applied sciences

Article

Speech Recognition for Task Domains with Sparse Matched Training Data

Byung Ok Kang *, Hyeong Bae Jeon and Jeon Gue Park

Electronics and Telecommunications Research Institute, Daejeon 34129, Korea; hbjeon@etri.re.kr (H.B.J.); jgp@etri.re.kr (J.G.P.)

* Correspondence: bokang@etri.re.kr; Tel.: +82-42-860-5684

Received: 8 August 2020; Accepted: 3 September 2020; Published: 4 September 2020

Abstract: We propose two approaches to handle speech recognition for task domains with sparse matched training data. One is an active learning method that selects training data for the target domain from another general domain that already has a significant amount of labeled speech data. This method uses attribute-disentangled latent variables. For the active learning process, we designed an integrated system consisting of a variational autoencoder with an encoder that infers latent variables with disentangled attributes from the input speech, and a classifier that selects training data with attributes matching the target domain. The other method combines data augmentation methods for generating matched target domain speech data and transfer learning methods based on teacher/student learning. To evaluate the proposed method, we experimented with various task domains with sparse matched training data. The experimental results show that the proposed method has qualitative characteristics that are suitable for the desired purpose, it outperforms random selection, and is comparable to using an equal amount of additional target domain data.

Keywords: automatic speech recognition; sparse training data; deep neural network; active learning; transfer learning

1. Introduction

Deep neural networks (DNN) have been widely adopted and applied to traditional pattern recognition applications, such as speech and image recognition. DNN-based acoustic models have significantly improved speech recognition performance [1,2]. Initially, studies focused on acoustic models based on the deep neural network-hidden Markov model (DNN-HMM). More recently, end-to-end speech recognition, which completely replaces HMM with DNN, has become the focus. It has been adopted for many commercialized speech recognition systems [3,4]. DNN-based acoustic models, especially end-to-end models, use more parameters than conventional HMM-based models and require massive amounts of training data for high performance.

However, for some tasks, collecting large amounts of speech data is difficult. Non-native speech recognition is an example. It is difficult to collect extensive speech databases from non-native speakers, compared to native speakers. The number of non-native speakers of a language is usually much smaller than that of native speakers. Furthermore, due to inaccurate pronunciation and lack of language fluency, it often costs more to obtain transcription data for non-native speakers than for native speakers. In addition, some applications such as call center recording speech recognition limit the collection of large amounts of speech data due to policies such as personal information security. The shortage of training data matched to applied task domains causes degradation in speech recognition accuracy. This problem tends to be more serious in end-to-end automatic speech recognition (ASR).

To handle these difficulties, various approaches have been studied. The most representative research approaches, which are also widely used in commercial ASR services, are shared models, domain adaptation mechanisms, and semi-supervised learning approaches such as self-training and multi-task learning. Approaches that share model parameters or phone-sets have mostly been used for speech recognition tasks involving languages with low resources [5–7]. Domain adaptation uses a well-trained source-domain model to adapt to the target domain with matched target data [8–11]. Semi-supervised learning approaches focus on joint learning with labeled and unlabeled speech data. To learn with unlabeled speech data, self-training approaches mainly focus on generating transcriptions for unlabeled speech data using a pre-trained ASR system. Research has been conducted to obtain reliable confidence measures among the generated transcriptions [12,13]. The semi-supervised learning approaches based on multi-task learning focus on linearly combining the supervised cost function of a deep classifier with the unsupervised cost function of a deep auto-encoder, and then minimizing the combination of costs [14,15].

These methods can be classified into methods that do, or do not require transcribed data from the target task domain. Shared models and domain adaptation methods are in the former category; semi-supervised learning approaches based on self-training and multi-task learning belong to the latter. Domain adaptation mechanisms are widely applied due to their ability to improve stability and performance. However, to achieve satisfactory performance improvement, most domain adaptation approaches require a considerable amount of domain speech data with transcription. The semi-supervised learning approaches based on self-training and multi-task learning do not have additional costs for the transcription of target domain speech data, but they have the drawback of requiring a considerable amount of un-transcribed speech data from the target domain.

In this work, we focus on the problem of constructing a speech recognition system with a stable performance for domains where it is difficult to collect large amounts of matched speech data. To handle this problem, we propose two approaches. The first method actively selects training data for a target domain from the training data of another domain that already has a significant amount of labeled speech data by using attribute-disentangled latent variables. The second is a method that combines data augmentation methods for the target domain of sparse matched speech data and a transfer learning method based on teacher/student learning.

The remainder of this paper is organized as follows. In Section 2, we briefly review the research areas related to our proposed methods. Section 3 describes our proposed approach in detail. Section 4 explains the experimental setting, and Section 5 presents the experimental results. Finally, Section 6 concludes this paper and discusses future work.

2. Related Work

2.1. End-to-End Speech Recognition

End-to-end (E2E) speech recognition systems consisting of a single integrated neural network model that is trained through input speech and output transcription have recently been proposed. Such systems have been applied to many commercialized speech recognition services. In traditional speech recognition, the acoustic model is trained through several steps. In addition, in order to combine the acoustic model, pronunciation dictionary, and the separately trained language model, weighted finite-state transducers (WFST) are used to find the most probable path and recognize speech. This process is cumbersome and requires prior knowledge of speech recognition to understand each role. By contrast, the end-to-end model does not need several training steps and each step of its structure is easy to understand.

A typical end-to-end speech recognition model uses a connectionist temporal classification (CTC) method [16–18]. This method uses a recursive neural network to infer text strings directly from input speech features. Similar to a Gaussian mixture model-hidden Markov model (GMM-HMM), character posterior probability is estimated in every frame, and the estimated character string attempts to determine the optimal path. However, a different approach using a speech recognition model based on a sequence-to-sequence (seq2seq) model has been proposed. It has achieved significant improvements in performance in the field of machine translation [19–22]. This model consists of a recursive network encoder and decoder. The encoder calculates the output for every frame from the input speech features. The decoder calculates which frame is paying attention to the encoder output, and estimates the final character string using the encoder value as an input according to the degree of attention. This seq2seq-based speech recognition model shows performance that is comparable to other end-to-end models, and ongoing research is being conducted.

To verify the proposed method in this study, we used ESPnet, a Python-based open source platform for public end-to-end speech recognition [23].

2.2. Active Learning

Active learning is a field of machine learning that allows a model to select its own training data. It aims to achieve the desired level of performance for the target task with less new labeled training data [24]. For this purpose, it is important to select training data with a significant amount of information. Studies have been conducted on various confidence scores to measure the informativeness of new training samples. The most common method in automatic speech recognition is the least confidence (LC) sampling technique, in which training samples with the least certainty are considered the most informative from the model's perspective [25]. The LC for a sequence model can be obtained as follows:

$$\varnothing^{LC}(\mathbf{x}) = 1 - P(\mathbf{y}^{*} \middle| \mathbf{x}; \theta), \tag{1}$$

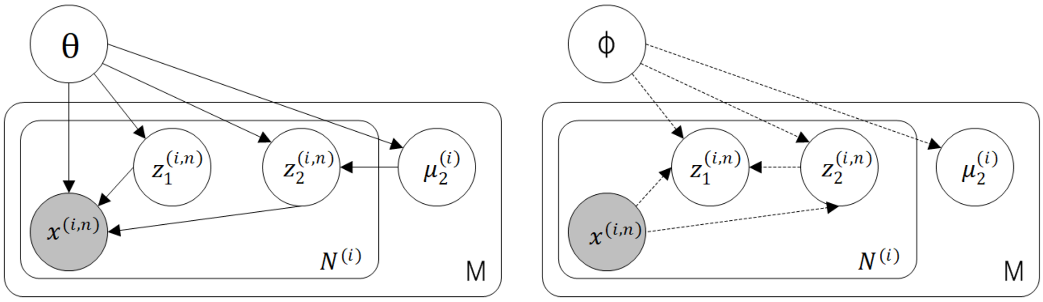
where $\mathbf{y}^{*}$ is the most likely label sequence, $\mathbf{x}$ is an observation sequence, and θ represents the model parameters. This is a method based on uncertainty sampling, which has already been proposed in the field of machine learning. Uncertainty sampling is a method of selecting new training samples predicted by the underlying model with the lowest confidence [26].

The drawback of LC-based sampling is that it can suffer from the problem of sampling bias and be particularly biased towards a specific group of speakers [27]. Conventional studies of speech recognition using active learning mainly focus on the problem of selecting data to be transcribed from unlabeled target domain speech data. They mainly target ASR of languages with low resources. The method proposed in the paper differs from conventional active learning techniques. It selects training data for a target domain with low resources from the training data of another domain that already has an immense amount of labeled speech data.

2.3. Disentangled Representation Learning in Speech Processing

Representation learning is a method that learns representations of input data, with the main aim of yielding abstract and useful representations for tasks such as classification. In the past, research focused on feature engineering to create representations that support machine learning algorithms. However, representations based on deep learning are now being widely studied [28]. Among the many applications of deep representation learning is disentangled representation. This method separates each feature into narrowly defined variables and encodes them into separate dimensions [29]. Assuming that the data is generated from independent factors of variation, disentanglement enables these factors to be captured by different independent variables in the representation, which yields a concise abstract representation of the data [30].

Several recent works have leveraged variational autoencoders (VAEs) to learn disentangled representations of sequence data such as speech, video, and text. Hsu et al. [31] proposed a novel factorized hierarchical VAE, which learns disentangled and interpretable latent representations from speech data by explicitly modeling the multi-scaled information with a factorized hierarchical graphical model. Speech data inherently contains information at multiple scales such as noise, channel, speaker, prosody, and phonetic content. These are independent factors operating at different time scales. For instance, noise, channel, and speaker identity affect the sequence level and tend to have a smaller amount of variation within an utterance compared to the variation between utterances. The sequence level is a time series of speech data that is the object of model training and speech recognition. It can be an utterance or a phrase. However, the phonetic content affects the segmental level and tends to have a similar amount of variation within and between utterances. The segment level is a small unit that comprises a sequence and is composed of frames of a certain length. Factorized hierarchical VAEs consist of an inference model and a generative model that learns a disentangled representation of a latent sequence variable and a latent segment variable, which have properties that change at the utterance and segment level, respectively. Figure 1 is a graphical illustration of factorized hierarchical VAEs [31].



(**a**) Generative Model (**b**) Inference Model

Figure 1. Graphical illustration of the factorized hierarchical variational autoencoders (VAEs). Grey nodes denote the observed variables, and white nodes are the hidden variables.

In Figure 1, $\left\{x^{(i,n)}\right\}_{n=1}^{N^{(i)}}$ is a sequence of $N^{(i)}$ observed variables, where $x^{(i,n)}$ is the n-th segment of the i-th sequence and $N^{(i)}$ is the number of segments for the i-th sequence. As Figure 1 illustrates, the following generation process is considered. First, an s-vector μ_2 is drawn from a prior distribution $p_\theta\left(\mu_2^{(i)}\right)$ for each i-th sequence. Second, $N^{(i)}$ independent and identically distributed $\left\{z_2^{(i,n)}\right\}_{n=1}^{N^{(i)}}$ and $\left\{z_1^{(i,n)}\right\}_{n=1}^{N^{(i)}}$ are drawn from a sequence-dependent prior distribution $p_\theta\left(z_2^{(i)}\middle|\mu_2^{(i)}\right)$ and sequence-independent distribution, $p_\theta\left(z_1^{(i)}\right)$, respectively; Finally, $N^{(i)}$ independent and identically distributed $\left\{x^{(i,n)}\right\}_{n=1}^{N^{(i)}}$ are drawn from a conditional distribution $p_\theta\left(x^{(i)}\middle|z_1^{(i)}, z_2^{(i)}\right)$.

2.4. Teacher-Student Learning for Domain Adaptation

The performance of ASR degrades significantly when there is a mismatch between the training and real test environments. The most intuitive and commonly used solution is to adaptively train a well-trained source domain model to target domain data [8,32]. However, the disadvantage of domain adaptation is that it requires a considerable amount of labeled target domain data. Thus, it is costly and not suitable for domains where it is difficult to collect large amounts of data.

To handle these issues, teacher-student (T/S) learning was proposed for the domain adaptation of DNN-HMM-based acoustic models [33]. In teacher-student learning, the Kullback-Leibler (KL) divergence between the posterior distributions of the teacher and student networks, given parallel source and target domain data as input, is minimized by updating the model parameters of the student network. The KL divergence between the source and target distributions is as follows:

$$D_{KL}\left(P_T(s|\mathbf{x}_{src}) \parallel P_S\left(s\middle|\mathbf{x}_{tgt}\right)\right) = \sum\nolimits_f \sum\nolimits_i P_T\left(s_i\middle|\mathbf{x}_{src,f}\right) \log\left(\frac{P_T\left(s_i\middle|\mathbf{x}_{src,f}\right)}{P_S\left(s_i\middle|\mathbf{x}_{tgt,f}\right)}\right), \tag{2}$$

where s_i indicates a state with index i and f is the frame index. $P_T\left(s_i \middle| \mathbf{x}_{src,f}\right)$ and $P_S\left(s_i \middle| \mathbf{x}_{tgt,f}\right)$ indicate the posterior distribution of the teacher and student networks, respectively, while $\mathbf{x}_{src,f}$ and $\mathbf{x}_{tgt,f}$ are the source and target inputs to the teacher and student networks.

T/S-based training using soft labels of the teacher network output showed improved results compared to conventional cross-entropy training, which directly uses the hard label in the target domain [33–35]. However, to obtain high performance, T/S learning for domain adaptation requires parallel sequences of source and target domain data, which consist of real data and the simulated pair generated from one domain to the other. These are synchronized frame-by-frame. Examples of parallel source/target data in previous T/S learning studies include real clean speech and simulated noisy speech [33], real adult speech and simulated child speech [33], enhanced clean speech and real noisy speech [34] and real close-talk speech and simulated far field speech [35]. The target domains in the studies are noisy speech, child speech, and far-field speech recognition, respectively. The results of previous studies suggest that the performance of T/S learning tends to be influenced by the quality of the simulated data. For example, the experimental results of T/S learning for the noisy speech domain were shown to outperform conventional training; however, in the child speech domain, only limited performance improvements were obtained.

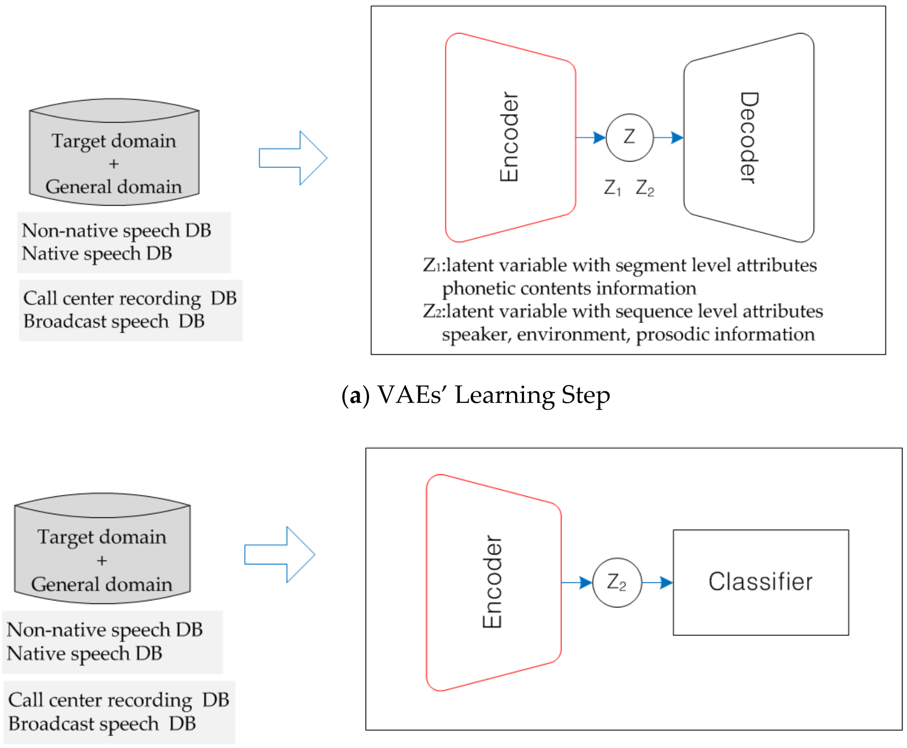
3. Proposed Methods

3.1. Active Learning Using Latent Variables with Disentangled Attributes

In this section, we describe the proposed method of actively selecting training data for a target domain using attribute-disentangled latent variables. As described in the Introduction, there are tasks that require large amounts of speech data that are difficult to collect. These include non-native speech recognition and call center recording speech recognition. The proposed active learning method is meant to effectively obtain training data close to the acoustic characteristics of these target domains from other domains, such as native speech recognition and broadcast speech recognition. In these other domains, we can obtain relatively large amounts of labeled speech data, although the overall acoustic characteristics, such as speaker and channel/noise environments, are somewhat different.

For this purpose, we designed an integrated system consisting of factorized hierarchical VAEs, [31] with an encoder that infers latent variables with disentangled attributes from the input speech, and a DNN-based classifier that selects training data with attributes matching the target domain using the preceding encoder output as input [36]. Each input speech subject to ASR contains various acoustic attributes. Among these attributes, phonetic content information changes at segmental level time scales. Additionally, environmental information such as channel/noise; speaker information such as individual speaker identity and gender/age; and prosodic information such as accent, tone, and speaking rate, change at sequence level time scales. VAEs that infer latent variables with disentangled attributes, and a DNN-based classifier are trained using a small amount of the speech data obtained from the target domain and other general domains. Here, the VAEs are trained by an unsupervised learning method using speech input/output. The classifier is trained via supervised learning by inputting attribute-disentangled latent variables that are the output of the preceding VAEs and outputting a label indicating whether the input is from the target domain or other general domains.

Figure 2 illustrates the learning stage of an integrated system consisting of the VAEs and classifier using the example training data. In the VAEs' learning step (a), VAEs, which generate latent variables with disentangled attributes, are trained through the process of inferring the attribute-disentangled latent variables in the encoder and generating the speech again in the decoder. For example, the VAEs can use paired non-native and native speech data as input/output training data. In the classifier learning step (b), the classifier is trained by supervised learning, using latent variables with sequence-level attributes of the training data with label information identifying whether it is the target domain or general domain.



(**a**) VAEs' Learning Step

(**b**) Classifier Learning Step

Figure 2. Graphical illustration of the learning stage of an integrated system consisting of the VAEs and classifier. The paired non-native and native speech data, and the paired call center recording and broadcast speech data, are presented as example training data.

Finally, the data selection step, which involves the selection of training speech data with the desired attributes matched for the target domain is shown in Figure 3.

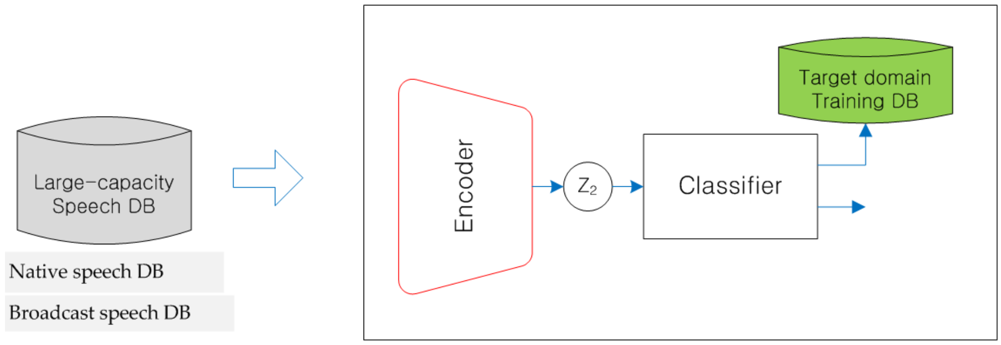


Figure 3. Graphical illustration of the data selection step.

3.2. *Teacher/Student-Based Transfer Learning Using Augmented Training Data*

In this section, we describe the proposed method, which consists of data augmentation of the target domain with sparse matched speech data and transfer learning based on teacher/student learning. For the domain with sparse matched training data, where it is difficult to collect large amounts of speech data, we propose a transfer learning method based on T/S learning. This method uses a considerable amount of labeled speech data from a general domain and augmented speech data converted from the general domain to provide the acoustic characteristics of the speaker, channel, and noise environment similar to the target domain.

The transfer learning method proposed in this paper is as follows. First, as in the VAE learning step of Figure 2a, VAEs that generate latent variables with disentangled attributes are trained using a paired target domain and general domain speech data as input/output training data. Thereafter, data augmentation is performed through the process that is illustrated in Figure 4. The speech data for training in the target domain can be augmented significantly using pre-trained encoders that generate latent variables with disentangled attributes. The encoders maintain the phonetic content attributes from the large-scale speech database (DB) and substitute other attributes, such as environmental channel/noise factors and speaking style factors from the target domain speech DB.

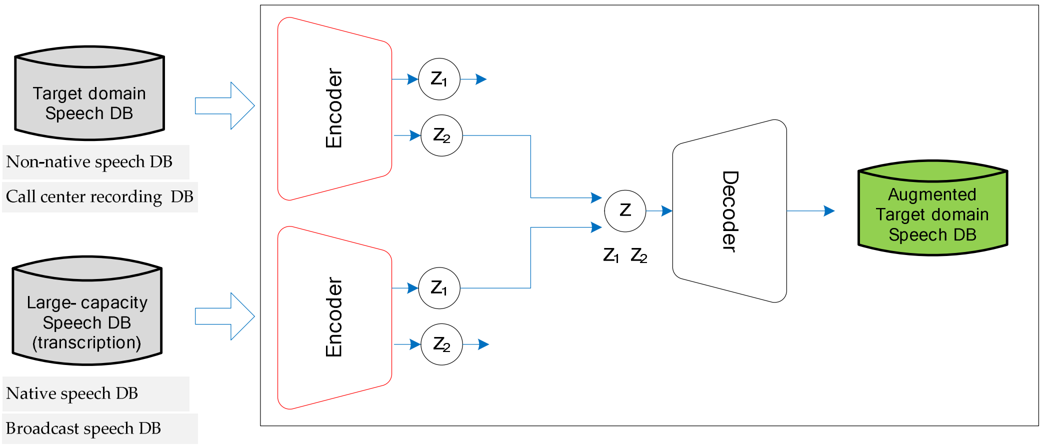


Figure 4. Graphic illustration of the data augmentation step of the proposed method.

Figure 5 illustrates the proposed teacher/student learning method, which uses training data augmented from the previous step.

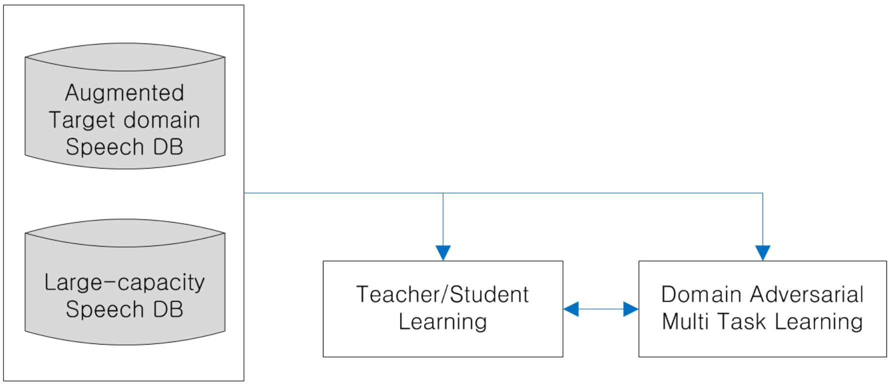

Augmented Target domain Speech DB
Large−capacity Speech DB
Teacher/Student Learning
Domain Adversarial Multi Task Learning

Figure 5. Graphic illustration of the proposed teacher/student learning method.

In Figure 5, T/S-based transfer learning is performed by inputting augmented target domain speech data and a large-scale speech DB, and sharing transcription information between them. The proposed method is a T/S-based transfer learning system with an elaborate speech recognition system trained with a large-scale speech DB as a teacher system, and a speech recognition system for the target domain as a student system. Learning is performed using the large-scale speech DB and augmented target domain speech DB as inputs, with shared labels to each teacher/student system. Simultaneously, multi-task learning, with a loss function learned in a direction insensitive to domain variations, is performed through a domain adversarial multi-task module that performs domain classification by inputting each deep feature obtained from the middle layer of the final system.

4. Experimental Settings

In this section, we describe the corpora and the detailed architecture of the end-to-end ASR system used for the experiments.

4.1. Corpus Descriptions

In order to verify the proposed methods, ASR for non-native Korean speech and Korean call center recording were used as the main task domains with sparse matched training data. The Korean broadcast speech corpus was used for the large-scale speech DB. The AMI meeting corpus [37] and certain other sources were used for the functional verification of the integrated systems. Table 1 shows the summarized characteristics of the corpora used for the experiments.

Table 1. Summarization of the corpora used for the experiments.

Corpus	Description of Purpose	Duration (h)	Number of Speakers
AMI Meeting	Functional verification of the integrated systems	100	200
Non-native Korean speech	Task domain with the sparse matched training DB	520	830
Korean call center recording	Task domain with the sparse matched training DB	1000	> 100
Korean broadcast speech	Large-scale labeled speech DB	14,000	> 1000

- AMI Meeting corpus (http://groups.inf.ed.ac.uk/ami/corpus)

The AMI meeting corpus consists of 100 h of annotated recordings of planned meetings. For each meeting, four participants have a free conversation in English, and simulate a project meeting on product design. The meetings last approximately 30 min each, and multiple microphones are used to simultaneously record conversations in different environments. Of the available microphone channels, we used the individual headset microphone (IHM) channel for clean close-talking speech and the single distant microphone (SDM) channel for far-field noisy speech.

- Non-native Korean speech corpus

We used 520 h of the in-house non-native Korean speech corpus gathered from Korean language education providers. This corpus was spoken by 830 non-native Korean speakers and recorded via PC microphone and smartphone channels. This non-native Korean speech corpus has been collected for approximately five years from Korean language education providers using speech technology for non-native speakers residing in Korea [38]. The level of Korean for each non-native speaker varies from beginner to advanced and all speech data was transcribed by humans.

- Korean call center recording corpus

The Korean call center recording corpus used in this study contains approximately 1000 h of conversations recorded at operating call centers. This information was provided with personal information deleted [39]. It is mainly composed of conversations between agents and customers. The speech recognition performance of the database is degraded due to its acoustic characteristics including the overall fast speaking rate and inaccurate pronunciation. Therefore, a considerable amount of data is required to improve the performance, but it represents a domain in which training data collection is difficult due to problems such as privacy protection.

- Korean broadcast speech corpus

The Korean broadcast audio speech corpus was used for the large-scale speech DB for the general domain as an opposite to the domain with sparse matched data. The broadcast audio data was easily accessible in large amounts. It contains speech data uttered by various speakers in diverse noise environments; therefore, it could be used to build a large-scale speech DB for improving ASR performance in the general domain. We used 14,000 h of the Korean broadcast speech corpus, with reliable transcriptions extracted from multi-genre broadcast raw audio data with inaccurate subtitle timestamps through the method proposed in [40]. In this study, all speech data was down-sampled to 8 kHz, the sampling rate of the Korean call center recording corpus.

4.2. Detailed Architecture of the End-to-End ASR

Each utterance of the training speech data was converted into 80-dimensional mel filter bank (MFB) features using the Kaldi toolkit [41]. Each frame of the feature vectors was computed with a 25ms window size and 10ms shift. One segment of the proposed method consists of five consecutive frames. For an end-to-end ASR system, we used ESPnet, an open source platform [23] to train the end-to-end model parameters in all of the experiments. The encoder network was represented by 2 blocks of VGG layer [42], followed by 5 layers with 1024 units of bidirectional long short-term memory (BLSTM). The decoder network consisted of 2 layers with 1024 units of long short-term memory (LSTM). A location-aware attention mechanism was used. Learning was performed so that cross entropy and connectionist temporal classification (CTC) loss was optimized.

5. Experimental Results

5.1. Experiments for Active Learning Using Latent Variables with Disentangled Attributes

In this experiment, we evaluated the performance of the proposed active learning method using latent variables with disentangled attributes. We implemented the designed integrated system, consisting of an encoder that infers the latent variables and a DNN-based classifier that selects training data with attributes matching the target domain, using the preceding encoder output as input.

First, for the functional verification of the integrated systems, we measured the performance of the classifier using the AMI meeting corpus. The factorized hierarchical VAEs were trained using 265k utterances from the SDM and IHM channels of the AMI meeting corpus. The classifier was trained using 10k utterances from the same channels. The classification performance of300 utterances from the SDM and IHM corpus was evaluated and the results are shown in Table 2. In general, when the utterances are longer, the classification performance is better. However, as an input vector, the sequence-level latent variable outperforms mel-frequency cepstral coefficients (MFCC) [43] and log-mel filter banks [44]. Indeed, its classification results remain robust even for short utterances.

Table 2. Classification error rate for different utterance times in seconds.

Feature	**Dimension**	**Above 1 s**	**Above 2 s**	**Above 5 s**	**Above 10 s**
MFCC	40	4.7	4.0	2.0	1.7
LogMel	40	2.7	3.0	2.7	2.2
Sequence Level Latent variable(Z_2)	32	1.0	0.3	0.0	0.0

We then evaluated the proposed active learning method for the non-native Korean speech recognition task, which is a task domain with sparse matched training data. The proposed integrated system consisted of VAEs with an encoder that infers latent variables with disentangled attributes and a classifier that actively selects training data with attributes matching the target domain. The system was trained using 520 h of both non-native Korean speech corpus and Korean broadcast speech corpus. Figure 6 shows the result obtained by visualizing the spatial distribution of sequence-level latent variables in the acoustic space through t-distributed stochastic neighbor embedding (t-SNE), which is the encoder output of trained VAEs with non-native Korean speech and native Korean speech as input.

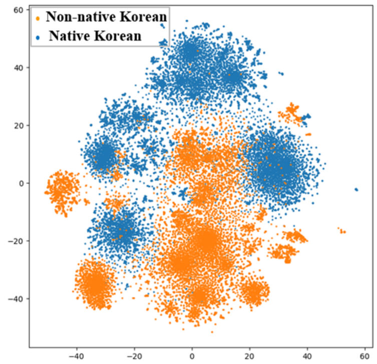


Figure 6. Spatial distribution of sequence-level attributes in the acoustic space of native/non-native Korean speech.

Table 3 shows the speech recognition performance of the models trained by the speech DB selected by the proposed active learning method. All models in Table 3 involve an end-to-end ASR system, whose detailed architecture was described in Section 4.2. We trained the baseline model using 200 h of non-native Korean speech DB, and then compared two versions of the model. One model was trained using 500 h of additional native Korean speech DB selected by the proposed active learning method, and one was trained by equal amounts of speech DB selected randomly. The evaluation set was comprised of 515 utterances recorded for assessing a Korean tutoring service for foreigners. As shown, there is a significant improvement with the proposed active learning method.

Table 3. Comparison of syllable error rate (%) on the non-native Korean speech recognition task, comparing the models trained using the speech database (DB) selected by the proposed active learning method and a random selection method.

Training DB	Error Rate
200 h non-native Korean	19.1
+500 h randomly selected native Korean	17.1
+500 h actively selected native Korean	16.3

We applied the proposed active learning method to the call center recording speech recognition task, which is another task domain with sparse matched training data. Figure 7 shows the result obtained by visualizing the spatial distribution of call center recordings, broadcast speech, selected DB, and unselected DB through t-SNE. In Figure 7, the selected and unselected DBs are both sampled from broadcast speech. Among them, the selected DB was selected by the proposed active learning method, while the unselected DB is the remaining part of the DB that was not selected. As Figure 7 shows, the selected and unselected DBs are both located in the same space as the broadcast speech. The selected DB is distributed in the space closer to the call center recording DB, and the unselected DB is distributed farther away from the call center recording DB.

Table 4 shows the speech recognition performance of the models trained by the speech DB selected by the proposed active learning method. We trained the baseline model using 500 h of Korean call center recordings, and then compared two versions. One model was trained using 1000 h of additional Korean broadcast speech selected by the proposed active learning method, and one was trained using an equal amount of the speech DB, which was selected randomly. The evaluation set was comprised of 329 utterances from one of the ASR tasks for Korean call center recording.

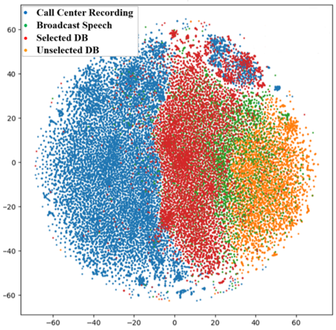


Figure 7. Spatial distribution of sequence-level attributes in the acoustic space of call center recordings, broadcast speech, unselected DB, and selected DB.

Table 4. Comparison of syllable error rate (%) on the call center recording speech recognition task, comparing the models trained using the speech DB selected by the proposed active learning method and the random selection method.

Training DB	Error Rate
500 h Korean call center recording	21.1
+1000 h randomly selected Korean broadcast speech	19.4
+1000 h actively selected Korean broadcast speech	18.5

5.2. Experiments on Teacher/Student-Based Transfer Learning Using Augmented Training Data

In this experiment, we evaluated the performance of transfer learning based on teacher/student learning and the use of augmented training data. First, we implemented the designed integrated system consisting of VAEs, which generate latent variables with disentangled attributes, and the following module, which generates the augmented target domain data using the paired target domain and general domain speech data as input data, as shown in Figure 4.

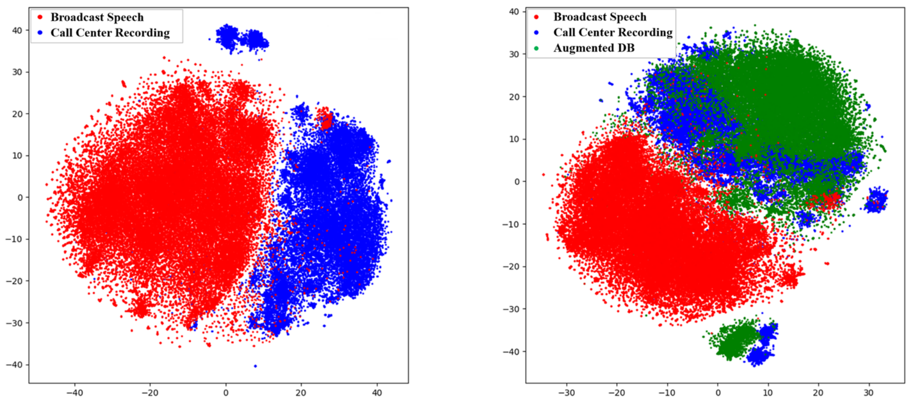
We evaluated the proposed data augmentation method (described in Figure 4) for the ASR task with the AMI meeting corpus. Table 5 shows the speech recognition performance of the proposed method. We trained three models using different training DB: IHM-only, SDM-only, and an augmented DB using the proposed method to add to SDM. For each IHM-only and SDM-only DB, we used the standard train/dev/eval data partitions of the AMI meeting corpus. The AugSDM was generated using the proposed method by maintaining the phonetic contents of IHM and substituting channel attributes from the random sample of SDM. As shown in Table 5, the model trained using only IHM suffers from a sharp degradation in performance due to the large mismatch in channel conditions between IHM and SDM. The matched SDM training and evaluation condition shows improved results, and even further performance improvement was obtained by adding the AugSDM generated by the proposed method.

Table 5. Comparison of character error rate (%) on the AMI meeting corpus task, for models trained using individual headset microphone (IHM) only, single distant microphone (SDM) only, and SDM + AugSDM for the evaluation and development set of SDM.

Training DB	Eval	Dev
IHM	55.9	51.1
SDM	35.0	32.3
SDM + AugSDM	32.5	29.9

Similar to the experimental results presented in Table 5, Manohar et al. [45] proposed a teacher-student learning approach for unsupervised domain adaptation and reported the results for domain adaptation from AMI-IHM speech to AMI-SDM speech. They used an architecture with time-delayed neural network layers interleaved with LSTM for their experiment. As we used the end-to-end ASR system architecture for our experiment, directly comparing the performance of our study with that of the previous study is difficult. However, when the relative degrees of improvement in performance over the baseline model are compared, our proposed method provides better results than that used in the previous study.

We then applied the proposed data augmentation method described in Figure 4 to the call center recording speech recognition task, a task domain with sparse matched training data. Figure 8 shows the spatial distribution of sequence-level attributes in the acoustic space for call center recordings, broadcast speech, and the augmented DB, which are visualized through t-SNE. Figure 8a shows that the call center recording and broadcast speech DBs have somewhat distinct spatial distributions. This is due to the difference in the channel/noise levels and other environmental factors. Figure 8b adds the augmented DB distribution to the space visualized in Figure 8a. The augmented DB was generated using the proposed method by maintaining the phonetic contents of broadcast speech and substituting sequence-level attributes from a random sample of the call center recording DB. As Figure 8b shows, the augmented DB distribution mainly overlaps that of the call center recording DB.



(**a**) Original distribution (**b**) Distribution with augmented DB added

Figure 8. Spatial distribution of sequence-level attributes in the acoustic space of the call center recordings, broadcast speech, and augmented DB.

Table 6 shows the performance results of the ASR task on the Korean call center recording. AugCall refers to the augmented DB generated using the proposed method. It can be seen that the model trained using additional 500 h of AugCall augmented by the proposed method showed improved results, compared to the model trained using equal amounts of additional Korean broadcast speech. The AugCall model is also comparable to the model trained using additional target domain data with the equal amount.

Table 6. Comparison of syllable error rate (%) for the call center recording speech recognition task, examining the models trained using AugCall, using additional Korean broadcast speech, and using an equal amount of call center recordings.

Training DB	Error Rate
500 h Korean call center recording	21.1
+500 h Korean broadcast speech	19.6
+500 h AugCall augmented by the proposed method	18.5
1000 h Korean call center recording	18.5

6. Conclusions

In this paper, we addressed speech recognition tasks where it is difficult to collect large amounts of labeled speech data. Domain adaptation and semi-supervised learning methods are representative approaches that are used in both academic research and application services. Domain adaptation approaches require a considerable amount of domain speech data with transcription. Semi-supervised learning approaches, on the other hand, require less costly transcription, but require a significant amount of unlabeled data from the target domain and are not effective if a pre-trained model is not matched to the target domain.

We focused on handling the speech recognition problem for task domains with sparse matched training data, and proposed an active learning method that selects training data for the target domain from another domain that already has a significant amount of labeled speech data. We also proposed a transfer learning method based on teacher/student learning combined with data augmentation. The experimental results show that the proposed method outperforms random selection and is comparable to using equal amounts of additional target domain data.

In the future, we will implement and verify the proposed transfer learning system shown in Figure 5, and integrate the augmentation method that we have verified in the experiment described in this paper.

Author Contributions: Conceptualization, B.O.K. and H.B.J.; methodology, B.O.K.; software, B.O.K.; validation, B.O.K. and H.B.J.; formal analysis, B.O.K. and H.B.J.; investigation, B.O.K.; resources, B.O.K.; data curation, B.O.K.; writing–original draft preparation, B.O.K.; writing–review and editing, B.O.K.; visualization, B.O.K.; supervision, B.O.K. and H.B.J.; project administration, J.G.P.; funding acquisition, J.G.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & Communications Technology 191 Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (2019-0-2019-0-00004, Development of 192 semi-supervised learning language intelligence technology and Korean tutoring service for foreigners).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.; Abdel-rahman, M.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
2. Dahl, G.E.; Yu, D.; Deng, L.; Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 30–42. [CrossRef]
3. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J. Deep speech 2: End-to-end speech recognition in English and mandarin. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–28 June 2016; pp. 173–182.
4. Graves, A.; Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In Proceedings of the International Conference on Machine Learning, Beijing, China, 22–24 June 2014; pp. 1764–1772.
5. Niesler, T. Language-dependent state clustering for multilingual acoustic modelling. *Speech Commun.* **2007**, *49*, 453–463. [CrossRef]
6. Huang, J.-T.; Li, J.; Yu, D.; Deng, L.; Gong, Y. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 7304–7308.
7. Kang, B.O.; Kwon, O.W. Combining multiple acoustic models in GMM spaces for robust speech recognition. *IEICE Trans. Inf. Syst.* **2016**, *99*, 724–730. [CrossRef]
8. Sun, S.; Zhang, B.; Xie, L. An unsupervised deep domain adaptation approach for robust speech recognition. *Neurocomputing* **2017**, *257*, 79–87. [CrossRef]
9. Asami, T.; Masumura, R.; Yamaguchi, Y.; Masataki, H.; Aono, Y. Domain adaptation of dnn acoustic models using knowledge distillation. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5185–5189.
10. Meng, Z.; Li, J.; Gong, Y.; Juang, B.-H. Adversarial teacher-student learning for unsupervised domain adaptation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5949–5953.

11. Meng, Z.; Li, J.; Gaur, Y.; Gong, Y. Domain adaptation via teacher-student learning for end-to-end speech recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Sentosa, Singapore, 14–18 December 2019; pp. 268–275.
12. Wang, L.; Gales, M.J.; Woodland, P.C. Unsupervised training for Mandarin broadcast news and conversation transcription. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, HI, USA, 15–20 April 2007; Volume 4, pp. 4–353.
13. Yu, K.; Gales, M.; Wang, L.; Woodland, P.C. Unsupervised training and directed manual transcription for 220 LVCSR. *Speech Commun.* **2010**, *52*, 652–663. [CrossRef]
14. Ranzato, M.; Szummer, M. Semi-supervised learning of compact document representations with deep networks. In Proceedings of the 25th International Conference on Machine learning. ACM, Helsinki, Finland, 5–9 July 2008; pp. 792–799.
15. Dhaka, A.K.; Salvi, G. Sparse autoencoder based semi-supervised learning for phone classification with limited annotations. In Proceedings of the GLU 2017 International Workshop on Grounding Language Understanding, Stockholm, Sweden, 25 August 2017; pp. 22–26.
16. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
17. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Chen, J.; Chrzanowski, M.; et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In Proceedings of the International Conference on Machine Learning, Lille, France, 11 July 2015.
18. Audhkhasi, K.; Saon, G.; Tüske, Z.; Kingsbury, B.; Picheny, M. Forget a Bit to Learn Better: Soft Forgetting for CTC-Based Automatic Speech Recognition. In Proceedings of the 2019 Interspeech, Graz, Austria, 15 September 2019; pp. 2618–2622.
19. Chorowski, J.; Bahdanau, D.; Cho, K.; Bengio, Y. End-to-end continuous speech recognition using attention-based recurrent NN. First results. *arXiv* **2014**, arXiv:1412.1602. Available online: https://arxiv.org/abs/1412.1602 (accessed on 8 August 2020).
20. Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-Based Models for Speech Recognition. In *Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 577–585.
21. Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakel, P.; Bengio, Y. End-to-end attention-based large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4945–4949.
22. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shangai, China, 20–25 March 2016; pp. 4960–4964.
23. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, E.Y.S.; Heyman, J.; Wiesner, M.; Ochiai, T.; et al. ESPnet: End-to-End Speech Processing Toolkit. *arXiv* **2018**, arXiv:1804.00015. Available online: https://arxiv.org/abs/1804.00015 (accessed on 8 August 2020).
24. Settles, B. *Active Learning Literature Survey*; Technical Report; University of Wisconsin-Madison Department of Computer Sciences: Madison, WI, USA, 2009.
25. Settles, B.; Craven, M. An Analysis of Active Learning Strategies for Sequence Labeling Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; Association for Computational Linguistics: Stroudsburg, PA, USA, 2008; pp. 1070–1079.
26. Lewis, D.D.; Catlett, J. Heterogeneous uncertainty sampling for supervised learning. In Machine Learning Proceedings, Proceedings of the Eleventh International Conference, New Brunswick, NJ, USA, July 10–13, 1994; Elsevier: Amsterdam, The Netherlands, 2017; pp. 148–156.
27. Dasgupta, S.; Hsu, D. Hierarchical sampling for active learning. In Proceedings of the 25th International Conference on Machine Learning ACM, Helsinki, Finland, 5–9 July 2008; pp. 208–215.
28. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv* **2020**, arXiv:2001.00378. Available online: https://arxiv.org/abs/2001.00378 (accessed on 8 August 2020).
29. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]

30. Tschannen, M.; Bachem, O.; Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv* **2018**, arXiv:1812.05069. Available online: https://arxiv.org/abs/1812.05069 (accessed on 8 August 2020).
31. Hsu, W.N.; Zhang, Y.; Glass, J. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017; pp. 1878–1889.
32. Meng, Z.; Chen, Z.; Mazalov, V.; Li, J.; Gong, Y. Unsupervised adaptation with domain separation networks for robust speech recognition. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017.
33. Li, J.; Seltzer, M.L.; Wang, X. Large-scale domain adaptation via teacher-student learning. In Proceedings of the 2017 Interspeech, Stockholm, Sweden, 20 August 2017.
34. Watanabe, S.; Hori, T.; Roux, J.L.; Hershey, J. Student—Teacher network learning with enhanced features. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.
35. Li, J.; Zhao, R.; Chen, Z. Developing far-field speaker system via teacher-student learning. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
36. Kang, B.O.; Park, J.G. Active Learning using Latent Variables with Independent Attributes for Speech Recognition. In Proceedings of the Winter Annual Conference of KICS, Kangwon, Korea, 5–7 February 2020.
37. Carletta, J.; Ashby, S.; Bourban, S.; Flynn, M.; Guillemot, M.; Hain, T.; Kadlec, J.; Karaiskos, V.; Kraijaj, W.; Lathoud, G.; et al. The AMI meeting corpus: A pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 28–39.
38. Oh, Y.R.; Park, K.; Jeon, H.B.; Park, J.G. Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition. *ETRI J.* **2020**. [CrossRef]
39. Kang, B.O.; Jeon, H.B.; Song, H.J.; Han, R.; Park, J.G. Long Short-Term Memory RNN based Speech Recognition System for Recording Data. In Proceedings of the Winter Annual Conference of KICS, Kangwon, Korea, 17–19 January 2018.
40. Bang, J.U.; Choi, M.Y.; Kim, S.H.; Kwon, O.W. Automatic construction of a large-scale speech recognition database using multi-genre broadcast data with inaccurate subtitle timestamps. *IEICE Trans. Inf. Syst.* **2020**, *103*, 406–415. [CrossRef]
41. Povey, D.; Ghoshal, A.; Boulianne, G.; Goel, N.; Hannemann, M.; Qian, Y.; Schwarz, P.; Stemmer, G. The Kaldi Speech Recognition Toolkit. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Hawaii, HI, USA, 11–15 December 2011.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. Available online: https://arxiv.org/abs/1409.1556 (accessed on 8 August 2020).
43. Picone, J.W. Signal modeling techniques in speech recognition. *Proc. IEEE* **1993**, *81*, 1215–1247. [CrossRef]
44. Mogran, N.; Bourlard, H.; Hermansky, H. Automatic speech recognition: An auditory perspective. In *Speech Processing in the Auditory System*; Springer: New York, NY, USA, 2004; pp. 309–338.
45. Manohar, V.; Ghahremani, P.; Povey, D.; Khudanpur, S. A teacher-student learning approach for unsupervised domain adaptation of sequence-trained ASR models. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 250–257.