UNSUPERVISED SUBMODULAR SUBSET SELECTION FOR SPEECH DATA

Kai Wei *Yuzong Liu* *Katrin Kirchhoff* *Jeff Bilmes*

Department of Electrical Engineering, University of Washington Seattle, WA 98195, USA

ABSTRACT

We conduct a comparative study on selecting subsets of acoustic data for training phone recognizers. The data selection problem is approached as a constrained submodular optimization problem. Previous applications of this approach required transcriptions or acoustic models trained in a supervised way. In this paper we develop and evaluate a novel and entirely unsupervised approach, and apply it to TIMIT data. Results show that our method consistently outperforms a number of baseline methods while being computationally very efficient and requiring no labeling.

Index Terms— speech processing, automatic speech recognition, machine learning

1. INTRODUCTION

Automatic speech recognition (ASR) systems are nowadays trained on abundant speech data, which can be easily collected in the field, e.g., via mobile applications. However, a large amount of acoustic training data goes hand in hand with increased computational demands and with increased effort for labeling the data. More importantly, the larger the training data set, the more redundant it tends to be. It has been observed numerous times that the performance curve with respect to the amount of training data often shows "diminishing returns": a smaller performance gain is achieved when adding new data to a larger pre-existing data set than when adding it to a smaller pre-existing set [1]. Therefore, it is critical to select the most informative and representative subset of a large data set, to maximize the potential benefits to be gained from additional data while minimizing resource requirements. We call this problem data subset selection problem.

It can be further categorized according to whether transcriptions of the training data are available (*supervised* data subset selection) or not (*unsupervised* data subset selection). Supervised data subset selection is of most interest if the desire is to shorten system development time, by tuning parameters on a small and representative subset of data that can be processed much faster yet produces results of the same quality as the original data set.

The unsupervised data subset selection scenario is applicable when a new corpus of speech data has been collected (e.g., from a new language or dialect) but has not yet been transcribed. The available budget for transcribing or annotating a small set of speech for bootstrapping purposes may be limited. Ideally, data subset selection methods should identify the subset that fits the budget and at the same time yields the maximum amount of information about the entire data set. Additionally, the selection method should require *low resources*, i.e., it should not rely on existing resources such as an already-trained ASR system for the language in question.

In this paper, we develop a novel unsupervised speech data subset selection methodology based on submodular functions. We first discuss relevant background literature (Section 2) before presenting our submodular approach to data subset selection (Section 3). Section 4 provides details about data and systems, and Section 5 gives experimental results on TIMIT data.

2. BACKGROUND

Previous approaches to selecting untranscribed acoustic data have mostly relied on batch active learning, where a subset of untranscribed training data is chosen for additional human transcription to update an existing trained system. In [2, 3, 4] items are selected in a greedy fashion according to their utility scores, measured as the confidence scores assigned by an existing word recognizer. Similar to the confidence-based approaches, [5] proposes to select unlabeled utterances such that maximum lattice entropy reduction can be achieved over the whole dataset. In [6], two criteria (informativeness and representativeness) are used to subselect acoustic data. The informativeness score of an utterance is the entropy of its N-best word hypothesis, decoded by an existing word recognizer. The representative score of an utterance with respect to a data pool is calculated as its average TF-IDF similarity with all other utterances in the pool. Similar to the active learning approaches, this method requires an already-trained ASR system with reasonably high performance. Low-resource methods proposed for speech data selection include [7]; however, this approach has only been investigated for the supervised data subset selection scenario, i.e., it assumes that transcriptions of the data are available. Selection is performed such as to maximize the entropy of the distribution over linguistic units (words, phones) in the subselected set. It should also be noted that all these methods select data in a greedy fashion but do not have any optimality guarantee in terms of the objective being optimized.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Hence, the resulting selection, in terms of its objective criterion, can in the worst case be quite poor. Another class of approaches formulates the problem as a constrained submodular maximization problem [8, 9]. [8] introduced this approach, and considers both the supervised and unsupervised data subset selection problem, and instantiates submodular objectives using a Fisher-kernel-based similarity measure. Similarly, [9] is also supervised and employs methods of computing similarity measures between speech utterances that go beyond direct acoustic characteristics. Our companion paper [10] addresses the supervised selection problem in the case of large-vocabulary ASR.

In the present work, we focus on the unsupervised scenario and present different submodular functions for this problem, in particular a novel two-level submodular function.

3. SUBMODULAR DATA SELECTION

Background on Submodularity: Discrete optimization is important to many areas of speech technology and recently an ever growing number of problems have been shown to be expressible as submodular function maximization [8, 9, 11]. A submodular function [12] is defined as follows: given a finite set $V = \{1, 2, \ldots, n\}$ and a discrete set function $f : 2^V \rightarrow \mathbb{R}$ that returns a real value for any subset $S \subseteq V$, f is submodular whenever $f(A) + f(B) \geq f(A \cup B) + f(A \cap B), \forall A, B \subseteq V$. Defining $f(j|S) \triangleq f(j \cup S) - f(S)$, an equivalent definition of submodularity is $f(j|S) \geq f(j|T), \forall S \subseteq T$. That is, the incremental gain of j decreases when the set in which j is considered grows from S to T. A submodular function f is monotone non-decreasing if $f(j|S) \geq 0, \forall j \in V \setminus S, S \subseteq V$. Submodular functions indeed exhibit favorable properties for discrete optimization problems. Although NP-hard, (constrained) submodular maximization admits constant factor approximation algorithms [13, 14]. For example, the problem of maximizing a monotone submodular function subject to a cardinality constraint can be approximately solved by a simple greedy algorithm [13] with worst approximation factor $(1 - e^{-1})$. This is the best possible outcome in polynomial time unless $P = NP$ [15]. Submodularity can be further exploited to accelerate a greedy implementation such that it has almost linear time complexity [16]. The same scalable greedy algorithm can easily generalize to approximately solve the problem of monotone submodular maximization under knapsack constraints, with similar theoretical guarantees [17]. We employ the same scalable greedy algorithm in our approach for speech data subset selection.

Problem formulation: Suppose we have a set of speech utterances $V = \{1, 2, \ldots, N\}$. Consider a monotonically non-decreasing submodular set function $f : 2^V \rightarrow \mathbb{R}$, which maps each subset $S \subseteq V$ to a real number that represents the value $f(S)$ of subset S. The speech data subset selection problem, then, can be viewed as maximizing the value $f(S)$ of S such that the cost of the selected subset S does not exceed a given budget. Mathematically, the problem can be formulated as monotone submodular function maximization under a knapsack constraint:

$$\max_{S \subseteq V, c(S) \leq B} f(S) \tag{1}$$

where B is a budget on the amount (or cost) of speech data to be selected and $c(S) = \sum_{j \in S} c(j)$ measures the amount (or cost) of speech contained in a subset S of the whole corpus, with $c(j)$ being the length of the utterance $j \in V$. The same problem formulation was applied in prior investigations of submodular speech data subset selection [9].

In [9], the **facility location**, i.e.,

$$f_{\text{fac}}(S) = \sum_{i \in V} \max_{j \in S} w_{ij}, \tag{2}$$

was applied as the submodular objective, where $w_{ij} \geq 0$ indicates the similarity between utterances i and j. The similarity measure w_{ij} was computed by kernels derived from discrete representations of the acoustic utterances i and j.

The facility location function usually leads to a highly representative solution, but, in some cases, the solution might still posses redundancy. In this work, we utilize a "**diversity reward**" function, first proposed in [17], to penalize redundancy by rewarding diversity. The function takes the following form:

$$f_{\text{div}}(S) = \sum_{n=1}^{K} \sqrt{\sum_{j \in P_n \cap S} \left(\sum_{i \in V} \frac{w_{ij}}{|V|} \right)} \tag{3}$$

where $P_1, \ldots, P_K$ is a partition of the set V into K (disjoint) blocks. f_{div} is monotone submodular. Maximizing f_{div} encourages selecting items from different blocks and leads to more diverse and less redundant selections. In order to select a subset that is both representative and non-redundant, we can mix both objectives f_{fac} and f_{div} together, training off between representation and non-redundancy, as in the following objective:

$$f_{\text{fac+div}}(S) = (1 - \lambda) f_{\text{fac}}(S) + \lambda f_{\text{div}}(S) \tag{4}$$

where $1 \geq \lambda \geq 0$ is a trade-off coefficient.

Both the facility location function and the diversity reward function are graph-based so a pair-wise similarity graph is required to instantiate them. Even with highly optimized data structures, efficient computation of similarity measures, and graph approximations, graph construction can become computationally prohibitive when $|V|$ is big (e.g., millions, or greater). An alternative class of submodular functions is **feature-based**, defined as:

$$f_{\text{fea}}(S) = \sum_{u \in \mathcal{U}} g(m_u(S)) \tag{5}$$

where $g()$ is a non-negative monotone non-decreasing concave function (e.g., the square root function), $\mathcal{U}$ is a set of

features, the modular feature function $m_u(S) = \sum_{j \in S} m_u(j)$ is a non-negative score for feature u in a set S, with $m_u(j)$ measuring the degree of feature u present in utterance $j \in S$. Maximizing this objective naturally encourages diversity and coverage of the features within the chosen set of elements. In the context of speech data subset selection, $\mathcal{U}$ can take various forms including triphones, words, phonemes, acoustically derived measures, etc. Moreover, there are many different ways to define the relevance score $m_u(s)$. One simple way might be to define it as the amount of feature u contained within utterance s. A more sophisticated measure utilizes term frequency inverse document frequency (TF-IDF) normalized counts, i.e. $m_u(s) = \text{TF}_u(s) \times \text{IDF}_u$, where $\text{TF}_u(s)$ is the count of feature u in s, and IDF_u is the inverse document count of the feature u (each utterance is a "document").

One issue with the feature-based functions is that they represent interactions between different items in the whole set V, but cannot represent interactions between different features or sets of features, meaning that information within one feature $u \in \mathcal{U}$ might be partially redundant with another feature $u' \in \mathcal{U}$, $u' \neq u$. A solution to this issue is to use a novel construct we call a **two-layer feature-based submodular functions**. Let $\mathcal{U}^1$ be a set of features, $\mathcal{U}^2$ be a set of meta-features, where $|\mathcal{U}^1| = d_1, |\mathcal{U}^2| = d_2$ and $d_1 > d_2$. Between $\mathcal{U}^1$ and $\mathcal{U}^2$, we define a weight matrix W of dimension $d_2 \times d_1$. Entries in W measure the interactions between the lower-level features in $\mathcal{U}^1$ and corresponding meta-features in $\mathcal{U}^2$. The two-layer feature-based submodular function takes the following form:

$$f_{\text{2-fea}}(S) = \sum_{u_2 \in \mathcal{U}^2} g_1\Big(\sum_{u_1 \in \mathcal{U}^1} W(u_2, u_1) g_2(m_{u_1}(S)) \Big) \quad (6)$$

where $g_1()$ and $g_2()$ are non-negative monotonically non-decreasing concave functions, $m_{u_1}(S)$ takes the same form and interpretation as in the feature-based submodular function. $W(u_2, u_1)$ is the entry in the weight matrix W that measures the interaction between the feature $u_1 \in \mathcal{U}^1$ and the feature $u_2 \in \mathcal{U}^2$. The submodularity of $f_{\text{2-fea}}(S)$ follows from Theorem 1 in [18].

4. DATA AND SYSTEMS

We evaluate our approach on subselecting the TIMIT corpus for phone recognizer training. The sizes of the training, development and test data (without the `sa` sentences) are 4620, 200 and 1144 utterances, respectively. Preprocessing extracts 39-dimensional MFCC features every 10 ms, with a window of 25.6 ms. Speaker mean and variance normalization were applied. A 3-state monophone HMM phone recognizer is trained for all 48 monophones. The HMM state output distributions are modeled by diagonal-covariance Gaussian mixtures. Performance is evaluated by phone accuracy, collapsing the 48 classes into 39 for scoring purposes, following standard practice [19]. Since we focus on acoustic modeling only we avoid the use of a phonetic language model. The goal of this work is not to achieve the highest phone accuracy possible; we care most about the *relative* performance of the different subset selection methods, especially on small data subsets.

We compare our approach to two different baseline selection methods, random selection and the histogram-entropy method from [7]. Following the selection of subsets (1%, 2.5%, 5%, 10%, 20%, 30% and 40% of the data, measured as percentage of non-silence speech frames) we construct the random baseline by randomly sampling 100 data sets of the desired sizes, training phone recognizers, and averaging the results. The histogram-entropy baseline systems are constructed such that utterances are selected to maximize the entropy of the histogram over phones in the selected subset [7]. The phone labels are derived from the true transcriptions.

5. EXPERIMENTS

In the first set of experiments we tested the performance of *supervised* submodular data selection, i.e., the transcriptions of the training data were used. We tested the objective functions f_{fac}, $f_{\text{fac+div}}$, and f_{fea} defined in Equations 2, 4, and 5, respectively. The similarity matrix used to instantiate f_{fac} and f_{div} was computed using a gapped string kernel [20] on a phone tokenization of the acoustic utterances. The phone tokenization was generated by a simple bottom-up monophone recognizer trained on the TIMIT transcriptions. In addition we partitioned the whole data set into 64 clusters using k-means clustering, in order to instantiate the diversity function f_{div}. The parameter λ was optimized on the development set for all subset sizes. The set of features $\mathcal{U}$ in the feature-based submodular function f_{fea} is the set of all phone trigrams obtained from a forced-alignment of the transcriptions. The feature score function $m_u(S)$ is instantiated to the sum of the TF-IDF weighted normalized counts of the feature u in the set S. The concave function $g()$ is the square root function. Figure 1 shows the performance of the average random baseline, histogram-entropy baseline and submodular systems with different objectives. The histogram-entropy baseline outperforms the random baseline at all percentages, except at 1%. Submodular selection with f_{fea} outperforms the two baselines at all subset sizes except at 10%. Submodular selection with f_{fac} or $f_{\text{fac+div}}$ yielded significantly ($p < 0.05$) better results than both baselines, especially at small subset sizes. The diversity term is indeed helpful for most subset sizes. Interestingly, selection by $f_{\text{fac+div}}$ at 2.5% even beat the random baseline at 5%, meaning that in this case the same performance could be achieved by only using half of the training data, if selected wisely. In the 10% case, we increase the number of random runs from 100 to 1000, which helps to span the random selection space better. In this case, the best out of 1000 random runs is very close to the histogram-entropy baseline, but is still significantly outperformed by the submodular methods with f_{fac} and $f_{\text{fac+div}}$.

In our second set of experiments we tested the performance

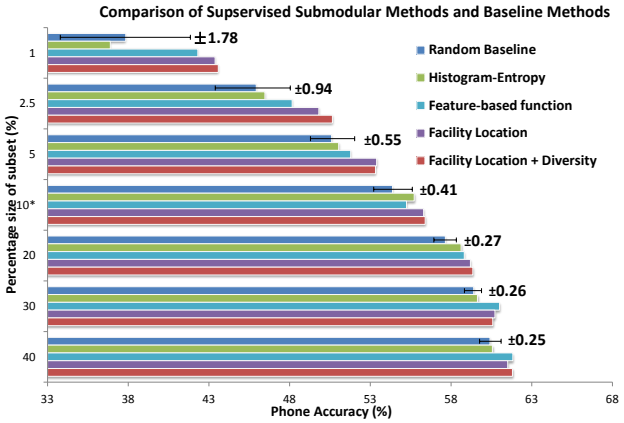

Fig. 1. Phone accuracy for different subset sizes; each block of bars lists, from top to bottom: random baseline (error bars indicate the best and the worst performance out of 100 random runs, except at 10%, the best and the worst performance is out of 1000 random runs; the numbers on top of error bars indicate the standard deviation of all runs), histogram-entropy baseline, f_{fea}, f_{fac}, $f_{\text{fac+div}}$.

of unsupervised submodular selection, which does not use the transcriptions. The different submodular objectives were f_{fac} and $f_{\text{2-fea}}$ (Equations 2 and 6, respectively). The similarity matrix in f_{fac} was again computed by a gapped string kernel on tokenized utterances. In contrast to the supervised case, however, we utilized an HMM trained *in an unsupervised way* to produce the tokenization. This unsupervised model had 40 HMM states and 64 Gaussian mixture components per state and was trained on all of the training data. To instantiate the two-layer feature based function $f_{\text{2-fea}}$, we constructed the set of meta features $\mathcal{U}^2$ as the set of tri-states extracted from the sequences of HMM state labels. Each tri-state $u_2 \in \mathcal{U}^2$ was distinguished by the dominating Gaussian component index at the middle state; its corresponding lower-level features were constructed as the tri-state with all possible Gaussian component indices in the middle state. We constrained interactions between lower-level features and meta features to the case where both features shared the same tri-state, i.e., $W(u_1, u_2) = 1$ if u_1 and u_2 shared the same tri-state, and $W(u_1, u_2) = 0$ otherwise.

In addition to using an unsupervised HMM as the tokenizer we also tried using a k-component single-state unsupervised GMM. This model converted acoustic utterances into sequences of indices representing the dominant Gaussian component at each frame. A 512-component GMM was used to generate the set of low-level features $\mathcal{U}^1$, and a 32-component GMM was used to generate meta features $\mathcal{U}^2$. In both cases, we generate features as the Gaussian mixture indices for two consecutive frames (bigrams). The weight $W(u_1, u_2)$ was set to the co-occurrence count of features $u_1 \in \mathcal{U}^1$ and $u_2 \in \mathcal{U}^2$ in the training set. In both instantiations of $f_{\text{2-fea}}$, the concave functions $g_1()$ and $g_2()$ were set to the square root function, and the feature score function $m_u(S)$ was the sum of TF-IDF normalized feature counts.

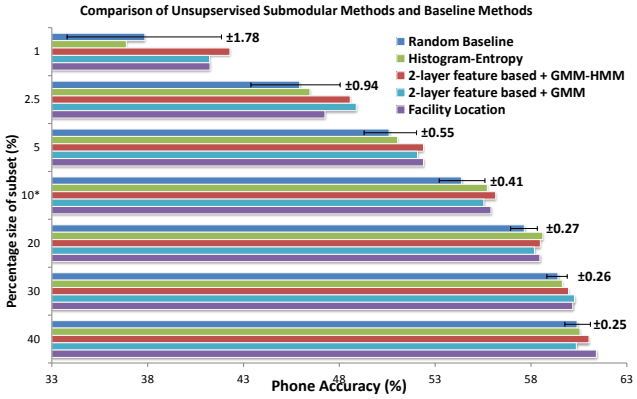

Fig. 2. Phone accuracy for different subset sizes; each block of bars lists, from top to bottom: random baseline (same as in Figure 1), histogram-entropy baseline, $f_{\text{2-fea}}$+ GMM-HMM, $f_{\text{2-fea}}$+GMM, f_{fac}.

Figure 2 shows the performance of the three unsupervised submodular selection methods described above. They all significantly ($p < 0.05$) outperform the random baseline for all subset sizes except for $f_{\text{2-fea}}$+GMM at 40%. The improvement is more evident for small subset sizes (1%, 2.5%, 5%). In general, these unsupervised methods yield a performance comparable to that of the histogram-entropy baseline, which is a supervised method. In particular, $f_{\text{2-fea}}$+GMM-HMM outperforms the histogram-entropy baseline at all subset sizes, except for 20%.

6. CONCLUSIONS

We have explored the problem of submodular speech data subset selection from two new angles: (a) we have tested novel submodular objectives (feature-based and two-level feature-based functions), which do not require similarity graphs; (b) we have extended this approach to a scenario where transcriptions of the training data are not available. In both the supervised and unsupervised scenario the submodular selection methods outperformed the baseline methods. Future work will extend these investigations to larger systems and data sets, in particular data sets that are acoustically more diverse.

Acknowledgment

This material is based on research sponsored by Intelligence Advanced Research Projects Activity (IARPA) under agreement number FA8650-12-2-7263.

7. REFERENCES

[1] Roger K Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners.," in *Proceedings of Interspeech*, 2003.

[2] Dilek Hakkani-Tur, Giuseppe Riccardi, and Allen Gorin, "Active learning for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, vol. 4, pp. IV–3904.

[3] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[4] Thomas Kemp and Alex Waibel, "Unsupervised training of a speech recognizer using TV broadcasts," in *ICSLP*, 1998, vol. 98, pp. 2207–2210.

[5] Balakrishnan Varadarajan, Dong Yu, Li Deng, and Alex Acero, "Maximizing global entropy reduction for active learning in speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 4721–4724.

[6] Nobuyasu Itoh, Tara N Sainath, Dan Ning Jiang, Jie Zhou, and Bhuvana Ramabhadran, "N-best entropy based data selection for acoustic modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4133–4136.

[7] Yi Wu, Rong Zhang, and Alexander Rudnicky, "Data selection for speech recognition," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, pp. 562–565.

[8] Hui Lin and Jeff A. Bilmes, "How to select a good training-data subset for transcription: Submodular active selection for sequences," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, UK, September 2009.

[9] Kai Wei, Yuzong Liu, Katrin Kirchhoff, and Jeff Bilmes, "Using document summarization techniques for speech data subset selection," in *North American Chapter of the Association for Computational Linguistics/Human Language Technology Conference (NAACL/HLT-2013)*, Atlanta, GA, June 2013.

[10] Kai Wei, Yuzong Liu, Katrin Kirchhoff, Chris Bartels, and Jeff Bilmes, "Submodular subset selection for large-scale speech training data," in *Proceedings of ICASSP*, Florence, Italy, 2014.

[11] Yuzong Liu, Kai Wei, Katrin Kirchhoff, Yisong Song, and Jeff Bilmes, "Submodular feature selection for high-dimensional acoustic score spaces," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[12] J. Edmonds, "Submodular functions, matroids and certain polyhedra," *Combinatorial structures and their Applications*, 1970.

[13] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[14] N. Buchbinder, M. Feldman, J. Naor, and R. Schwartz, "A tight (1/2) linear-time approximation to unconstrained submodular maximization," *In FOCS*, 2012.

[15] U. Feige, "A threshold of ln n for approximating set cover," *Journal of the ACM (JACM)*, vol. 45, no. 4, pp. 634–652, 1998.

[16] M. Minoux, "Accelerated greedy algorithms for maximizing submodular set functions," *Optimization Techniques*, pp. 234–243, 1978.

[17] Hui Lin and Jeff Bilmes, "A class of submodular functions for document summarization," in *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT-2011)*, Portland, OR, June 2011.

[18] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," *In ACL*, 2011.

[19] K-F Lee and H-W Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 37, no. 11, pp. 1641–1648, 1989.

[20] J. Rousu and J. Shawe-Taylor, "Efficient computation of gapped substring kernels on large alphabets," *Journal of Machine Learning Research*, vol. 6, no. 2, pp. 1323, 2006.