INCREMENTAL SEMI-SUPERVISED LEARNING FOR MULTI-GENRE SPEECH RECOGNITION

Banriskhem Khonglah[1]*, Srikanth Madikeri*[1]*, Subhadeep Dey*[1]*, Hervé Bourlard*[1]*,*
Petr Motlicek[1]*, Jayadev Billa*[2]
[1]Idiap Research Institute, Martigny, Switzerland
[2] Information Sciences Institute, University of Southern California
{banriskhem.khonglah, srikanth.madikeri, sdey, herve.bourlard, petr.motlicek}@idiap.ch
jbilla@isi.edu

ABSTRACT

In this work, we explore a data scheduling strategy for semi-supervised learning (SSL) for acoustic modeling in automatic speech recognition. The conventional approach uses a seed model trained with supervised data to automatically recognize the entire set of unlabeled (auxiliary) data to generate new labels for subsequent acoustic model training. In this paper, we propose an approach in which the unlabelled set is divided into multiple equal-sized subsets. These subsets are processed in an incremental fashion: for each iteration a new subset is added to the data used for SSL, starting from only one subset in the first iteration. The acoustic model from the previous iteration becomes the seed model for the next one. This scheduling strategy is compared to the approach employing all unlabeled data in one-shot for training. Experiments using lattice-free maximum mutual information based acoustic model training on Fisher English gives 80% word error recovery rate. On the multi-genre evaluation sets on Lithuanian and Bulgarian relative improvements of up to 17.2% in word error rate are observed.

Index Terms— semi-supervised learning, incremental training, multi-genre speech recognition

1. INTRODUCTION

Semi-supervised learning (SSL) is often employed on large amounts of unlabelled data, to train automatic speech recognition (ASR) system for low-resource languages. Typically, a seed model trained with labelled data is used to generate (hard/soft) labels on the unlabelled (auxiliary) set. The acoustic model is then trained with the newly generated labels along with that of the supervised data. In this paper, we describe our efforts for the MATERIAL program[1], where the training data consists of only conversational speech and the evaluation data consists of three genres: conversational speech, news and topical broadcast (CS, NB and TB, respectively). Moreover, unlike the training data, majority of test data belong to NB and TB. This is both a multi-genre (in terms of speaking style and content) and a multi-bandwidth condition. Such domain adaptation problems can be addressed with SSL using data collected from the web [1–3].

In the conventional approach to SSL, labels for the entire untranscribed set are generated with a seed model trained on manually labelled data. In [4–12], a subset of unlabelled data is chosen based on confidence scores since it is difficult to accurately determine the quality of the labels generated. In [13], the labels from the best path are used along with the frame-level posteriors as weights for the loss function during subsequent training. In the aforementioned approaches, the entire unlabelled data is decoded only once with a seed model trained using manually labeled data. In [14, 15], multiple systems (or outputs) are used to obtain better labels. In [16], interleaved training by continuously updating the model used to generate labels was shown to be effective. The latest model is used in each sub-epoch to provide transcriptions for the next batch of data (typically processing 25'000 hours of data). However, the previously seen data is completely ignored since the entire training was run for only one epoch (given that it was trained on 1 million hours).

Clearly, the performance of the seed model determines the quality of labels generated on the unseen data. Results from [16] also show that it is possible to use parts of unlabelled set to improve the seed models. We thus propose to constantly update the labels with better seed models by running the SSL process on portions of unsupervised data. More specifically, first, the entire unlabelled set is divided into subsets of same size. The SSL process is performed for a num-

The research is based upon the work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via AFRL Contract #FA8650-17-C-9116. 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 the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The work is also partially supported by the ROXANNE H2020 project (http://www.roxanne-euproject.org), under grant agreement No 833635.

[1]https://www.iarpa.gov/index.php/research-programs/material

ber of iterations equal to the number of splits generated. For each iteration i, the first i subsets are chosen and the latest model replaces the seed model for SSL. After each iteration, we expect to obtain more precise labels on the unlabelled data which is yet to be seen, compared to the one generated with the seed model.

In this paper, the standard SSL technique for lattice free maximum mutual information (LF-MMI) based models is used [13]. We evaluate our method on 3 data-sets: the Fisher English data used as a frequent ASR setup to compare SSL with baseline approaches, and on two MATERIAL data-sets: Lithuanian and Bulgarian. We show that a simple heuristic of splitting the unsupervised data in chunks of size comparable to the amount of supervised data helps considerably, compared to one-shot training with all unlabelled data.

The rest of the paper is organized as follows: Section 2 describes the semi-supervised acoustic model training for ASR using LF-MMI. The incremental semi-supervised training is described in section 3. Results of the experiments on Fisher English and MATERIAL data-sets are given in section 4. The conclusion is given in section 5.

2. SEMI-SUPERVISED TRAINING USING LFMMI

Current state-of-the-art hybrid ASR systems employ LF-MMI training as it provides an efficient way to perform sequence discriminative training on GPUs. During training, the MMI objective function is optimized along with cross entropy function as a regularizer. The alignments for LF-MMI training are obtained from a hidden Markov model-Gaussian mixture model (HMM-GMM) system, which are then used to create numerator graphs. The denominator graph is a composition of the HMM states with context-dependency tree, the lexicon and a phone language model (LM).

A simple approach to semi-supervised training in the LF-MMI framework is to generate 1-best output as transcription for the unlabelled data. In [13], this approach is extended by using posteriors in the 1-best path in the lattices generated during decoding as frame weights. The 1-best path is used as a numerator graph during semi-supervised training, where the supervised and unsupervised data are combined together.

There are numerous works that extend this simple strategy for semi-supervised learning with LF-MMI. The original work was demonstrated on in-domain data-sets (conversational speech). Nevertheless, similar techniques have also been employed to use a seed model trained on out-of-domain data followed by SSL on in-domain data [1,17,18]. The latter is often obtained from data crawled from the web. As the collection of untranscribed data is uncontrolled, subset selection is done based on confidence measures. The confidence scores obtained from the LF-MMI system are often sparse and different techniques have been proposed to get informative measures for data selection or weighting. In addition, LM data augmentation and adaptation has been shown to be effective.

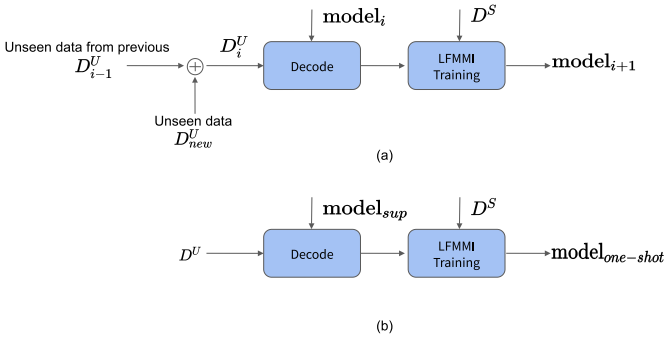

Fig. 1. (a) Incremental training at the i^{th} iteration (b) One shot training. Superscript S and U indicate the supervised and unsupervised part

In our work, we target broadcast data; in particular, news and topical broadcast (NB and TB, respectively). The seed model is trained with several tens of hours (h), depending on the language. This is also accompanied by augmenting data for the LM and expanding the lexicon to address OOV issues and reduce deletion errors on the unlabelled data. We consider SSL with 400h of data obtained from the web for both target languages (as they are the focus of MATERIAL evaluations in October 2019). Note that our systems for the evaluations used at least 4 times more data.

3. INCREMENTAL SEMI-SUPERVISED TRAINING

Typically, the SSL process involves decoding the unlabelled data with a seed model trained on manually transcribed data. The performance of the seed model on the unseen data is critical. In order to improve the quality of transcriptions produced for the unlabelled data, we propose a simple method to generate and update labels for unlabelled data without any change to the core SSL framework being employed. The motivation for this method is based on the observation that SSL can improve the acoustic model even with limited amounts of unlabelled data. Thus, we divide the entire unlabelled data-set into several equal-sized parts and begin SSL training with only one part. While there exist many ways to divide the data, in this work we have considered closely matching the amount of supervised data to our split-size.

Enumerating each split from $1 \ldots n$, we run n iterations of SSL (a version of [13] which is available in Kaldi). In the i^{th} iteration, splits $1 \ldots i$ are used as the unlabelled set for SSL. As shown in Figure 1, in each iteration we use the previous model as the seed for a new iteration of SSL training from scratch. In our experiments, we did not observe continuing the new iteration from the final acoustic model of the previous iteration to be better than our approach. The data used for each iteration includes the supervised set, all the portions of unsupervised set used in the last iteration and one unused sub-

set for the current iteration. In doing so, we are continuously improving the seed model on the domain of the unlabelled data. We note that this data scheduling strategy, however, is computationally intensive since it involves multiple decodes of the data.

4. EXPERIMENTAL RESULTS

The experiments are performed on Fisher English and two languages from the MATERIAL program. The former is a standard database used extensively in semi-supervised experiments.

In our experiments, we consider two baselines: (1) a system that does not use any unlabelled data, and (2) a one-shot SSL baseline, following [13] that uses 1-best output on the entire unlabelled set generated from the seed model from (1).

4.1. Fisher English Setup

This setup strictly follows the Kaldi recipe [19]. For the unsupervised acoustic data, a random subset of speakers (250h) was chosen. The language model required for decoding was trained on the remaining 1250h of transcripts. The seed model was trained using a 50h subset of data chosen from the corpus. Dev and test sets are used to report the results. Each test set is approximately 5h long.

Initially, a GMM-HMM system was trained using the supervised data in order to provide alignments for the LF-MMI training of the neural network which will then act as the seed model. The overall system is a time delay neural network-hidden Markov model (TDNN-HMM) and consists of 7 hidden layers of TDNN along with 745 hidden units in each layer [20]. Online i-vectors [21] of 100 dimensions are appended to MFCC features of 40 dimensions at the input. The i-vector extractor is trained using a combination of the supervised and the unsupervised data. The context-dependent decision tree is trained using the statistics of the supervised data. The phone LM was trained using the phone sequences of supervised and the unsupervised data with more weight given to the phone sequences corresponding to the supervised data (supervised data=1.5 and unsupervised data=1).

4.2. Fisher English Results

The results on Fisher English are presented in Table 1. One-shot SSL improves the system performance in WER relatively by 11% and 10.6 % on the dev and test set, respectively. The 250h of data is 3-way speed perturbed. The proposed incremental method of training is applied as follows: the 250h of data selected for SSL is 3-way speed perturbed and split into 5 parts so that in each iteration the amount of new unlabelled data is similar to that of supervised data. Our proposed method already outperforms the one-shot SSL after 3 iterations. Overall, this method of incremental training gives a relative improvement of 4 % and 4.6 % in terms of WER on the dev and test sets, respectively, over one-shot SSL. The WER recovery rate (WRR) is **78** % and **80** % on the dev and test sets respectively [13].

Table 1. *Performance using TDNN-HMM system on the dev and test data of Fisher English data-set in terms of WER (%). (Sup: supervised, Unsup: unsupervised)*

System	dev	test
Sup	21.8	21.5
Unsup 250 (one-shot SSL)	19.4	19.2
Incremental Unsupervised (*proposed*)		
Unsup 50	20.6	20.4
Unsup 100	19.7	19.2
Unsup 150	**19.1**	**19.0**
Unsup 200	**18.8**	**18.5**
Unsup 250	**18.6**	**18.3**
Oracle 250	17.7	17.5

4.3. MATERIAL Data-set Setup

The unsupervised data is collected from YouTube [22]. A total of 400 hours each for Lithuanian and Bulgarian are considered for the experiments. The results are reported on two sets: the dev set, which is part of the official Babel release, and the IARPA MATERIAL Analysis Pack 1 (Analysis) [17,22]. The dev set consists of only CS while the Analysis set contains the three domains: CS, NB and TB. The broadcast data consists of audio files at 44.1 kHz and 48 kHz sampling rate, described in Table 2.

Trigram LMs for both Lithuanian and Bulgarian used 15M and 28M sentences of text respectively; the majority of this text was mined from the web. Using these web crawl text based LMs resulted in significant degradation on CS, with a 3.1% absolute increase in WER. To address this we linearly interpolated two LMs: one that uses all text data and one that uses only transcripts from the training set. All our results on dev and CS use the interpolated LM.

The GMM-HMM system is trained as in the Fisher English case for generating the alignments. The overall system used for this experiment is based on the time delay neural network factorization-hidden markov model (TDNN-F-HMM) [23] and the number of hidden layers for the TDNN-F used is 16. This system gives better baseline performance than the TDNN-HMM system on the Babel data-sets by a significant margin. The 100-dimensional i-vectors are also used in this experiment and the extractor for these i-vectors is trained in a multilingual fashion with a total of 18 languages. The context dependent decision tree and the phone LM are trained in the same way as Fisher English.

Table 2. *Statistics of BABEL target languages used for testing. Note that the Eval sets mentioned refer to the dev set in the official BABEL release. All durations are calculated prior to silence removal.*

Parameter	Lithuanian	Bulgarian
Vocabulary (words)	630k	530k
LM perplexity	650	408
Supervised data (hours)	69	58
Unsupervised data (hours)	400	400
dev (hours)	17.8	18
Analysis (CS,NB,TB in hours)	1.4, 3.8, 6.3	6.5, 3.9, 10.6

Table 3. *Performance using TDNN-F-HMM system on the dev and Analysis test data of Babel Lithuanian Data in terms of WER (%). (CS: Conversational speech, NB: News Broadcast, TB: Topical Broadcast, Sup: supervised, Unsup: unsupervised). Note that dev and CS part of the analysis are decoded using an interpolated LM.*

System ↓	dev	CS	NB	TB
Sup	43.4	42.3	33.4	34.2
Unsup 400 (one shot)	43.5	43.3	24.3	25.9
Incremental Unsupervised				
Unsup 100	44.4	44.3	25.3	27.4
Unsup 200	**42.9**	**42.8**	**22.9**	**25.3**
Unsup 300	**43.2**	**42.7**	**21.9**	**24.0**
Unsup 400	**42.6**	**42.1**	**20.1**	**22.8**

4.4. MATERIAL Data Results

The results on Lithuanian and Bulgarian are presented in Table 3 and 4. While the performance of the one-shot training improves for NB and TB domains compared to the case of using only supervised data, the performance on CS degrades. This is because of data imbalance since the unsupervised data, which consists of speech related to TB and NB, is 8 times larger than supervised data. Using incremental training at steps of 100 h of data at a time improves the system further. The improvement over one-shot training can be observed after using only 200 h of data. The improvements are observed with respect to NB and TB, although after 400 h of incremental training, improvement for CS is also obtained. At the end of 400 h of incremental training, the relative improvement in WER for dev, CS, NB and TB are 2.1%, 2.7%, 17.2% and 11.9 %, respectively, for Lithuanian over SSL with one-shot training. We obtained similar gains for Bulgarian as well. The relative improvement in WER for CS, NB and TB are 6.9%, 9.9% and 10.3% over the baseline. For the same baseline, a relative improvement of 1% is obtained on the dev set.

Table 4. *Performance using TDNN-F-HMM system on the dev and analysis test data of Babel Bulgarian Data in terms of WER (%). (CS: Conversational speech, NB: News Broadcast, TB: Topical Broadcast, Sup: supervised, Unsup: unsupervised)*

System ↓	dev	CS	NB	TB
Sup	40.4	42.5	21.6	32.2
Unsup 400 (one shot)	37.4	41.6	15.1	23.2
Incremental Unsupervised				
Unsup 100	39.8	41.9	16.0	24.8
Unsup 200	38.2	**40.3**	**14.6**	**22.2**
Unsup 300	37.5	**39.8**	**13.8**	**21.2**
Unsup 400	**37**	**38.7**	**13.6**	**20.8**

5. CONCLUSION

This paper proposed an effective data scheduling strategy for SSL. The results shown on Fisher English, and two MATERIAL languages, indicate that the method consistently outperforms the baseline of one-shot training on all the three databases. Relative improvements up to 17.2 % were observed on multi-genre domains in the evaluation set. This demonstrates the benefit of the proposed SSL method which helps to obtain better alignments after every iteration, instead of using the unlabelled data at once to generate the alignments.

6. REFERENCES

[1] Andrea Carmantini, Peter Bell, and Steve Renals, "Untranscribed web audio for low resource speech recognition," *Proc. Interspeech*, pp. 226–230, 2019.

[2] Deblin Bagchi and William Hartmann, "Learning from the best: A teacher-student multilingual framework for low-resource languages," in *Proc. International conference on acoustics, speech and signal Processing (ICASSP)*. IEEE, 2019, pp. 6051–6055.

[3] Jeff Ma, Spyros Matsoukas, Owen Kimball, and Richard Schwartz, "Unsupervised training on large amounts of broadcast news data," in *Proc. International conference on acoustics speech and signal processing (ICASSP)*. IEEE, 2006, vol. 3, pp. III–III.

[4] 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.

[5] Karel Veselỳ, Mirko Hannemann, and Lukáš Burget, "Semi-supervised training of deep neural networks," in

Proc. Workshop on automatic speech recognition and understanding (ASRU). IEEE, 2013, pp. 267–272.

[6] Samuel Thomas, Michael L Seltzer, Kenneth Church, and Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2013, pp. 6704–6708.

[7] Pengyuan Zhang, Yulan Liu, and Thomas Hain, "Semi-supervised dnn training in meeting recognition," in *Proc. Spoken language technology workshop (SLT)*. IEEE, 2014, pp. 141–146.

[8] Frantiseli Grezl and Martin Karafiát, "Semi-supervised bootstrapping approach for neural network feature extractor training," in *Proc. Workshop on automatic speech recognition and understanding (ASRU)*. IEEE, 2013, pp. 470–475.

[9] Petr Motlicek, David Imseng, Blaise Potard, Philip N. Garner, and Ivan Himawan, "Exploiting foreign resources for dnn-based asr," *EURASIP Journal on Audio, Speech, and Music Processing*, , no. 2015:17, June 2015.

[10] David Imseng, Blaise Potard, Petr Motlicek, Alexandre Nanchen, and Hervé Bourlard, "Exploiting untranscribed foreign data for speech recognition in well-resourced languages," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014, pp. 2322 – 2326, IEEE.

[11] David Imseng, Petr Motlicek, Philip N. Garner, and Hervé Bourlard, "Impact of deep mlp architecture on different acoustic modeling techniques for under-resourced speech recognition," in *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Dec. 2013.

[12] Ajay Srinivasamurthy, Petr Motlicek, Mittul Singh, Youssef Oualil, Matthias Kleinert, heiko Ehr, and Hartmut Helmke, "Iterative learning of speech recognition models for air traffic control," in *Proceedings of Interspeech 2018*. ISCA, Sept. 2018, pp. 3519–3523.

[13] Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur, "Semi-supervised training of acoustic models using lattice-free mmi," in *Proc. International conference on acoustics, speech and signal Processing (ICASSP)*. IEEE, 2018, pp. 4844–4848.

[14] Yu Wang, Xie Chen, Mark JF Gales, Anton Ragni, and Jeremy HM Wong, "Phonetic and graphemic systems for multi-genre broadcast transcription," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5899–5903.

[15] Sibo Tong, Apoorv Vyas, Philip N Garner, and Hervé Bourlard, "Unbiased semi-supervised lf-mmi training using dropout," *Proc. Interspeech*, pp. 1576–1580, 2019.

[16] Sree Hari Krishnan Parthasarathi and Nikko Strom, "Lessons from building acoustic models with a million hours of speech," in *Proc. International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 6670–6674.

[17] Anton Ragni and Mark JF Gales, "Automatic speech recognition system development in the" wild"," in *Proc. Interspeech*, 2018, pp. 2217–2221.

[18] Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models," in *Proc. Spoken language technology workshop (SLT)*. IEEE, 2018, pp. 250–257.

[19] Daniel Povey et al., "The kaldi speech recognition toolkit," in *Proc. Workshop on automatic speech recognition and understanding (ASRU)*. IEEE Signal Processing Society, 2011.

[20] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015.

[21] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 55–59.

[22] Elizabeth Boschee et al., "Saral: A low-resource cross-lingual domain-focused information retrieval system for effective rapid document triage," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019, pp. 19–24.

[23] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks.," in *Proc. Interspeech*, 2018, pp. 3743–3747.