SEMI-SUPERVISED DNN TRAINING IN MEETING RECOGNITION

Pengyuan Zhang[1]*, Yulan Liu*[2]*, Thomas Hain*[2]

[1]Key Laboratory of Speech Acoustics and Content Understanding, IACAS, Beijing, China
[2]Speech and Hearing Research Group, The University of Sheffield, Sheffield, UK

pzhang@hccl.ioa.ac.cn, acp12yl@sheffield.ac.uk, t.hain@dcs.shef.ac.uk

ABSTRACT

Training acoustic models for ASR requires large amounts of labelled data which is costly to obtain. Hence it is desirable to make use of unlabelled data. While unsupervised training can give gains for standard HMM training, it is more difficult to make use of unlabelled data for discriminative models. This paper explores semi-supervised training of Deep Neural Networks (DNN) in a meeting recognition task. We first analyse the impact of imperfect transcription on the DNN and the ASR performance. As labelling error is the source of the problem, we investigate two options available to reduce that: selecting data with fewer errors, and changing the dependence on noise by reducing label precision. Both confidence based data selection and label resolution change are explored in the context of two scenarios of matched and unmatched unlabelled data. We introduce improved DNN based confidence score estimators and show their performance on data selection for both scenarios. Confidence score based data selection was found to yield up to 14.6% relative WER reduction, while better balance between label resolution and recognition hypothesis accuracy allowed further WER reductions by 16.6% relative in the mismatched scenario.

Index Terms: semi-supervised acoustic model training, confidence selection, deep neural networks

1. INTRODUCTION

Acoustic models for state-of-the-art speech recognition systems are typically trained on several hundred hours of task dependent training data. Increasing the amount of training data consistently improves performance, especially with discriminative learning techniques. However, due to domain specificity, there are low resource scenarios where annotated training data can be especially expensive to obtain, and the acoustic model has to be trained either with much less transcribed training data or with mismatched data [1, 2, 3]. For this reason, much effort has been devoted to unsupervised and semi-supervised acoustic modelling [4, 5, 6].

Recently, acoustic modelling based on the Deep Neural Networks (DNNs) has gained popularity with the consistent improvement in recognition performance over earlier Neural Network based front-ends (e.g. [7]). DNNs are either deployed as the front-end for standard Hidden Markov Model based on Gaussian Mixture Models (HMM-GMMs), or in a hybrid form to directly estimate state level posteriors. As noted in several publications [8, 9, 10, 11], DNNs show general Word Error Rate (WER) improvements on the order of 10-30% relative across a variety of small and large vocabulary tasks when compared with HMM-GMMs built on classic features (e.g. MFCC, PLP).

Semi-supervised training of acoustic models for ASR was successfully implemented and tested in several key domains. One typical set of examples are the self-training methods [12, 13, 14, 15]. In this method the transcribed data is used to construct a seed model, with which to decode the untranscribed data at the second stage. The most reliable transcriptions are selected based on confidence measures, and are used for further optimization of the acoustic models.

This paper investigates semi-supervised training of DNNs for the purpose of extracting features, which are then used to train standard HMM-GMM acoustic models. In the first stage a seed DNN and the corresponding seed HMM-GMMs are trained on a small amount of transcribed data. Both the seed DNN and HMM-GMM are improved with the benefit from transcription of data with the seed system, by retraining both the discriminative front-end feature extraction and the acoustic models. Confidence based data selection can also be employed to discard unreliable data.

Discarding unreliable data is one way to address the issues of semi-supervised learning. A second option is to use model training methods that are more robust to labelling errors. Discriminative models such as DNNs suffer in particular from such errors. One unique option for DNN based front-ends is to reduce the label resolution. This allows the first-pass recognition to have higher accuracy which in turn may yield more useful transcription data.

This paper focuses on both improving confidence score estimation as well as on exploring the effect of label resolution. DNNs trained with phone-class, monophone, monophone state and triphone state targets are compared.

The rest of this paper is organized as follows. Section §2 reviews semi-supervised training. Section §3 starts with system structure used in §3.1, and the baseline experiments

with a limited amount of transcribed training data in §3.2. Section §3.3 presents improved performance with the self-training and proposes a novel confidence measure. Section §4 concludes the paper.

2. SEMI-SUPERVISED TRAINING

2.1. Previous work on semi-supervised DNN training

Existing work on semi-supervised training often starts with a baseline recogniser trained on either limited or slightly mismatched, but transcribed data. This baseline recogniser serves as the seed decoder to generate recognition hypothesis for untranscribed training data. The segment based hypothesis transcription is then evaluated according to its reliability. The segments with most reliable transcriptions are combined with the limited fully transcribed training data to optimize the baseline system - in our case both the DNN front-end and the HMM-GMM acoustic models. An optimal data selection strategy is key to semi-supervised training. Typically this is performed by evaluating the hypothesis transcription reliability using confidence measurement, given prior models. In [3], the reliability evaluation is realized with the ASR-based word confidence scores and the MLP posteriogram-based phoneme occurrence confidence. Confidence scores can be calculated at frame level and word level, but are often expressed at utterance level [13, 14] for the convenience of data selection and the reliability of confidence estimates.

2.2. Confidence measures for data selection

There are several strategies for confidence score estimation. One strategy [3] used the posteriors generated by DNNs. Each hypothesis word from the seed decoding is mapped to a set of constituent phonemes according to a pronunciation lexicon. The DNN posteiors corresponding to the phonemes in the hypothesis utterances are selected according to the lexical mapping for the utterance's posteriogram representation, by averaging over the posteriorgram of composing words. For an hypothesis word w_i, the frame based posteriors corresponding to the hypothesis phoneme are binarized using a set threshold to indicate the phoneme's presence or absence. The average occurrence ratio of the constituent phonemes in the hypothesized time span $[t_s, t_e]$ (or time indices) along a Viterbi path is then used as the confidence score $C_{occ}(w_i, t_s, t_e)$ for hypothesis word w_i. The assumption behind this strategy is that if a phoneme is hypothesized correctly, it is likely that all its constituent frames will be present in the posteriogram, hence leading to a high average occurrence count [3]. With w_i as the i-th word in the hypothesis utterance, the count-based word level confidence score is thus

$$C_{occ}(w_i, t_s, t_e) = \frac{c_{occ}(t_s, t_e)}{N_{w_i}} \tag{1}$$

where c_{occ} is the total count of phoneme occurrences in word w_i in the hypothesized interval (t_s, t_e), and N_{w_i} is the total number of frames in the word w_i. The utterance level score is the average of all word level scores in the utterance, i.e.

$$C_{occ}(u) = \frac{1}{K} \sum_{i=1}^{K} C_{occ}(w_i, t_s, t_e) \tag{2}$$

where K is number of words in the hypothesis utterance u.

This confidence measure can be modified to emphasize more on the DNN posteriors. The DNN posterior based phoneme sequence for each utterance can be generated by finding the phoneme of highest posterior estimation, i.e. the DNN softmax output value, for each frame. This DNN posterior based phoneme sequence is then compared with the aligned hypothesis from the seed decoder HMM-GMMs. A frame-based phoneme occurrence count is accumulated into c_{agr}, when both hypothesis sequences agree on a phoneme for a specific frame. The confidence score of the i-th word w_i in hypothesis utterance u is thus

$$C_{agr}(w_i, t_s, t_e) = \frac{c_{agr}(t_s, t_e)}{N_{w_i}} \tag{3}$$

The utterance level confidence score is calculated as the arithmetic mean of the word level confidence scores, i.e.

$$C_{agr}(u) = \frac{1}{K} \sum_{i=1}^{K} C_{agr}(w_i, t_s, t_e) \tag{4}$$

where K is number of words in the hypothesis utterance u.

A further modification can be made by utilising the frame based posteriors directly rather than counting the number of tokens in agreement. Denote p_k as the DNN posterior estimate value at frame k on the phoneme hypothesized by the seed HMM-GMMs decoder of that frame. The word level confidence score represents the acoustic confidence of DNNs on the hypothesis word from the seed decoder, by accumulating the log posterior on each frame corresponding to that hypothesis word w_i

$$C_{pos}(w_i, t_s, t_e) = \frac{1}{t_e - t_s + 1} \sum_{k=t_s}^{t_e} (\log p_k) \tag{5}$$

Again the utterance level confidence score is taken as the arithmetic mean of the word level confidence scores, i.e.

$$C_{pos}(u) = \frac{1}{K} \sum_{i=1}^{K} C_{pos}(w_i, t_s, t_e) \tag{6}$$

2.3. Target resolution and accuracy

In supervised training of DNN front-end based ASR system, our prior experimental work indicates that training targets

Table 1. Information about the dataset used in two semi-supervised training scenarios. (Dur.: duration; #Utt.: number of utterances; #Words: number of words.)

Dataset	Dur.	#Utt.	#Words	Corpus	Scenario
acntrain	15.8h	12876	152876	AMI	S1
acotrain	72.0h	60297	710850	AMI	S1
icsi	10.0h	7268	126487	ICSI	S2
acftrain	87.8h	73173	863726	AMI	S2, S1
acftest	6.1h	4633	54820	AMI	S1, S2

of higher resolution improves the overall recognition performance [11] even for challenging tasks. However in unsupervised and semi-supervised training of the DNN front-end, a higher resolution of imperfect hypothesis targets results in a DNN biased to the errors. The hypothesis on targets of higher resolution such as triphone states is more vulnerable to noise than the targets of lower resolution, such as monophone states or monophones. Thus in the semi-supervised training of DNN front-end, it can be expected that in this setting a compromise between the target resolution and target accuracy has to be found. The experiments investigate the optimal balance between these two factors.

3. EXPERIMENTS

3.1. Data and system configuration

Experiments are performed using the headset recording data from two meeting corpora AMI [16] and ICSI [17]. Two typical scenarios in semi-supervised training are investigated. In the first scenario (S1) only a very limited amount of data is transcribed and available in training. In the second scenario (S2) no transcribed data on the target corpus (AMI) is available at all for training, however there is limited transcribed data of the same domain (ICSI).

For scenario 1 (S1), 15.8 hours non-overlapping speech from the AMI corpus are selected as the seed training set with transcriptions *acntrain*, and 6.1 hours of data composed of both overlapping and non-overlapping speech is chosen as the test set *acftest*. The remaining 71.9 hours data in *acftrain* (as defined in [11]) are retained to simulate the untranscribed data set *acotrain*.

For scenario 2 (S2), 10 hours of speech data was carefully selected from the ICSI corpus, to cover most speakers present in the corpus while excluding the digit-reading part. This serves as the transcribed out-of-corpus training set *icsi.nodgt-10h* (or *icsi* in short in this paper). The whole AMI corpus is considered untranscribed for this scenario. All recognition evaluation is performed on the same test set *acftest* with S1. Table 1 lists the statistic details of the dataset used.

DNN front-end configurations follow the setup used in [11], with the same topology being used in all experiments: 368 dimensional input is composed of the compressed log Mel-filterbank features from 31 adjacent frames with global mean and variance normalization; 3 hidden layers are composed with 1745 nodes each, followed by a 26 dimensional bottleneck layer, and the output layer of slightly varied dimension depending on training targets. The 26 dimensional linear BottleNeck (BN) features are concatenated with standard 39 dimensional PLP features to train PLP-BN HMM-GMMs with Single Pass Retraining (SPR) from the corresponding PLP models, followed by 8 iterations of Baulm-Welch re-estimation (denoted with '*spr*').

Further, the triphone states in PLP-BN HMM-GMMs are re-clustered to roughly 4000 states, and HMM-GMM parameters are optimized with maximum likelihood criterion in a standard HTK mixup procedure (denoted with '*xwrd*').

Table 2. %Word Error Rate (WER) on test set *acftest* with the baseline system in two scenarios. Scoring is performed with *sclite* in the NIST scoring toolkit *sctk 2.4.8*. "*#state*" refers to number of states

Feature	Data	Training	#states	%WER
PLP	acntrain	xwrd	1259	35.5
PLP+BN	acntrain	spr	1259	30.5
PLP+BN	acntrain	xwrd	1262	30.5
PLP	icsi	xwrd	1259	46.9
PLP+BN	icsi	spr	1259	41.4
PLP+BN	icsi	xwrd	1521	40.4

3.2. Baseline experiments

Table 2 shows the recognition performance in S1: having limited amounts of transcribed data *acntrain* from AMI corpus; and S2: having limited amount of transcribed data *icsi* from the ICSI corpus. The number of triphone states is kept to be approximately the same after re-clustering. All DNNs are trained using triphone state targets at this stage.

The results show a significant performance degradation when HMM-GMM models and DNN front-ends are not trained on corpus specific data. In fact the difference remains similar before and after application of DNN front-ends.

3.3. Confidence score

For S1, the seed models for decoding are trained on *acntrain* and used to obtain the hypothesis on *acotrain*. The confidence scores C_{occ}, C_{agr} and C_{pos} are calculated. Utterances from *acotrain* are then selected based on the confidence scores. Figure 1 shows the performance of the three confidence computation methods for data selection. The figure displays the word error rate (WER) of data chosen as a function of the number of words selected. This is preferred to representation as the percentage of segments chosen, as it better reflects the amount of training data obtained. Confidence based data selection requires the WER curves to be as low as possible.

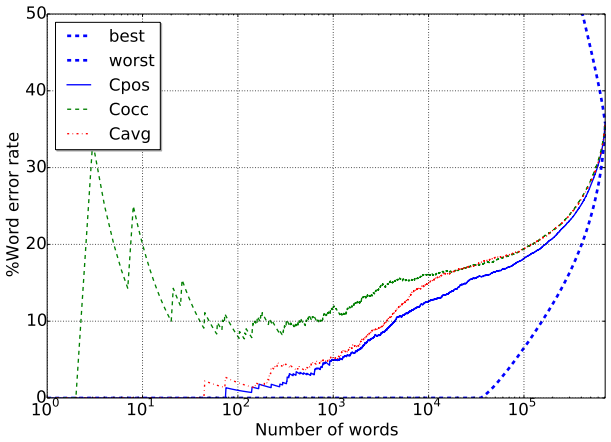

Fig. 1. Comparison of hypothesis utterances WER selected using different confidence scores.

The best and worst possible curves are also shown in the figure. The "best" curve refers to selecting segments with lowest segment level WER, while the "worst" curve refers to always selecting segments with highest segment level WER.

The results indicate that both confidence scores C_{agr} and C_{pos} are more effective than C_{occ}, as introduced in [3] in filtering out unreliable hypothesis utterances. Using soft decisions given by posterior scores rather than binary occurrence counts, C_{pos} performs consistently better than C_{agr}. Therefore in all further experiments, hypothesis selection is based on C_{pos}.

Table 3 shows the improvement in recognition performance in both scenarios brought by adding the best 70% of *acotrain* hypothesis data to *acntrain*, based on C_{pos} score data selection. In S2, adding hypothesis data to HMM-GMM training alone gives considerable gains, benefiting from less mismatch between training data and test data. Further WER reduction in S2 is achieved by adding task-matched hypothesis data to DNN training as well, contributing to an overall 14.6% relative WER reduction from seed system baseline. However in S1, less improvement is observed from adding hypothesis data. Adding hypothesis data to DNN training degraded the performance a little bit over adding hypothesis data to HMM-GMM training, especially when the added all data without selection.

3.4. Label resolution and accuracy

As shown in previous sections, DNN training is susceptible to label errors and confidence based hypothesis data selection help to reduce this problem. As outlined in Section §2.3, changing the label resolution could give a better balance in DNN training with imperfect hypothesis data. Table 4 illustrates how the target label accuracy changes with label resolution in the hypothesis transcriptions, in both scenario S1 and scenario S2.

Table 4 shows that the label accuracy improves by reduction in hypothesis label resolution. Notably the improvement in S1 is significantly larger without the out-of-corpus mismatch between seed system and test data in S2. For example the frame accuracy for monophone labels ("M") is 31% relatively higher than that for triphone state ("TS") in S1, while only 18% relatively higher in S2.

Figure 2 shows the recognition performance when changing hypothesis target label resolution in DNN front-end training. In all cases, for both scenarios, the use of monophone class ("MC") gives the worst recognition performance as the label precision is too low. In S1 the best performance is observed when using triphone states as DNN training targets in all cases. In S2, the best performance for seed system with manual transcription data is observed when using triphone states as DNN training targets, while with hypothesis data added the best performance is achieved using monophone as DNN training targets. Recall that Table 4 shows that the frame accuracy in S1 is much higher than the frame accuracy in S2, because of a task-matched seed system. In S2, by reducing the label resolution from triphone state level to monophone level, the recognition performance is improved up to 1.4% absolutely. The influence from mismatch between the seed system and the test data is reduced with a better balance between label accuracy and label resolution. With 70% data selected, 16.6% WER reduction was achieved relatively com-

Table 3. %WER change when adding manual transcription ("ref"), all hypothesis data ("100% hyp") and selected hypothesis data ("70% hyp") in DNN and HMM-GMM training. All DNNs are trained on triphone state targets; hypothesis data is selected based on C_{pos}.

Scenario	Data	%WER
S1	seed system	**30.5**
	100% ref→HMM-GMM	*26.3*
	100% ref→DNN&HMM-GMM	*24.3*
	100% hyp→HMM-GMM	29.2
	70% hyp→HMM-GMM	**29.1**
	100% hyp→DNN&HMM-GMM	29.8
	70% hyp→DNN&HMM-GMM	29.3
S2	seed system	**40.4**
	100% ref→HMM-GMM	*29.2*
	100% hyp→HMM-GMM	35.5
	70% hyp→HMM-GMM	35.8
	100% hyp→DNN&HMM-GMM	35.1
	70% hyp→DNN&HMM-GMM	**34.5**

Table 4. %Frame accuracy (FAC) on all hypothesis data at different label levels. TS: triphone states; MS: monophone states; M: monophone; MC: monophone class (8 in total).

	Seed data	hyp data	TS	MS	M	MC
S1	acntrain	acotrain	60.1	63.7	78.9	84.9
S2	icsi	acftrain	45.7	51.0	54.1	63.6

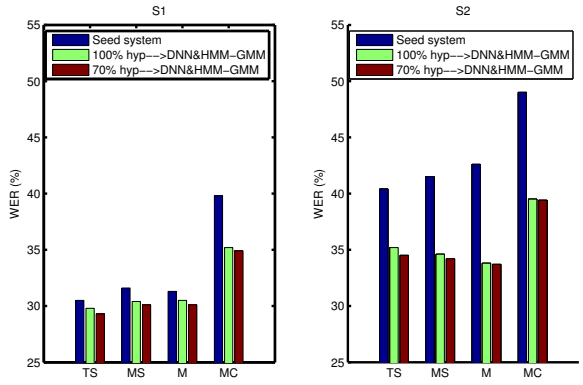

Fig. 2. %WER change when decreasing hypothesis label resolution in DNN front-end training: (left) Scenario 1; (right) Scenario 2.

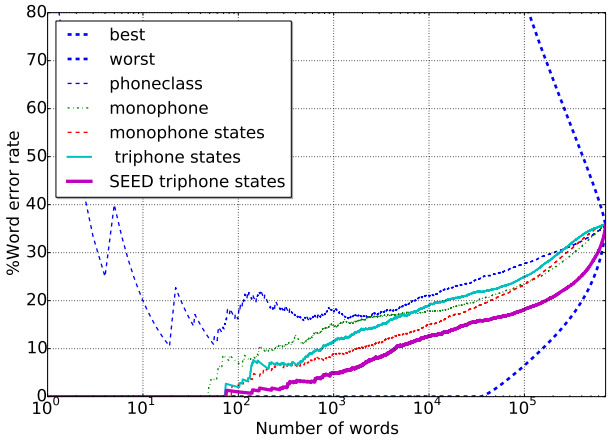

Fig. 3. Updating the confidence score based selection using DNN and HMM-GMMs trained with AMI seed transcription and all hypothesis transcription by the AMI seed system (S1)

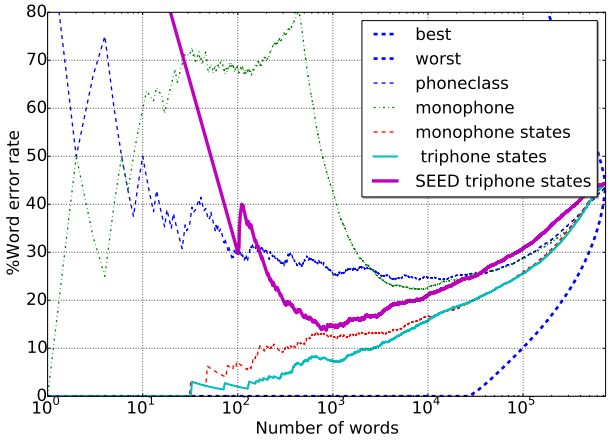

Fig. 4. Updating the confidence score based selection using DNN and HMM-GMMs trained with ICSI seed transcription and all hypothesis transcription by the ICSI seed system (S2)

pared to the seed system. While in S1, as the frame accuracy of hypothesis labels at different levels is generally high, reducing the label resolution from triphone state to monophone state and monophone degraded the performance slightly.

3.5. Updating the confidence scores

As shown in Table 1 and Figure 2, adding hypothesis data to seed system improved the recognition performance due to the improvement in both DNN front-end and HMM-GMMs. In addition, using monophone targets gave more robust performance than using triphone state targets in DNN training. After the updates the models should be more accurate and hence may serve again as better confidence predictors. Figure 3 (S1) and Figure 4 (S2) compare the confidence scores for different label resolution after retraining on all candidate data in each scenario. The figures also include the confidence curves used at the start, obtained from seed models.

In S1, the seed system gives better selection than the updated DNN models. However, monophone state (MS) targets outperform triphone state (TS) targets after the model update. This is consistent with the fact that the higher label resolution may suffer more from the noisy data in training. In S2 with corpus mismatch, the updated DNNs signficantly outperform the seed models. In contrast to S1, monophone state (MS) based selection performs worse than triphone state (TS) based selection for small amounts of data, and equally well for larger amounts. This implies that domain mismatch is a greater problem for DNNs than label noise.

Correlation analysis between confidence scores revealed that scores based on phone classes (PC) were most correlated with scores estimated on monophone (M) level, while those on monophone (M) level were most correlated on monophone state (MS) level, and so forth. While phone class (PC) and monophone (M) based confidence scores show a correlation coefficient of 0.92, the ρ value for phone class (PC) and triphone state (TS) based scores is only 0.33. Overall the best correlation with segment based WER is achieved with MS level based confidence scores, however only with $\rho = 0.29$.

4. CONCLUSION

In this paper, semi-supervised training for two scenarios has been explored. In S1 a limited amount of transcribed data with matching test conditions is available. In S2, a limited amount of transcribed data of a different corpus is available.

We proposed new posterior-based confidence scores and showed that they outperform existing techniques, for the purpose of selecting hypothesis utterances of relatively high reliability. Gain from data selection for retraining of DNNs and HMM-GMMs is shown by experiments in both scenarios. We observed significant WER improvements of up to 5.9% absolute, compared to seed system.

It was further shown that lowering label resolution in

DNN training brings more robust performance across tasks. In scenario S2 the WER can yield an overall WER reduction of 16.6% relative. There is potential gain by further iterations of hypothesis selection, DNN label switching plus DNN-HMM-GMM retraining. However we found that this is scenario-dependent, or data dependent.

5. ACKNOWLEDGEMENT

The authors would like to acknowledge the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustic, Chinese Academy of Sciences and the Natural Speech Technology (NST) project for the support and sponsorship on the international research collaboration.

6. REFERENCES

[1] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115 – 129, 2002.

[2] Scott Novotney and Richard Schwartz, "Analysis of low-resource acoustic model self-training," 2009.

[3] Samuel Thomas, Michael L Seltzer, Kenneth Church, and Hynek Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6704–6708.

[4] H. Y. Chan and P.C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, May 2004, vol. 1, pp. I–737–40 vol.1.

[5] L. Chen, L. Lamel, and J. Gauvain, "Lightly supervised acoustic model training using consensus networks," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, May 2004, vol. 1, pp. I–189–92 vol.1.

[6] M. Gibson and T. Hain, "Correctness-adjusted unsupervised discriminative acoustic model adaptation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 10, pp. 2648–2656, Dec 2012.

[7] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV–757–IV–760.

[8] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *J. Mach. Learn. Res.*, vol. 10, pp. 1–40, June 2009.

[9] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *INTERSPEECH*, 2011, pp. 437–440.

[10] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 285–290.

[11] Yulan Liu, Pengyuan Zhang, and Thomas Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *ICASSP2014 - Speech and Language Processing (ICASSP2014 - SLTC)*, Florence, Italy, May 2014.

[12] Scott Novotney, Richard Schwartz, and Jeff Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4297–4300.

[13] Karel Veselý, Mirko Hannemann, and Lukas Burget, "Semi-supervised training of deep neural networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 267–272.

[14] Frantisek Grzl and Martin Karafit, "Semi-supervised bootstrapping approach for neural network feature extractor training.," in *ASRU*. 2013, pp. 470–475, IEEE.

[15] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 368–373.

[16] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," vol. 3869, pp. 28–39, 2006.

[17] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters, "The ICSI meeting corpus," 2003, pp. 364–367.