

Different active learning approaches based on model change have been evaluated, in the context of MAP adaptation of acoustic models of GMM–HMM systems. The results have shown that this type of selection method can perform better than the random selection method.

This method proved to work efficiently in the GMM-HMM modeling framework because model units and distances are well defined and theoretically sound. In another experiment we applied GSD on tandem DNN configuration and arrived at the same qualitative conclution. The GSD metric provides important information and this is extracted from the statistical model rather than the training data. With detailed consideration, this concept will be adapted in a deep learning framework in future studies.

REFERENCES

[1] Phil C Woodland, "Speaker adaptation for continuous density HMMs: A review," in *Proc. of ITRW on Adaptation Methods for Speech Recognition*, 2001.

[2] Meng Wang and Xian-Sheng Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 2, pp. 10, 2011.

[3] G. Riccardi and D. Hakkani-Tür, "Active learning: Theory and applications to automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 504–511, 2005.

[4] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech and Language*, vol. 24, no. 3, pp. 433–444, 2010.

[5] Yuzo Hamanaka, Koichi Shinoda, Sadaoki Furui, Tadashi Emori, and Takafumi Koshinaka, "Speech modeling based on committee-based active learning," in *Proc. of ICASSP*, 2010, pp. 4350–4353.

[6] Ali Raza Syed, Andrew Rosenberg, and Ellen Kislal, "Supervised and unsupervised active learning for automatic speech recognition of low-resource languages," in *ICASSP*. IEEE, 2016, pp. 5320–5324.

[7] Burr Settles, "From theories to queries: Active learning in practice," *Active Learning and Experimental Design W*, pp. 1–18, 2011.

[8] Jean-Luc Gauvain and Chin-Hui Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 2, pp. 291–298, 1994.

[9] Sebastian Thrun, "Exploration in active learning," *Handbook of Brain Science and Neural Networks*, pp. 381–384, 1995.

[10] Li Deng and Xiao Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. on Audio, Speech and Language Proc.*, 2013.

[11] Bernhard Steffen, Falk Howar, and Maik Merten, "Introduction to active automata learning from a practical perspective," in *Formal Methods for Eternal Networked Software Systems*, pp. 256–296. Springer, 2011.

[12] David A Cohn, Zoubin Ghahramani, and Michael I Jordan, "Active learning with statistical models," *Journal of artificial intelligence research*, 1996.

[13] Sanjoy Dasgupta and Daniel Hsu, "Hierarchical sampling for active learning," in *Proc. of ICML*, 2008, pp. 208–215.

[14] Greg Schohn and David Cohn, "Less is more: Active learning with support vector machines," in *ICML*. Citeseer, 2000, pp. 839–846.

[15] Hong-Kwang Jeff Kuo and Vaibhava Goel, "Active learning with minimum expected error for spoken language understanding," in *Proc. of Interspeech*, 2005.

[16] Stephen H Shum, Najim Dehak, and James R Glass, "Limited labels for unlimited data: Active learning for speaker recognition," in *Proc. of Interspeech*, 2014.

[17] Dilek Hakkani-Tur, Giuseppe Riccardi, and Allen Gorin, "Active learning for automatic speech recognition," in *Proc. of ICASSP*, 2002.

[18] Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz, "Active learning for accent adaptation in automatic speech recognition.," in *Proc. of SLT Workshop*, 2012, pp. 360–365.

[19] Zixing Zhang and Björn Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition.," in *Proc. of Interspeech*, 2012.

[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 4, May 2011.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2015.

[22] T. Hain, L. Burget, J. Dines, P. N. Garner, A. El Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "The AMIDA 2009 meeting transcription system," in *Proc. Interspeech 2010*, 2010, pp. 358–361.