Speech Communication 132 (2021) 123–131

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Curriculum Learning based approaches for robust end-to-end far-field speech recognition☆

Shivesh Ranjan, John H.L. Hansen *

Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75080, USA

ARTICLE INFO

Keywords:
Curriculum learning
Far field
Speech recognition
End to end
Robust

ABSTRACT

Performance of Automatic Speech Recognition (ASR) systems is known to suffer considerable degradation when exposed to Far-Field speech data capture. Consequently, far-field ASR has received considerable attention in recent years. Motivated by our recent work using Curriculum Learning (CL) based strategies to improve Speaker Identification (SID) under noisy and degraded conditions, this study proposes a novel approach to improve far-field ASR using CL based approaches. Specifically, we propose using a CL based approach for training a Bidirectional Long Short Term Memory (BLSTM) based ASR network trained using the Connectionist Temporal Classification (CTC) objective function. We initiate the training with comparatively easier near-field data, and include more diverse (difficult) far-field data progressively in the later stages of training. These proposed approaches are shown to significantly outperform the baseline BLSTM ASR system, and offer relative reductions in WERs of up to +7.3% and +10.1% for the dev and eval sets of the AMI far-field voice capture corpus.

1. Introduction

A unique attribute of the human learning process is our ability to grasp a difficult concept by relating it to an already assimilated and comparatively easier concept. This human learning mechanism has served as a strong motivation behind a distinct category of *Curriculum Learning* (CL) based algorithms in Machine Learning (Bengio et al., 2009). The use of CL based training strategies have been explored for a variety of applications such as: to train richer architectures for Deep Neural Networks (DNN) based Automatic Speech Recognition (ASR) systems (Amodei et al., 2016), to improve noise robustness of automatic speech recognition (ASR) (Braun et al., 2017), to improve the performance of Speaker Identification (SID) systems in the presence of severe noise and channel degradations among others (Ranjan et al., 2017; Ranjan and Hansen, 2018). In this current study, we propose a novel framework to improve the performance of end-to-end deep Bidirectional Long Short Term Memory Network (BLSTM) based ASR systems trained using the Connectionist Temporal Classification (CTC) objective function on far-field voice captured speech.

ASR systems experience significant degradation in accuracy when exposed to far-field speech utterances due to a variety of causes such as reverberation, background noise, and multiple concurrently active acoustic sources (Swietojanski et al., 2014). Connectionist Temporal Classification (CTC) based ASR systems have previously been explored in Graves et al. (2013). Unlike conventional Deep Neural Networks (DNN) based ASR systems, the CTC based ASR systems can be trained in an end-to-end manner without requiring the alignments from an HMM–GMM system for training. This simplifies the ASR pipeline, albeit with some degradation in accuracy compared to conventional DNN based ASR systems (Miao et al., 2015).

Several DNN based approaches have been proposed to improve ASR accuracy for far-field speech. In Miao and Metze (2015), bottleneck (BNF) features from a near/far field classifier were used together with regular acoustic features to improve far-field speech recognition. In the same work, the authors also proposed a novel Distance Adaptive training strategy, where BNF features from the near/far field classifier were used to learn a distance-normalized feature space.

The use of DNNs for far-field ASR was also investigated in Swietojanski et al. (2013), where a significant reduction in WER was observed by employing feature-level concatenation for the multiple channels of the far-field data. Furthermore, the use of multi-style training using each of the multiple array channels was also investigated. In Swietojanski et al. (2014), the use of Convolutional Neural Networks (CNNs) for improved far-field speech recognition was investigated. Specifically, the authors explored using weight sharing, and cross channel pooling to improve far-field ASR.

☆ This project was supported by the National Science Foundation (NSF) under Grant Award #1918032 (PI: Hansen), and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

* Corresponding author.
E-mail addresses: shivesh.ranjan@utdallas.edu (S. Ranjan), john.hansen@utdallas.edu (J.H.L. Hansen).

https://doi.org/10.1016/j.specom.2021.06.003
Received 4 July 2020; Received in revised form 13 May 2021; Accepted 2 June 2021
Available online 18 June 2021

In Qian et al. (2016), several approaches to improve far-field speech recognition were examined. The authors investigated jointly training a network to perform both recognition and dereverberation. Furthermore, the use of model sharing, where some parameters of the far-field model were made close to the near-field model by minimizing the mean squared error (MSE) between the outputs of some hidden layers was also considered. Additionally, an environment code aware training was also proposed in the same work, wherein BNF features extracted from a network trained to map far-field features to near-field were used as auxiliary inputs to the ASR system.

In Peddinti et al. (2015a), the use of Time Delay Neural Networks (TDNN) for far-field speech recognition was explored. Significant reduction in WER was observed by using acoustic features and i-Vectors with a TDNN architecture. Using Room Impulse Responses (RIRs) to create additional pre-training simulated far-field data for training the networks was investigated in Peddinti et al. (2015b), Ko et al. (2017). In Ko et al. (2017), the performance using augmented simulated data improved by adding point noise sources.

An approach combining the networks for speech enhancement and speech recognition was proposed in Ravanelli et al. (2017). The authors explored using a network of DNNs, wherein modules for speech enhancement and recognition were jointly trained. In this approach, the speech enhancement loss function also included contributions from the DNN for ASR. This resulted in overall improvements to far-field speech recognition.

A student–teacher transfer learning based approach to far-field speech recognition was proposed in Kim et al. (2017). In that proposed *Bridgenet* architecture, the authors used a knowledge bridge between a teacher network trained on near-field speech, to facilitate dereverberation by the student network trained on far-field speech data. This approach was motivated by knowledge distillation based techniques (Hinton et al., 2015) where a larger teacher network is used to train a student network with improved generalization ability.

To improve recognition performance on reverberant speech, in Giri et al. (2015), a multi-task learning based approach was adopted that both classified speech senones and performed feature enhancement. In addition, in the same work, the use of *room aware* features that characterized reverberation was also explored. The room aware features were input as auxiliary information to the network. In Himawan et al. (2015), a DNN was used to transform the far-field speech features to those corresponding to near-field utterances.

Several beamforming based approaches that utilize audio captured using a microphone array have also been proposed to improve the recognition accuracy of far-field speech (Khoubrouy and Hansen, 2016; Kumatani et al., 2012; Marino and Hain, 2011). In Seltzer et al. (2004), array processing was carried out in a way that maximized the likelihood of generating correct hypotheses. A tutorial on various array processing based techniques for far-field speech recognition is presented in Kumatani et al. (2012). Also, an investigative study on the effectiveness of beamforming techniques for ASR in meeting scenarios was also presented in Marino and Hain (2011).

Curriculum Learning (CL) based approaches have also been explored to improve far-field speech recognition: in Chang et al. (2019), Signal-to-Noise Ratio (SNR) was used as a curriculum criterion to train a sequence-to-sequence (seq2seq) architecture. The work in Braun et al. (2017) had originally explored using SNR to improve ASR accuracy in noisy conditions. In Zhang et al. (2020), several CL criterion such as SNR, gender, and duration of the utterance were explored to improve end-to-end multi-talker ASR.

This current study proposes a novel approach to far-field speech recognition using Curriculum Learning (CL). Specifically, we focus on improving ASR accuracy using train and test data captured using a single distant microphone (SDM). To this end, we employ the SDM dataset of the AMI corpus (Carletta, 2006; Hain et al., 2008; Renals and Swietojanski, 2017). All CL based approaches require a suitable difficulty criterion. In Braun et al. (2017), SNR was used for assigning difficulty while training the system for ASR. In Ranjan et al. (2017), Ranjan and Hansen (2018), SNR and the difference of Signal-to-Noise-Distortion and Noise-Distortion (SND-ND) values were investigated for devising CL based strategies for improving Speaker Identification (SID) in the presence of severe noise and channel distortions. In Marchi et al. (2018), an LSTM network was used to obtain speaker embeddings using a CL based approach.

In this current study, we explore the use of distance information (near vs. far) as a choice for a difficulty criterion for training ASR systems. We note that this is different from other CL based approaches to improve ASR that have used SNR, gender, and utterance-duration as difficulty criteria (Braun et al., 2017; Chang et al., 2019; Zhang et al., 2020). We note that the motivation for using distance guided CL is somewhat similar to using SNR as the difficulty criterion. However, since far-field speech recognition is challenging due to several factors such as reverberation, background noise, and multiple concurrently active acoustic sources (Swietojanski et al., 2014), the use of distance information for guiding CL can be viewed as encapsulating all these factors jointly (including SNR). The key contributions of this study are:

1. We propose novel CL based approaches to improve the performance of far-field speech recognition where the training and test data have been captured using a Single Distant Microphone (SDM). Specifically, we propose 3 related CL based approaches: CL Data Hop (CL-DH), CL Data Merge (CL-DM), and CL Data Hop and Merge (CL-DHM) to improve recognition accuracy of far-field captured speech.
2. We investigate the use of end-to-end CTC based BLSTM acoustic models for far-field speech recognition.
3. We report results of our proposed CL based approaches on both *dev* and *eval* sets of the AMI distant speech corpus.

In all three CL based approaches: CL-DH, CL-DM and CL-DHM, training is initiated with data that is deemed to be comparatively easier. However, in CL-DH, only the difficult data is used toward the end of training. In our proposed CL-DM based approach for far-field ASR, the training set is augmented with more difficult data in the final stages of training while also utilizing the easier data. Finally, the CL-DHM based approach combines both data hoping and merging, by moving to only difficult data from easy data for training, and then including both easy and difficult data to obtain the final models.

2. Bi-directional LSTM architecture for ASR

Multiple bi-directional LSTM layers can be stacked to obtain a deep architecture. Fig. 1 shows the architecture of the LSTM unit used in this work (Miao et al., 2015). In Fig. 1, $\odot$ are the multiplicative gates, and the non-linearities indicated in the figure are hyperbolic tangents, while x_t refers to the input speech feature frame. Also appearing in Fig. 1, i_t, o_t, f_t, c_t are respectively the outputs of the input gate, output gate, forget gate, and the LSTM memory cell. The figure also shows peephole connections from the memory cell c_t to the various gates which are used to learn output timings (Miao et al., 2015). The various parameters of the LSTM unit shown in Fig. 1 can be evaluated as provided in Miao et al. (2015),

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i), \tag{1}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f), \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{cx}x_t + W_{ch}h_{t-1} + b_c), \tag{3}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_t + b_o), \tag{4}$$

$$h_t = o_t \odot c_t. \tag{5}$$

For Eqs. (1) through (5), the matrices $W_{kx}, k \in \{i, f, c, o\}$ represent the input connections with the corresponding units. Similarly, $W_{lh}, l \in$

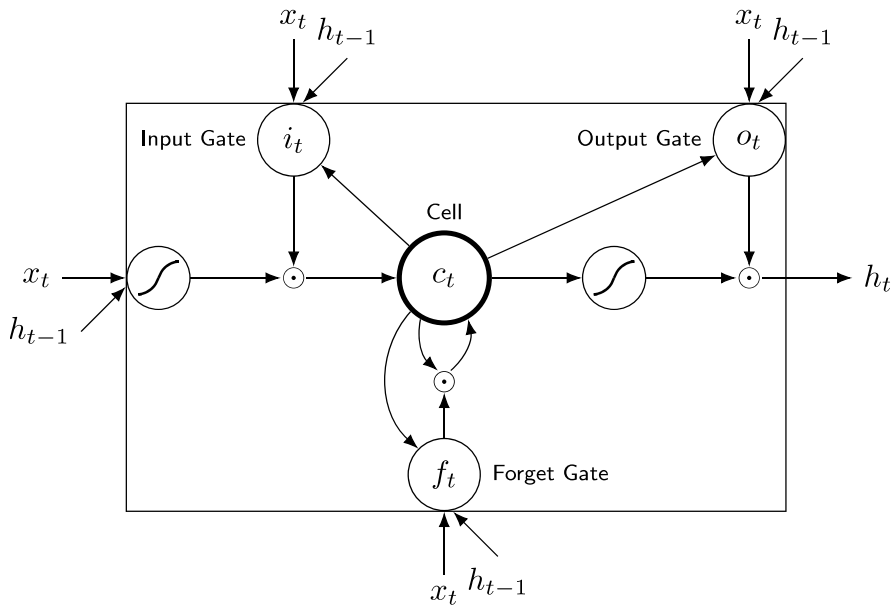

Fig. 1. Structure of an LSTM unit.

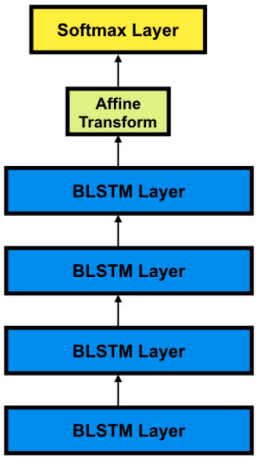

Fig. 2. Architecture of the deep BLSTM network used for ASR in this study. Four bi-directional LSTM layers are stacked together, followed by an affine transform, and a softmax layer respectively.

$\{i, f, c, o\}$ denote the connections between the previous cell states and the units, and $W_{mc}, m \in \{i, f, o\}$ represent the diagonal weight matrices used in the peephole connections. Additionally, σ and ϕ denote the sigmoid and hyperbolic tangent nonlinearities respectively. Also, $b_n, n \in \{i, f, c, o\}$ are biases of the respective units. For ASR applications, several LSTM layers can be stacked together to form a deep architecture, followed by an affine transform layer, and a softmax layer corresponding to the phonemes at the output layer. Finally, the network can be trained with back-propagation though time (BPTT) (Miao et al., 2015). Fig. 2 shows the architecture of the deep BLSTM network used in this study.

As can be seen in Fig. 2, four BLSTM layers are stacked together to form a deep architecture. The network is trained using the CTC objective function (Graves et al., 2006). This allows us to train the network in an end-to-end fashion without requiring alignments from an already trained HMM–GMM or DNN based ASR system.

3. Curriculum learning based approaches for far-field speech recognition

Here, we introduce three CL based approaches for far-field speech recognition: CL-DH, CL-DM and CL-DHM. The key step in all three approaches is to divide the training data into subsets based on a suitable difficulty criterion, and introduce more difficult data sequentially to the training algorithm in the later stages. A critical requirement of any CL based learning algorithm is measure for assigning the difficulty of the training examples, which is introduced next.

3.1. Assigning difficulty to the training data

In Braun et al. (2017), SNR values of the utterances were used to assign difficulty metric for training DNNs for ASR. For SID applications, the use of SNR information and SND - ND values were investigated in Ranjan and Hansen (2018), Ranjan et al. (2017). In Marchi et al. (2018), domain information of the query was used as the curriculum information for training a SID system to extract speaker embeddings for speaker verification.

In this study, we propose to use the distance information of the microphone to subject/speaker employed to capture the audio, that is, far vs. near for assigning difficulty of the training data. Since near-field speech is comparatively easier to work with than far-field for ASR systems, as evident by consistent higher accuracy, we argue that distance information (i.e., near vs. far field) can be used as a suitable difficulty metric in formulating a CL based training approach for far-field ASR. It is well known that ASR systems experience significant degradation in accuracy when exposed to far-field speech utterances due to a variety of causes such as reverberation, background noise, and multiple concurrently active acoustic sources (Swietojanski et al., 2014). We argue that far-field data is comparatively *harder* to train on due to the underlying causes behind degradation. Thus, microphone distance (i.e., near-field or far-field) can be a suitable hardness measure for training ASR systems.

3.2. CL-DH "Data-Hop" training

In contrast to traditional training of a BLSTM network trained with the CTC objective function, CL-DH operates differently in the manner in how it uses training data to estimate the model parameters. More

specifically, the training data is first arranged into two subsets based on the difficulty criterion (i.e., near vs. far-field sets). Next, the deep BLSTM network is trained for ASR using a CTC loss. However, unlike traditional training where all available data is included, CL-DH uses only the easy (i.e., near-field) data in the first training phase. After the network has been trained on easy data for a predetermined number of iterations, more difficult (i.e., far-field) data is progressively included in training. At this point, the network trained on the easy data is used as the initial network when training on far-field data. This gradual shift to more difficult data and diversity leads to better estimation of the deep BLSTM ASR network parameters compared to the traditional training approach.

It has been suggested that CL based approaches can operate in a manner similar to continuation methods (Bengio et al., 2009; Allgower and Georg, 2012), where initialization of the optimization set-up with a simpler, more smoothed version of a non-convex function may lead the parameter estimation algorithm to a dominant (and better) overall minimum of the function. In the CL-DH paradigm, this translates to improved estimation of the deep BLSTM network parameters by letting the CTC loss to obtain more robust network parameters compared to a traditional CTC loss based BLSTM training approach. Moreover, it has also been argued that CL based techniques can also facilitate improved regularization during network training, thus leading to better model parameters that are more robust to unseen data (Bengio et al., 2009). It is hypothesized that our proposed CL-DH approach may also benefit from increased generalization by employing CL driven network training. The various steps of CL-DH algorithm are presented in Algorithm 1.

Algorithm 1: CL-DH "Data-Hop" for robust far-field speech recognition.

Partition the training data d_{all} into 2 distinct subsets d_{NF}, and d_{FF} corresponding to near and far-field data respectively.
Initialize the deep BLSTM network parameters randomly.
Initialize an empty current dataset $d_{cu} \leftarrow \emptyset$
Choose a suitable value for n_k, the number of iterations per new data subset.
for *k=NF:FF* **do**
  $d_{cu} \leftarrow d_k$
  for *l=1:*n_k **do**
    Use the current deep BLSTM network and d_{cu} to obtain updated network parameters using the CTC loss and BPTT.
  end
end
Use the deep BLSTM network trained after the last iteration as the final network for far-field ASR.

3.3. CL-DM "Data-Merge" training

Our proposed CL-DM algorithm differs from the previously presented CL-DH algorithm in the manner in which the second, and more difficult metric based set containing far-field data gets included in training. Specifically, after updating the deep BLSTM network with a pre-determined number of iterations using the near-field data, both near and far-field data are included in second stage training. Thus, the second stage of CL-DM is similar to a traditional multi-condition training approach using both near and far-field data.

We hypothesize that initiating the training with comparatively easier near-field data, and subsequently adding far-field data would lead to a more robust deep BLSTM network parameter set compared to traditional multi-condition training using both near and far-field data. Algorithm 2 presents the details of our proposed CL-DM approach.

Algorithm 2: CL-DM "Data-Merge" for robust far-field speech recognition.

Partition the training data d_{all} into 2 distinct subsets d_{NF}, and d_{FF} corresponding to near and far-field data respectively.
Initialize the deep BLSTM network parameters randomly.
Initialize an empty current dataset $d_{cu} \leftarrow \emptyset$
Choose a suitable value for n_k, the number of iterations per new data subset.
for *k=NF:FF* **do**
  $d_{cu} \leftarrow d_{cu} \cup d_k$
  for *l=1:*n_k **do**
    Use the current deep BLSTM network and d_{cu} to obtain updated network parameters using the CTC loss and BPTT.
  end
end
Use the deep BLSTM network trained after the last iteration as the final network for far-field ASR.

3.4. CL-DHM "Data Hop & Merge" training

Next, we combine the essential steps of CL-DH and CL-DM approaches in what we term the "CL Data Hop and Merge" (CL-DHM) strategy. Here, training is first initiated with the easiest subset, next, as in CL-DH, only the more difficult data is used for several iterations to train the ASR model. Lastly, the training uses both easy and difficult data to obtain the final model, similar to the method that was carried out for the CL-DM strategy. This approach was also inspired by a recent CL based approach used for SID, where training was carried out respectively on easy, difficult, and then both easy and difficult data combined (Marchi et al., 2018). Algorithm 3 presents our proposed CL-DHM data hop and merge based approach for robust far-field ASR.

Algorithm 3: CL-DHM "Data Hop & Merge" for robust far-field speech recognition.

Partition the training data d_{all} into 2 distinct subsets d_{NF}, and d_{FF} corresponding to near and far-field data respectively.
Initialize the deep BLSTM network parameters randomly.
Initialize an empty current dataset $d_{cu} \leftarrow \emptyset$
Choose suitable values for n_k, the number of iterations per new data subset d_k.
for *k=NF:FF:NF* $\cup$ *FF* **do**
  $d_{cu} \leftarrow d_k$
  for *l=1:*n_k **do**
    Use the current deep BLSTM network and d_{cu} to obtain updated network parameters using the CTC loss and BPTT.
  end
end
Use the deep BLSTM network trained after the last iteration as the final network for far-field ASR.

3.5. Baseline BLSTM CTC loss based ASR systems

The baseline systems corresponding to the CL-DH, CL-DM and CL-DHM based approaches are formulated using deep BLSTM networks trained with the CTC loss. For a one-to-one comparison, for CL-DH, the corresponding baseline is a deep BLSTM network trained with far-field data. Similarly, for the CL-DM approach, the corresponding baseline system is trained on both near and far-field data together in a standard multi-condition set up. The same baseline as that used for comparing the CL-DM approach is used for benchmarking the CL-DHM systems. More details about the various set-ups are presented in Section 5. Next, we consider the far-field speech corpus.

Table 1
Composition of the train, dev and eval subsets of the AMI Corpus. Only far-field (SDM) versions of dev and eval data have been used for evaluations reported in this study.

Data set	Number of utterances
train	108221
dev	13059
eval	12612

4. AMI corpus

For experiments reported in this study, we have used the AMI corpus (Carletta, 2006) which has approximately 100 h of speech recorded in meeting scenarios. The audio was captured using both near-field and far-field microphones (using 8 microphone circular array). For the far-field scenario, we only use the data from a Single Distant Microphone (SDM) for both training and testing. The near-field data, referred to as the Individual Headset Microphone (IHM) in the corpus, consists of the same utterances as those found in the SDM set. The AMI corpus has been widely used for far-field ASR, and continues to be used in several other speech applications such as speech enhancement (Mirsamadi and Hansen, 2019; Trinh and Mandel, 2020; Grezes et al., 2020).

To assess performance, the entire corpus is divided into *train*, *dev*, and *eval* subsets. Separate versions of these subsets exist for the near-field (IHM) and far-field (SDM) cases to train and test the corresponding near/far field utterances. Table 1 presents the data partition of the train, dev and eval subsets of the AMI corpus used in this study. For the experiments in this study, performance is only reported for far-field (SDM) versions of the dev and eval data. Only for training, near-field (IHM) and far-field (SDM) data have been used as indicated in the corresponding training descriptions.

The entire train set corresponds to approximately 75 h of speech data. Furthermore, to carry out BLSTM training, approximately 10% of the train set of the AMI corpus was randomly set aside as a cross-validation set to monitor the training.

5. Experiments, results & discussion

5.1. Baseline deep BLSTM CTC loss based ASR systems

For the baseline systems corresponding to the CL-DH and CL-DM based approaches, we have used the Eesen toolkit (Miao et al., 2015). Specifically, for the baseline corresponding to the CL-DH based approach, referred to as system B-1 in this study, 40-dim log-Mel filterbank features together with delta and delta–delta are used as input to the deep BLSTM network. Thus, the final input to the network is 120-dim. We use the Switchboard recipe available with Eesen, albeit with some modifications, to train the baseline systems. The language model was trained using both Switchboard and AMI text materials. The baseline system corresponding to the CL-DH approach has 4 BLSTM layers with 320 cells per layer in each direction. The last layer is a softmax layer with 46 outputs preceded by an affine transform layer. Unless otherwise stated, the same architecture has been used in all experiments reported in this study. We kept the learning rate the same as in the Switchboard recipe, with an initial value of 0.00004. As mentioned previously, approximately 10% of the training data was set aside as a held-out set to monitor the training. The maximum number of iterations was set at 20.

For the baseline corresponding to the CL-DH approach, referred to as system B-1 in this work, the deep BLSTM network is trained with far-field (SDM) data only. We note that while the baseline system B-1 is trained on far-field data only, the corresponding CL-DH system is first exposed to near-field, and subsequently to far-field data. Our motivation for comparing CL-DH with B-1 baseline is only to show that performance can be improved by moving to difficult data (FF) from easier data (Near-field) in the CL paradigm.

Table 2
Description of the baseline systems corresponding to the CL-DH approach (system B-1), CL-DM and CL-DHM approaches (system B-2).

Baseline system	No of BLSTM layers	Mem. cells per dir.	BLSTM trained on
B-1	4	320	FF
B-2	4	320	NF + FF

The Eesen recipe stops training when the relative improvement falls below a predefined threshold. For the baseline corresponding to the CL-DM and CL-DHM approaches, the deep BLSTM network was trained with both near and far field data in every iteration. Thus, this baseline is a standard multi-condition trained network with both near and far-field data. Table 2 highlights the architectural details of the two baseline systems B-1 and B-2 used in this study.

5.2. CL-DH based far-field ASR systems

For the CL-DH based far-field ASR system, first the BLSTM network is trained with the comparatively easier near-field (IHM) data for a certain number of iterations as outlined in Alg. 1. Next, the BLSTM network obtained using the near-field (IHM) data is used as the initial network for training the final network with the more difficult far-field (SDM) data. As can be observed in Alg. 1, the number of iterations n_k used for training with a particular data set (i.e., near-field or far-field) can have an impact on performance of the final ASR system. This issue is examined in more detail in Section 5.5.

5.3. CL-DM based far-field ASR systems

For the CL-DM based far-field ASR system, first the BLSTM network is trained with the comparatively easier near-field (IHM) data for a predetermined number of iterations as outlined in Alg. 2. Next, the BLSTM network obtained using the near-field IHM data is used as the initial network for training the final BLSTM ASR network with both easy (IHM) and difficult (SDM) data. As can be observed in Alg. 2, the number of iterations n_k used for training with a particular data set (i.e., near-field or near-field + far-field) can have an impact on the performance of the final ASR system. This is investigated in more detail in Section 5.5.

5.4. CL-DHM based far-field ASR systems

For the CL-DHM based far-field ASR system, first the BLSTM network is trained with comparatively easier near-field (IHM) data for a predetermined number of iterations as outlined in Alg. 3. Next, the BLSTM network obtained using the near-field IHM data is used as the initial network for training with a few iterations on only the more difficult far-field (SDM) data. In the last step, the final BLSTM ASR network is trained with both easy (IHM) and difficult (SDM) data. As can be observed in Alg. 3, the number of iterations n_k used for training with a particular dataset also impacts on the performance of the final ASR system, which is considered in the next section.

5.5. Determining optimal number of iterations for CL based approaches

For the CL-DH based far-field ASR systems, we experimented with several values of n_k from Alg. 1 for training using easy near-field IHM data. For the second stage of both CL-DH and CL-DM based approaches, we did not vary the number of iterations and made use of the EESEN recipe's typical stopping criterion for terminating the training. To this end, Table 3 shows WERs obtained on the dev subset of the AMI corpus using the CL-DH based approach for different values of n_k. We investigated three values for the number of iterations on the easy (near-field) IHM data: 3, 6, and 9. Thus, these set of experiments were aimed

Table 3
WER obtained by our proposed CL-DH based approach on dev set of the AMI corpus for varying iteration count on the easy (near-field) data.

No. of iterations, n_k on easy (NF) data	WER (%) on dev
3	65.5
6	65.5
9	64.4

Table 4
WER obtained by our proposed CL-DM based approach on dev set of AMI corpus for varying iteration count on the easy (near-field) data.

No. of iterations, d_k on easy (NF) data	WER on dev
3	61.4
6	60.9
9	61.1

at exploring the sensitivity to the number of iterations in the first step of our 2-step training strategy for CL-DH.

As can be seen from Table 3, increasing the number of iterations from 3 to 9 had a positive impact on performance of our proposed CL-DH based approach as evident by a reduction in WER on the dev set. However, no reduction in WER was observed when increasing the number of iterations from 3 to 6, suggesting that our proposed CL-DH based approach is not very sensitive to the number of iterations on the easy data set as long as we employ at least a few iterations on the easier dataset in stage 1. We also observed that increasing the number of iterations beyond 9 did not offer any significant improvement compared to the CL-DH based approach trained with 9 iterations. We report performance for the two CL-DH based approaches CL-DH1 and CL-DH2 trained with 6 and 9 iterations respectively.

Next, we examine the effect of the number of iterations on the CL-DM based approaches. To this end, Table 4 also shows WERs on the dev set of the AMI corpus, for three CL-DM based systems with 3, 6, and 9 iterations respectively on the easy (near-field) data

From results in Table 4, it is observed that increasing the number of iterations n_k on the easy (near-field) IHM data in the first step of our proposed CL-DM based approach has a positive impact on performance when the number of iterations is increased from 3 to 6. This is substantiated by a small reduction in WER on the dev set of the AMI corpus for the corresponding BLSTM networks. However, there is a slight loss in performance as marked by a small increase in WER on the dev set, when the number of iterations is further increased to 9. This points out that increasing the number of iterations beyond a certain value may not offer any additional improvements. For the CL-DM experiments reported in this study corresponding to the CL-DM approach, we report the results of two systems: CL-DM1 and CL-DM2 with 6 and 9 iterations respectively on the easy (near-field) IHM data. We did not observe any additional improvements by increasing the number of iterations on the easy, near-field data beyond 9.

We also investigated the effect of the number of iterations on the difficult data on our proposed CL-DHM based approach. To this end, Table 5 shows results on the dev set of the AMI corpus obtained by the CL-DHM systems with 3, 6 and 9 iterations on the difficult (far-field, SDM) data. For all three systems, the network was first trained with 9 iterations on the easy (IHM) data before being trained only on difficult data in the 2nd stage. All systems were trained on both near and far-field data in the last step to obtain the final models.

As can be observed from Table 5, our proposed CL-DHM based approach is not very sensitive to the number of iterations on the difficult data in the second step. However, there is a slight degradation marked by a small increase in WER, as the number of iterations is increased to 9 from 3. The performance for all three systems is generally

Table 5
WER obtained by our proposed CL-DHM based approach on dev set of the AMI corpus for varying iterations on the difficult (far-field) data.

No. of iterations, d_k on difficult (far-field) data	WER on dev
3	61.4
6	61.7
9	61.7

Table 6
Comparison of WER on the dev and eval sets of the AMI corpus obtained by the CL-DH1 and CL-DH2 systems trained with 6 and 9 iterations respectively on the easy (near-field) data, and the far-field data, compared against the baseline system B-1 trained with far-field data. For the baseline, and for CL approaches (in the last stage of training), maximum number of iterations was kept at 20, but in practice, the training was observed to stop before 20 iterations due to the stopping criterion used.

ASR system	No. iter. (easy data)	Max No. Iter.	WER	
			DEV	EVAL
Baseline B-1	NA	20	66.6	71.0
CL-DH1	6	26	65.5	69.1
CL-DH2	9	29	64.4	68.0

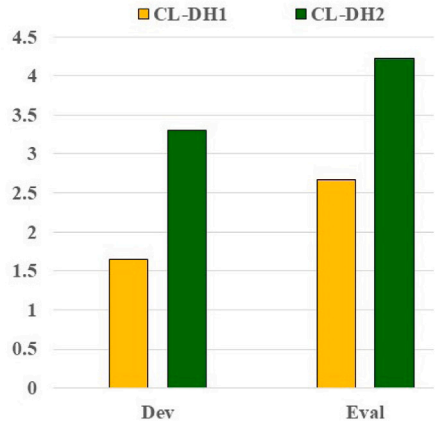

Fig. 3. Relative reduction in WER for our proposed CL-DH approach based systems: CL-DH1 and CL-DH2 compared against the baseline system B-1 (trained on far-field data) on the dev and eval sets of the AMI corpus.

similar. These results also indicate that for practical considerations, it is reasonable to use 3 iterations on the difficult data for our proposed CL-DHM based approach.

We observed that for all our proposed CL based approaches, the loss function decreased faster in comparison to the corresponding baseline systems. This is in line with other works that have reported faster reduction in loss function when CL is used (Bengio et al., 2009; Amodei et al., 2016).

5.6. Results & discussion

We first investigate performance of our proposed CL-DH based approaches for far-field speech recognition. To this end, Table 6 shows performance of our two CL-DH systems: CL-DH1 (with 6 iterations on the easy, NF data) and CL-DH2 (with 9 iterations on the easy, NF data) compared against the baseline system B-1 trained on only the far-field data. The table shows results for both dev and eval subsets of the AMI corpus.

Fig. 3 shows the relative reduction in WER achieved by our proposed CL-DH approach. Specifically, the WER reduction achieved by the CL-DH1 and CL-DH2 systems compared against the baseline system B-1 (trained on only far-field data) is shown.

Table 7
Comparison of WER on the dev and eval sets of the AMI corpus obtained by the CL-DM1 and CL-DM2 systems trained with 6 and 9 iterations respectively on the easy (near-field) data, and the far-field data, compared against a baseline system trained with both near and far-field data. For the baseline, and for CL approaches (in the last stage of training), maximum number of iterations was kept at 20, but in practice, the training was observed to stop before 20 iterations due to the stopping criterion used.

ASR system	No. iter. (easy data)	Max No. iter.	WER	
			DEV	EVAL
Baseline B-2	NA	20	65.7	73.2
CL-DM1	6	26	60.9	66.1
CL-DM2	9	29	61.1	65.9

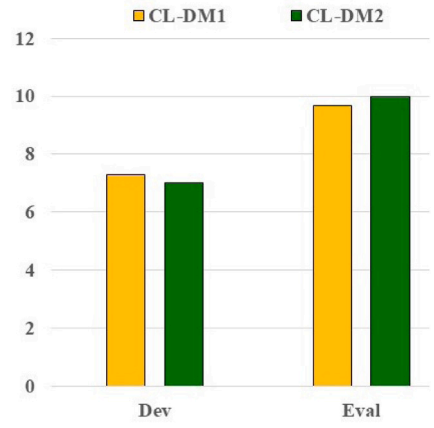

Fig. 4. Relative reduction in WER for our proposed CL-DM approach based systems: CL-DM1 and CL-DM2 compared against the baseline system B-2 (trained on near and far-field data) on the dev and eval sets of the AMI corpus.

Table 8
Comparison of WER on the dev and eval sets of the AMI corpus obtained by the CL-DHM1 and CL-DHM2 systems trained with 6 and 9 iterations respectively on the easy (near-field) data, and difficult (far-field) data, compared against a baseline system trained with both near and far-field data. For the baseline, and for CL approaches (in the last stage of training), maximum number of iterations was kept at 20, but in practice, the training was observed to stop before 20 iterations due to the stopping criterion used.

ASR system	No. iter. (difficult data)	Max No. Iter.	WER	
			DEV	EVAL
Baseline B-2	NA	20	65.7	73.2
CL-DHM1	6	35	61.7	65.9
CL-DHM2	9	26	61.7	65.8

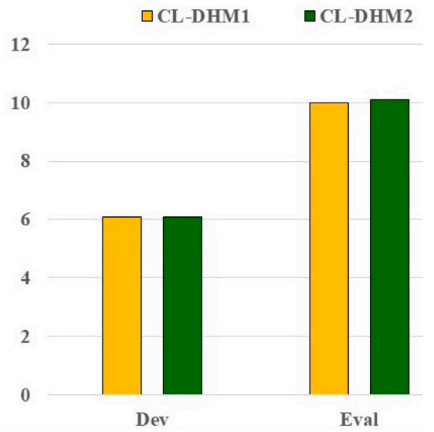

Fig. 5. Relative reduction in WER for our proposed CL-DHM approach based systems: CL-DHM1 and CL-DHM2 compared against the baseline system B-2 (trained on near and far-field data) on the dev and eval sets of the AMI corpus.

As seen in Fig. 3, CL-DH2 (trained with 9 iterations on the easy, near-field data in the first stage), consistently outperforms the CL-DH1 system (trained with 6 iterations on the near-field data) on both dev and eval subsets of the AMI corpus. The reductions in WER for the eval set are more compared to the dev set, with CL-DH1 and CL-DH2 offering relative reduction in WER of +2.67% and +4.22% respectively.

Performance of our proposed CL-DM based approaches are investigated in Table 7. To this end, WERs of CL-DM1 and CL-DM2 systems are compared against a baseline Bidirectional LSTM based ASR system, B-2, trained with both near and far-field data. The results are shown for both dev and eval sets of the AMI corpus.

Comparing results of the baseline systems B-1, B-2 from Tables 5 and 6 respectively, we observe that including both near and far-field data in training improves performance on dev set of the AMI corpus, as expected. However, a noticeable loss in performance on the eval set is marked by an increase in WER as experienced by the baseline system B-2.

Fig. 4 shows the relative reductions in WER achieved by our proposed CL-DM approaches. Specifically, WER reduction achieved by CL-DM1 and CL-DM2 systems compared against the baseline system B-2 (trained on near and far-field data) is shown.

As seen in Fig. 4, CL-DM2 (trained with 9 iterations on the easy, near-field) data in the first stage, achieves a similar reduction in WER as the CL-DM1 system (trained with 6 iterations on the near-field data) on both dev and eval subsets of the AMI corpus. This is different from what was observed for the CL-DH based systems, where CL-DH2 outperformed CL-DH1 system consistently. On dev set of AMI corpus, CL-DM1 and CL-DM2 reduced the WER by +7.30% and +7.00% respectively. The relative reductions in WER for the eval set of the AMI corpus were +9.69% and +9.97% respectively for the CL-DM1 and CL-DM2 systems. Similar to what was observed for the CL-DH approach, compared to the dev set of the AMI corpus, the relative reductions in WER are higher for the eval set. Overall, these results demonstrate the effectiveness of our proposed CL-DM approach at improving the ASR accuracy of the CTC trained deep BLSTM network.

Table 8 shows performance of our proposed CL-DHM based approaches: CL-DHM1 (trained with 6 iterations) and CL-DHM2 (trained with 9 iterations) respectively on dev and eval sets of the AMI corpus. These results can be compared against the multi-condition baseline system B-2, trained on near-field and far-field data combined.

From Table 8, it is observed that our proposed CL based approaches CL-DHM1 and CL-DHM2, are both effective in reducing WERs compared to a traditional multi-condition baseline. Also, the improvement in performance achieved with the CL-DH approach is similar to what was observed for the CL-DM approach (from Table 7).

Fig. 5 shows the relative reduction in WER achieved by our proposed CL-DHM approach. Specifically, the reductions in WER achieved by the CL-DHM1 and CL-DHM2 systems compared against the baseline system B-2 (trained on near and far-field data) are shown.

As seen, both CL-DHM2 (trained with 9 iterations on difficult far-field data in step 2) and CL-DHM1 (trained with 6 iterations on far-field data in step 2), achieve similar reductions in WER as the CL-DM based systems on both dev and eval subsets of the AMI corpus, compared to baseline system B-2. On the dev set of the AMI corpus, both CL-DHM1 and CL-DHM2 reduced WER by +6.08% each. The relative reductions in WER for the eval set of the AMI corpus are +9.97% and +10.10% for the CL-DHM1 and CL-DHM2 systems respectively.

Fig. 6 shows the cross-validation (CV) token accuracy of the baseline B-2 compared against our proposed CL-DM1 and CL-DM2 systems. The initialization phase (on NF data) of the two CL based systems are not shown, since the system B-2 initiates training with entire data.

As can be seen from Fig. 6, the baseline system's training saturates quicker and stops at iteration 11 with 62.46% classification accuracy.

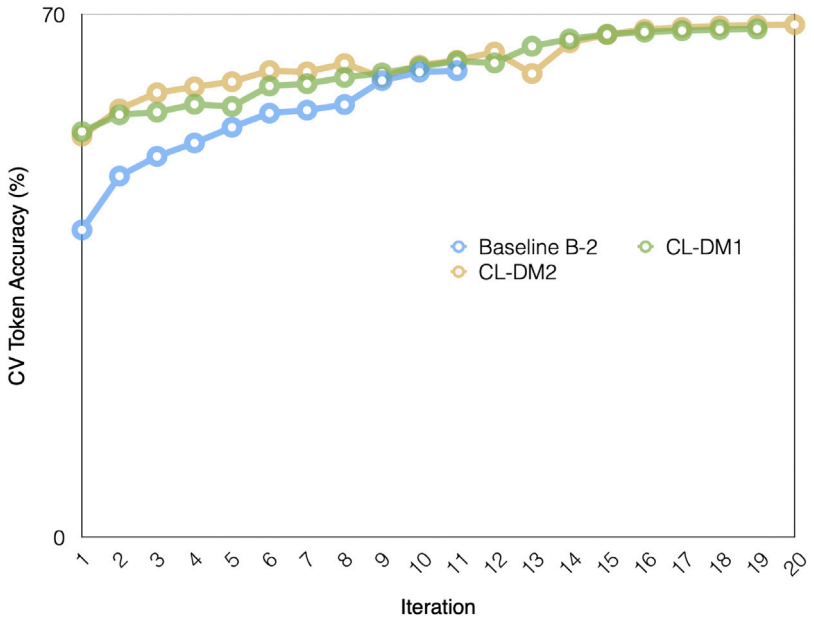

Fig. 6. Cross-validation (CV) token accuracy (%) of the baseline B-2 compared against our proposed CL based systems CL-DM1 (initialized with 6 iterations on NF data) and CL-DM2 (initialized with 6 iterations on NF data). With the same hyper-parameters and training setup, the system B-2 saturated, and the corresponding training stopped at iteration 11, while CL-DM1 and CL-DM2 stopped at iteration 19 and 20 respectively.

However, the two CL based systems CL-DM1 and CL-DM2 are able to achieve 68.04% and 68.62% classification accuracy respectively as the training were able to continue for more iterations before saturating. We hypothesize that our CL based systems achieve higher accuracy due to better local minima by using distance as a difficulty criterion in training.

6. Conclusion

This study has proposed novel Curriculum Learning based approaches for robust far-field speech recognition using CTC trained deep bidirectional LSTM networks. The proposed CL based algorithms operate by initializing training with easier data and gradually include more difficult data as training progresses. We hypothesized this approach of gradually increasing training data diversity leads to better estimation of the model parameters for far-field speech recognition. Our proposed CL-DH and CL-DM based approaches were first trained on the comparatively easier near-field data, and then on far-field, and a combination of near and far-field data respectively.

We also proposed combining the elements of CL-DH and CL-DM algorithms in the CL-DHM approach, wherein the training progressed with including easy (near-field), to difficult (far-field), to both easy and difficult data in training. We reported performance of our proposed CL-DH, CL-DM and CL-DHM approaches on both dev and eval sets of the AMI corpus. All three CL based approaches were shown to significantly outperform the corresponding CTC trained BLSTM ASR baseline networks.

On the dev set of the AMI corpus, our proposed CL-DM based approach offered a relative reduction of +7.3% in WER compared to a standard multi-condition baseline system trained on both near and far-field data. Compared to the CL-DHM approach, the CL-DM strategy gave comparable, albeit superior, performance on the dev set of the AMI corpus. On the eval set of the AMI corpus, our proposed CL-DHM approach reduced the WER by +10.10% (relative) compared to the baseline multi-condition system. However, on the eval set of the AMI corpus, the performance of both CL-DM and CL-DHM approaches were generally similar, with both offering close to +10% relative reduction in WER.

While we have shown our CL-based approaches to be effective for training deep BLSTM based ASR systems trained with the CTC objective function, CL based approaches have also been shown to work for encode–decoder architectures in Speech Translation (Kano et al., 2017). Therefore, we believe our proposed CL based approaches can also work on more advanced attention based encoder–decoder architectures such as ESPnet (Watanabe et al., 2018), and other advanced approaches using transformers (Karita et al., 2019). We also note that while the AMI corpus has near-field and far-field versions of the same data, this may not be the case for many real-life datasets. In such cases, our proposed CL based approaches can still be used provided there is some access to near-field data from another corpus. We note that matched near and far-fields data for the same utterance is not a requirement of our CL based algorithms. Furthermore, similar CL based approaches can also be devised using other criterion such as Word Error Rate (WER). In such a CL guided approach, training data can be assigned WER bins, and a CL based approach can then use WER guided bins in training, starting with lowest WER data.

Moreover, CL based approaches have also been used for improved speaker recognition in scenarios where multiple layers of difficulty are present in the data (Ranjan and Hansen, 2018; Marchi et al., 2018). Similar strategies can also be explored to improve far-field ASR that could potentially utilize multi-layered difficulty-graded data.

This study has, therefore, demonstrated the merits of adopting curriculum learning based strategies for training deep BLSTM based ASR systems trained with the CTC objective function. Future work will explore using our proposed CL-based approaches to improve performance of ASR systems under other domain mismatch scenarios. Another direction of future work would be to extend the proposed approaches using advanced architectures such as Conformer (Gulati et al., 2020) and ESPnet (Watanabe et al., 2018). The proposed approaches presented in this work can also be extended for use in other scenarios such as the Chime challenges, where data from both binaural and array microphones are available (Barker et al., 2018). Another avenue for future work could be to explore the use of automated curriculum learning based approaches that can learn from data without a human-assigned difficulty criterion for improving far-field speech recognition (Graves et al., 2017; Narvekar et al., 2020).

CRediT authorship contribution statement

Shivesh Ranjan: Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft. **John H.L. Hansen:** Supervision, Conceptualization, Methodology, Resources, Writing - review & editing, Project administration, Funding acquisition.

References

Allgower, E.L., Georg, K., 2012. Numerical continuation methods: an introduction. 13, Springer.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al., 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. Int. Conf. Mach. Lear. 173–182.

Barker, J., Watanabe, S., Vincent, E., Trmal, J., 2018. The fifth'chime'speech separation and recognition challenge: dataset, task and baselines. ISCA Interspeech 2018.

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. ICML 41–48.

Braun, S., Neil, D., Liu, S.-C., 2017. A curriculum learning method for improved noise robustness in automatic speech recognition. In: 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, pp. 548–552.

Carletta, J., 2006. Announcing the AMI meeting corpus. ELRA Newsl. 11 (1), 3–5.

Chang, X., Zhang, W., Qian, Y., Le Roux, J., Watanabe, S., 2019. MIMO-speech: End-to-end multi-channel multi-speaker speech recognition. 2019 IEEE ASRU 237–244.

Giri, R., Seltzer, M.L., Droppo, J., Yu, D., 2015. Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning. IEEE ICASSP 2015 5014–5018.

Graves, A., Bellemare, M.G., Menick, J., Munos, R., Kavukcuoglu, K., 2017. Automated curriculum learning for neural networks. In: International Conference on Machine Learning. PMLR, pp. 1311–1320.

Graves, A., Fernández, S., Gomez, F., Schmidhuber, J., 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. ICML 369–376.

Graves, A., Mohamed, A.-R., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. IEEE ICASSP 2013 6645–6649.

Grezes, F., Ni, Z., Trinh, V.A., Mandel, M., 2020. Combining spatial clustering with LSTM speech models for multichannel speech enhancement. ArXiv Preprint ArXiv: 2012.03388.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al., 2020. Conformer: Convolution-augmented transformer for speech recognition. ISCA Interspeech 2020.

Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Van Leeuwen, D., Lincoln, M., Wan, V., 2008. The 2007 AMI (DA) system for meeting transcription. Springer, pp. 414–428,

Himawan, I., Motlicek, P., Imseng, D., Potard, B., Kim, N., Lee, J., 2015. Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition. IEEE ICASSP 2015 (EPFL-CONF-207946).

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. ArXiv Preprint ArXiv:1503.02531.

Kano, T., Sakti, S., Nakamura, S., 2017. Structured-based curriculum learning for end-to-end english-Japanese speech translation. ISCA Interspeech 2630–2634..

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N.E.Y., Yamamoto, R., Wang, X., et al., 2019. A comparative study on transformer vs RNN in speech applications. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 449–456.

Khoubrouy, S.A., Hansen, J.H.L., 2016. Microphone array processing strategies for distant-based automatic speech recognition. IEEE Signal Process. Lett. 23 (10), 1344–1348.

Kim, J., El-Khamy, M., Lee, J., 2017. Bridgenets: Student-teacher transfer learning based on recursive neural networks and its application to distant speech recognition. IEEE ICASSP 5719–5723.

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017. A study on data augmentation of reverberant speech for robust speech recognition. IEEE ICASSP 2017 5220–5224.

Kumatani, K., McDonough, J., Raj, B., 2012. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. IEEE Signal Process. Mag. 29 (6), 127–140.

Marchi, E., Shum, S., Hwang, K., Kajarekar, S., Sigtia, S., Richards, H., Haynes, R., Kim, Y., Bridle, J., 2018. Generalised discriminative transform via curriculum learning for speaker recognition. IEEE ICASSP 2018.

Marino, D., Hain, T., 2011. An analysis of automatic speech recognition with multiple microphones. ISCA Interspeech 1281–1284.

Miao, Y., Gowayyed, M., Metze, F., 2015. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. IEEE ASRU 2015 167–174.

Miao, Y., Metze, F., 2015. Distance-aware DNNs for robust speech recognition. ISCA Interspeech 2015.

Mirsamadi, S., Hansen, J.H., 2019. Multi-domain adversarial training of neural network acoustic models for distant speech recognition. Speech Commun. 106, 21–30.

Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M.E., Stone, P., 2020. Curriculum learning for reinforcement learning domains: A framework and survey. J. Mach. Learn. Res. 21 (181), 1–50.

Peddinti, V., Chen, G., Manohar, V., Ko, T., Povey, D., Khudanpur, S., 2015b. Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms. IEEE ASRU 2015 539–546.

Peddinti, V., Chen, G., Povey, D., Khudanpur, S., 2015a. Reverberation robust acoustic modeling using i-vectors with time delay neural networks. ISCA Interspeech 2015.

Qian, Y., Tan, T., Yu, D., 2016. An investigation into using parallel data for far-field speech recognition. IEEE ICASSP 2016 5725–5729.

Ranjan, S., Hansen, J.H.L., 2018. Curriculum learning based approaches for noise robust speaker recognition. IEEE/ACM Trans. Audio, Speech, & Lang. Process. (TASLP) 26 (1), 197–210.

Ranjan, S., Misra, A., Hansen, J.H.L., 2017. Curriculum learning based probabilistic linear discriminant analysis for noise robust speaker recognition. In: ISCA INTERSPEECH 2017. ISCA, pp. 3717–3721.

Ravanelli, M., Brakel, P., Omologo, M., Bengio, Y., 2017. A network of deep neural networks for distant speech recognition. IEEE ICASSP 2017 4880–4884.

Renals, S., Swietojanski, P., 2017. Distant speech recognition experiments using the AMI corpus. New Era for Robust Speech Recognition 355–368.

Seltzer, M.L., Raj, B., Stern, R.M., et al., 2004. Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Trans. Speech Audio Process. 12 (5), 489–498.

Swietojanski, P., Ghoshal, A., Renals, S., 2013. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. IEEE ASRU 2013 285–290.

Swietojanski, P., Ghoshal, A., Renals, S., 2014. Convolutional neural networks for distant speech recognition. IEEE Signal Process. Lett. 21 (9), 1120–1124.

Trinh, V.A., Mandel, M.I., 2020. Large scale evaluation of importance maps in automatic speech recognition. ISCA Interspeech 1166–1170.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., Ochiai, T., 2018. Espnet: End-to-end speech processing toolkit. ISCA Interspeech 2207–2211.

Zhang, W., Chang, X., Qian, Y., Watanabe, S., 2020. Improving end-to-end single-channel multi-talker speech recognition. IEEE/ACM Trans. Audio, Speech, & Lang. Process. 28, 1385–1394.