INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Generalized Distillation Framework For Speaker Normalization

Neethu Mariam Joy, Sandeep Reddy Kothinti, S. Umesh, Basil Abraham

Indian Institute of Technology-Madras, India
ee11d009@ee.iitm.ac.in, sandeep.kothinti@gmail.com, {umeshs,ee11d032}@ee.iitm.ac.in

Abstract

Generalized distillation framework has been shown to be effective in speech enhancement in the past. We extend this idea to speaker normalization without any explicit adaptation data in this paper. In the generalized distillation framework, we assume the presence of some "privileged" information to guide the training process in addition to the training data. In the proposed approach, the privileged information is obtained from a "teacher" model, trained on speaker-normalized FMLLR features. The "student" model is trained on un-normalized filterbank features and uses teacher's supervision for cross-entropy training. The proposed distillation method does not need first pass decode information during testing and imposes no constraints on the duration of the test data for computing speaker-specific transforms unlike in FMLLR or *i*-vector. Experiments done on Switchboard and AMI corpus show that the generalized distillation framework shows improvement over un-normalized features with or without *i*-vectors.

Index Terms: Distillation, teacher-student, FMLLR, i-vector, Switchboard, AMI

1. Introduction

In recent years, deep neural networks (DNN) have replaced the traditional Gaussian mixture model-hidden Markov model (GMM-HMM) acoustic models in majority of automatic speech recognition (ASR) applications. Several studies have shown that DNNs are inherently invariant to speaker and environmental variations [1][2][3][4]. This is due to the ability of DNNs to learn higher level representation of features which are inherently less prone to speaker and environment variations [5]. However, additional improvement in recognition performance can be obtained by applying speaker normalization or adaptation techniques to DNN models [6][7][8].

Vocal tract length normalization (VTLN) or feature space maximum likelihood linear regression (FMLLR) [9][10] estimated with a GMM-HMM can be used for speaker-normalizing the input features given to a DNN. Also, speaker-specific features can be given along with un-normalized features as input during DNN training. This allows the network to learn about the speaker normalization on its own. This requires two sets of features to be fed into the DNN: one for phonetic discrimination and another for speaker characterization. This framework is referred to as speaker-aware training (SaT) . Speaker codes [8][11] and *i*-vectors [12][13][14][15] can be used as additional speaker-related feature for neural networks.

In this paper, we propose using generalized distillation framework to transfer knowledge about speaker normalization to a DNN model trained on un-normalized features. This concept of *machine-teaching-machines* was first proposed in [16][17]. It incorporates an intelligent teacher into machine learning which provides additional information about each feature-label pair during training. This method, referred to as learning using privileged information, is capable of building a better classifier than those built on regular features alone. In [18], Hinton proposed a framework for distilling knowledge in neural networks, where a simple machine learns a complex task by learning from a complicated machine. *Generalized distillation* framework combines the strengths of both the aforementioned techniques [19]. The *teacher* which has access to privileged information behaves like the complicated machine in distillation process. The *student* is a simple network learned through distillation process from the teacher and is used for testing when privileged information is unavailable.

In [20], teacher-student framework was used to leverage untranscribed speech data to train the student network for obtaining a better approximation of the function learned in the teacher network. In a recent study, generalized distillation framework was used for noise robust speech recognition [21]. Here the clean speech was used to train the teacher and the student model was learned from noisy speech using the corresponding clean speech information from the teacher. Our paper extends this concept to speaker normalization, especially in real-world ASR applications, where there is very limited adaptation data and test utterance duration is small.

In the proposed generalized distillation framework for speaker normalization, the teacher network is learned from speaker-normalized speech, which acts as the privileged information. The student network is learned from un-normalized speech and is guided through the learning process by the teacher which has access to the speaker normalized version of the same speech. The proposed method finds applications in scenarios where only one (possibly short) test utterance is available for decoding and the test speaker's identity is unknown. Each test utterance is hence treated as though coming from different speakers and cannot be aggregated to do speaker-level normalization. Using the privileged information supplied by the teacher network, the student network learns to speaker-normalize the test speech utterance.

The paper is organized as follows. Section 2 describes in detail the proposed generalized distillation framework for speaker-normalization. The details of the experiments done on Switchboard and AMI corpus are given in section 3. A detailed analysis of the results of these experiments are discussed in section 4. Section 5 summarizes the findings and highlights the major contributions of the paper.

2. Proposed Method

Generalized distillation framework combines Hinton's distillation method [18] and Vapnik's privileged information learning [16][17] to provide a methodology for training compact models from complex models. In this framework, the basic idea is to incorporate an intelligent teacher to generate easier targets for training student models. Training a student model to imitate the teacher model improves generalizing ability of the student model.

http://dx.doi.org/10.21437/Interspeech.2017-874

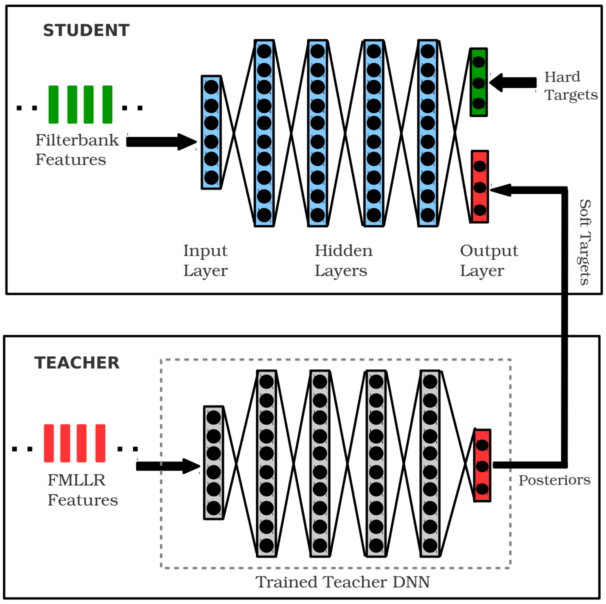


Figure 1: *Proposed Generalized Distillation Framework for speaker normalization*

In this paper, we propose to use generalized distillation method for building acoustic models which can provide speaker normalization without any explicit adaptation data or first pass decode information. The concept is illustrated in Figure 1. A DNN model trained on FMLLR features is the teacher. The student DNN model is trained on filterbank features. Posterior distribution for each data point in the train data provides information about competing classes. This helps the student model to learn the inherent structure of the train data like speaker variations, environment variations etc. Since the teacher model in the proposed framework is trained on speaker normalized FMLLR features, the student model will learn to normalize for speaker variations and use this knowledge during decoding phase. The algorithm for the proposed approach is as follows:

Algorithm 1 Training of student DNN in the proposed generalized distillation framework for speaker normalization

```
1. Train teacher DNN with FMLLR features.
2. Pass FMLLR features through teacher DNN to generate output posterior probabilities.
3. Prune these posterior probabilities and retain top 50 candidates. These are the soft targets ($s_i$) for student DNN.
4. Combine these soft targets with GMM-HMM alignments (hard targets $y_i$) for supervising the student DNN.
5. Objective function for training student DNN model on filterbank features ($\mathbf{x}_i$) is,
```

$$f_s = \arg\min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} [(1-w) * l\left(y_i, \sigma\left(f\left(\mathbf{x}_i\right)\right)\right) + w * l\left(s_i, \sigma\left(f(\mathbf{x}_i)/T\right)\right)] \quad (1)$$

where, T is temperature factor and $w \in [0,1]$ is the weight parameter.

Temperature T is a scaling parameter which is applied on the logit units of the output layer of both the teacher and student DNNs. Temperature applied on teacher model controls the peakiness of the softmax outputs and increasing T will make targets softer. Output layer contains two softmax activation functions, one applied directly on logit units and other one applied on scaled logit units. Since softmax outputs, soft and hard targets are probability distributions, cross-entropy criterion is used as objective function during parameter estimation. Imitation factor w controls the extent of imitation by the student. Since there is no direct measure of finding the generalizing ability of the teacher, optimal values of T and w should be found empirically. In this paper, we used $w = 0.5$ and $T = 1$.

2.1. FMLLR vs Proposed Approach

The proposed method can overcome the following requirements of FMLLR transform estimation during decoding.

- First pass transcription during decode
- GMM-HMM for FMLLR test feature extraction
- Atleast 30sec of data for robust FMLLR estimation

However, the student DNN, once trained, requires only plain filterbank features while decoding. Hence, it is ideal for online real-world ASR applications where test utterances are short and FMLLR estimation is not feasible.

2.2. *i*-vector vs Proposed Approach

i-vectors are another alternative to generate speaker adapted acoustic models and they can be generated even for short utterances. But, these speaker adapted models still require *i*-vector computation during decoding. When *i*-vectors are computed on a per-utterance level during decoding, like in a practical ASR system, degradation in performance is observed. The proposed method does not impose a constraint like this. Our experiments show that the proposed approach performs significantly above DNNs trained on *i*-vector appended filterbank features.

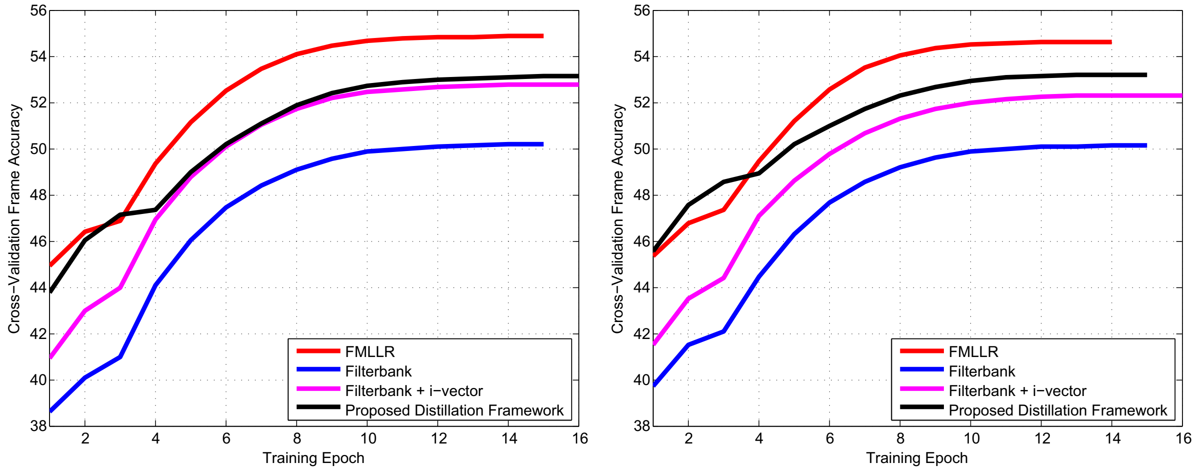
3. Experimental Details

Experiments were conducted on Switchboard–1 Release 2 telephone speech corpus [22] and AMI corpus [23][24], the details of which are given below.

Switchboard corpus comprises of 2400 two–sided telephone conversations collected from 543 speakers from all over the United States. The experiments were conducted on 33–hour (30K utterances) and 110–hour (100K utterances) subset of the train data as mentioned in the Kaldi recipe. HUB5 English evaluation dataset [25] of conversational telephonic speech with 2.1 hours of audio was used for decoding. A 4-gram language model built by adding Fisher English corpus [26] to the entire Switchboard train data was used as the decode language model.

AMI meeting corpus[1] consists of 100 hours of meeting recordings. The recordings include close-talking and far-field microphones, individual and room-view video cameras, and output from a slide projector and an electronic white board. The meetings were recorded in English using three different rooms with different acoustic properties, and include mostly non-native speakers. Acoustic models were trained using Kaldi toolkit on close-talking independent headset microphone (ihm) data alone as per the full ASR split into training, development and evaluation sets mentioned in

[1] http://groups.inf.ed.ac.uk/ami/corpus/overview.shtml



(a) *Switchboard-110hour*

(b) *AMI corpus*

Figure 2: *Cross-validation data frame accuracy for various DNNs trained on Switchboard-110hour and AMI corpus*

http://groups.inf.ed.ac.uk/ami/corpus/datasets.shtml. A 3-gram language model built from the entire ihm AMI train data was used for decoding purposes.

Using a frame length of 25ms and frame shift of 10ms, we extracted 13-dimensional Mel frequency cepstral coefficients (MFCC) for each frame. First and second order derivatives were augmented to this feature to get a 39-dimensional vector, which is then mean normalized at speaker level. Nine consecutive frames of MFCC are then spliced together and projected to a 40-dimensional vector using linear discriminant analysis (LDA) and further diagonalized by maximum likelihood linear transformation (MLLT). These are then speaker-normalized via FMLLR transform computed on a per speaker basis. Additionally, 40-dimensional log Mel filterbank features were also extracted for every frame. Delta and acceleration components were augmented to it and was then cepstral mean normalized (CMN) at speaker level to form 120-dimensional feature.

3.1. Details of Baseline DNNs

We extracted 120-dimensional filterbank (40 +Δ + $\Delta\Delta$) and 40-dimensional FMLLR features to train DNN models using frame length of 25ms and frame shift of 10ms. Kaldi toolkit [27] was used for building GMM-HMM and DNN-HMM models. All the DNNs had 6 hidden layers with sigmoid activation functions and 2048 neurons in each layer. The number of units in the output softmax layer equals the number of context-dependent states in phonetic decision tree. The frame-level alignment information was taken from the DNN trained on FMLLR features.

Additionally, *i*-vectors were augmented to the input filterbank features to supply speaker information. An M-dimensional *i*-vector was extracted for all train and test utterances from a full-covariance GMM with 512-mixture components trained on LDA features, where $M = \{40, 100\}$ for Switchboard-33hour and 110hour respectively. For AMI, 100-dimensional *i*-vector was extracted for all train and test utterances from a diagonal-covariance GMM with 1024-mixture components trained on FMLLR features as mentioned in Kaldi recipe.

Prior to training, the entire training data was randomized at frame level. Layer-wise RBM pretraining was used for initializing DNN parameters. Stochastic gradient descent with minibatch size of 256 frames and learning rate of 0.008 is used for supervised training. Over a context window of 11 frames (± 5), the input features were stacked and fed to the DNN. Early stopping was used to avoid over fitting. Weighted finite state transducer (WFST) based graph generated for GMM-HMM model was used for decoding the DNN models by scaling DNN posteriors with class priors computed from alignments. Both speaker-wise and utterance-wise CMN versions of the test data was used for decoding to compare the effectiveness of the model for real time applications.

3.2. Details of Distillation DNN

In our experiments, we considered DNN model trained on FMLLR features as teacher DNN model. The student DNN model is trained on filterbank features. The architecture of these models is as explained in section 3.1. The soft targets for the student model is generated by passing FMLLR features through the teacher model with T values applied on the softmax function. Once posteriors were extracted from the teacher model, a threshold posterior probability was used to prune the soft targets. The values below the threshold were made zero. The concept of threshold was used to reduce the number of target posteriors, which in turn reduces the number of memory reads and overall training time. The threshold was chosen in such a manner that at most 50 target posteriors remained. For training student DNN, tied-state targets obtained from frame level DNN alignments and pruned soft targets were appended and multitask optimization was performed. Multiple imitation factors and temperatures have to be tried empirically to find the best possible combination of $w = 0.5$ and $T = 1$.

4. Results & Discussion

Figure 2 shows the frame accuracy on cross-validation data for baseline and distillation DNNs in Switchboard-110hour and AMI corpus. It shows that the proposed distillation DNN, although trained using just the filterbank features, is closer to that of DNN trained on *i*-vector appended filterbank features. This

Table 1: *Word Error Rate (%) of various DNN acoustic models for Switchboard 33-hour and 110-hour with 4-gram language model tested against HUB5 English evaluation dataset*

Features	**Switchboard-33hour**				**Switchboard-110hour**			
	Speaker-wise Normalization		**Utterance-wise Normalization**		**Speaker-wise Normalization**		**Utterance-wise Normalization**	
	swbd	**swbd+ callhm**	**swbd**	**swbd+ callhm**	**swbd**	**swbd+ callhm**	**swbd**	**swbd+ callhm**
Filterbank	21.0	27.6	22.7	29.7	17.4	23.6	18.9	25.4
Filterbank + i-vector	20.3	27.0	22.6	29.7	17.3	23.9	19.2	26.2
FMLLR	**18.4**	**25.0**	28.3	38.0	**15.8**	**21.8**	24.7	33.2
Generalized Distillation Framework Teacher = FMLLR; Student = Filterbank	19.8	26.4	**21.5**	**28.4**	16.8	22.8	**18.3**	**24.6**

**swbd*: Switchboard component of HUB5, *swbd+callhm*: Entire HUB5 comprising of both Switchboard and CallHome components

Table 2: *Word Error Rate (%) of various DNN acoustic models for AMI corpus with 3-gram language model*

Features	**Speaker-wise Normalization**		**Utterance-wise Normalization**	
	dev	**eval**	**dev**	**eval**
Filterbank	27.5	29.7	31.0	33.3
Filterbank + i-vector	**26.6**	28.0	31.5	34.1
FMLLR	26.8	**27.5**	33.7	36.6

Generalized Distillation
Teacher = FMLLR;
Student = Filterbank
	26.9	28.7	**30.1**	**32.2**

*Only independent headset microphone data is used for training

is due to the supervision from the soft targets generated from DNN trained on FMLLR features which transfers knowledge about speaker normalization to the distillation DNN.

Tables 1 and 2 provides the word error rate (WER) of the various DNN acoustic models mentioned in sections 3.1 and 3.2 for Switchboard-33hour and 110hour and AMI corpus respectively. The results are compartmentalized into two: speaker-wise normalization and utterance-wise normalization. Speaker-wise normalization refers to the scenario where the normalization operations like CMN, FMLLR and *i*-vector are computed on a per-speaker basis for the test utterances and vice-versa for utterance-wise normalization. The following observations can be made:

- The proposed generalized distillation framework trained using filterbank features, learns about speaker normalization information from DNN trained on FMLLR features and mimics this knowledge during test without the requirement of explicit speaker information via *i*-vectors or FMLLR transformed test features.
- The proposed method performs significantly above DNN trained on filterbank features and *i*-vector augmented filterbank features in the case of Switchboard. This can be verified from Figure 2 where the cross validation data frame accuracy of the proposed method is above the DNN trained on *i*-vector appended filterbank features. However, for AMI corpus, the *i*-vectors are extracted from GMM trained from FMLLR features and get this added benefit which aids them in performing at par with FMLLR features as shown in Table 2. The proposed method in this case performs better than DNN trained on filterbank features and not on their *i*-vector appended counterparts. The proposed method, however, has the advantage of not computing speaker-related features during decoding phase.
- In speaker-wise normalization scenario, the proposed method degrades on an average of 1.5% absolute in WER compared to DNN trained on FMLLR features. This is to be expected as in generalized distillation framework, the student model can perform only at par or below that of the teacher model.
- In utterance-wise normalization case, the proposed method performs well above all other DNNs. As only a single, possibly short utterance is available to compute the FMLLR transform or *i*-vector, the estimation of these features degrades. But the proposed method, having imposed no such constraints on the test utterance, requires just the filterbank features for decoding. Thus makes the distillation framework of speaker normalization ideal for real-world applications.
- The afore mentioned observations are consistent with increase in the amount of train data (33hour to 110hour) and across different speech corpora (Switchboard, AMI).

5. Conclusion

In this paper, we explored the use of generalized distillation framework for speaker normalization. The student DNN trained on filterbank features learns from a teacher DNN trained on FMLLR features. In decoding phase, the student model requires only filterbank features as input and will be able to approximate the speaker-normalizing function learned from the teacher DNN. The proposed method thus overcomes the limitations of FMLLR transforms like dependency on GMM-HMM model, first pass transcription and constraint on the length of utterance for transform estimation. Unlike in FMLLR and *i*-vector, the proposed approach does not require estimation of speaker-related features during decoding. Experiments done on Switchboard-33hour, 110hour and AMI corpus shows that the proposed approach performs significantly above DNNs trained on filterbank features and their *i*-vector appended counterparts, when speaker-wise normalization is done on test utterances. The proposed approach performs significantly above DNNs trained on FMLLR, filterbank features with or without *i*-vectors when utterance-wise normalization is done.

6. References

[1] M. L. Seltzer, D. Yu, and Y. Wang, "An Investigation of Deep Neural Networks for Noise Robust Speech Recognition," in *Proc. ICASSP*, 2013, pp. 7398–7402.

[2] F. Seide, G. Li, X. Chen, and D. Yu, "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription," in *Proc. ASRU*, 2011, pp. 24–29.

[3] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation," in *Proc. EUROSPEECH*, 1995.

[4] H. Liao, "Speaker Adaptation of Context Dependent Deep Neural Networks," in *Proc. ICASSP*, 2013, pp. 7947–7951.

[5] D. Yu, M. L. Seltzer, J. Li, J. Huang, and F. Seide, "Feature Learning in Deep Neural Networks - A Study on Speech Recognition Tasks," in *Proc. ICLR*, 2013.

[6] 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. SLT*, 2012, pp. 366–369.

[7] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-Divergence Regularized Deep Neural Network Adaptation for Improved Llarge Vocabulary Speech Recognition," in *Proc. ICASSP*, 2013, pp. 7893–7897.

[8] O. Abdel-Hamid and H. Jiang, "Fast Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition Based on Discriminative Learning of Speaker Code," in *Proc. ICASSP*, 2013, pp. 7942–7946.

[9] S. P. Rath, D. Povey, K. Vesley, and J. H. Cernocky, "Improved Feature Processing for Deep Neural Networks," in *Proc. INTERSPEECH*, 2013, pp. 109–113.

[10] S. H. K. Parthasarathi, B. Hoffmeister, S. Matsoukas, A. Mandal, N. Strom, and S. Garimella, "fMLLR Based Feature-Space Speaker Adaptation of DNN Acoustic Models," in *Proc. INTERSPEECH*, 2015.

[11] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct Adaptation of Hybrid DNN/HMM Model for Fast Speaker Adaptation in LVCSR based on Speaker Code," in *Proc. ICASSP*, 2014, pp. 6339–6343.

[12] G. Saon, H. Soltau, D. Nahamoo, and M. A. Picheny, "Speaker Adaptation of Neural Network Acoustic Models using I-Vectors," in *Proc. ASRU*, 2013, pp. 55–59.

[13] M. Rouvier and B. Favre, "Speaker Adaptation of DNN-based ASR with I-Vectors: Does It Actually Adapt Models to Speakers?" in *Proc. INTERSPEECH*, 2014, pp. 3007–3011.

[14] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-Vector Based Speaker Adaptation of Deep Neural Networks for French Broadcast Audio Transcription," in *Proc. ICASSP*, 2014, pp. 6334–6338.

[15] S. Garimella, A. Mandal, N. Strom, B. Hoffmeister, S. Matsoukas, and S. H. K. Parthasarathi, "Robust I-Vector based Adaptation of DNN Acoustic Model for Speech Recognition," in *Proc. INTERSPEECH*, 2015.

[16] V. Vapnik and A. Vashist, "A New Learning Paradigm: Learning using Privileged Information," *Neural Networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[17] V. Vapnik and R. Izmailov, "Learning using Privileged Information: Similarity Control and Knowledge Transfer," *Journal of Machine Learning Research*, vol. 16, pp. 2023–2049, 2015.

[18] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *CoRR*, 2015.

[19] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying Distillation and Privileged Information," in *Proc. ICLR*, 2016.

[20] R. Price, K. Iso, and K. Shinoda, "Wise Teachers Train Better DNN Acoustic Models," *EURASIP Journal on Audio, Speech and Music Processing*, vol. 2016, p. 10, 2016.

[21] K. Markov and T. Matsui, "Robust Speech Recognition using Generalized Distillation Framework," in *Proc. INTERSPEECH*, 2016.

[22] J. Godfrey and E. Holliman, "Switchboard-1 Release 2 LDC97S62," *Linguistic Data Consortium*, 1993.

[23] 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," in *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, 2005, pp. 28–39.

[24] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, D. van Leeuwen, M. Lincol, and V. Wan, *The 2007 AMI(DA) System for Meeting Transcription*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 414–428.

[25] L. D. Consortium, "2000 HUB5 English Evaluation Speech LDC2002S09," *Web Download. Philadelphia: Linguistic Data Consortium*, 2002.

[26] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher English Training Speech Part 1 Speech LDC2004S13," *Linguistic Data Consortium*, 2004.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, 2011.