IDIAP RESEARCH REPORT

SEMI-SUPERVISED LEARNING WITH SEMANTIC KNOWLEDGE EXTRACTION FOR IMPROVED SPEECH RECOGNITION IN AIR TRAFFIC CONTROL

Ajay Srinivasamurthy Petr Motlicek Ivan Himawan
Gyorgy Szaszak Youssef Oualil Hartmut Helmke

Idiap-RR-21-2017

SEPTEMBER 2017

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Semi-supervised Learning with Semantic Knowledge Extraction for Improved Speech Recognition in Air Traffic Control

Ajay Srinivasamurthy[1], *Petr Motlicek*[1], *Ivan Himawan*[1],
György Szaszák[2], *Youssef Oualil*[2], *Hartmut Helmke*[3]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Spoken Language Systems Group, Saarland University (UdS), Saarbrücken, Germany
[3]Institute for Flight Guidance, German Aerospace Center (DLR), Braunschweig, Germany

ajay.srinivasamurthy@idiap.ch, petr.motlicek@idiap.ch, ivan.himawan@idiap.ch,
gszaszak@lsv.uni-saarland.de, youalil@lsv.uni-saarland.de, hartmut.helmke@dlr.de

Abstract

Automatic Speech Recognition (ASR) can introduce higher levels of automation into Air Traffic Control (ATC), where spoken language is still the predominant form of communication. While ATC uses standard phraseology and a limited vocabulary, we need to adapt the speech recognition systems to local acoustic conditions and vocabularies at each airport to reach optimal performance. Due to continuous operation of ATC systems, a large and increasing amount of untranscribed speech data is available, allowing for semi-supervised learning methods to build and adapt ASR models. In this paper, we first identify the challenges in building ASR systems for specific ATC areas and propose to utilize out-of-domain data to build baseline ASR models. Then we explore different methods of data selection for adapting baseline models by exploiting the continuously increasing untranscribed data. We develop a basic approach capable of exploiting semantic representations of ATC commands. We achieve relative improvement in both word error rate (23.5%) and concept error rates (7%) when adapting ASR models to different ATC conditions in a semi-supervised manner.

Index Terms: Speech Recognition, Air Traffic Control, Semi-supervised learning

1. Introduction

Air Traffic Control (ATC) involves spoken language communication between aircraft pilots and air traffic controllers, who guide aircraft to navigate safely in air and at airports. The intensive use of spoken language in ATC is natural and hence preferred in some ways, but it also hampers the introduction of higher levels of automation. Introduction of ASR (Automatic Speech Recognition) into ATC systems is an enabler for different levels of automation, reducing the efforts of air traffic controllers leading to significant gains in terms of reduced human effort and saved flight times.

Recently, Assistant based Speech Recognition (ABSR)[1] [1] that combines ASR with a radar based assistant system has been shown to be useful. An ABSR generates context information to reduce the search space for the speech recognizer and can reduce controller's workload by a factor of three [2], in addition to significant fuel savings [3] resulting from shorter flight times and increased operational efficiency. However, extending ABSR to real world operational environments is challenging for many reasons. To build robust ASR systems for each operating ATC environment, transcribed speech data is necessary, obtaining which is time and resource consuming. Owing to its global nature, ATC uses standardized English vocabulary and phraseology for communication. However, local variations in each ATC area exist due to local runways, waypoints, airlines, acoustic conditions, local English accents and the occasional use of local language words. Further, some of the local conditions (airlines, runways, waypoints) can also change over time and hence the ASR systems need regular maintenance. Due to continuous operation of ATC systems, an increasing amount of (untranscribed) speech and radar data is generated and is archived for flight safety reasons. The MALORCA[2] project has been constituted to address these issues and automate re-learning, adaptation and customization of ASR systems to new ATC environments. The main goal is to continuously update the ASR models in an unsupervised/semi-supervised manner by utilizing increasing amounts of speech data, while exploiting local acoustic, language and semantic constraints. In addition, data from other modalities such as radar can be used, which provide a context for the commands issued by the controllers to pilots.

ASR systems built to a specific domain ensure the best performance. However, in their absence, adapting out-of-domain (OOD) ASR models to a specific domain has been explored [4, 5]. In aviation, ASR is a known technology used with considerable success in training simulators [6]. Applying ASR to ATC domain has been previously explored [7], but the use of untranscribed data is a new challange. Semi-supervised learning methods [8] can be used to utilize the untranscribed data to improve and build domain specific ASR systems. A "first iteration" ASR built with limited training data can be used to automatically transcribe raw audio data, thus generating approximate transcriptions that can be used as additional training data. Data selection for semi-supervised learning [9] from such automatic transcripts then becomes a central task, where different confidence measures at frame, word and sentence level have been used and several methods have been proposed [10, 11, 12]. Semantics based confidence measures have received some attention in specific tasks related to spoken dialogue systems [13, 14]. However, the variation in semantics across different application domains of ASR motivates the need for domain specific semantic confidence measures.

In this paper, we explore the tasks associated with automatic deployment and adaptation of ASR models to a new ATC environment. We use a limited amount of transcribed data available from Vienna ATC area while also utilizing additional

[1]AcListant®: http://www.aclistant.de

[2]MAchine Learning Of speech Recognition models for Controller Assistance: http://www.malorca-project.de/

OOD data. We propose data selection methods to choose suitable training data from untranscribed speech from Vienna ATC area and discuss directions for further improvement of semi-supervised learning methods. ATC communication has a limited vocabulary with strong semantic restrictions. The goal of such communication is to ensure that the necessary commands from controllers to pilots are conveyed through spoken language. The commands are hence primary, while the exact spoken text is of secondary importance. Any improvements to an ASR system in such an application should be geared towards improving the accuracy of command recognition. Hence measures and approaches that can work with command semantics in addition to the commonly used phone and word levels are preferred. We also explore such methods in this paper.

2. Semi-supervised Learning: Methods

In this paper, (1) we build base ASR models using limited in-domain data from Vienna ATC area and out-of-domain data, (2) the base ASR models are then supervised-adapted to Vienna ATC area. Subsequently, (3) the ASR models are used for further semi-supervised learning experiments. We start by first describing the datasets used in the experiments.

2.1. Datasets

The speech data used in this paper has been recorded from Vienna approach sector and feeder controller. A part of the speech data is transcribed, with text and command transcriptions. The availability of a partial set of transcriptions provides us the right opportunity to explore semi-supervised learning methods to utilize the complete untranscribed data. Vienna ATC continuously records speech data and hence can provide increasing amounts of (untranscribed) speech data. At the moment, this data is not publicly available. The speech content of the dataset is similar to other publicly available ATC domain datasets such as the LDC ATC dataset [15] and ATCOSIM dataset [16].

While additional data from Vienna approach is expected, presently the dataset has over 20 hours of speech data from 46 different controllers (speakers). All the data was recorded from operational ATC environments in the second half of 2016 at a sampling rate of 8kHz. The data has been segmented into short utterances containing only a few (upto 5) controller commands (most utterances have just one command). A command from a controller is repeated by the pilots (readback), but the pilot replies are not recorded and stored since they are not relevant. While all recordings have speaker labels, only a part of the dataset is annotated by professional air traffic controllers with text and command transcripts using an in-house annotation tool.

For training the base ASR models, we use about 5 hours of transcribed data, which we term as VDev1. The transcriptions include text transcriptions of the speech utterance, along with a transcription of the command that the speech utterance conveys to the pilots. For testing, we use about 2 hours of transcribed data with 6 speakers, termed as VTest dataset. About 9 hours of untranscribed data termed as VDev2 is used for semi-supervised learning of models. The three datasets are disjoint and do not share any speakers across them, as described in Table 1.

Since the amount of transcribed data available from Vienna approach is limited, we utilize other available transcribed resources to train the ASR system. We hypothesize that the use of standard English datasets is useful for seed training an acoustic model. We pool 150 hours of speech data from the publicly available LIBRISPEECH [17], ICSI [18], AMI [19] and TED-LIUM [20] datasets, which have been extensively used for recognition of conversational speech. The speech data and accompanying transcripts (called MEGA) are used in conjunction with training data from Vienna approach.

Dataset	Source	Dur. (hr)	Speakers
VDev1	Vienna approach	5.1	13
VDev2	Vienna approach	9.1	24
VTest	Vienna approach	1.9	6
MEGA	LIBRISPEECH, AMI, ICSI, TED-LIUM	150	1043

Table 1: *Datasets, showing the source, duration and speakers*

2.2. Dictionary, Acoustic and Language models

We add all the possible in-domain words associated with Vienna ATC area (e.g. airlines and waypoints) to the standard CMU-Sphinx dictionary[3] to form an extended pronunciation dictionary for use with both acoustic and language models. There are hence no out-of-vocabulary words during training or testing.

DNN/HMM (Deep Neural Network Hidden Markov Model) acoustic models are the state of the art in speech recognition acoustic modeling. As reliable training of DNNs require significant amount of labeled data, we add the 150 hour MEGA dataset to the limited Vienna VDev1 dataset. We use the combined data to train a Gaussian Mixture Model based GMM/HMM acoustic model (AM). Using the state level alignments of the combined data using the GMM/HMM model, we train a DNN/HMM acoustic model (called the DNN-BASE).

The DNN-BASE acoustic model is then adapted to Vienna ATC domain using the VDev1 dataset. To adapt, we start from the DNN-BASE model, and first reinitialize and randomize the weights of the last layer of the DNN. The architecture and weights of the other layers are unchanged. We then retrain the entire network using VDev1 training dataset to obtain supervised-adapted DNN (DNN-SA). This way of reinitializing the last layer and retraining the complete network was found to be effective for supervised adaptation using in-domain data.

For decoding a test utterance, we use a trigram language model (LM) built using the transcripts of VDev1 (vocabulary size $\approx$ 700 words) to ensure that an in-domain Vienna specific language model is used. Together with the language model, the ASR system using DNN-BASE and DNN-SA with the trigram language model from VDev1 is called ASR-BASE and ASR-SA, respectively.

The standard vocabulary and phraseology used in ATC is an argument to construct a rule based Context-Free Grammar (CFG) that can be used to build a Vienna specific language model. However, in practice, the controllers often deviate from standards, and hence an N-gram statistical language model is used instead for recognition, while a CFG is used for concept extraction, as further described next.

2.3. Concept and Command extraction

The output from an ASR system is a sequence of words as spoken by the controller. We however then need to extract the controller command that the sequence contains. From the controller utterances we extract concepts and commands. Concepts include all meaningful words or expressions which are related to the controller command and the required action of the aircraft. Concepts basically include (i) the callsign composed of an airline identifier (International Civil Aviation Organization airline code) and a flight number, (ii) the command word or ex-

[3] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

pression itself, and (iii) the command attributes (usually target values for some flight parameters). This sequence of concepts forms a command. For example, the following utterance "*hello lufthansa eight echo kilo, start reduce your speed to two two zero knots*" contains the following concepts:

- DLH8EK (*lufthansa eight echo kilo* - callsign)
- REDUCE (*reduce* - command word)
- 220 (*two two zero* - speed attribute)

The command in its semantic form is hence: DLH8EK REDUCE 220. In order to extract the concepts from the utterance, we use a CFG that models controller phraseology. Normally phraseology is highly standardized, i.e. the controllers are fairly bound on how they formulate a command. All possible command words or expressions have an entry in the CFG, for modelling standard phraseology and often used deviation forms. Each semantic slot for the command is tagged in the CFG, and hence, transducing [21] a transcript hypothesis from the ASR over the CFG results in an XML tagged version as follows (using the previous example): *hello* <callsign> <airline> *lufthansa* </airline> <flightnumber> *eight echo kilo* </flightnumber> </callsign> *start* <commands> <command="reduce"> *reduce your speed to* <speed> *two two zero* </speed> *knots* </command> </commands>. If transductions fail (due to a deviation in phraseology not modelled by the CFG or due to ASR errors), the command extractor returns "NO_CALLSIGN" if the callsign is missed, and "NO_CONCEPT", if the command word or the command attribute could not be recovered.

Thus, given a speech utterance by a controller, we obtain a plain text hypothesis (sequence of words as they were spoken), an XML tagged version of hypothesis (tagged with semantic concepts), and the command hypothesis.

2.4. Semi-supervised learning

Semi-supervised learning aims to exploit the untranscribed data available in VDev2 dataset to improve the ASR models. Starting with the supervised-adapted ASR-SA system, the approach we use in this paper consists of three steps: transcript generation, data selection, and semi-supervised training.

2.4.1. Transcript generation

First we use the system adapted to VDev1 (ASR-SA) to generate the text and command transcripts for the data in VDev2. These automatically generated transcripts are used for further experiments.

2.4.2. Data selection

The automatically generated transcripts along with speech in VDev2 can be used as training data. However, these transcripts might have errors and those should be excluded from training, which is a problem often termed as data selection. Data selection is done by assigning confidence scores to ASR outputs, so that high confidence transcripts (and corresponding utterances) can be selected for further experiments. We explore two different data selection strategies, one that uses word level confidences and another that uses concept and command level confidences. Both data selection methods aim to utilize automatically transcribed data to provide additional training resources.

Word confidence: A logistic regression model is built with word-lattice derived features using the VDev1 transcribed data. The features include the posterior probability of a word obtained from Minimum Bayes Risk (MBR) decoding [22], word length, competing words, and frames per character ratio. The trained logistic regression model is applied with the same features extracted from the decoding word-lattices of VDev2 and output confidences (ranging from 0 to 1) per word are obtained. Utterance level confidence is then obtained as the average word-confidence of the words in the output. The utterance-confidence values are sorted and a threshold is used to select high confidence data into a subset VDev2-W of the automatically transcribed VDev2 dataset.

System	Training dataset	#Sen.	WER (%)	CER (%)
ASR-DEV1	VDev1	2143	12.3	38.6
ASR-BASE	MEGA + VDev1	3861	13.3	41.4

Table 2: *Baseline results on evaluation with VTest dataset and using an LM built with VDev1 transcripts, showing the number of senones (#Sen.), Word (WER) and Concept (CER) error rate.*

Concept confidence: Since the output commands are more relevant than the plain text ASR hypotheses, a data selection method that can incorporate a confidence measure based on output command hypothesis is preferred. We hypothesize that an accurate ASR output would result in an accurate command hypothesis generated by the command extractor. In case the command extractor is unable to decipher a valid command from the ASR output, it implies an erroneous automatic transcription. We base our data selection method on this premise, and exclude all automatic transcriptions that contain NO_CALLSIGN or NO_CONCEPT (and hence indicate the failure of the command extractor to extract a meaningful and valid concept and command hypothesis) as output command, to obtain a subset VDev2-C. Note that a valid output from the command extractor does not always imply an accurate command hypothesis. Nevertheless, we observed that command recognition is mostly accurate when the command extractor does not return NO_CONCEPT/NO_CALLSIGN. Without ground truth command transcripts, we explore this method as a first step towards command semantics based data selection.

2.4.3. Semi-supervised training

With either data selection methods, we combine VDev1 with the selected subset of VDev2 (either VDev2-W or VDev2-C) and their automatically obtained transcripts to form a larger adaptation dataset. Based on our previously published ideas, we explore adapting either the AM [23, 24], LM [25], or both using this adaptation dataset. To adapt only the AM, similar to training the DNN-SA, we reinitialize the last layer of DNN-BASE model and retrain the complete network with the adaptation dataset, while using the LM built with only VDev1. To adapt only the LM, we use the supervised-adapted DNN-SA acoustic model with a 3-gram LM built with the adaptation dataset. To adapt both AM and LM, we adapt DNN-BASE with the combined dataset and use a 3-gram LM built with the adaptation dataset. The ASR systems with semi-supervised adaptation using word and concept based confidences are termed ASR-SSA-W and ASR-SSA-C, respectively.

2.5. Evaluation measures

The most relevant metric of performance for ATC applications is at the command semantics level. However, since the ASR system outputs hypothesis at both word level and command level, we report the commonly used Word Error Rate (WER) and the Concept Error Rate (CER). For the CER, we discard all the semantically irrelevant words with respect to the command type from the output text hypothesis and match only the con-

System	Selection method	Adaptation dataset (Duration)	WER (%)			CER (%)		
			AM	LM	AM+LM	AM	LM	AM+LM
ASR-SA	—	VDev1 (5.1 hr)	10.0	—	—	37.5	—	—
ASR-SSA-none	None	+ VDev2 (9.1 hr)	9.6	9.8	9.6	36.6	37.3	36.9
ASR-SSA-W	Word	+ VDev2-W (7.2 hr)	9.6	9.8	**9.4**	36.8	36.7	37.0
ASR-SSA-C	Concept	+ VDev2-C (7 hr)	9.8	9.8	9.5	37.1	36.1	**35.9**

Table 3: *Results on evaluation with VTest dataset for supervised (ASR-SA) and semi-supervised methods (ASR-SSA-none, ASR-SSA-W, ASR-SSA-C), showing the selection method, adaptation dataset used, AM, LM or AM+LM adaptation, and the measures Word (WER) and Concept (CER) error rate. All acoustic models have 3861 senones at the output. Since the default LM is built with VDev1, LM adaptation is not applicable to ASR-SA. The best WER and CER is marked in bold.*

cepts, by treating the command word and its attribute together. For example, supposing a ground truth transcript of DLH8EK REDUCE_230 and a hypothesis of DLH8EK REDUCE_220, the CER is 50%, since the callsign is correctly hypothesized while command attribute is wrong. Owing to its inclusion of semantics, CER is a stricter measure than the WER.

3. Experiments

The speech recognition experiments are done using the Kaldi speech recognition toolkit [26].

3.1. Experimental setup

The GMM/HMM acoustic model is trained in a conventional fashion and consists of ≈3900 senones. In all the training cases, 50K Gaussians were added to the GMM/HMM model using diagonal covariance matrices. As input features, we applied 13 dim MFCCs accompanied with their delta and acceleration coefficients (39 dim feature vector), along with fMLLR transforms for speaker adaptive training. For the DNN/HMM model, the DNN comprises 4 layers: 351 dim input layer (9 stacked MFCC vectors with a context of 4 frames around the centered frame), hidden layers of 1200 nodes and output layer trained to discriminate among senones to estimate senone posterior probabilities. The DNN is trained to minimize frame-level cross entropy. To establish baselines, we additionally train smaller GMM/HMM (consisting of ≈2100 senones) and DNN/HMM acoustic models with VDev1 without utilizing the OOD data.

Starting from the DNN-BASE model, the supervised-adapted DNN-SA is trained as described in Section 2.2, with the same architecture. The semi-supervised methods follow the process described in Section 2.4.3, adapting either the AM, LM or both, using word confidence (ASR-SSA-W) or concept confidence (ASR-SSA-C). An average word confidence threshold of 0.95 is used for utterance selection, selecting (from VDev2) 7.2 hours of speech into VDev2-W. Command confidence based selection retains 7 hours of speech in VDev2-C data subset. In order to compare the performance of data selection, we also report results with no data selection (i.e. using all of VDev2), termed as ASR-SSA-none.

3.2. Results

We report results only with DNN/HMM acoustic models since they provided a better performance than the GMM/HMM counterparts. The baseline results are shown in Table 2 while the results of supervised and semi-supervised (AM, LM) adaptation are shown in Table 3. Both tables show the evaluation results on VTest dataset, with the baseline supervised training with only VDev1 (ASR-DEV1), VDev1 combined with MEGA (ASR-BASE), supervised adaptation of DNN-BASE with VDev1 (ASR-SA) and the two semi-supervised methods ASR-SSA-W (word confidence) and ASR-SSA-C (concept confidence), in addition to ASR-SSA-none (no data selection). While using only VDev1 to build smaller models seems to perform better in baselines in Table 2, the use of MEGA dataset helps building generalizable larger models that outperform with supervised adaptation as seen from ASR-SA (WER: 10.0%, which is an 18.7% relative decrease compared to 12.3% WER of ASR-DEV1).

Table 3 shows that the addition of automatically transcribed data for training is useful and improves performance over ASR-SA in all cases. It also shows the advantage of AM and LM adaptation, while adapting both AM and LM leads to better WER. The results also indicate that AM adaptation is marginally better than LM adaptation to improve WER, while such an observation does not extend to CER.

The ASR system built without data selection (ASR-SSA-none) shows a 4% relative improvement in WER over ASR-SA, while further data selection methods provide marginal improvement. The best performing WER of 9.4% (6% relative improvement over ASR-SA) is with AM+LM adaptation using word confidence based data selection (ASR-SSA-W), while the best performing CER of 35.9% (relative 4% improvement over ASR-SA, with 35 more concepts correctly hypothesized in total) is with AM+LM adaptation using concept confidences for data selection (ASR-SSA-C). This indicates that concept confidence measures help to achieve lower CER, while word confidence measures improve WER.

4. Conclusions

We built domain specific ASR models for controller pilot communication for Vienna approach by utilizing 150 hours of OOD data and adapting with 5 hours of in-domain transcribed data. We proposed data selection methods using word level and concept level confidences to benefit from cheaply available untranscribed in-domain data. This complemented transcribed in-domain data, enabling an adaptation of both acoustic and language models. Exploiting OOD data, plus complementing transcribed data with untranscribed in-domain data through data selection gives a relative reduction of WER by 23.5% (using word confidences) and CER by 7% (using concept confidences), when compared to using only in-domain transcribed data (ASR-DEV1, WER: 12.3%, CER: 38.6%). In the future, we will explore using additional amounts of untranscribed data for data selection. We also plan to integrate additional semantic information and other modalities such as radar data to develop improved training (such as transfer learning, sequence training with concept error metrics) and data selection methods for semi-supervised learning.

Acknowledgements: The work presented was conducted as a part of MALORCA project, funded by the SESAR Joint Undertaking (Grant Agreement No. 698824), under EU Horizon 2020 Research and Innovation programme. The authors thank Austro Control, Vienna for providing data from Vienna approach.

5. References

[1] H. Helmke, J. Rataj, T. Mühlhausen, O. Ohneiser, H. Ehr, M. Kleinert, Y. Oualil, M. Schulder, and D. Klakow, "Assistant-based speech recognition for ATM applications," in *Proc. of 11th USA/Europe Air Traffic Management Research and Development Seminar (ATM 2015)*, Jun. 2015.

[2] H. Helmke, O. Ohneiser, T. Mühlhausen, and M. Wies, "Reducing Controller Work-load with Automatic Speech Recognition," in *Proc. of the IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, Sacramento, USA, 2016.

[3] H. Helmke, O. Ohneiser, J. Buxbam, and C. Kern, "Increasing ATM Efficiency with Assistant Based Speech Recognition," in *Proc. of the 13th USA/Europe Air Traffic Management Research and Development Seminar*, Seattle, USA, 2017 (to appear).

[4] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. of the IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 366–369.

[5] P. C. Woodland, X. Liu, Y. Qian, C. Zhang, M. J. Gales, P. Karanasou, P. Lanchantin, and L. Wang, "Cambridge university transcription systems for the multi-genre broadcast challenge," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 639–646.

[6] D. Schäfer, "Context-sensitive speech recognition in the air traffic control simulation," Ph.D. dissertation, University of Armed Forces, Munich, 2001.

[7] H. D. Kopald, A. Chanen, S. Chen, E. C. Smith, and R. M. Tarakan, "Applying automatic speech recognition technology to air traffic management," in *Proc. of the IEEE/AIAA 32nd Digital Avionics Systems Conference (DASC)*, 2013.

[8] J. Glass, "Towards unsupervised speech processing," in *Proc. of the 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. IEEE, 2012, pp. 1–4.

[9] T. Drugman, J. Pylkkönen, and R. Kneser, "Active and Semi-Supervised Learning in ASR: Benefits on the Acoustic and Language Models," in *Proc. of Interspeech*, 2016, pp. 2318–2322.

[10] D. Imseng, B. Potard, P. Motlicek, A. Nanchen, and H. Bourlard, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2322–2326.

[11] R. Zhang and A. I. Rudnicky, "A new data selection approach for semi-supervised acoustic modeling," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[12] S. Li, Y. Akita, and T. Kawahara, "Semi-supervised Acoustic Model Training by Discriminative Data Selection from Multiple ASR Systems' Hypotheses," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1520–1530, Sep. 2016.

[13] R. Sarikaya, Y. Gao, M. Picheny, and H. Erdogan, "Semantic confidence measurement for spoken dialog systems," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 534–545, 2005.

[14] S. S. Pradhan and W. H. Ward, "Estimating semantic confidence for spoken dialogue systems," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2002.

[15] J. Godfrey, "Air Traffic Control Complete LDC94S14A," DVD, 1994. [Online]. Available: http://catalog.ldc.upenn.edu/LDC94S14A

[16] K. Hofbauer, S. Petrik, and H. Hering, "The ATCOSIM Corpus of Non-Prompted Clean Air Traffic Control Speech," in *Proc. of the 6th International Conference on Language Resources and Evaluation (LREC)*, 2008.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[18] 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 *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2003.

[19] J. Carletta, "Announcing the AMI meeting corpus," *The ELRA Newsletter*, vol. 11, no. 1, pp. 3–5, 2006.

[20] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks," in *Proc. of the 9th International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 3935–3939.

[21] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *Proc. of the International Conference on Implementation and Application of Automata*. Springer, 2007, pp. 11–23.

[22] H. Xu, D. Povey, L. Mangu, and J. Zhu, "An improved consensus-like method for Minimum Bayes Risk decoding and lattice combination," in *Proc. of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4938–4941.

[23] P. Motlicek, "Automatic Out-of-Language Detection Based on Confidence Measures Derived from LVCSR Word and Phone Lattices," in *Proc. of the 10th Annual Conference of the International Speech Communication Association*, 2009.

[24] D. Imseng, B. Potard, P. Motlicek, A. Nanchen, and H. Bourlard, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 2322 – 2326.

[25] G. Lecovre, J. Dines, T. Hain, and P. Motlicek, "Supervised and unsupervised web-based language model domain adaptation," in *Proc. of Interspeech*, 2012, pp. 131–134.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. of the IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.