SUBMODULAR SUBSET SELECTION FOR LARGE-SCALE SPEECH TRAINING DATA

Kai Wei[⋆] *Yuzong Liu*[⋆] *Katrin Kirchhoff*[⋆] *Chris Bartels*[†] *Jeff Bilmes*[⋆]

[⋆]Department of Electrical Engineering, University of Washington Seattle, WA, USA
[†]SRI International, Menlo Park, CA, USA

ABSTRACT

We address the problem of subselecting a large set of acoustic data to train automatic speech recognition (ASR) systems. To this end, we apply a novel data selection technique based on constrained submodular function maximization. Though NP-hard, the combinatorial optimization problem can be approximately solved by a simple and scalable greedy algorithm with constant-factor guarantees. We evaluate our approach by subselecting data from 1300 hours of conversational English telephone data to train two types large-vocabulary speech recognizers, one with Gaussian mixture model (GMM) based acoustic models, and another based on deep neural networks (DNNs). We show that training data can be reduced significantly, and that our technique outperforms both random selection and a previously proposed selection method utilizing comparable resources. Notably, using the submodular selection method, the DNN system using only about 5% of the training data is able to achieve performance on par with the GMM system using 100% of the training data — with the baseline subset selection methods, however, the DNN system is unable to accomplish this correspondence.

Index Terms— speech processing, automatic speech recognition, machine learning, large-scale systems

1. INTRODUCTION

Present-day automatic speech recognition (ASR) systems are trained on vast amounts of acoustic data. Although larger training data sets often lead to gains in system performance, there are well-known problems associated with ever-increasing data sets: First, larger data sets place greater demands on available resources, such as storage and CPU cycles. Second, existing software infrastructure often needs to be modified to be able to process ever-larger data sets, which requires developer time and expertise. Third, and most importantly, the gains in system performance achieved by increasing training data sets are often sublinear: after an initial increase the gains become smaller, a phenomenon known as diminishing returns. This is because additional data may be noisy, irrelevant to the task at hand or, most probably, fully or partially redundant with already existing data. An aggravating factor is that many statistical learning procedures (e.g., the Expectation-Maximization algorithm typically used for training Gaussian Mixtures Models (GMM), or the back-propagation algorithm for training neural networks) typically process training data sets repeatedly. Having unnecessary and redundant data thus results in wasted computational resources. Therefore, a critical goal is to develop methods to select an informative and representative subset of a large data set that retains as many of the benefits of the large data set as possible, while simultaneously minimizing resource requirements.

Data subset selection can be conducted for several different scenarios that frequently occur in the field of speech processing: (a) identifying a smaller subset of the data that fits a given budget but provides as much information as the original large data set; (b) selection for adaptation, where the goal is to tune a system to a known development or test set; (c) data selection for human annotation (e.g., batch active learning), which is of interest when developing ASR systems for new languages or dialects whose audio data has not yet been transcribed. In this work we investigate the first scenario: given a large amount of acoustic data (> 1000 hours) our goal is to select a subset of the data that fits a given budget (maximum number of hours of speech) but loses as little information as possible. We, in particular, are interested in studying the following question: given a drastic reduction in training set size (from one to two orders of magnitude), what is the smallest degree of information loss possible? The intended benefit is to significantly shorten experimental turn-around time – often, systems need to be trained repeatedly with different model configurations or parameters. If this could be done on a small subset, more experimentation could be performed within a given time period, and systems could be more highly and accurately tuned.

Our proposed approach is based on submodular function maximization. Submodular functions, often used in economics, operations research, or (more recently) machine learning, have the property of diminishing returns. Certain subclasses of submodular functions can be optimized easily with theoretical performance guarantees. The optimization algorithms, moreover and very importantly, are scalable to very large data sets. Submodular data subset selection was presented in [1, 2] for a small-scale phonetic recognition task on TIMIT. A companion paper [3] also utilizes small-scale phonetic recognition (again TIMIT) but in a purely unsupervised fashion (no training labels are used). In this paper we significantly extend this approach and apply it to *large-vocabulary* speech recognition under two different acoustic modeling paradigms: hidden Markov models with Gaussian mixture output distributions (HMM-GMMs), and the more recent and state-of-the-art deep neural network (DNN) approach. In this paper, our approach is supervised (i.e., we utilize information derived from the utterances' word-level transcriptions). Moreover, we introduce and utilize a type of submodular function based on a feature representation of speech utterances that scales to very large data sets.

In the following sections, we first summarize related previous work (Section 2) and then explain our submodular data selection approach (Section 3). Section 4 then describes data sets and systems used for our experiments. Section 5 provides experimental results, and Section 6 concludes.

2. RELATED WORK

Most approaches to data subset selection in speech recognition are based on the framework of active learning, where additional training data is chosen to update an already existing system [4, 5, 6]. Under this approach the utility of each sample in the training data set is equivalent to the confidence score given by an existing ASR system. All samples are then ranked according to their utility score, and

the samples with highest scores are selected. The drawback of this approach is that it requires a fully-trained speech recognizer with reasonable performance that repeatedly iterates over the training set. In [7], data selection at different levels (e.g., utterance level, phone level, and frame-level) is performed for the purpose of discriminative training of acoustic models. The selection criterion is the average phone accuracy of utterances. In [8], they propose a method for data subset selection that meets the criteria of both informativeness and representativeness. The informative score of an utterance is computed as the entropy of its N-best word hypotheses produced by a baseline ASR system. The representative score of an utterance with respect to a data pool is calculated as the average TF-IDF similarity with all other utterances in the pool. Like the active selection methods, this approach requires a word recognizer. In [9], they propose a method to subselect acoustic training data based on the transcriptions of the data. Their objective is to select a subset of the data that results in a maximum-entropy distribution over linguistic units (e.g., phones or words) in the set. The entropy of the distribution is computed from the true transcriptions; thus, this method does not require an existing system but does require gold-standard annotations.

While the above methods select data using a greedy algorithm, in general there is not a guarantee that such an algorithm applied to the above objectives has any quality assurance. As we will see below, the greedy algorithm has theoretical approximation guarantees with respect to our objectives when they are formulated as monotone submodular functions. We also will argue that submodularity is a natural model for this problem.

3. SUBMODULARITY & DATA SUBSET SELECTION

Submodular functions have traditionally been used in economics and operations research; recently, they have also become popular in machine learning as they represent natural models of many real-world combinatorial selection problems.

A *submodular* function [10] is defined as follows: Given a finite set $V = \{1, 2, \ldots, n\}$, a discrete set function $f : 2^V \rightarrow \mathbb{R}$ that returns a real value for any subset $S \subseteq V$ is submodular if

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B), \ \forall A, B \subseteq V. \quad (1)$$

Defining $f(j|S) \triangleq f(j \cup S) - f(S)$, then an equivalent definition of submodularity is $f(j|S) \geq f(j|T), \forall S \subseteq T$. That is, the incremental gain of adding item j to the set decreases when the set in which j is considered grows from S to T. A submodular function f is *monotone non-decreasing* if

$$f(j|S) \geq 0, \forall j \in V \setminus S, S \subseteq V \quad (2)$$

We say that f is *normalized* if $f(\emptyset) = 0$. Powerful guarantees exist for subtypes of monotone submodular function maximization. Though NP-hard, the problem of maximizing a monotone submodular function subject to a cardinality constraint can be approximately solved by a simple greedy algorithm [11] with a worst-case approximation factor $(1 - e^{-1})$. This is also the best solution obtainable in polynomial time unless P=NP [12]. The greedy algorithm starts with the empty set $S_0 = \emptyset$, and at each iteration i, adds the element s_i that maximizes the conditional gain $f(s_i|S_{i-1})$ with ties broken arbitrarily, (i.e. finding $s_i \in \text{argmax}_{e \in V \setminus S_{i-1}} f(e|S_{i-1})$), and then updates with $S_i \leftarrow S_{i-1} \cup \{s_i\}$. The algorithm stops after k iterations, where k is the cardinality constraint. There is also an accelerated greedy implementation with an almost linear time complexity [13]. This is one of the key reasons why the approach for data subset selection is efficient and can scale to very large data sets.

3.1. Large Scale Speech Data Subset Selection

Prior investigations of submodular speech data subset selection are found in [1, 2], albeit only for very small datasets. Handling large-scale data sets requires different objectives.

Given a set of N utterances $V = \{1, 2, \ldots, N\}$ we must construct a non-decreasing submodular set function $f : 2^V \rightarrow \mathbb{R}$, mapping each subset $S \subseteq V$ to a real number that represents the value of subset S. We can formulate the problem of selecting the best subset S given some budget (maximum number of hours of speech) as monotone submodular function maximization under a knapsack constraint:

$$\max_{S \subseteq V, c(S) \leq B} f(S) \quad (3)$$

where B is the budget and $c(S) = \sum_{j \in S} c(j)$ is a cost function that measures the amount of speech contained in subset S, with $c(j)$ being the length of the utterance $j \in V$. Notice that the cardinality constraint is a special case of the knapsack constraint where $c(j) = 1, \forall j \in V$. The same scalable greedy algorithm described above can be easily generalized to approximately solve Problem (3) with similar theoretical guarantee [14]. The function $f(S)$ may take various forms. In [2] two functions were used, the first of which is the **facility location** function, defined as:

$$f_{\text{fac}}(S) = \sum_{i \in V} \max_{j \in S} w_{ij} \quad (4)$$

where $w_{ij} \geq 0$ indicates the similarity between utterance i and j. The similarity measure w_{ij} is computed by kernels derived from discrete representations of the acoustic utterance i and j. More specifically, a tokenizer is run over the acoustic signal that converts it into a sequence of discrete labels. Then a TF-IDF kernel or string kernel is used to compute the pair-wise similarity between the sequences of discrete labels of two speech utterances. A second function can be called **saturated coverage**, and is defined as follows:

$$f_{\text{sat}}(S) = \sum_{i \in V} \min\{C_i(S), \alpha C_i(V)\} \quad (5)$$

where $C_i(S) = \sum_{j \in S} w_{ij}$ and $0 \leq \alpha \leq 1$ is a saturation coefficient. We refer to both of these functions as graph-based submodular functions since a pair-wise similarity graph is required, i.e., w_{ij} needs to be computed for all $i \in V$ and $j \in V$. This has a time complexity of $O(|V|^2)$ and a memory complexity of $O(|V|^2)$. In our task of large-scale speech data subset selection, the whole speech corpus is segmented into about 1.3 million individual segments; thus $|V| \approx 1.3e7$. Even with highly optimized data structures, efficient computation of similarity measures, and nearest neighbor graph approximation, graph construction presents a computational challenge (avoidable, as we will soon see) for the application of such graph-based submodular objectives to large-scale speech data subset selection.

We now introduce an alternative class of submodular functions that avoid the use of a pair-wise similarity graph. We call these **feature-based submodular** functions:

$$f_{\text{fea}}(S) = \sum_{u \in \mathcal{U}} g(m_u(S)) \quad (6)$$

where $g()$ is a non-negative monotone non-decreasing concave function, $\mathcal{U}$ is a set of features, and $m_u(S) = \sum_{j \in S} m_u(j)$ is a non-negative score for feature u in set S, with $m_u(j)$ measuring the degree to which utterance $j \in S$ possesses feature u. Maximizing

this objective naturally encourages diversity and coverage of the features within the chosen set of elements. We note that Equation (6) is a sum of concave functions over modular functions, and is easily shown to be submodular [15] — our novel take on this is to have each term in the sum be based on a "feature" of the objects being scored. The feature based submodular functions are convenient for applications in speech processing since speech objects can often be described by a variety of phonetic or prosodic feature labels (e.g. phonemes, triphones, words, syllables, tones, etc.). Feature-based submodular functions, therefore, have the ability to leverage much of the important work on both knowledge- and data-driven feature engineering that has been available in speech processing. In our work, $\mathcal{U}$ is the set of triphones over frame labels that are derived from the word transcriptions via a forced Viterbi alignment of a trained system. The function $g()$ is the square root function. The score $m_u(s)$ is the count of feature u in element s, normalized by term frequency-inverse document frequency (TF-IDF), i.e., $m_u(s) = \text{TF}_u(s) \times \text{IDF}_u(s)$, where $\text{TF}_u(s)$ is the count of feature u in s, and $\text{IDF}_u = \log(\frac{|V|}{d(u)})$ is the inverse document count of the feature u with $d(u)$ being the number of utterances that contain the feature u (each utterance is considered a "document").

4. DATA AND SYSTEMS

We evaluate our approach on selecting subsets from 1300 hours of conversational English telephone data from the Switchboard, Switchboard Cellular, and Fisher corpora. We train a separate ASR system on each resulting subset, where the sizes are chosen to be significantly smaller than the whole (namely, 1%, 5%, 10% and 20% of the whole non-silence training data). The development and test data sets are unchanged, and are the 2001 and 2002 NIST Rich Transcription development sets, with 2.2 hours and 6.3 hours of acoustic data, respectively. Two different ASR systems were used for our experiments, distinguished by their acoustic modeling approach. The first is SRI's DECIPHER system (Stolcke et al., 2000). The preprocessing component extracts 13-dimensional mel-frequency cepstral coefficients (MFCCs) along with their 1st, 2nd, and 3rd order derivatives. The resulting 52-dimensional feature vectors are mean and variance normalized, and reduced to 39 dimensions by heteroscedastic Linear Discriminant Analysis (HLDA) [16]. The features are then discriminatively transformed using feature minimum phone error training (fMPE) [17]. Acoustic models consist of three-state left-to-right HMMs with GMMs as output probability distributions. Each GMM represents a decision-tree clustered cross-word triphone state. GMMs are first estimated using the maximum likelihood criterion and are used to generate phone lattices, which are utilized for minimum phone error training (MPE) [18] to create the final models. During decoding, a first-pass search is performed using a bigram language model. A recognition pass using maximum-likelihood linear regression (MLLR) speaker-adapted acoustic models generates a set of lattices. Finally, these lattices are rescored with a trigram language model to generate the final output.

The second system was also developed at SRI and utilizes DNNs as acoustic models and Kaldi [19] as the decoder. The inputs to the DNN consist of 15 frames of 40 dimensional Mel-scaled filter-bank outputs. The DNN targets are decision-tree clustered triphones with approximately 3750 targets for data subsets and 7800 for the full set.[1] The number of layers in each network was tuned based on the development set word error rate. As before, a first pass search is performed during decoding using a bigram language model, and the resulting lattices are rescored using a trigram language model. The language model (LM) is the same in both systems and consists of an interpolation of various genre-specific LMs trained on meeting transcriptions, spontaneous telephone speech, broadcast news, and web data selected to match the transcribed data. The interpolation weights are optimized on a held-out set of meeting data.

[1]To clarify this point, the process of training a system using a subset (rather than all) of the training data was optimized in an attempt to get the best performance possible for any given amount of available data. Included in this process was the number of decision-tree clustered triphones for a given subset size.

	1%	5%	10%	20%	all
Rand	52.1± 1.5	38.2±0.2	35.1±0.3	34.4±0.2	31.0
HE (words)	49.6	36.5	34.8	N/A	
HE (3-phones)	47.5	37.6	34.2	N/A	
SM (3-phones)	**47.5**	**35.7**	**33.3**	**32.6**	

Table 1. Word error rates for the HMM-GMM system, for subsets of various sizes chosen by the random (Rand), histogram-entropy (HE), and the submodular (SM) selection method. The histogram-entropy results for the 20% condition are not available due to that objective's saturation after 10%.

	1%	5%	10%	20%	all
Rand	43.7±0.5	34.3±0.9	31.5±0.5	29.8±0.2	26.0
HE (3-phones)	42.8	33.9	31.3	N/A	
SM (3-phones)	**41.1**	**31.8**	**29.3**	**28.2**	

Table 2. Word error rates for the DNN system, for the random, histogram-entropy (HE) and the submodular (SM) training data subset selection methods.

5. EXPERIMENTS AND RESULTS

5.1. Baselines

We compare our submodular data selection approach against two different baseline methods: a random sampling baseline, and the "histogram-entropy" based method described in [9], which is not submodular but assumes a comparable level of existing resources. For the random baseline we randomly sample data sets at the appropriate size (1%, 5%, 10%, or 20%), train different ASR systems for each set, and average their word error rates.

For the histogram-entropy baseline, the objective is to select a subset of the data that results in a maximum-entropy distribution over linguistic units (e.g., triphone states or words) in the set. We implemented two variants of the baseline: (a) using the words from the true word transcriptions as phonetic units, as described in [9], and (b) using the triphone state labels from a forced alignment of the transcriptions to the acoustic data. For our experiments the fully-trained Decipher system was used for the forced alignment; note, however, that it is in principle also possible to use different acoustic models, or to perform an unsupervised tokenization of the acoustic signal. Also note that for the histogram-entropy method the objective criterion (maximum entropy of the distribution over phonetic units in the data) may saturate (i.e., no further increase is possible) before the budget constraint is reached. We found that this was the case for the 20% subset, i.e., the entropy saturates before 20% of the data has been selected. To reach 20%, more data would have to be added randomly, which would render the method largely equivalent to the random selection baseline. Results are unavailable for the histogram-entropy at 20% level.

5.2. Submodular method

All experiments were conducted with the feature-based submodular function (Equation (6)). Several experiments were run with different instantiations of feature sets (words, triphones, triphone HMM state ids, and n-grams thereof), as well as different ways of normalizing feature counts and different concave functions. The best results obtained with submodular selection were based on TF-IDF weighted counts of triphones as the $m_u(s)$ scores, where the triphone annotation was also based on a forced-alignment of the word transcriptions to the signal. For each system and subset, the complexity of the acoustic model (number of initial clusters for bootstrapping the acoustic model, and the number of leaves in the decision tree used for state clustering) was optimized on the development set. Optimizing the number of parameters is important since data sets with greater inherent complexity can in theory support more parameters in the acoustic model, whereas inherently redundant data sets might lead to poorly trained models if too many parameters are utilized.

Table 1 shows the results for the HMM-GMM system. The random results are an average over four independent subsets for each size percentage, and are shown as mean $\pm$ standard_deviation. We see that our proposed method outperforms all baseline systems under all conditions.

Results for the DNN system are shown in Table 2. In this case, owing to the longer training time, the random results are an average of three subsets for each size percentage. The histogram-entropy baseline and the submodular systems followed the same design as in the previous set of experiments. For all systems, the number of parameters (layers in the deep neural network) was optimized on the development set; it varied between three and six (with most systems having five layers). The number of hidden units was 1200 in each case. Here, too, the results shows that the submodular method clearly outperforms both baseline methods in all cases.

Comparing Table 1 to Table 2, moreover, shows an interesting trend. The submodular selected subset at 5% achieves a WER of 31.8% with the deep model system, while the HMM-GMM system using 100% of the training data achieves a WER of 31.0%. That is, the deep system using the "right" 5% of the training data almost matches the HMM-GMM system using all of the data. Note that the deep system is unable to do that, however, using the baseline methods for choosing 5% of the training data. Also, the deep system using only 10% of the submodular subselected data achieves a result (i.e., 29.3%) that is strictly better than the GMM system at 100% of the training data (i.e., 31.0%). These results, therefore, offer evidence of the combined power of a properly chosen subset of the training data (using a submodular function) and a modern speech recognition system (deep model based). Moreover, with such a large reduction (at 5%), the time spent training a deep system is approximately $20\times$ faster, which could lead to many more model variants being investigated in the same amount of time.

6. CONCLUSIONS

We have presented a framework for subselecting large-scale acoustic data based on submodular function optimization. Different from prior submodular work on speech data subset selection where the construction of a similarity graph is required, we proposed an alternative feature-based submodular objective that performs well but does not require the construction of an $O(N^2)$ similarity graph. For both acoustic modeling approaches investigated here, Gaussian mixture and deep neural network, our best submodular function leads to better performance than either baseline selection scheme. Results seem particularly good for the deep system, which is encouraging given the recent success deep neural networks have had improving speech recognition systems.

Future work will concentrate on subselecting large complex speech data using an unsupervised model to tokenize the data and generate feature labels, thus eliminating the need for transcriptions. Moreover, while in this work we concentrated on finding the best subset of training data at drastically reduced sizes (from one to two orders of magnitude), future work will also concentrate on the task of finding the smallest training data subset that loses no information relative to the whole data set. We also plan to work on extending the strictly unsupervised methods mentioned in our companion paper [3] to the large-vocabulary setting.

7. 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.

8. REFERENCES

[1] 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.

[2] 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.

[3] Kai Wei, Yuzong Liu, Katrin Kirchhoff, and Jeff Bilmes, "Unsupervised submodular subset selection for speech data," Florence, Italy, 2014.

[4] 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.

[5] 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.

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

[7] Shih-Hung Liu, Fang-Hui Chu, Shih-Hsiang Lin, Hung-Shin Lee, and Berlin Chen, "Training data selection for improving discriminative training of acoustic models," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 2007, pp. 284–289.

[8] 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.

[9] 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.

[10] Satoru Fujishige, "Submodular systems and related topics," *Mathematical Programming at Oberwolfach II*, pp. 113–131, 1984.

[11] 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.

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

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

[14] 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.

[15] P. Stobbe and A. Krause, "Efficient minimization of decomposable submodular functions," in *NIPS*, 2010.

[16] Nagendra Kumar and Andreas G Andreou, *Investigation of silicon auditory models and generalization of linear discriminant analysis for improved speech recognition*, Ph.D. thesis, Johns Hopkins University, 1997.

[17] Daniel Povey, Brian Kingsbury, Lidia Mangu, George Saon, Hagen Soltau, and Geoffrey Zweig, "fMPE: Discriminatively trained features for speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, vol. 1, pp. 961–964.

[18] Daniel Povey and Philip C Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, vol. 1, pp. I–105.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.