Semi-supervised Training of Acoustic Models Leveraging Knowledge Transferred from Out-of-Domain Data

Tien-Hong Lo[1] and Berlin Chen[1,2]
[1]National Taiwan Normal University, Taipei, Taiwan
[2] Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan
E-mail: {teinhonglo, berlin}@ntnu.edu.tw

***Abstract*— More recently, a novel objective function of discriminative acoustic model training, namely lattice-free MMI (LF-MMI), has been proposed and achieved the new state-of-the-art in automatic speech recognition (ASR). Although LF-MMI shows excellent performance in a wide array of ASR tasks with supervised training settings, there is a dearth of work on investigating its effectiveness in the scenario of unsupervised or semi-supervised training. On the other hand, semi-supervised (or self-training) of acoustic model suffers from the problem that it is hard to estimate a good model when only a limited amount of correctly transcribed data is made available. It is also generally acknowledged that the performance of discriminative training is vulnerable to correctness of speech transcripts employed for training. In view of the above, this paper explores two novel extensions to LF-MMI. The first one is to distill knowledge (acoustic training statistics) from a large amount of out-of-domain data to better estimate the seed models for use in semi-supervised training. The second one is to make effective selection of the untranscribed target domain data for semi-supervised training. A series of experiments conducted on the AMI benchmark corpus demonstrate the gains from these two extensions are pronounced and additive, which also reveals their effectiveness and viability.**

I. INTRODUCTION

For many practical situations, in-domain speech training data without correct transcripts are much easier to collect than those with transcripts when building ASR systems. In the face of such situations, how to leverage limited available transcribed in-domain data in conjunction with a large amount of untranscribed (or imperfectly transcribed) in-domain data and/or a large amount of publicly available transcribed out-of-domain data becomes a topic of central concern. To this end, two orthogonal but complementary research directions are worthy of exploration: 1) transfer learning that distills knowledge from an outside domain that has sufficient and inexpensive data equipped with orthographic transcripts, and 2) semi-supervised training that makes effective use of a collection of untranscribed in-domain data.

Transfer learning, which manages to transfer knowledge from one domain to another, is inspired by the fact that humans has the ability to make clever use of knowledge learned beforehand to tackle new problems in a better and efficient manner. Transfer learning is also known by a variety of other names [1, 2, 3], including multi-task learning, inductive learning, cumulative learning, and among others. On a separate front, semi-supervised training addresses the issue that the transcribed data may be too few to build a good statistical model (e.g. classifier or recognizer), by making use of a small amount of transcribed data and a large amount of untranscribed data. In the context of ASR, most of the previous studies on semi-supervised training of acoustic models belong to the so-called self-training, which generally consists of two stages [4, 5, 6]. In the first stage, a prototype ASR system is constructed by training its (seed) acoustic model with limited labeled (or supervised) data. In the second stage, the prototype ASR system is employed to generate the transcripts of a large amount of unlabeled data, which in turn can be treated as augmented training data for estimating a refined acoustic model. However, the automatic transcripts are error-prone such that some mechanisms based on ASR confidence-based filtering can be applied to filter out unreliable data, which can be conducted at either the frame-level [7], word-level [8] or utterance-level [5, 7, 9].

More recently, a novel objective function of discriminative acoustic model training, namely LF-MMI, has been proposed and achieved the new state-of-the-art for ASR. Although LF-MMI shows excellent performance in a wide array of ASR tasks with supervised training settings, there has been little work on investigating its effectiveness in the scenario of unsupervised or semi-supervised training. On the other hand, semi-supervised of acoustic model suffers from the problem that it is hard to estimate a good model when only a limited amount of correctly transcribed data is made. It is also generally acknowledged that the performance of discriminative training is vulnerable to correctness of speech transcripts employed for training. In view of the above, this paper explores two novel extensions to LF-MMI [11, 12, 13]. The first one is to transfer knowledge (acoustic training statistics) from a large amount of out-of-domain data to better estimate the seed models for use in semi-supervised training. The second one is to make effective selection of the untranscribed target domain data for semi-supervised training. We also evaluate whether the benefits from the two extensions can add together. Our code is open sourced.[1]

[1]https://github.com/teinhonglo/ami-s5b-semi

II. RELATED WORK

A. Transfer learning

A number of studies have been conducted to develop transfer learning methods for use in speech and language processing with various settings; a comprehensive summary of recent related attempts can be found in [14]. For example, a simple linear input network (LIN) has been employed in [15] which try to adapt the acoustic model to a new domain by adjusting the network parameters, especially for speaker adaptation. This work has motivated many follow-up studies, such as feature space discriminant linear regression (fDLR) [16] and linear transform using linear hidden networks (LHN) at various stages within the component neural network of an acoustic model [17]. More recently, LHN-based adaptation and multitask-based adaptation of deep neural network (DNN) based acoustic models were compared for ASR in [18]. Notably, an appealing multitasking architecture has been successfully designed and developed for multilingual acoustic model training [19, 20]. For transfer learning, it was found that not only the amount of training data but also the relatedness of the tasks was found to be important for its practical effectiveness [21, 22].

B. Semi-supervised training

For acoustic modeling, semi-supervised training (sometimes called self-training) is developed to lessen the need for large volumes of transcribed training speech, which plays a crucial role when building an ASR system for a resource-scarce task or reconfiguring it for a new domain. It is common practice to utilize a small amount of orthographically-transcribed training speech utterances to build an initial acoustic model which is then employed to generate automatic transcripts for a large amount of untranscribed training speech utterances. The orthographically- and automatically-transcribed training speech are then used in conjunction to estimate a better acoustic model. However, empirical evidence in the literature indicates that discriminating training algorithms (including LF-MMI) are sensitive to the accuracy of transcripts of training speech [11, 12, 13]. Therefore, previous studies had largely focused on designing confidence-based filters to select automatically transcribed training speech segments that are expected to be useful for semi-supervised training of acoustic models [4, 5, 6]. For example, a frame-level confidence filter is adopted in the discriminant training process to preserve important training speech segments [23]. In addition, an utterance-level confidence-based filter in combination with one-best results was investigated in [24]. Further, the authors in [25] attempted to use ASR lattice posteriors [26, 27] meanwhile retaining the whole lattices of automatically-transcribed training speech segments for semi-supervised training. Our approach proposed in this paper bear a close resemblance in spirit to that proposed in [25], with the key distinction that our re-trained acoustic model is not based on random initialization but instead on top of a pre-trained model which is powered by weight transfer or multitask learning, stemming from transfer learning.

III. TRAINING CRITERION FOR LOW-RESOURCE ASR

For supervised training, the objective of LF-MMI can be expressed as the summation of the conditional log-likelihoods of the reference (orthographic) transcripts of training utterances given their acoustic feature vector sequences [28]:

$$\mathcal{F}^{\text{LFMMI}} = \sum_u \log P(S_u|O_u, \lambda) \tag{1}$$

where, S_u and is the reference transcript of utterance u, and $P(S_u|O_u, \lambda)$ is the probability of S_u given the acoustic feature vector sequence O_u of utterance u and the model parameter λ. However, for semi-supervised training, the automatic transcripts of training speech utterances are not necessarily correct. Therefore, we can rewrite the formula as follows:

$$\mathcal{F}^{\text{SemiLFMMI}} = \sum_u \alpha_{S'_u} \log \frac{P(O_u|S'_u, \lambda_{\text{seed}})P(S'_u)}{\sum_{S''_u} P(O_u|S''_u, \lambda_{\text{seed}})P(S''_u)} \tag{2}$$

where S'_u is top ranking automatic transcript of utterance u, S'' belongs to all possible automatic transcripts and λ_{seed} denotes the seed model. The weighting factor $\alpha_{S'_u}$ controls the contribution of S''_u to the training and can be calculated through the following formula:

$$\alpha_{S'_u} = \begin{cases} 1, & u \in \text{transcribed data} \\ P\left(O_u \middle| S'_u, \lambda_{\text{seed}}\right), & u \in \text{untranscribed data} \end{cases} \tag{3}$$

Furthermore, the optimization of Equation (2) can be broken down into two problems: 1) how to improve the "quality" of the seed model? and 2) how to determine the weighting factor? We will defer the tackling of the latter one to the next section. The former one hinges on several factors, including initialization, model structure, training criteria, and more; any kind of adjustment will affect the performance of the model.

A. Improving the seed model

In this paper, we seek to leverage the conception of transfer learning to improve the accuracy of the seed model. To this end, we use the weight transfer strategy advocated in [29]. A common practice of weight transfer is to conduct a two-stage training process by first freezing the low-level layers and train task-specific layers at the first stage, and then fine-tuning the whole neural network at the second stage using a smaller learning rate [22]. In contrast, the method proposed in [29] tries to train the transferred layers with a smaller learning rate, while simultaneously training the task-specific layers with a larger learning rate, with a single-stage training setup.

B. The weighting factor of the untranscribed data

For semi-supervised training, we use the recipe proposed in [25]. Unlike the traditional semi-supervised training methods that adopt the frame-level, word-level or utterance-level confidence scores generated by the seed model to solicit possibly correct automatically transcribed data for training, the method proposed in [25] uses the entire lattice pertaining to a training utterance as the supervision. In concordance with the fundamental procedure for semi-supervised training, we first use the LF-MMI based acoustic model trained with a small amount of orthographically transcribed speech data or trained with transfer learning to decode those in-domain data

Table 1: AMI Corpus (Semi-supervised setup)

	Train (supervised)	Train (unsupervised)	Dev.	Eval.
Hrs.	16	62	8.71	8.97
Utts.	20,000	88,104	13,059	12,612

without orthographic transcripts to obtain the lattice of alternative pronunciations for a speech utterance. In turn, the word lattices of speech utterances are converted to their phone lattices as depicted in [10]. Finally, we take the resulting phone lattices as the supervision and perform the standard LF-MMI training on it by weighting the top ranking hypotheses with their corresponding posterior scores.

$$\mathcal{F}^{\mathrm{NCE}} = \sum_{u} \sum_{S'_u} P(S'_u|O_u, \lambda_{\mathrm{seed}}) \log P(S'_u|O_u, \lambda_{\mathrm{seed}})$$

$$= -\sum_{u} H(S'_u|O_u, \lambda_{\mathrm{seed}}) \quad (4)$$

where $P(S'_u|O_u, \lambda_{seed})$ is lattice posterior score of a training speech utterance without its corresponding orthographic transcript, and set to be 1 otherwise. We can refer to Equation (4) as the negative conditional entropy (NCE), denoted by $-H\left(S'_u \middle| O_u, \lambda_{seed}\right)$, of the transcript S'_u under the condition that the model parameter λ_{seed} and its corresponding acoustic feature vector sentence O_u are given [17, 18, 28]. Therefore, Equation (4) can naturally account for the quality of the top ranking transcripts when conducting the LF-MMI based discriminative training in a semi-supervised manner, which is anticipated to improve the training performance without resource to the confidence filter. Note here that our method can be viewed as a substantial extension of the one that was proposed in [25], the key distinction that retaining the initialized parameters of the seed model and limiting the learning rate are employed when updating the model parameters.

IV. Empirical Experiments

A. Experimental setup

We evaluate our proposed approach on the AMI meeting transcription database and task [31], while all experiments are conducted using Kaldi toolkit [30]. For the AMI database, the speech corpus consisted of recordings from the individual headset microphones (IHM), while a pronunciation lexicon of 50K words was used. Trigram language models were trained on the AMI training set. We set aside a randomly chosen subset of speech utterances (62 hours) from the corpus as the untranscribed data, while the others as the transcribed data. Table 1 shows some basic statistics of the AMI corpus in our experiments. The word error rate (WER) improvements from semi-supervised training are evaluated by Absolute Improvement (AI) and WER Recovery Rate (WRR) [32]:

$$\mathrm{WRR} = \frac{BaslineWER - SemisupWER}{BaselineWER - OracleWER} \quad (5)$$

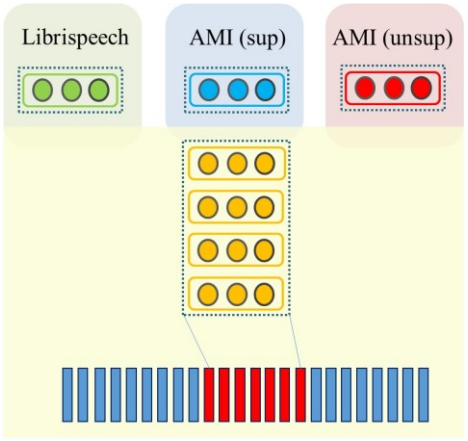

Figure 1: A schematic depiction of the overall training strategy.

1) Baseline approach

For acoustic modeling, our basic recipe is to first train a GMM-HMM acoustic model using the speech utterances with corresponding orthographic transcripts and use the prior distribution of the senones obtained from the GMM components, in conjunction with the LF-MMI training objective, to build the time-delay neural network (TDNN) acoustic model [10], which is in turn taken as the seed model. The speech feature vectors are 40 MFCC coefficients extracted in 25 ms long windows every 10 ms, augmented with 100-dimensional i-vectors [33] for speaker adaptation of TDNN. To rule out the effect of the i-vector extractor, we trained the i-vector extractor combining both transcribed and untranscribed datasets. Furthermore, for comparison purposes, we only used statistics obtained from the untranscribed dataset to train the context-dependent decision tree. The phone language model used for creating the denominator FST was estimated using phone sequences gathered from both transcribed and untranscribed datasets as in [29]. We, however, give a slightly higher weight to the phone sequences extracted from the transcribed data set. Figure 1 is a schematic depiction of the overall training strategy which consists of three steps. In the first step (involving the yellow and green blocks), we obtain the acoustic model trained on Librispeech. In the second step (involving the yellow and blue blocks), The knowledge learned from Librispeech is transferred to the AMI through the additional use of a 16-hour supervised data (AMI). In the third step (involving the yellow, blue and red blocks), we retain the parameters of the seed model and use the unsupervised data to further enhance the acoustic model.

2) Transfer learning for estimating the seed model

We used the Librispeech dataset [34], which contains about 1,000 hours of audiobook speech recordings, as the orthographically transcribed out-of-domain data. For acoustic modeling of the Librispeech dataset, the number of hidden layers for the TDNN architecture was set to 6, which is the same as the baseline system for the AMI dataset. The overall

Table 2. WER (%) and WRR results of Semi-supervised training

Supervision	Dev.	Eval.	WRR
Baseline	27.2	27.8	-
NW	26.2	26.8	24%
TOP	26.0	26.2	33%
LS	25.5	25.7	45%
WLS	**24.7**	**25.3**	**60%**
Oracle	23.5	23.1	-

Table 3. Weight transfer in conjunction with semi-supervised training

Supervision	Dev	Eval	WRR	AI
Baseline(AMI)	27.2	27.8	-	-
TF (LIB2AMI)	24.8	25.2	60%	2.4%
TF (LIB2AMI) + LS	24.5	24.6	70%	2.9%
TF (LIB2AMI) + WLS	24.2	24.3	78%	3.2%
TF (LIB2AMI) + WLS_ES	**23.8**	**23.8**	**88%**	3.7%
Oracle	23.5	23.1	-	-

WER of the acoustic model trained by Librispeech, pertaining to its four test sets (Dev, Dev_Other, Test and Test_Other), are 3.72, 9.90, 4.18, and 10.37, respectively. This is in parallel with the previous results in the literature, revealing that the TDNN-based acoustic model can give very good ASR performance when sufficient training data (e.g., 1,000 hours) is made available. Despite this, for many new ASR application domains or tasks, we are always facing the issue of lacking orthographically transcribed training data.

In the experiments of transfer learning, we will re-initialize an affine layer instead of training an extra layer as the transferred layer and train the transferred layer with a smaller learning rate while training the task-specific layers with a larger learning rate in the single-stage training. Put another way, the difference between the training approaches stated in Section A and Section B is the initialization of the seed acoustic model and ASR retraining criterion for the semi-supervised setup.

B. Experimental results

1) Semi-supervised training using lattices as supervision

In the first set of experiments, we assess the performance level of adding the lattice posterior scores (i.e., weighting factors) and using lattices (with a beam size set to 4) as supervision for semi-supervised LF-MMI training. The corresponding results are shown in Table 2, where the acoustic model of "Baseline" was trained with the 16-hour training utterances equipped with their orthographic transcripts. "NW" refers to no weighting, for which the acoustic model was trained by simple putting 16-hour orthographically transcribed and 62 untranscribed datasets together as the training corpus. "TOP" is similar to "NW" by using the top-one recognition hypothesis of an untranscribed training utterance as its automatic transcript except that a weight factor is additionally included in the LF-MMI training formula, as previously shown in Equation (2). "LS" denotes that the top-M recognition hypotheses (M is equal to 4) of an untranscribed training utterance as taken as its automatic transcripts, with their respective weight factors additionally included in the LF-MMI training formula. "Oracle" stands for the condition that the all training speech utterances of AMI (amounted to 72 hours) were equipped with their corresponding orthographic transcripts.

As can be seen from Table 2, when the traditional two-stage semi-supervised training is directly carried out, the WRR can reach a moderate performance level of 24%, thanks to the well-trained seed model by LF-MMI, which is achieved with only 16 hours of training utterances. The use of lattice posterior scores as the weighting factors (denoted by TOP) can further increase WRR to 33%, which also manifests that the weighting factors (or the notion of minimizing negative conditional entropy) can effectively work in conjunction with LF-MMI to assist semi-supervised training. Therefore, from now on, unless otherwise stated, all results for the subsequent experiments of semi-supervised training will make use of lattice posterior scores. LS that capitalizes on the lattices for supervision by using the top-*M* recognition hypotheses of an untranscribed training utterance as its automatic transcripts achieves better WRR than TOP. Lastly, WLS is the extension of the LS, which retains the parameters of the seed and limits the learning rate during the process of semi-supervised training, achieves the best WRR of 60 %. This result also confirms the usefulness of retaining more recognition hypotheses for untranscribed training utterances which can benefit the semi-supervised acoustic model training.

2) Weight transfer for semi-supervised training

In the second set of experiments, we turn to evaluate the utility of applying weight transfer from an outside domain (namely, Librispeech) to the AMI task with the semi-supervised training setup; the corresponding results are shown in Table 3. The acoustic model of "Baseline (AMI)" was trained merely with 16-hour orthographically transcribed utterances of AMI. "TF (LIB2AMI)" is intended to pre-train the acoustic model with Librispeech, followed by using the aforementioned 16-hour utterances of AMI to transfer knowledge from Librispeech to AMI through weight transfer. We then can obtain the automatic transcripts of the rest 62-hour AMI untranscribed utterances for the subsequent semi-supervised training setups, namely, "TF (LIB2AMI) + LS", "TF (LIB2AMI) + WLS" and "TF (LIB2AMI) + WLS_ES." They all are counterparts of "LS" shown in Table 2. For "TF (LIB2AMI) + WLS", its seed acoustic model was pre-trained with the Librispeech corpus, in contrast to "TF (LIB2AMI) + LS" that adopted random initialization. Further, "TF (LIB2AMI) + WLS_ES" additionally employed the early stopping scheme in ration to "TF (LIB2AMI) + LS." Several observations can be drawn from Table 3. First, the system "TF (LIB2AMI)" that used Librispeech as the source for model pre-training as well as the 16-hour orthographically transcribed utterances of AMI for subsequent model training, can deliver superior performance than the best system shown in Table 2. This also indicates that the knowledge of out-of-domain data can be simply transferred to a target domain via weight transfer, whose effect is more significant than including untranscribed data for semi-supervised training. Second, semi-supervised training using the Librispeech dataset as the source for model pre-training can further benefit

semi-supervised training ("TF (LIB2AMI) + WLS" vs. "TF (LIB2AMI) + LS"). Finally, our best system ("TF (LIB2AMI) + WLS_ES") can achieve the highest WRR of 88%.

V. CONCLUSIONS

In this paper, we have explored two novel extensions to the standard TDNN-based acoustic modeling with the LF-MMI training criterion. The first one is to distill knowledge (acoustic training statistics) from a large amount of out-of-domain data to better estimate the seed models for use in semi-supervised training. The second one is to make effective selection of the untranscribed target domain data for semi-supervised training. Empirical experiments conducted on the AMI benchmark corpus have demonstrated the effectiveness and viability of our approach. As to future work, we are interested in applying our approach to the sequence-to-sequence ASR paradigms, including CTC (connectionist temporal classification), attention model, and their fusion. We also plan to explore more sophisticated semi-supervised training, transfer learning and their combination with disparate discriminative acoustic model training criteria.

ACKNOWLEDGMENT

This research is supported in part by the National Science Council, Taiwan, under Grant Number MOST 107-2634-F-008-004- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan, and Grant Numbers MOST 105-2221-E-003 -018 -MY3, MOST 107-2221-E-003 -013 -MY2 and MOST 108-2221-E-003-005-MY3.

REFERENCES

[1]. S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[2]. J. Lu et al, "Transfer learning using computational intelligence: a survey," *Knowledge-Based Systems*, vol. 80, pp. 14–23, 2015.

[3]. Y. Bengio et al., "Deep learning of representations for unsupervised and transfer learning." *ICML Workshop*, vol. 27, pp. 17–36, 2012.

[4]. K. Vesely et al., "Semi-supervised training of deep neural networks," in *Proc. ASRU*, 2013.

[5]. F. Grezl et al., "Semi-supervised bootstrapping approach for neural network feature extractor training," in *Proc. ASRU*, 2013.

[6]. P. Zhang et al., "Semi-supervised DNN training in meeting recognition," in *Proc. SLT*, 2014.

[7]. K. Vesely et al., "Semi-supervised training of Deep Neural Networks," in *Proc. ASRU*, 2013.

[8]. S. Thomas et al., "Deep neural network features and semisupervised training for low resource speech recognition," in *Proc. ICASSP*, 2013.

[9]. P. Zhang et al., "Semi-supervised DNN training in meeting recognition," in *Proc. SLT*, 2014.

[10]. D. Povey et al., "Purely sequence-trained neural networks for ASR Based on Lattice-Free MMI," in *Proc. Interspeech*, 2016.

[11]. L. Mathias et al., "Discriminative training of acoustic models applied to domains with unreliable transcripts." in *Proc. ICASSP*, 2005.

[12]. K. Yu et al., "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 25, no. 2–3, pp. 652–663, 2010.

[13]. X. Cui et al., "Multi-view and multi-objective semi-supervised learning for large vocabulary continuous speech recognition," in *Proc. ICASSP*, 2011.

[14]. D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Proc. APSIPA*, pp. 1225–1237, 2015.

[15]. J. Neto et al., "Speaker adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. Eurospeech*, pp. 2171–2174, 1995.

[16]. K. Yao et al., "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. SLT*, pp. 366–369, 2012.

[17]. R. Gemello et al., "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.

[18]. Z. Huang et al., "Rapid adaptation for deep neural networks through multi-task learning," in *Proc. Interspeech*, 2015.

[19]. G. Heigold et al., "Multilingual acoustic models using distributed deep neural networks," in *Proc. ICASSP*, pp. 8619–8623, 2013.

[20]. J.-T. Huang et al., "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, pp. 7304–7308, 2013.

[21]. F. Grezl et al., "Study of large ´data resources for multilingual training and system porting," *Procedia Computer Science*, vol. 81, pp. 15–22, 2016.

[22]. R. Sahraeian and D. V. Compernolle, "Using weighted model averaging in distributed multilingual DNNs to improve low resource ASR," *Procedia Computer Science*, vol. 81, pp. 152–158, 2016.

[23]. S.-H. Liu et al., "Investigating data selection for minimum phone error training of acoustic models," in *Proc. ICME*, 2007.

[24]. S. Walker et al., "Semi-supervised model training for unbounded conversational speech recognition," *arXiv*, 2017.

[25]. V. Manohar et al., "Semi-supervised training of acoustic models using lattice-free MMI," in *Proc. ICASSP*, 2018.

[26]. Y. Grandvalet et al., "Semi-supervised learning by entropy minimization," in *Proc. NIPS*, 2005.

[27]. J.-T. Huang et al., "Semi-supervised training of Gaussian mixture models by conditional entropy minimization," in *Proc. Interspeech*, 2010.

[28]. A. Nadas, "A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 4, pp. 814–2817,1983.

[29]. P. Ghahremani et al., "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in *Proc. ASRU*, pp. 279–286, 2017.

[30]. I. McCowan et al., "The AMI meeting corpus," in *Proc. ICMTBR*, 2005.

[31]. D. Povey et al., "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[32]. J. Ma and R. Schwartz, "Unsupervised versus supervised training of acoustic models," in *Proc. Interspeech*, 2008.

[33]. N. Dehak et al., "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[34]. V. Panayotov et al., "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.