This is a repository copy of *Data-Selective Transfer Learning for Multi-Domain Speech Recognition*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/92451/

Version: Published Version

Proceedings Paper:
Doulaty Bashkand, M., Saz, O. and Hain, T. (2015) Data-Selective Transfer Learning for Multi-Domain Speech Recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 16th Annual Conference of the International Speech Communication Association, 06-10 Sep 2015, Dresden, Germany. ISCA (International Speech Communication Association) , pp. 2897-2901.

Data–selective Transfer Learning for Multi–Domain Speech Recognition

Mortaza Doulaty, Oscar Saz, Thomas Hain

Speech and Hearing Group, University of Sheffield, Sheffield, UK

{mortaza.doulaty, o.saztorralba, t.hain}@sheffield.ac.uk

Abstract

Negative transfer in training of acoustic models for automatic speech recognition has been reported in several contexts such as domain change or speaker characteristics. This paper proposes a novel technique to overcome negative transfer by efficient selection of speech data for acoustic model training. Here data is chosen on relevance for a specific target. A submodular function based on likelihood ratios is used to determine how acoustically similar each training utterance is to a target test set. The approach is evaluated on a wide–domain data set, covering speech from radio and TV broadcasts, telephone conversations, meetings, lectures and read speech. Experiments demonstrate that the proposed technique both finds relevant data and limits negative transfer. Results on a 6–hour test set show a relative improvement of 4% with data selection over using all data in PLP based models, and 2% with DNN features.

Index Terms: data selection, transfer learning, negative transfer, speech recognition

1. Introduction

As Automatic Speech Recognition (ASR) systems improve their accuracy, new applications and domains become the target of research. Automatic transcription of speech with unknown origin is a challenging task, which is related to access to so–called "found data", such as media and historical audio archives. For this to be feasible, ASR has to produce an accurate output for whichever the conditions contained in the target data (e.g. interviews, distant recordings, telephone conversations, etc). Training acoustic models for an unknown domain, e.g. YouTube recordings, can be infeasible if the origin of the target speech can not be properly assessed, and the loss of accuracy can be large due to wrong modelling decisions. Another option is to train an acoustic model on a large amount of data from multiple domains, although this is not guaranteed to give the most optimal results.

Maximum Likelihood Estimation (MLE) of Gaussian Mixture Model (GMM) parameters of a Hidden Markov Model (HMM) is still a standard approach to train acoustic models in ASR, either with perceptually–based features like Perceptual Linear Prediction (PLP) features [1], or with Deep Neural Network (DNN) based features [2] in tandem configuration. However, MLE has two well–known requirements: first, model correctness is assumed; and second the amount of training data is required to be infinite [3]. None of the above are valid in standard situations in ASR, although systems are sometimes trained with many years of speech data (e.g [4]). However, adding more data does not guarantee that the performance of the system will improve, and even if it does, the gains become smaller and smaller [5]. A further effect, negative transfer, is found in several examples, which indicates that knowledge acquired for a task can have a negative performance effect in another task

[6]. As a result, being able to select informative training data remains an important task.

This paper studies positive and negative transfer in ASR in a multi–domain scenario. The work proposes to use submodular functions based on acoustic similarity between the target test set and training data, in which positive transfer will be exploited to improve performance across domains, while reducing the impact of negative transfer at the same time. Submodular functions have been successfully used before to select data in semisupervised training and active learning for ASR tasks [7, 8]. However, here we show that these can also be used to select acoustically matching data in an un–supervised manner.

This paper is structured as follows: Section 2 provides a review of data selection techniques for ASR, and Section 3 introduces the proposed approach for data selection. Section 4 describes the experimental setup, followed by results and analysis in Section 5. The final Section 6 summarises and concludes the paper.

2. Data selection for ASR

Data selection for ASR has mostly been studied for minimal representative data selection [5, 8, 9, 7, 10, 11, 4, 12, 13]. Here the objective is, given a large pool of training data, to find a subset of data such that a model set trained with that data will achieve comparable performance to a model set trained with all the data. This line of work is related to active learning, where the aim is to select a subset for manual transcription with the least budget [14, 15, 16], and with unsupervised and semi–supervised learning techniques, where the overall objective is to select a subset of the training set with the most reliable available transcripts [17, 18, 19].

Two techniques are typically used for selecting data: *uncertainty sampling* [20], where the scores from an existing model are used to choose or reject data; and *query by committee* [21], where votes of distinctly trained models are used [7]. For uncertainty sampling two types of scores have been explored. Confidence scores are used to select data with the most reliable transcriptions, as in semi–supervised training [17, 4], or to select data for manual transcription in active learning [15, 14]. Entropy–based methods aim to pick data that, for instance, fits a uniform distribution of target units (phonemes, words, etc), resulting in maximum entropy [7, 10, 9] or having a similar distribution to a target set [12, 13, 19].

The use of submodular functions has been proposed to tackle the effect of the diminishing returns, when adding more data to a training set [8, 5, 7]. A submodular function is defined as any function $f : 2^\Omega \to \mathbb{R}$ that fulfils

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T), \forall S, T \subseteq \Omega \quad (1)$$

With submodular functions the problem of data selection turns into a submodular maximisation problem, where the ob-

jective is to find a subset S from the complete training set Ω so that any new subset T added to S will not increase the value of the submodular function f:

$$\operatorname*{argmax}_{S \subseteq \Omega} \{ f(S) | f(S \cup T) < f(S),\ T \subseteq \Omega \setminus S \} \qquad (2)$$

Finding S is an NP–hard problem [22, 8] and greedy solutions are proposed where the subset S is increased iteratively by the item $s \in \Omega$ that maximises the value of f when added to S as in Equation 3.

$$s = \operatorname*{argmax}_{s \in \Omega \setminus S} \{ f(S \cup \{s\}) \} \qquad (3)$$

The set S is obtained when either the optimal S is found $(f(S) > f(S \cup \{s\}))$, or a budget N is reached $(|S| \leq N)$.

If the function f is a normalised monotone submodular function, then the simple greedy algorithm provides a good approximation of the optimal solution [23, 22, 7]

Several functions f can be found in the literature to perform data selection for ASR tasks, including facility location functions, saturated coverage functions [24, 8], diversity reward functions [5] or graph cut functions [7].

3. Likelihood ratio data selection

To decide whether data bears resemblance to a training set, one can opt for a classification approach that identifies an item to be suitable or not. Here we make use of the Likelihood Ratio (LR) between a GMM trained on the target data (Θ_{tgt}), and a GMM trained on the complete training set (Θ_Ω). The total LR of an utterance in the training set $LR(\mathcal{O}), \mathcal{O} \in \Omega$ of length T frames is defined as the geometric mean of the frame–based LR values of the target data model Θ_{tgt} and the background model Θ_Ω, assuming frame independence.

$$LR(\mathcal{O}) = \frac{1}{T} \sum_{t=1}^{T} \frac{p(\mathcal{O}_t | \Theta_{tgt})}{p(\mathcal{O}_t | \Theta_\Omega)} \qquad (4)$$

One can define a modular function [22] based on the accumulated LRs of all utterances included in a subset $S \subseteq \Omega$ in the following form:

$$f_{LR}(S) = \sum_{\mathcal{O} \in S} \Big(LR(\mathcal{O}) \Big). \qquad (5)$$

Modular functions are a special case of submodular functions [22] where the greater than or equal sign in Equation 1 changes to the equal sign. This way, the proposed function f_{LR} is submodular as well. And since all of the values for LR are non–negative, and therefore any sum of these numbers, as constituted by the function f, the function is necessarily monotonic with expanding sets ($A \subseteq B \subseteq \Omega, f(A) \leq f(B)$). If a submodular function is non–decreasing and normalised ($f(\emptyset) = 0$), then the greedy solution obtained by Equation 3 is no worse than the optimal value by a constant fraction $(1 - 1/e)$ [23]. Thus the subset S (greedy solution) can be used as the training set. The stopping criterion for adding more data to this subset S is based on a "budget", in the form of a maximum amount of hours of speech to be used.

4. Experimental setup

To evaluate the proposed approach in a multi–domain ASR task, a data set combining 6 different types of data was chosen from the following sources:

- Radio (RD): BBC Radio4 broadcasts on February 2009.
- Television (TV): Broadcasts from BBC on May 2008.
- Telephone speech (CT): From the Fisher corpus[1] [25].
- Meetings (MT): From AMI [26] and ICSI [27] corpora.
- Lectures (TK): From TedTalks [28].
- Read speech (RS): From the WSJCAM0 corpus [29].

A subset of 10h from each domain was selected to form the training set (60h in total), and 1h from each domain was used for testing (6h in total). The selection of the domains aims to cover the most common and distinctive types of audio recordings used in ASR tasks.

Two types of acoustic features were used: first, 13 PLP features plus first and second derivatives for a total of 39–dimensional feature vectors; and second, a 65–dimensional feature vector concatenating the 39 PLP features and 26 bottleneck (BN) features extracted from a 4–hidden–layer DNN trained on the full 60 hours of data. 31 adjacent frames (15 frames to the left and 15 frames to the right) of 23 dimensional log Mel filter bank features were concatenated to form a 713–dimensional super vector; Discrete Cosine Transform (DCT) was applied to this super vector to de–correlate and compress it to 368 dimensions and then fed into the neural network. The network was trained on 4,000 triphone state targets and the 26 dimensional bottleneck layer was placed before the output layer. The objective function used for training was frame–level cross–entropy and the optimisation was performed with stochastic gradient descent using the backpropagation algorithm. DNN training was performed with the TNet toolkit [30] and more details can be found at [31]. For both types of features, MLE–based GMM–HMM models were trained using HTK [32] with 5–state cross-word triphones and 16 gaussians per state. The language model was based on a 50,000–word vocabulary and was trained by combination of component language models for each of the 6 domains. The interpolation weights were tuned using an independent development set.

4.1. Baseline results

Table 1 presents results using both types of acoustic features. These results show the large variety in performance among domains, from 17–18% for read speech and radio broadcasts to 51% for television broadcasts. The use of DNN front–ends provides a 25% relative improvement in performance against PLP features; which is consistent across domains and follows results previously seen in the literature [33].

Table 1: WER (%) of models trained on full set

Features	RD	TV	CT	MT	TK	RS	Total
PLP	18.4	51.1	46.6	44.0	34.1	17.3	**36.0**
PLP+BN	13.3	42.0	33.5	32.2	23.5	13.0	**26.8**

5. Results

An initial set of experiments was conducted to identify and measure negative transfer in ASR tasks, and an evaluation of the proposed data selection technique was performed.

---

[1]All of the telephone speech data was up–sampled to 16 kHz to match the sampling rate of the rest of the data.
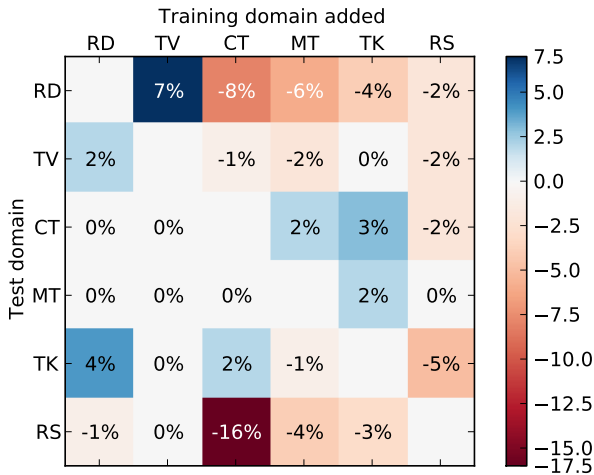
Figure 1: *Relative WER improvement by adding cross–domain data to in–domain models*

Figure 2: *WER improvement with budget–based data selection*

Figure 3: *Types of data selected for a 10–hour budget*

5.1. Evaluation of negative transfer

Six different domain–dependent MLE models were trained from the 10 hours of training data for each domain (in all of the experiments PLP features were used, unless stated otherwise). Each of these models was then used to decode the complete test set. The results in Table 2 show that in–domain results (when the train and test data match based on manually labelled domains) are not greatly different from those obtained with a model trained on 60–hour training set. Instead, cross–domain scores (train and test are mismatched) result in considerable performance decreases everywhere.

Table 2: WER (%) with domain specific acoustic models using PLP features)

Domain	RD	TV	CT	MT	TK	RS	Total
RD	**19.1**	55.1	72.1	57.2	50.7	24.9	47.8
TV	26.5	**52.9**	77.3	63.8	52.1	35.2	52.5
CT	82.3	90.1	**44.4**	71.9	67.9	86.6	72.6
MT	44.9	72.3	69.2	**44.0**	51.1	41.1	54.7
TK	39.8	62.8	69.3	56.1	**35.1**	55.4	53.6
RS	29.9	66.2	84.1	67.2	68.9	**16.9**	57.4

A second set of experiments was performed with models trained on 20 hours of data, using data from every possible pair of domains, for a total of 30 new acoustic models. Figure 1 shows the results in terms of relative improvement and degradation over the results of the 10–hour in–domain models. The rows of Figure 1 represent the testing domain and the columns represent the domain that was added in training to the data of the domain of the row. Positive values (blue squares) mark the existence of positive transfer, such as adding TV data to Radio data (7% improvement) or adding Radio data to Lecture data (4% improvement). But negative values (red squares) mark negative transfer, like adding Telephone data to Read speech (16% loss) or adding Read speech to Lecture data (5% loss).

These results showed that positive and negative transfer occurred across domains, possibly due to similarities and differences in speech styles, acoustic channels and background conditions. However a rule–based optimisation of the best model for each target domain would require a complex and error–prone
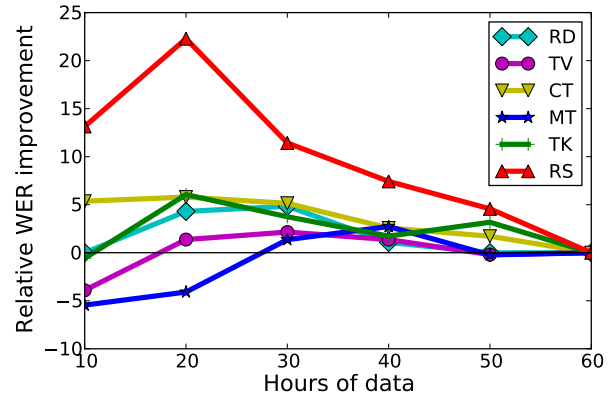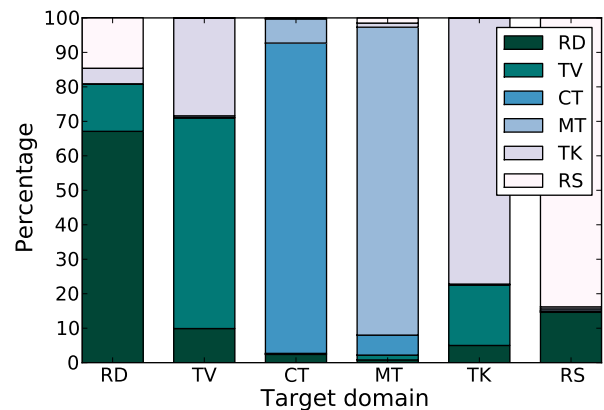
process. The next experiments aimed to evaluate how an automatic selection of training could exploit positive transfer, while restricting negative transfer.

5.2. Data selection based on budget

The data selection technique proposed in Section 3 was evaluated next. For each of the six target test domains, Gaussian Mixture Models (GMM) with 512 mixtures were trained ($\Theta_{tgt_{1:6}}$), and a background 512–mixture GMM (Θ_Ω) was trained from the complete training set of 60 hours. These GMMs were used to calculate the LR value for each training utterance ($LR(\mathcal{O})$) in order select the training data according to the acoustic similarity.

The first evaluation was performed using data selection based on budget. Five possible budgets of 10, 20, 30, 40 and 50 hours were designed for each test domain and the respective training data was chosen using the $f_{LR}(S)$ submodular function. Figure 2 shows relative improvement for each domain and budget against the results with the 60–hour model. The graphs show that all domains improve performance as the budget increases until a certain limit is reached, then negative transfer decreases the performance, converging to the WER achieved with the 60–hour trained model.

In order to observe which types of data were selected for each domain with the different budgets, Figure 3 presents the percentage of training data selected for each test domain with a 10–hour budget. While the majority of the data was chosen

from the same domain, some cross–domain data was also selected, indicating positive transfer between domains. This occurred, for instance, with TV and Read speech data towards Radio data; and Lecture data towards TV data.

5.3. Automatic decision on budget

An issue that can arise with the evaluated budget–based proposal is the fact that a decision on a budget has to be made, and as the results in Figure 2 suggest, the optimal budget varies across different domains. A method for deciding a budget for a given target domain was proposed by selecting only utterances whose likelihood–ratio is above a threshold defined as the mean of the highest–weighted mixture of a GMM fitted to the distribution of likelihood ratios. The use of the mixture with the highest weight avoids the influence of outliers in the distribution of the LR values.

The experiments with an automatic budget decision were performed for both types of features, PLP and PLP+BN. Table 3 presents the results for these experiments and compares them to the outcome of data selection based on a 30–hour budget, which was the best fixed budget from Figure 2. The results showed that the use of an automatically derived threshold improved the results obtained with a fixed budget for both types of features, indicating that the proposed method could estimate the right amount of data to select for each target domain.

Table 3: WER(%) using data selection

Method	RD	TV	CT	MT	TK	RS	Total
PLP features							
Budget–30h.	17.7	50.0	44.2	43.4	33.4	15.5	**34.9**
Auto. Decision	17.7	49.7	44.2	43.8	32.9	15.1	**34.7**
PLP+BN features							
Budget–30h.	13.0	41.5	32.6	32.1	22.5	12.1	**26.3**
Auto. Decision	12.7	41.4	32.5	32.3	22.4	11.8	**26.2**

The amount of data selected for each domain is presented in Table 4. This Table shows how Read speech and Conversational Telephone speech are the ones which benefited from a lower amount of training data (20 hours or less), while the rest of the domains preferred more data (from 30 to 40 hours). These values were consistent with the patterns of positive and negative transfer observed in Figure 2.

Table 4: Hours of data selected by automatic budget decision

Domain	RD	TV	CT	MT	TK	RS
Hours	41.2	35.8	21.9	35.6	31.4	17.1

6. Conclusion

In this paper, the effect of positive and negative transfer across widely diverse domains in ASR was explored. We confirmed that the use of more data in MLE–based acoustic models does not always provide increases in performance. A submodular function based on Likelihood Ratio was proposed and used to perform an informed and efficient selection of data for different target test sets. The evaluation of selection techniques based on budget and on automatic budget decision has achieved gains of 4% over a 60–hour MLE model for PLP features and 2% for PLP+BN features.

Previous works have shown that data selection techniques can result in data sets biased towards specific groups of phones or triphones [19]. A phonetic analysis of the data sets given by the likelihood ratio function used in this paper did not show any bias on phones in these data sets. The 60–hour training data used in this work was well balanced phonetically which limited the risk of phonetic biases in the selected data. In situations where the original training data might present less well distributed phonetic content, the proposed function should be complemented by a function that takes into account the resulting phone distribution of the data.

Future work should explore similar data selection techniques for other training criteria besides MLE. The presented methods are based on LR and hence well–suited for MLE, but other submodular functions will be required to cater for needs given by discriminative objective functions such as Minimum Phone Error training. Further work should also investigate data selection techniques for datasets larger than the one studied here, and in completely mismatched conditions and using different features that better describe the background's acoustic characteristics [34].

The technique presented in this paper can be used for building targeted models for "found speech data". The ability of using very diverse data sets to transcribe newly found sets of speech recorded in unknown conditions is especially necessary to deal with this type of data. Other tasks, such as the automatic transcription of multi–genre media archives might also potentially benefit from the advances achieved in this work.

7. Acknowledgements

8. Data Access Statement

The speech data used in this paper was obtained from the following sources: Fisher Corpus (LDC catalogue number LDC2004T19), ICSI Meetings corpus (LDC catalogue number LDC2004S02), WSJCAM0 (LDC catalogue number LDC95S24), AMI corpus (DOI number 10.1007/11677482_3), TedTalks data (freely available as part of the IWSLT evaluations), BBC Radio and TV data (this data was distributed to the NST project's partners with an agreement with BBC R&D and not publicly available yet).

The specific file lists used for training and testing in the experiments in this paper, as well as result files can be downloaded from `http://mini.dcs.shef.ac.uk/publications/papers/is15-doulaty`.

9. References

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[2] F. Grezl, M. Karafiát, S. Kontár, and J. Cernockỳ, "Probabilistic and bottle-neck features for LVCSR of meetings." in *Proceedings of ICASSP*, Hawaii, USA, 2007, pp. 757–760.

[3] X. Huang, A. Acero, and H. Hon, *Spoken language processing*. Prentice Hall: Englewood Cliffs, 2001.

[4] O. Kapralova, J. Alex, E. Weinstein, P. Moreno, and O. Siohan, "A big data approach to acoustic model training corpus selection," in *Proceedings of Interspeech*, Singapore, 2014, pp. 2083–2087.

[5] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Unsupervised submodular subset selection for speech data," in *Proceedings of ICASSP*, Florence, Italy, 2014.

[6] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich, "To transfer or not to transfer," in *NIPS 2005 Workshop on Transfer Learning*, vol. 898, 2005.

[7] H. Lin and J. Bilmes, "How to select a good training-data subset for transcription: Submodular active selection for sequences," in *Proceedings of Interspeech*, Brighton, UK, 2009.

[8] K. Wei, Y. Liu, K. Kirchhoff, C. Bartels, and J. Bilmes, "Submodular subset selection for large-scale speech training data," in *Proceedings of ICASSP*, Florence, Italy, 2014.

[9] Y. Wu, R. Zhang, and A. Rudnicky, "Data selection for speech recognition," in *Proceedings of ASRU*, Kyoto, Japan, 2007, pp. 562–565.

[10] R. Zhang and A. Rudnicky, "A new data selection approach for semi-supervised acoustic modeling," in *Proceedings of ICASSP*, Toulouse, France, 2006.

[11] A. Nagroski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," in *Proceedings of ASRU*, St. Thomas, US Virgin Islands, 2003, pp. 67–72.

[12] E. Gouvea and M. H. Davel, "Kullback-Leibler divergence-based ASR training data selection," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 2297–2300.

[13] O. Siohan and M. Bacchiani, "iVector-based acoustic data selection." in *Proceedings of Interspeech*, Lyon, France, 2013, pp. 657–661.

[14] G. Riccardi and D. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition." in *proceedings of Interspeech*, Geneva, Switzerland, 2003.

[15] G. Tur, R. Schapire, and D. Hakkani-Tür, "Active learning for spoken language understanding," in *Proceedings of ICASSP*, Hong Kong, 2003.

[16] B. Settles, "Active learning literature survey," University of Wisconsin, Madison, WI, USA, Tech. Rep., 2010.

[17] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[18] P. Lanchantin, P. J. Bell, M. J. Gales, T. Hain, X. Liu, Y. Long, J. Quinnell, S. Renals, O. Saz, and M. S. Seigel, "Automatic transcription of multi-genre media archives," in *Proceedings of SLAM Workshop*, Marseille, France, 2013.

[19] O. Siohan, "Training data selection based on context-dependent state matching," in *Proceedings of ICASSP*, Florence, Italy, 2014, pp. 3316–3319.

[20] X. Zhu, "Semi-supervised learning literature survey," University of Wisconsin, Madison, WI, USA, Tech. Rep., 2005.

[21] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of COLT Workshop*, Pittsburgh, PA, USA, 1992, pp. 287–294.

[22] A. Krause and D. Golovin, "Submodular function maximization," *Tractability: Practical Approaches to Hard Problems*, 2014.

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

[24] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Using document summarization techniques for speech data subset selection." in *Proceedings of HLT-NAACL*, Atlanta, USA, 2013, pp. 721–726.

[25] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text." in *Proceedings of LREC*, Lisbon, Portugal, 2004, pp. 69–71.

[26] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, W. Karaiskos, Vasilis Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A preannouncement," in *Proceedings of MLMI*, Bethesda, USA, 2006, pp. 28–39.

[27] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proceedings of ICASSP*, Hong Kong, 2003.

[28] R. W. N. Ng, M. Doulaty, R. Doddipatla, O. Saz, M. Hasan, T. Hain, W. Aziz, K. Shaf, and L. Specia, "The USFD spoken language translation system for IWSLT 2014," Lake Tahoe, USA, 2014.

[29] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A british english speech corpus for large vocabulary continuous speech recognition," in *Proceedings of ICASSP*, Detroit, USA, 1995.

[30] K. Vesely, L. Burget, and F. Grezl, "Parallel training of neural networks for speech recognition," in *Proceedings of Interspeech*, Makuhari, Japan, 2010.

[31] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Proceedings of ICASSP*, Florence, Italy, 2014.

[32] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book (for HTK version 3.4)," *Cambridge university engineering department*, vol. 2, no. 2, 2006.

[33] Z.-J. Yan, Q. Huo, and J. Xu, "A scalable approach to using dnn-derived features in gmm-hmm based acoustic modeling for lvcsr." in *Proceedings of Interspeech*, Lyon, France, 2013, pp. 104–108.

[34] O. Saz, M. Doulaty, and T. Hain, "Background–tracking acoustic features for genre identification of broadcast shows," Lake Tahoe NV, USA, 2014.