Training Data Selection for Acoustic Modeling via Submodular Optimization of Joint Kullback-Leibler Divergence

Taichi Asami, Ryo Masumura, Hirokazu Masataki, Manabu Okamoto, Sumitaka Sakauchi

NTT Media Intelligence Laboratories, NTT Corporation, Japan
{asami.taichi, masumura.ryo, masataki.hirokazu, okamoto.manabu, sakauchi.sumitaka}@lab.ntt.co.jp

Abstract

This paper provides a novel training data selection method to construct acoustic models for automatic speech recognition (ASR). Various training data sets have been developed for acoustic modeling. Each training set was created for a specific ASR application such that acoustic characteristics in the set, e.g., speakers, noise and recording devices, match those in the application. A mixture of such already-created training sets (an out-of-domain set) becomes a large utterance set containing various acoustic characteristics. The proposed method selects the most appropriate subset of the out-of-domain set and uses it for supervised training of an acoustic model for a new ASR application. The subset that has the most similar acoustic characteristics to the target-domain set (i.e., untranscribed utterances recorded by the target application) is selected based on the proposed joint Kullback-Leibler (KL) divergence of speech and non-speech characteristics. Furthermore, in order to select one of the many subsets in practical computation time, we also propose a selection algorithm based on submodular optimization that minimizes the joint KL divergence by greedy selection with guaranteed optimality. Experiments on real meeting utterances that use deep neural network acoustic models show that the proposed method yields better acoustic models than random or likelihood-based selection.

Index Terms: speech recognition, acoustic model, training data selection, Kullback-Leibler divergence, submodular optimization

1. Introduction

The recent improvements made to automatic speech recognition (ASR) performance is enabling various kinds of ASR applications. Each application is used in an application-specific acoustic environment. For example, voice search is used through smartphone devices by the general public in various noisy conditions, call center applications are used through headset microphones by well-trained speakers (i.e., operators) in a large operator room, and meeting recognition involves desktop or headset microphones by general business people in a meeting room. This encourages us to construct application-specific acoustic models for each target acoustic environment, and many kinds of acoustic training sets have been developed for various applications.

However, the human-transcription of spoken utterances for supervised acoustic model training incurs very high cost. In order to reduce the cost of acoustic model construction for a new application, we address training data selection for acoustic models. The goal is to select the best subset of utterances from the large (> 1000 hours) out-of-domain utterance set created by combining a wide variety of already-constructed training sets. The selected subset should have acoustic characteristics that are most similar to the characteristics of the target-domain set, which is a small (< 10 hours) utterance set recorded by the target application prior to actual use. While the out-of-domain set has already been transcribed, we assume that the target-domain set is not transcribed for cost saving. If the selected subset is good match, it can be used for supervised training of the acoustic model for the target application.

There are two issues with data selection: defining similarity between the selected subset and the target-domain set, and construction of a subset selection algorithm. Since acoustic characteristics are generated by many factors such as speakers, speaking styles, noise, recording devices etc., an appropriate similarity measure that reflects the effects of all factors is required. Even if a similarity measure is defined, an efficient subset selection algorithm is required since selecting an optimal subset from a huge out-of-domain utterance set is basically a large scale combinatorial optimization problem. To address these issues, we propose a novel framework for data selection, which formulates the similarity measure based on Gaussian mixture models (GMMs) and Kullback-Leibler (KL) divergence that takes account of all factors of acoustic characteristics, and optimizes the similarity measure by an efficient submodular optimization algorithm.

We propose a new similarity measure that separately considers speech characteristics and non-speech characteristics. The speech segments in utterances exhibit variability only in phonemes, but the non-speech segments exhibit extremely broad variability of all sounds other than phonemes, i.e. background noise. To model the background noise in detail, the similarity between the selected subset and the target-domain set is formulated as joint KL divergence based on the speech GMM and the non-speech GMM trained on the target-domain set. By using the non-speech GMM independently, the variability of the non-speech segments can be explicitly modeled.

The training set is selected so that the posterior probability distributions of the Gaussian components of the two GMMs in the selected set approach the target posterior distributions, i.e., the mixture weights of the two GMMs. We also propose an efficient subset selection algorithm that minimizes the joint KL divergence. The proposed method replaces the joint KL divergence minimization problem with an approximately equivalent submodular maximization problem [1, 2]. Submodular maximization can be efficiently solved by much faster variants of a greedy algorithm with guaranteed optimality [3, 4].

Several kinds of training data selection methods have been investigated. Siohan et al. proposed a simple selection algorithm with i-vector distribution similarity; it repeatedly picks an
utterance at random from the out-of-domain set and adds it to
the training set if this reduces the KL divergence of the i-vector distribution between the selected set and the target set [5]. Since i-vector [6] is designed for speaker recognition, similarity based on i-vector is effective in selecting utterances in terms of speaker characteristics, but the factors other than speakers are not explicitly considered. Moreover, the simple random picking algorithm has no guarantee of optimality. Submodular-based data selection methods have also been studied [7, 8, 9].

Wei et al. proposed a method of summarizing speech data set; it reduces the size of large speech data set so as to keep the phoneme balance while maximizing a submodular function based on phoneme TF-IDF features [7, 8]. Since the problem is data set size reduction, this method assumes that the original data set has uniform acoustic characteristics, i.e. each subset always has the same acoustic characteristics as the original set, and so focuses on phoneme balance and ignoring the other factors. In contrast, the problem addressed in this paper assumes that utterances in the original (out-of-domain) set have various acoustic characteristics, and the target characteristics are unknown. Shinohara showed that the KL divergence of phoneme frequency distribution can be optimized by maximizing a specific submodular function, and that a phoneme-balanced sentence set can be selected from a large text corpus [9]. We extend this knowledge about KL divergence optimization to joint KL divergence and use it for utterance set selection. Note that active learning techniques [10, 11] address a similar but different problem. Active learning methods aim to select utterances from the target-domain set, which should be transcribed and added to the training set. In contrast, our approach selects utterances from the out-of-domain set and does not require additional transcription of the utterances in the target-domain set.

This paper is organized as follows. Section 2 details the similarity measure that takes account of both speech and non-speech characteristics. Section 3 describes our efficient subset selection algorithm based on submodular maximization; it optimizes the similarity measure defined in Section 2 with guaranteed optimality. Conditions and results of speech recognition experiments on real meeting data are presented in Section 4, and Section 5 concludes this paper.

2. Joint KL divergence formulation of speech and non-speech similarity

The selected training set should have similar speech (i.e. speakers, speaking styles and channel distortion of speech) characteristics and non-speech (i.e. noise) characteristics to the target-domain set. Our formulation of a similarity (distance) measure that takes account of both characteristics is described below.

Both speech and non-speech characteristics of the target-domain set are separately modeled by GMMs. GMMs capture total structure of an acoustic feature space by Gaussian components, which represent all factors of the acoustic characteristics of the utterance set. First, the case of one GMM is considered. Mixture weights of Gaussian components of the GMM trained on the target-domain set are treated as the target distribution of the components that the selected training set should have. Thus, we use KL divergence from the mixture weight distribution to the posterior probability distribution on selected set $U$ as the objective function of $U$, which should be minimized:

$$
\hat{D}(U) = \sum_{i=1}^{M} w_i \log \frac{w_i}{\gamma_i(U)},
$$

(1)

where $M$ is the number of Gaussian components in the GMM trained on the target-domain set, $w_i$ is the mixture weight corresponding to the $i$-th Gaussian component, and $\gamma_i(U)$ is the posterior probability of the $i$-th component computed for selected utterance set $U$. Posterior $\gamma_i(U)$ can be calculated as the averaged frame-wise posterior probability of the $i$-th component:

$$
\gamma_i(U) = \frac{\sum_{u \in U} f_i(u)}{\sum_{u \in U} T_u},
$$

(2)

$$
f_i(u) = \sum_{j=1}^{M} \sum_{k=1}^{M} P_i(O_{sk}) \cdot P_k(O_{si}),
$$

(3)

where $T_u$ is the number of frames in utterance $u$, $O_{si}$ is the acoustic feature of the $t$-th frame in $u$, $P_i(O_{si})$ is the p.d.f. of $O_{si}$ computed by the $i$-th Gaussian component. Note that we use frequency notation, $f_i$, for Eq. (3) since $f_i(u)$ can be viewed as the occurrence frequency of the $i$-th component in $u$.

The proposed method extends the KL divergence $\hat{D}(U)$ to joint KL divergence of speech and non-speech GMMs to model the speech and non-speech characteristics in detail. Speech and non-speech frames in utterances can be separated by a voice activity detection (VAD) technique. A speech GMM and a non-speech GMM are trained on acoustic features extracted from the speech and non-speech frames in the target-domain set, respectively. A standard 38 dimensional acoustic feature vector (MFCC, ΔMFCC, ΔΔMFCC, Δpower and ΔΔpower) is used.

The utterances in the out-of-domain set can also be separated into speech and non-speech frames; let $U^{(s)}(U^{(n)})$ be the speech and non-speech frames of selected utterance set $U$, respectively. The proposed objective function is the sum of the KL divergence of the speech and non-speech GMMs:

$$
D(U) = \sum_{i=1}^{M^{(s)}} w_i^{(s)} \log \frac{w_i^{(s)}}{\gamma_i^{(s)}(U)} + \sum_{j=1}^{M^{(n)}} w_j^{(n)} \log \frac{w_j^{(n)}}{\gamma_j^{(n)}(U)},
$$

(4)

where $M^{(s)}/M^{(n)}$ are the numbers of Gaussian components of speech/non-speech GMMs trained on the target-domain set, $w_i^{(s)}/w_j^{(n)}$ are the mixture weights of $i$-th/$j$-th-component in the speech/non-speech GMMs, and $\gamma_i^{(s)}(U)/\gamma_j^{(n)}(U)$ are the averages of the frame-wise posteriors of the $i$-th/$j$-th-component in the two GMMs on the speech/non-speech frames of the selected set $U$, which are computed on $U^{(s)}(U^{(n)})$. $D(U)$ becomes small when the speech and non-speech characteristics in $U$ are simultaneously close to those in the target-domain set.

Subset $U$ should be selected from the out-of-domain set so that $D(U)$ is minimized under a size constraint, $|U| \leq K$, or a cost constraint, $C(U) \leq L$, where $|U|$ is the number of utterances in $U$, and $C(U)$ is the total length (i.e. the number of frames) of $U$.

3. Utterance selection algorithm based on submodular maximization

Finding the optimal subset that minimizes $D(U)$ from the huge out-of-domain utterance set is a large scale combinatorial optimization problem. Thus an efficient selection algorithm that yields an approximately optimal subset in practical computation time is required.

In order to achieve efficient optimization, we redefine the
objective function as follows:

\[
\hat{D}(U) = \sum_{i=1}^{M(s)} w^s_i \log \hat{f}^s_i(U) + \sum_{j=1}^{M(n)} w^n_j \log \hat{f}^n_j(U),
\]

where \(\hat{f}^s_i(u)\) and \(\hat{f}^n_i(u)\) are the component frequencies (i.e. the sum of the frame-wise posteriors) as in Eq. (3) but computed on the speech/non-speech frames of utterance \(u\). As described in [9], maximizing the weighted sum of logarithms of non-decreasing frequency functions is approximately equivalent to minimizing the corresponding KL divergence between the frequency distributions and weights (i.e. target distribution). Thus, maximizing the first term of Eq. (5) is equivalent to minimizing the first term of Eq. (4), and maximizing the second term of Eq. (5) is equivalent to minimizing the second term of Eq. (4). Therefore, maximizing \(\hat{D}(U)\) is equivalent to minimizing \(D(U)\).

The aim of this redefinition of the objective function is to yield sub-optimal solutions of large scale combinatorial optimization problems by application of an efficient greedy selection algorithm. \(\hat{D}(U)\) is a submodular function as is proven hereinafter. Submodular functions can be maximized by the greedy algorithm with guaranteed good lower bound of accuracy, \(1 - 1/e\) [2]. Furthermore, much faster variants of the greedy algorithm, lazy evaluation for the size constraint \([3]\) and cost-effective lazy forward (CELF) for the cost constraint \([4]\), can be used for maximizing the submodular functions.

The submodularity of \(\hat{D}(U)\) can be easily proven. It has already been proven in [9] that the first and second terms in Eq. (5) are submodular. Thus, \(D(U)\) is the sum of two submodular functions. The sum of two submodular functions is also submodular as described in Section 2 of [12]. Therefore, \(\hat{D}(U)\) is submodular.

Algorithm 1 shows the proposed training data selection algorithm with size constraint that jointly optimizes speech and non-speech characteristics of the selected utterance set. Lines 01-03 train the speech and non-speech GMMs, lines 04-09 prepare speech/non-speech component frequencies and mixture weights for calculating \(\hat{D}(U)\). Note that we use vector notation \(f^s(u), f^n(u), w^s,\) and \(w^n\) to represent the component frequencies and the mixture weights of a series of Gaussian components. Lines 10-15 select the subset of the out-of-domain set by greedy submodular maximization. Line 12 can be efficiently computed by lazy evaluation [3].

The size constraint in Line 11 can be easily replaced by a cost constraint, \(C(U) \leq L\), by replacing 11-15 with the CELF algorithm [4] and giving \(L\) as the input instead of \(K\). The cost constraint and the CELF algorithm version was used for all experiments described in Section 4.

4. Experiments

4.1. Data

We assumed that the target application is a meeting recognition system, and the experiments were conducted on utterances recorded in real meetings at our company. 1901 utterances (5.0 hours) of 5 (4 males and 1 female) speakers in 3 meetings were recorded and transcribed as the test set. Untranscribed 2731 utterances (7.5 hours) by anonymized (mix of males and females) speakers in 5 meetings were recorded as the target-domain set for training data selection. The 5 meetings of the target-domain set differed from the 3 meetings of the test set. The utterances were recorded by headset microphones, but specific non-stationary noises were recorded in addition to the statements in the meetings; sounds of shuffling papers and laptop key strokes, and laughter, cough, and back-channel of neighboring people.

The out-of-domain set consisted of 920K utterances (1100 hours), and was created by mixing various manually transcribed training sets including real data of voice search application, call center recordings, Corpus of Spontaneous Japanese (academic presentations) \([13]\), and Japanese Newspaper Article Sentences (reading newspapers) \([14]\).

All utterances were recorded with 16kHz sampling rate and 16bit resolution, and segmented by about 0.5 seconds of silence. Thus, all utterances have non-speech frames at the beginning and the end. Speech and non-speech frames in each utterance were separated by the energy-based VAD that is used for speaker recognition \([15]\).

4.2. Experimental conditions

The parameters of the proposed method are the number of speech Gaussian components \(M^s\), the number of non-speech Gaussian components \(M^n\), and the cost constraint \(L\). The number of speech components \(M^s\) and the number of non-speech components \(M^n\) were fixed to \(M^s = M^n = 256\) in all experiments. Several variations of training set size (10h, 50h, 100h, 200h, 300h and 500h) were investigated by changing the value of the cost constraint \(L\). In addition, the case where the entire 1100 hours out-of-domain set was used as the training set was also examined.

The proposed selection method was compared to two other selection methods: Random selection and simple GMM likelihood-based selection. GMM likelihood-based selection sorts the utterances in the out-of-domain set in descending order according to the likelihood computed by the speech GMM.
with 256 components, the same as the speech GMM used in the proposed method. The utterances with higher likelihood were selected so as to satisfy the cost constraint. GMM likelihood-based selection is equivalent to utterance-by-utterance selection where the utterances that have the closest posterior distribution to the target GMM are selected; note that this provides no guarantee of the optimality of the distribution in the whole selected subset. The proposed submodular optimization of objective function on the selected subset can be evaluated by comparing it to GMM likelihood-based selection.

Moreover, in order to evaluate the effectiveness of speech/non-speech separation, the same selection algorithms, but using only one GMM, were tested. Two types of GMM, a speech GMM with 256 components trained on $U^{(s)}_t$ and a global GMM with 512 components trained on $U^{(s)}_t \cup U^{(n)}_t$, were compared to the proposed method using two (speech and non-speech) GMMs.

Hybrid deep neural network and hidden Markov model (DNN-HMM) acoustic models were trained on each selected utterance set. The input of the DNN was a 418 dimensional vector formed by concatenating 11 consecutive (center and 5 previous and 5 succeeding) frames of 38 dimensional acoustic feature vectors described in Section 2. The number of hidden layers was 8, the number of units in each hidden layer was 2048, and the number of units in the output layer (i.e. the number of triphone HMM states) was 3072. The activation function of the hidden layers was sigmoid and that of the output layer was soft-max. The DNN was initialized by discriminative pre-training [16] on the entire out-of-domain set, 1100 hours. In each condition, this pre-trained DNN was fine-tuned by using the selected training set. Standard mini-batch stochastic gradient descent (SGD) with early stopping was used for fine-tuning. 1 hour of the validation set for early stopping was randomly split from each selected training set.

The same 3-gram language model was used for speech recognition in all conditions: it has 520K vocabulary size and was trained on various text corpora consisting of 2.3G words including meeting transcription. Decoding was performed by the WFST-based decoder VoiceRex [17, 18].

### 4.3. Results

Character error rates on the test set for all conditions are shown in Figure 1. “Random” means random selection, “Likelihood (speech256)” means simple likelihood-based selection as described in the previous section, “KLD (global512)” is the proposed selection algorithm but using one global GMM trained on $U^{(s)}_t \cup U^{(n)}_t$, “KLD (speech256)” is the proposed algorithm using one speech GMM trained on $U^{(s)}_t$, and “KLD (speech256+non-speech256)” is the complete version of the proposed method using the speech and non-speech GMMs. The lowest error rate, 30.7%, was achieved by selecting 200 hours of training utterances by “KLD (speech256+non-speech256).”

“KLD (speech256+non-speech256)” attained a lower error rate than “KLD (global512)” or “KLD (speech256).” Both “KLD (global512)” and “KLD (speech256+non-speech256)” use the same number of Gaussian components for modeling the target-domain set. The performance improvement yielded by explicitly separating speech and non-speech characteristics is confirmed by comparing the two conditions.

The difference between “KLD (speech256)” and “KLD (speech256+non-speech256)” is the use of the non-speech GMM. The consistent improvement obtained by “KLD (speech256+non-speech256)” indicates the effectiveness of the proposed joint KL divergence measure in separately taking account of both speech and non-speech characteristics in training data selection.

“KLD (speech256+non-speech256)” also achieved lower error rate than the other two selection methods, “Random” and “Likelihood (speech256),” and especially large improvements were yielded when the training set was small. Furthermore, “KLD (speech256)” outperformed “Likelihood (speech256)” despite the use of the same GMM. These results confirm that the proposed selection approach, which optimizes the distribution of Gaussian components for the whole selected subset is more appropriate than the likelihood measure that focuses on the distribution of each utterance.

Finally, training at 50, 100, 200 and 500 hours yielded higher “KLD (speech256+non-speech256)” performance than when the entire out-of-domain set was used as the training set (1100 hours). This means that better acoustic models for the target application can be constructed by the proposed data selection approach from relatively small amounts of utterances recorded in the target application without human-transcription.

### 5. Conclusion

We proposed a novel acoustic training data selection algorithm that extracts the most appropriate subset from a large out-of-domain utterance set such that the selected subset has acoustic characteristics most similar to those of the target-domain set.

The proposed method effectively takes account of various acoustic characteristics by separately modeling speech and non-speech characteristics by GMMs, and measuring similarity by the joint KL divergence of distributions of speech and non-speech Gaussian components. Moreover, we formulated a submodular objective function for minimizing the joint KL divergence and constructed an algorithm that offers efficient suboptimal subset selection from the huge out-of-domain set.

Speech recognition experiments on real meeting data showed effectiveness of the proposed joint KL divergence, which separately takes account of speech and non-speech characteristics, and the proposed selection algorithm, which employs submodular maximization techniques. We also confirmed that acoustic models that well match the acoustic environment of the target application can be constructed from the training data selected by the proposed method without human-transcription of the target-domain set.
6. References


