Using Document Summarization Techniques for Speech Data Subset Selection

Kai Wei*, **Yuzong Liu***, **Katrin Kirchhoff** , **Jeff Bilmes**
Department of Eletrical Engineering
University of Washington
Seattle, WA 98195, USA
{kaiwei,yzliu,katrin,bilmes}@ee.washington.edu

Abstract

In this paper we leverage methods from submodular function optimization developed for document summarization and apply them to the problem of subselecting acoustic data. We evaluate our results on data subset selection for a phone recognition task. Our framework shows significant improvements over random selection and previously proposed methods using a similar amount of resources.

1 Introduction

Present-day applications in spoken language technology (speech recognizers, keyword spotters, etc.) can draw on an unprecedented amount of training data. However, larger data sets come with increased demands on computational resources; moreover, they tend to include redundant information as their size increases. Therefore, the performance gain curves of large-scale systems with respect to the amount of training data often show “diminishing returns”: new data is often less valuable (in terms of performance gain) when added to a larger pre-existing data set than when added to a smaller pre-existing set (e.g.,(Moore, 2003)). Therefore it is of prime importance to develop methods for data subset selection. We distinguish two data subselection scenarios: (a) a priori selection of a data set before (re-)training a system; in this case the goal is to subselect the existing data set as well as possible, eliminating redundant information; (b) selection for adaptation, where the goal is to tune a system to a known development or test set. While many studies have addressed the second scenario, this paper investigates the first: our goal is to select a smaller subset of the data that fits a given ’budget’ (e.g. maximum number of hours of data) but provides, to the extent possible, as much information as the complete data set. Additionally, our selection method should be a low-resource method that does not require an already-trained complex system such as an existing word recognizer.

This problem is akin to unsupervised data ’summarization’. In (Lin and Bilmes, 2009) a novel class of summarization techniques based on submodular function optimization were proposed for extractive document summarization. Interestingly, these methods can also be applied to speech data ’summarization’ with only small modifications. In the following sections we develop a submodular framework for speech data summarization and evaluate it on a proof-of-concept phone recognition task.

2 Related Work

Most approaches to data subset selection in speech have relied on “rank-and-select” approaches that determine the utility of each sample in the data set, rank all samples according to their utility scores, and then select the top N samples. In weakly supervised approaches (e.g.,(Kemp and Waibel, 1998; Lamel et al., 2002; Hakkani-Tur et al., 2002), utility is related to the confidence of an existing word recognizer on new data samples: untranscribed training data is automatically transcribed using an existing baseline speech recognizer, and individual utterances are selected as additional training data if they have low

[*] These authors are joint first authors with equal contributions.

Proceedings of NAACL-HLT 2013, pages 721–726,
Atlanta, Georgia, 9–14 June 2013.

confidence. These are active learning approaches suitable for a scenario where a well-trained speech recognizer is already available and additional data for retraining needs to be selected. However, we would like to reduce available training data ahead of time with a low-resource approach. In (Chen et al., 2009) individual samples are selected for the purpose of discriminative training by considering phone accuracy and the frame-level entropy of the Gaussian posteriors. (Itoh et al., 2012) use a utility function consisting of the entropy of word hypothesis N-best lists and the representativeness of the sample using a phone-based TF-IDF measure. The latter is comparable to methods used in this paper, though the first term in their objective function still requires a word recognizer. In (Wu et al., 2007) acoustic training data associated with transcriptions is subselected to maximize the entropy of the distribution over linguistic units (phones or words). Most importantly, all these methods select samples in a greedy fashion without optimality guarantees. As we will explain in the next section, greedy selection is near-optimal only when applied to monotone submodular functions.

3 Submodular Functions

Submodular functions (Edmonds, 1970) have been widely studied in mathematics, economics, and operations research and have recently attracted interest in machine learning (Krause and Guestrin, 2011). A submodular function is defined as follows: Given a finite ground set of objects (samples) $V = \{v_1, ..., v_n\}$ and a function $f : 2^V \rightarrow \mathbb{R}^+$ that returns a real value for any subset $S \subseteq V$, f is submodular if $\forall A \subseteq B$, and $v \notin B$, $f(A + v) - f(A) \geq f(B + v) - f(B)$. That is, the incremental "value" of v decreases when the set in which v is considered grows from A to B. Powerful optimization guarantees exist for certain subtypes of submodular functions. If, for example, the function is *monotone submodular*, i.e. $\forall A \subseteq B, f(A) \leq f(B)$, then it can be maximized, under a cardinality constraint, by a greedy algorithm that scales to extremely large data sets, and finds a solution guaranteed to approximate the optimal solution to within a constant factor $1 - 1/e$ (Nemhauser et al., 1978). Submodular functions can be considered the discrete analog of convexity.

3.1 Submodular Document Summarization

In (Lin and Bilmes, 2011) submodular functions were recently applied to extractive document summarization. The problem was formulated as a monotone submodular function that had to be maximized subject to cardinality or knapsack constraints:

$$\text{argmax}_{S \subseteq V} \{f(S) : c(S) \leq K\} \qquad (1)$$

where V is the set of sentences to be summarized, K is the maximum number of sentences to be selected, and $c(\cdot) \geq 0$ is sentence cost. $f(S)$ was instantiated by a form of **saturated coverage**:

$$f_{SC}(S) = \sum_{i \in V} \min\{C_i(S), \alpha C_i(V)\} \qquad (2)$$

where $C_i(S) = \sum_{j \in S} w_{ij}$, and where $w_{ij} \geq 0$ indicates the similarity between sentences i and j — $C_i : 2^V \rightarrow \mathbb{R}$ is itself monotone submodular (modular in fact) and $0 \leq \alpha \leq 1$ is a saturation coefficient. $f_{SC}(S)$ is monotone submodular and therefore has the previously mentioned performance guarantees. The weighting function w was implemented as the cosine similarity between TF-IDF weighted n-gram count vectors for the sentences in the dataset.

3.2 Submodular Speech Summarization

Similar to the procedure described above we can treat the task of subselecting an acoustic data set as an extractive summarization problem. For our *a priori* data selection scenario we would like to extract those training samples that jointly are representative of the total data set. Initial explorations of submodular functions for speech data can be found in (Lin and Bilmes, 2009), where submodular functions were used in combination with a purely acoustic similarity measure (Fisher kernel). In addition Equation 2 the **facility location** function was used:

$$f_{fac}(S) = \sum_{i \in V} \max_{j \in S} w_{ij} \qquad (3)$$

Here our focus is on utilizing methods that move beyond purely acoustic similarity measures and consider kernels derived from discrete representations of the acoustic signal. To this end we first run a tokenizer over the acoustic signal that converts it into a sequence of discrete labels. In our case we use a simple bottom-up monophone recognizer (without higher-level constraints such as a phone language model) that produces phone labels. We then use the hypothesized sequence of phonetic labels to compute two different sentence similarity measures: (a) cosine similarity using TF-IDF weighted phone n-gram counts, and (b) string kernels. We compare their performance to that of the Fisher kernel as a purely acoustic similarity measure.

TF-IDF weighted cosine similarity

The cosine similarity between phone sequences s_i and s_j is computed as

$$\mathrm{sim}_{ij} = \frac{\sum_{w \in s_i} \mathrm{tf}_{w,s_i} \times \mathrm{tf}_{w,s_j} \times \mathrm{idf}_w^2}{\sqrt{\sum_{w \in s_i} \mathrm{tf}_{w,s_i}^2 \,\mathrm{idf}_w^2}\sqrt{\sum_{w \in s_j} \mathrm{tf}_{w,s_j}^2 \,\mathrm{idf}_w^2}} \tag{4}$$

where tf_{w,s_i} is the count of n-gram w in s_i and idf_w is the inverse document count of w (each sentence is a "document"). We use $n = 1, 2, 3$.

String kernel

The particular string kernel we use is a gapped, weighted subsequence kernel of the type described in (Rousu and Shawe-Taylor, 2005). Formally, we define a sentence s as a concatenation of symbols from a finite alphabet Σ (here the inventory of phones) and an embedding function from strings to feature vectors, $\phi : \Sigma^* \rightarrow \mathcal{H}$. The string kernel function $\mathcal{K}(s,t)$ computes the distance between the resulting vectors for two sentences s_i and s_j. The embedding function is defined as

$$\phi_u^k(s) := \sum_{\mathbf{i}:u=s(\mathbf{i})} \lambda^{|i|} \quad u \in \Sigma^k \tag{5}$$

where k is the maximum length of subsequences, $|i|$ is the length of i, and λ is a penalty parameter for each gap encountered in the subsequence. $\mathcal{K}$ is defined as

$$\mathcal{K}(s_i, s_j) = \sum_u \langle \phi_u(s_i), \phi_u(s_j) \rangle w_u \tag{6}$$

where w is a weight dependent on the length of u, $l(u)$. Finally, the kernel score is normalized by $\sqrt{\mathcal{K}(s_i,s_i) \cdot \mathcal{K}(s_j,s_j)}$ to discourage long sentences from being favored.

Fisher kernel

The Fisher kernel is based on the vector of derivatives U_X of the log-likelihood of the acoustic data (X) with respect to the parameters in the phone HMMs $\theta_1, ..., \theta_m$ for m models, having similarity score:

$$\mathrm{sim}_{ij} = (\max_{i',j'} d_{i'j'}) - d_{ij}, \text{ where } d_{ij} = ||U_i' - U_j'||_1,$$

$$U_X^\theta = \triangledown_\theta \log P(X|\theta), \text{ and } U_X' = U_X^{\theta_1} \circ U_x^{\theta_2}, ..., \circ U_X^{\theta_m}.$$

4 Data and Systems

We evaluate our approach on subselecting training data from the TIMIT corpus for training a phone recognizer. Although this not a large-scale data task, it is an appropriate proof-of-concept task for rapidly testing different combinations of submodular functions and similarity measures. Our goal is to focus on acoustic modeling only; we thus look at phone recognition performance and do not have to take into account potential interactions with a language model. We also chose a simple acoustic model, a monophone HMM recognizer, rather than a more powerful but computationally complex model in order to ensure quick experimental turnaround time. Note that the goal of this study is not to obtain the highest phone accuracy possible; what is important is the relative performance of the different subset selection methods, especially on small data subsets.

The sizes of the training, development and test data are 4620, 200 and 192 utterances, respectively. Preprocessing was done by extracting 39-dimensional MFCC feature vectors every 10 ms, with a window of 25.6ms. Speaker mean and variance normalization was applied. A 16-component Gaussian mixture monophone HMM system was trained on the full data set to generate parameters for the Fisher kernel and phone sequences for the string kernel and TF-IDF based similarity measures.

Following the selection of subsets (2.5%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80% of the data, measured as percentage of non-silence speech frames), we train a 3-state HMM monophone recognizer for all 48 TIMIT phone classes on the resulting sets and evaluate the performance on the core test set of 192 utterances, collapsing the 48 classes into 39 in line with standard practice (Lee and Hon, 1989). The HMM state output distributions are modeled by diagonal-covariance Gaussian mixtures with the number of Gaussians ranging between 4 and 64, depending on the data size.

As a baseline we perform 100 random draws of the specified subset sizes and average the results.

The second baseline consists of the method in (Wu et al., 2007), where utterances are selected to maximize the entropy of the distribution over phones in the selected subset.

5 Experiments

We tested the three different similarity measures described above in combination with the submodular functions in Equations 2 and 3. The parameters of the gapped string kernel (i.e. the kernel order (k), the gap penalty (λ), and the contiguous substring length l) were optimized on the development set. The best values were $\lambda = 0.1, k = 4, l = 3$. We found that facility location was superior to saturated cover function across the board.

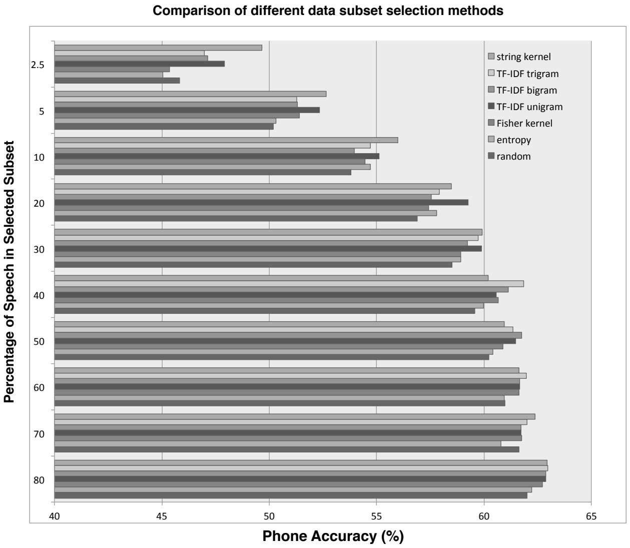

Figure 1: Phone accuracy for different subset sizes; each block of bars lists, from bottom to top: random baseline, entropy baseline, Fisher kernel, TF-IDF (unigram), TF-IDF (bigram), TF-IDF (trigram), string kernel.

Figure 1 shows the performance of the random and entropy-based baselines as well as the performance of the facility location function with different similarity measures. The entropy-based baseline beats the random baseline for most percentage cases but is otherwise the lowest-performing method overall. Note that this baseline uses the true transcriptions in line with (Wu et al., 2007) rather than the hypothesized phone labels output by our recognizer. The low performance and the fact that it is even outperformed by the random baseline in the 2.5% and 70% cases may be because the selection method encourages highly diverse but not very representative subsets. Furthermore, the entropy-based baseline utilizes a non-submodular objective function with a heuristic greedy search method. No theoretical guarantee of optimality can be made for the subset found by this method.

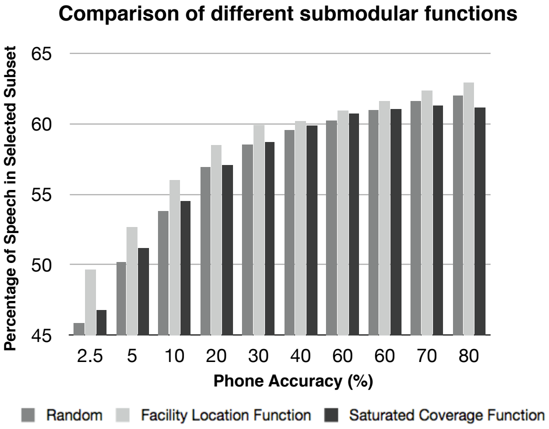

Figure 2: Phone accuracy obtained by random selection, facility location function, and saturated coverage function (string kernel similarity measure).

Among the different similarity measures the Fisher kernel outperforms the baseline methods but has lower performance than the TF-IDF kernel and the string kernel. The best performance is obtained with the string kernel, especially when using small training data sets (2.5%-10%). The submodular selection methods yield significant improvements ($p < 0.05$) over both the random baseline and over the entropy-based method.

We also investigated using different submodular functions, i.e. the facility location function and the saturated coverage function. Figure 2 shows the performance of the facility location (f_{fac}) and saturated coverage (f_{SC}) functions in combination with the string kernel similarity measure. The reason f_{fac} outperforms f_{SC} is that f_{SC} primarily controls for over-coverage of any element not in the subset via the α saturation hyper-parameter. However, it does not ensure that every non-selected element has good representation in the subset. f_{SC} measures the quality of the subset by how well each individual element outside the subset has a surrogate within the subset (via

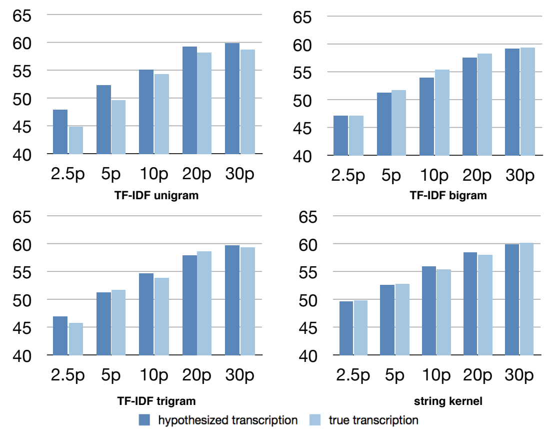

Figure 3: Phone accuracy for true vs. hypothesized phone labels, for string-based similarity measures.

the max function) and hence tends to model complete coverage better, leading to better results.

Finally we examined whether using hypothesized phone sequences vs. the true transcriptions has negative effects. Figure 3 shows that this is not the case: interestingly, the hypothesized labels even result in slightly better results. This may be because the recognized phone sequences are a function of both the underlying phonetic sequences that were spoken and the acoustic signal characteristics, such as the speaker and channel. The true transcriptions, on the other hand, are able to provide information only about phonetic as opposed to acoustic characteristics.

6 Discussion

We have presented a low-resource framework for acoustic data subset selection based on submodular function optimization, which was previously developed for document summarization. Evaluation on a proof-of-concept task has shown that the method is successful at selecting data subsets that outperform subsets selected randomly or by a previously proposed low-resource method. We note that the best selection strategies for the experimental conditions tested here involve similarity measures based on a discrete tokenization of the speech signal rather than direct acoustic similarity measures.

Acknowledgments

This material is based on research sponsored by Intelligence Advanced Research Projects Activity (IARPA) under agreement number FA8650-12-2-7263. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Intelligence Advanced Research Projects Activity (IARPA) or the U.S. Government.

References

B. Chen, S.H Liu, and F.H. Chu. 2009. Training data selection for improving discriminative training of acoustic models. *Pattern Recognition Letters*, 30:1228–1235.

J. Edmonds, 1970. *Combinatorial Structures and their Applications*, chapter Submodular functions, matroids and certain polyhedra, pages 69–87. Gordon and Breach.

G. Hakkani-Tur, G. Riccardi, and A. Gorin. 2002. Active learning for automatic speech recognition. In *Proc. of ICASSP*, pages 3904–3907.

N. Itoh, T.N. Sainath, D.N. Jiang, J. Zhou, and B. Ramabhadran. 2012. N-best entropy based data selection for acoustic modeling. In *Proceedings of ICASSP*.

Thomas Kemp and Alex Waibel. 1998. Unsupervised training of a speech recognizer using TV broadcasts. In *in Proceedings of the International Conference on Spoken Language Processing (ICSLP-98)*, pages 2207–2210.

A. Krause and C. Guestrin. 2011. Submodularity and its applications in optimized information gathering. *ACM Transactions on Intelligent Systems and Technology*, 2(4).

L. Lamel, J.L. Gauvain, and G. Adda. 2002. Lightly supervised and unsupervised acoustic model training. *Computer, Speech and Language*, 16:116 – 125.

K.F. Lee and H.W. Hon. 1989. Speaker-independent phone recognition using Hidden Markov Models. *IEEE Trans. ASSP*, 37:1641–1648.

Hui Lin and Jeff A. Bilmes. 2009. 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.

H. Lin and J. Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of ACL*.

R.K. Moore. 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of Eurospeech*, pages 2581–2584.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. 1978. An analysis of approximations for maximizing submodular functions-I. *Math. Program.*, 14:265–294.

J. Rousu and J. Shawe-Taylor. 2005. Efficien computation of of gapped substring kernels for large alphabets. *Journal of Machine Leaning Research*, 6:13231344.

Y. Wu, R. Zhang, and A. Rudnicky. 2007. Data selection for speech recognition. In *Proceedings of ASRU*.