Can We Simulate Generative Process of Acoustic Modeling Data? Towards Data Restoration for Acoustic Modeling

Ryo Masumura*, Yusuke Ijima*, Satoshi Kobashikawa*, Takanobu Oba*, Yushi Aono*
* NTT Media Intelligence Laboratories, NTT Corporation, Japan
E-mail: ryou.masumura.ba@hco.ntt.co.jp

***Abstract*—In this paper, we present an initial study on data restoration for acoustic modeling in automatic speech recognition (ASR). In the ASR field, the speech log data collected during practical services include customers' personal information, so the log data must often be preserved in segregated storage areas. Our motivation is to permanently and flexibly utilize the log data for acoustic modeling even though the log data cannot be moved from the segregated storage areas. Our key idea is to construct portable models that can simulate the generative process of acoustic modeling data so as to artificially restore the acoustic modeling data. Therefore, this paper proposes novel generative models called acoustic modeling data restorers (AMDRs), that can randomly sample triplets of a phonetic state sequence, an acoustic feature sequence, and utterance attribute information, even if original data is not directly accessible. In order to precisely model the generative process of the acoustic modeling data, we introduce neural language modeling to generate the phonetic state sequences and neural speech synthesis to generate the acoustic feature sequences. Experiments using Japanese speech data sets reveal how close the restored acoustic data is to the original data in terms of ASR performance.**

Index Terms: data restoration, acoustic models, automatic speech recognition, acoustic modeling data restorers, generative models

I. Introduction

Automatic speech recognition (ASR) technologies have accomplished remarkable progress in the last few years. Deep learning technologies have improved ASR performance, and ASR-based applications such as voice search have become familiar to customers. As practical ASR systems are being used in more various applications, log data of speech samples and their transcripts are rapidly accumulating. One concern is that the log data often includes customer's personal privacy information. Therefore, when ASR developers use such log data to improve ASR performance, special handling restrictions must be followed.

It is known that log data is extremely beneficial in constructing acoustic models because it is recorded in real environments [1], [2]. However, the log data often must be split and preserved in secured storage areas that are segregated from the online network for privacy protection. A key issue is that it is impossible to use certain company's log data sets together with other company's log data because the log data sets are often preserved in different company's secure repositories. In other words, we cannot jointly leverage various company's log data sets for improving ASR performance.

Our motivation is to flexibly utilize the log data for constructing acoustic models without accessing the original data in the secured sets. In speech processing fields, although several techniques have been examined for privacy protection, they are not suitable for use in acoustic modeling because they change the original contents [3]–[5]. Our key idea is to construct a portable model that can artificially generate acoustic modeling data similar to the original data. This enables us to train acoustic models permanently even if the original data is deleted. One concern is whether we can precisely simulate the generative process of acoustic modeling data or not because the quality of data generation directly affects the ASR performance of the acoustic models.

In this paper, we present an initial study of acoustic modeling data restoration based on generative modeling of acoustic modeling data. To this end, we propose novel generative models of acoustic modeling data, which we call acoustic modeling data restorers (AMDRs). AMDRs can simulate the generative process of the acoustic modeling data that can be used for constructing context-dependent phone-based deep neural network (DNN) acoustic models [6]–[8]. The AMDRs trained from acoustic modeling data sets can randomly generate triplets of a phonetic state sequence, an acoustic feature sequence, and utterance attribute information without using the original data sets.

In order to construct precise generative models, the AMDRs adopt neural network based modeling. We introduce long short-term memory recurrent neural network (LSTM-RNN) based language models that can take long-range context into consideration to model the generative process of phonetic state sequences [9], [10]. In addition, we introduce bidirectional LSTM-RNN (BLSTM-RNN) based neural speech synthesizers to model the generative process of acoustic feature sequences [11]–[13]. In fact, general neural speech synthesizers are composed by regression networks that produce acoustic features of speech on the basis of a minimum mean error criterion because speech synthesis applications aim to generate high-quality speech [12], [14]. Our aim, on the other hand, is to produce acoustic features that are effective for acoustic modeling in ASR. To generate the acoustic features needed to simulate real environments, the features must exhibit sufficient variation. Therefore, this paper examines not only regression

networks but also density networks [15] as they can randomly produce a variety of acoustic features.

Data restoration is closely related to data augmentation of acoustic modeling data [16]–[22]. Data augmentation is often conducted for low resource ASR fields. The main approach is to manipulate the original data using vocal tract length perturbation or elastic spectral distortion. In addition, model based transformation of the original data has been examined [23], [24]. These methods involve label-preserving transformation of the original data, so the original data must be accessed. To the best of our knowledge, this paper is the first study on acoustic modeling data generation in which the original data is not needed at all.

Our main contributions are summarized as follows.

- This paper is first work to describe data restoration for acoustic modeling. In section II, we define data restoration that is accomplished by constructing generative models of acoustic modeling data. In addition, we elucidate its goal.
- This paper proposes novel generative models of acoustic modeling data called AMDRs. AMDRs achieve acoustic modeling data generation simply by preserving model parameters. Section III details the modeling proposal and its training procedure.
- This paper is an initial study that leverages neural speech synthesis technologies for improving acoustic models although the speech synthesis technologies recently utilized for end-to-end ASR [25]–[28]. This paper introduces regression network based and density network based neural speech synthesis (see section III-D).
- Our experiments construct AMDRs from Japanese acoustic modeling data, and restore acoustic modeling data using trained AMDRs. For acoustic modeling, we introduce context-dependent phone-based convolutional long short-term memory fully-connected DNN (CLDNN) acoustic models [29], [30]. We reveal whether acoustic models constructed from restored data using AMDRs can yield ASR performance close to that possible with the original data. In addition, we demonstrate that AMDRs can improve ASR performance by leveraging them for data augmentation (see section IV).

II. Data Restoration for Acoustic Modeling

This section details data restoration for acoustic modeling in ASR. In acoustic modeling, data restoration is necessary to permanently and flexibly utilize acoustic modeling data for constructing acoustic models even if the original acoustic modeling data is not accessible.

Fig. 1 shows the relationships between data restoration and acoustic modeling. In order to achieve data restoration, a generative model that can artificially restore acoustic modeling data similar to the original acoustic modeling data is constructed from original acoustic modeling data. The generative model enables acoustic modeling data restoration by random sampling on the basis of its generative process. Our goal is to construct an acoustic model from the restored data that yields ASR performance comparable to that constructed from the original data.

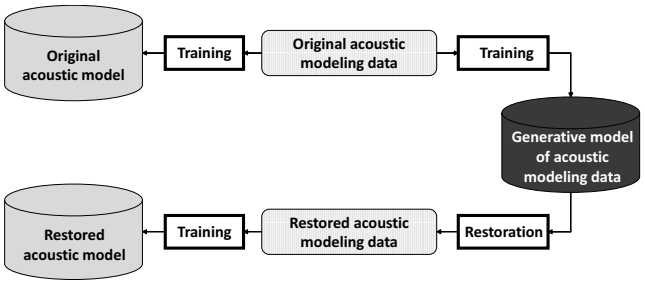

Fig. 1. Relationship between data restoration and acoustic modeling.

III. Generative Models of Acoustic Modeling Data

This section describes novel generative models of acoustic modeling data called acoustic modeling data restorers (AMDRs). This paper assumes ASR systems that use context-dependent phone-based DNN acoustic models [6]–[8]. The context-dependent phone-based DNNs that estimate phonetic states from acoustic features in a frame-by-frame manner are the most practical acoustic modeling approach used in recent ASR services.

We first define acoustic modeling data that is represented as sets of utterance-level data. The utterance-level acoustic modeling data is defined as a triplet of a phonetic state sequence, an acoustic feature sequence, and an attribute label. We define a phonetic state sequence and its corresponding acoustic feature sequence in an utterance as $S = \{s_1, \cdots, s_T\}$ and $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\}$, respectively. $s_t \in \mathcal{S}$ is the t-th frame's phonetic state index where $\mathcal{S}$ is a set of the phonetic states, and $\boldsymbol{x}_t$ is the t-th frame's acoustic feature. In addition, an utterance attribute label, which includes speech attributes such as speaker information and linguistic attributes such as topic information, is defined as $a \in \{\bar{a}_1, \cdots, \bar{a}_M\}$. The attribute label is needed to restore the acoustic modeling data but is unnecessary for acoustic modeling. In this case, an acoustic modeling data set $\mathcal{D}$ is defined as:

$$\mathcal{D} = \{(S^1, \boldsymbol{X}^1, a^1), \cdots, (S^N, \boldsymbol{X}^N, a^N)\}, \tag{1}$$

where $(S^n, \boldsymbol{X}^n, a^n)$ denotes the n-th utterance-level acoustic modeling data and N represents the number of utterances in the acoustic modeling data set.

A. Generative Process of Acoustic Modeling Data

In AMDR, generative process of acoustic modeling data is simulated as shown in Fig. 2. AMDRs assume that $\mathcal{D}$ is generated according to the following generative process.

1) For $n = 1, \cdots, N$:
   a) Sample an attribute label:

$$a^n \sim P(a^n|\boldsymbol{\theta}_{\mathtt{a}}), \tag{2}$$

   b) For $t = 1 \cdots, T^n$:

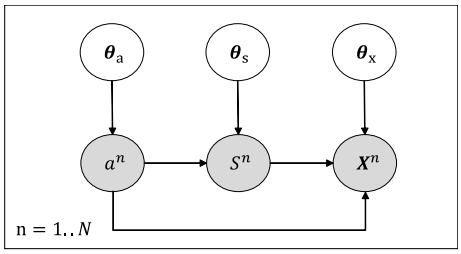

Fig. 2. Graphical models of AMDRs.

i) Sample a phonetic state:

$$s_t^n \sim P(s_t^n | s_1^n, \cdots, s_{t-1}^n, a^n, \boldsymbol{\theta}_{\mathtt{s}}), \tag{3}$$

c) For $t = 1, \cdots, T^n$:

i) Sample an acoustic feature:

$$\boldsymbol{x}_t^n \sim P(x_t^n | S^n, a^n, t, \boldsymbol{\theta}_{\mathtt{x}}), \tag{4}$$

where T^n is the length of the n-th utterance-level acoustic feature sequence and phonetic state sequence, and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{\mathtt{a}}, \boldsymbol{\theta}_{\mathtt{s}}, \boldsymbol{\theta}_{\mathtt{x}}\}$ is a model parameter of the AMDR. Thus, the phonetic state sequence is generated dependent on the attribute label and the acoustic feature sequence is generated dependent on the phonetic state sequence and the attribute label. In this case, the generative probability of $\mathcal{D}$ is formulated as:

$$\begin{aligned} P(\mathcal{D}|\boldsymbol{\Theta}) &= \prod_{n=1}^{N} P(\boldsymbol{X}^n, S^n, a^n | \boldsymbol{\Theta}) \\ &= \prod_{n=1}^{N} P(\boldsymbol{X}^n | S^n, a^n, \boldsymbol{\theta}_{\mathtt{x}}) P(S^n | \boldsymbol{\theta}_{\mathtt{s}}) P(a^n | \boldsymbol{\theta}_{\mathtt{a}}) \\ &= \prod_{n=1}^{N} P(a^n | \boldsymbol{\theta}_{\mathtt{a}}) \{ \prod_{t=1}^{T^n} P(\boldsymbol{x}_t^n | S^n, a^n, t, \boldsymbol{\theta}_{\mathtt{x}}) \} \\ &\qquad \{ \prod_{t=1}^{T^n} P(s_t^n | s_1^n, \cdots, s_{t-1}^n, a^n, \boldsymbol{\theta}_{\mathtt{s}}) \}. \end{aligned} \tag{5}$$

The AMDRs enable us to randomly produce acoustic modeling data simply by preserving $\boldsymbol{\Theta}$. The model parameter can be optimized from $\mathcal{D}$ using maximum likelihood estimation. The optimized parameter $\hat{\boldsymbol{\Theta}}$ is estimated by:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} P(\mathcal{D}|\boldsymbol{\Theta}). \tag{6}$$

In fact, the AMDRs have no latent variables and each component parameter is independent, so the maximum likelihood estimation can be split into:

$$\hat{\boldsymbol{\theta}}_{\mathtt{a}} = \underset{\boldsymbol{\theta}_{\mathtt{a}}}{\operatorname{argmax}} \prod_{n=1}^{N} P(a^n | \boldsymbol{\theta}_{\mathtt{a}}), \tag{7}$$

$$\hat{\boldsymbol{\theta}}_{\mathtt{s}} = \underset{\boldsymbol{\theta}_{\mathtt{s}}}{\operatorname{argmax}} \prod_{n=1}^{N} \prod_{t=1}^{T^n} P(s_t^n | s_1^n, \cdots, s_{t-1}^n, a^n, \boldsymbol{\theta}_{\mathtt{s}}), \tag{8}$$

$$\hat{\boldsymbol{\theta}}_{\mathtt{x}} = \underset{\boldsymbol{\theta}_{\mathtt{x}}}{\operatorname{argmax}} \prod_{n=1}^{N} \prod_{t=1}^{T^n} P(\boldsymbol{x}_t^n | S^n, a^n, t, \boldsymbol{\theta}_{\mathtt{x}}), \tag{9}$$

where $\hat{\boldsymbol{\Theta}} = \{\hat{\boldsymbol{\theta}}_{\mathtt{a}}, \hat{\boldsymbol{\theta}}_{\mathtt{s}}, \hat{\boldsymbol{\theta}}_{\mathtt{x}}\}$. Each optimization depends on individual model structures used to compute generative probabilities.

B. Generative Models of Attribute Labels

We model the generative probability of attribute labels using a categorical distribution. The generative probability of a is simply formulated as:

$$P(a|\boldsymbol{\theta}_{\mathtt{a}}) = \mathrm{Categorical}(a; \boldsymbol{\theta}_{\mathtt{a}}), \tag{10}$$

where $\mathrm{Categorical}()$ represents the categorical distribution that is a discrete probability distribution. Therefore, the attribute label can be randomly sampled from the categorical distribution. In this case, $\boldsymbol{\theta}_{\mathtt{a}}$ means $\{q_1, \cdots, q_M\}$ where q_m is the generative probability of the m-th label. The maximum likelihood estimation of q_m is defined as:

$$\hat{q}_m = \frac{c(\bar{a}_m, \mathcal{D})}{N}, \tag{11}$$

where $c(\bar{a}_m, \mathcal{D})$ means the frequency of $\bar{a}_m$ in $\mathcal{D}$.

C. Generative Models of Phonetic State Sequences

We model the generative probability of phonetic state sequences using the categorical distribution with neural language models. The neural language models can produce predicted probabilities $\boldsymbol{o}_t$ from $\{s_1, \cdots, s_{t-1}\}$ and a. The predicted probabilities correspond to the parameter of the categorical distribution. The generative probability of s_t is formulated as:

$$\begin{aligned} &P(s_t | s_1, \cdots, s_{t-1}, a, \boldsymbol{\theta}_{\mathtt{s}}) \\ &\quad = \mathrm{Categorical}(s_t; \boldsymbol{\Phi}(s_1, \cdots, s_{t-1}, a; \boldsymbol{\theta}_{\mathtt{s}})), \\ &\quad = \mathrm{Categorical}(s_t; \boldsymbol{o}_t), \end{aligned} \tag{12}$$

where $\boldsymbol{\Phi}()$ is the function of the neural language models. Therefore, the phonetic state sequences can be randomly sampled from the categorical distribution in which the parameter is incrementally updated by the neural language models. Fig. 3 shows the data generation procedure of phonetic state sequences.

In order to produce $\boldsymbol{o}_t$ using the neural language models, individual phonetic state indices and an attribute label are converted into continuous representations. The continuous representations of s_{t-1} and a are defined as:

$$\boldsymbol{s}_{t-1} = \mathtt{EMBED}(s_{t-1}; \boldsymbol{\lambda}_{\mathtt{s}}), \tag{13}$$

$$\boldsymbol{a} = \mathtt{EMBED}(a; \boldsymbol{\lambda}_{\mathtt{a}}), \tag{14}$$

where $\mathtt{EMBED}()$ is a linear transformational function to embed a symbol into a continuous vector. $\boldsymbol{\lambda}_{\mathtt{s}}$ and $\boldsymbol{\lambda}_{\mathtt{a}}$ are the trainable parameters. Next, the continuous representation of phonetic state and the continuous representation of attribute label are merged as:

$$\boldsymbol{e}_{t-1} = [\boldsymbol{s}_{t-1}{}^{\top}, \boldsymbol{a}^{\top}]^{\top}. \tag{15}$$

The merged representation is converted into a hidden representation that uses LSTM-RNN to summarize past context

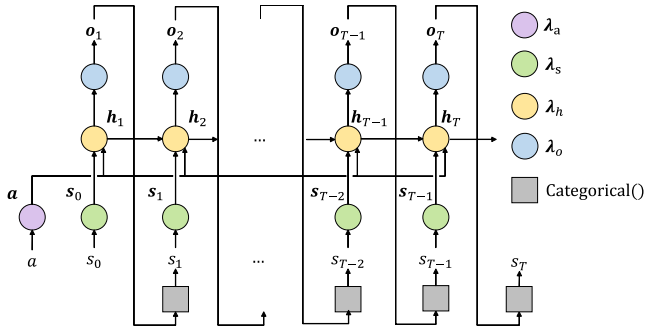

Fig. 3. Data generation procedure of phonetic state sequences.

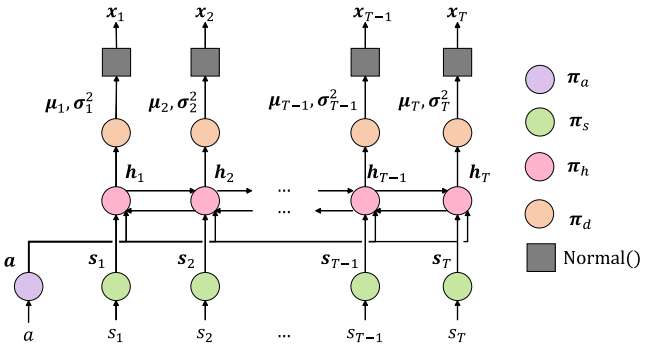

Fig. 4. Data generation procedure of acoustic feature sequences.

information. The hidden representation for estimating the t-th phonetic state is calculated as:

$$\begin{aligned} \boldsymbol{h}_t &= \texttt{LSTM}(\boldsymbol{e}_1, \cdots, \boldsymbol{e}_{t-1}; \boldsymbol{\lambda}_\texttt{h}) \\ &= \texttt{LSTM}(\boldsymbol{e}_{t-1}, \boldsymbol{h}_{t-1}; \boldsymbol{\lambda}_\texttt{h}), \end{aligned} \tag{16}$$

where $\texttt{LSTM}()$ is a function of the unidirectional LSTM-RNN layer, $\boldsymbol{\lambda}_\texttt{h}$ is the trainable parameter, $\boldsymbol{s}_0$ is a zero vector, and s_0 means an initial symbol. In the output layer, predicted probabilities are produced by:

$$\boldsymbol{o}_t = \texttt{SOFTMAX}(\boldsymbol{h}_t; \boldsymbol{\lambda}_\texttt{o}), \tag{17}$$

where $\texttt{SOFTMAX}()$ is a transformational function with softmax activation and $\boldsymbol{\lambda}_\texttt{o}$ is the trainable parameter. In this process, $\boldsymbol{\theta}_\texttt{s}$ corresponds to $\{\boldsymbol{\lambda}_\texttt{s}, \boldsymbol{\lambda}_\texttt{a}, \boldsymbol{\lambda}_\texttt{h}, \boldsymbol{\lambda}_\texttt{o}\}$. The maximum likelihood estimation of $\boldsymbol{\theta}_\texttt{s}$ is redefined as:

$$\hat{\boldsymbol{\theta}}_\texttt{s} = \underset{\theta_\texttt{s}}{\operatorname{argmin}} - \sum_{n=1}^{N} \sum_{t=1}^{T^n} \log \operatorname{Categorical}(s_t^n; \boldsymbol{o}_t^n), \tag{18}$$

where s_t^n and $\boldsymbol{o}_t^n$ are the t-th phonetic state index and the t-th predicted probabilities for the n-th acoustic modeling data, respectively. This optimization is followed by the back-propagation through time algorithm.

D. Generative Models of Acoustic Feature Sequences

We use neural speech synthesizers to model the generative probability of acoustic feature sequences. To this end, we introduce regression networks and density networks. In AMDRs, acoustic feature sequences are produced depending on S and a. In regression network based AMDRs, the generative probability of $\boldsymbol{x}_t$ is defined as:

$$\begin{aligned} P(\boldsymbol{x}_t|S, a, t, \boldsymbol{\theta}_\texttt{x}) &= \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\psi}_\texttt{r}(S, a, t; \boldsymbol{\theta}_\texttt{x}), \beta \boldsymbol{I}) \\ &= \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_t, \beta \boldsymbol{I}), \end{aligned} \tag{19}$$

where $\mathcal{N}()$ represents a normal distribution, $\boldsymbol{\psi}_\texttt{r}()$ is the regression network function, an$\boldsymbol{\mu}_t$ is the output of the regression networks and corresponds to a mean vector of the normal distribution. $\boldsymbol{I}$ is the diagonal identity matrix, and β is a hyper parameter. By limiting β to 0, $\boldsymbol{x}_t$ sampled from the normal distribution exactly matches $\boldsymbol{\mu}_t$.

On the other hand, in density network based AMDRs, mean vector $\boldsymbol{\mu}_t$ and diagonal variance vector $\boldsymbol{\sigma}_t^2$ for the normal distribution are produced using S and a. The generative probability of $\boldsymbol{x}_t$ is defined as:

$$\begin{aligned} &P(\boldsymbol{x}_t|S, a, t, \boldsymbol{\theta}_\texttt{x}) \\ &= \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\psi}_\texttt{d}^{(\boldsymbol{\mu})}(S, a, t; \boldsymbol{\theta}_\texttt{x}), \exp(\boldsymbol{\psi}_\texttt{d}^{(\boldsymbol{\sigma})}(S, a, t; \boldsymbol{\theta}_\texttt{x}))^2) \\ &= \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{z}_t^{(\boldsymbol{\mu})}, \exp(\boldsymbol{z}_t^{(\boldsymbol{\sigma})})^2) \\ &= \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2), \end{aligned} \tag{20}$$

where $\boldsymbol{\psi}_\texttt{d}^{(\boldsymbol{\mu})}()$ and $\boldsymbol{\psi}_\texttt{d}^{(\boldsymbol{\sigma})}()$ are density network functions, and $\boldsymbol{z}_t^{(\boldsymbol{\mu})}$ and $\boldsymbol{z}_t^{(\boldsymbol{\sigma})}$ are outputs of the density network. Therefore, the acoustic feature sequences can be randomly sampled from the normal distribution with the produced vectors. Fig. 4 shows the data generation procedure of acoustic feature sequences on the basis of the density networks.

In both networks, individual phonetic states and attribute label are converted into continuous representations, and merged representations are composed as in the LSTM-RNN language models. The merged representation of s_{t-1} and a is defined as:

$$\boldsymbol{s}_{t-1} = \texttt{EMBED}(s_{t-1}; \boldsymbol{\pi}_\texttt{s}), \tag{21}$$

$$\boldsymbol{a} = \texttt{EMBED}(a; \boldsymbol{\pi}_\texttt{a}), \tag{22}$$

$$\boldsymbol{e}_{t-1} = [\boldsymbol{s}_{t-1}{}^\top, \boldsymbol{a}^\top]^\top, \tag{23}$$

where $\boldsymbol{\pi}_\texttt{s}$ and $\boldsymbol{\pi}_\texttt{a}$ are the trainable parameters. The merged representation is converted into a hidden representation using BLSTM-RNNs. The hidden representation for estimating the t-th acoustic feature is calculated as:

$$\boldsymbol{v}_t = \texttt{BLSTM}(\boldsymbol{e}_1, \cdots, \boldsymbol{e}_T, t; \boldsymbol{\pi}_\texttt{v}), \tag{24}$$

where $\texttt{BLSTM}()$ is a function of the BLSTM-RNN layer and $\boldsymbol{\pi}_\texttt{v}$ is the trainable parameter.

In the output layer of the regression network, $\boldsymbol{\mu}_t$ is produced by:

$$\boldsymbol{\mu}_t = \texttt{LINEAR}(\boldsymbol{v}_t; \boldsymbol{\pi}_\texttt{r}), \tag{25}$$

where $\texttt{LINEAR}()$ is the linear transformational function and $\boldsymbol{\pi}_\texttt{r}$ is the trainable parameter. In this case, $\boldsymbol{\theta}_\texttt{x}$ corresponds to

$\{\boldsymbol{\pi}_{\mathtt{s}}, \boldsymbol{\pi}_{\mathtt{a}}, \boldsymbol{\pi}_{\mathtt{v}}, \boldsymbol{\pi}_{\mathtt{r}}\}$. In the regression networks, the maximum likelihood estimation of $\boldsymbol{\theta}_{\mathtt{x}}$ is redefined as:

$$\hat{\boldsymbol{\theta}}_{\mathtt{x}} = \underset{\boldsymbol{\theta}_{\mathtt{x}}}{\operatorname{argmin}} \sum_{n=1}^{N} \sum_{t=1}^{T^n} (\boldsymbol{x}_t^n - \boldsymbol{\mu}_t^n)^\top (\boldsymbol{x}_t^n - \boldsymbol{\mu}_t^n), \quad (26)$$

where $\boldsymbol{x}_t^n$ and $\boldsymbol{\mu}_t^n$ are, respectively, the t-th acoustic feature and the t-th predicted acoustic feature.

On the other hand, in the output layer of the density network, $\boldsymbol{z}_t^{(\boldsymbol{\mu})}$ and $\boldsymbol{z}_t^{(\boldsymbol{\sigma})}$ are produced by:

$$[\boldsymbol{z}_t^{(\mu)\top}, \boldsymbol{z}_t^{(\sigma)\top}]^\top = \mathtt{LINEAR}(\boldsymbol{v}_t; \boldsymbol{\pi}_{\mathtt{d}}), \quad (27)$$

where $\boldsymbol{\pi}_{\mathtt{d}}$ is the trainable parameter. In this case, $\boldsymbol{\theta}_{\mathtt{x}}$ corresponds to $\{\boldsymbol{\pi}_{\mathtt{s}}, \boldsymbol{\pi}_{\mathtt{a}}, \boldsymbol{\pi}_{\mathtt{v}}, \boldsymbol{\pi}_{\mathtt{d}}\}$. In the density networks, the maximum likelihood estimation of $\boldsymbol{\theta}_{\mathtt{x}}$ is redefined as:

$$\hat{\boldsymbol{\theta}}_{\mathtt{x}} = \underset{\boldsymbol{\theta}_{\mathtt{x}}}{\operatorname{argmin}} - \sum_{n=1}^{N} \sum_{t=1}^{T^n} \log \mathcal{N}(\boldsymbol{x}_t^n; \boldsymbol{\mu}_t^n, \boldsymbol{\sigma}_t^{n2}), \quad (28)$$

where $\boldsymbol{x}_t^n$, $\boldsymbol{\mu}_t^n$, and $\boldsymbol{\sigma}_t^{n2}$ are, respectively, the t-th acoustic feature, the t-th predicted mean vector, and the t-th predicted variance vector for the n-th utterance.

IV. Experiments

A. Setups

Our experiments used the Corpus of Spontaneous Japanese (CSJ) [31], which we divided into a training data set (Train) and three evaluation data sets (Eval 1–3). Train was used for training AMDRs and context-dependent phone based DNN acoustic models for ASR while the Eval 1–3 were used in ASR evaluations. We used lecture IDs, which include topic and gender information, as the attribute labels. Details of the data sets are shown in Table I.

In the work reported in this paper, context-dependent phone based CLDNN acoustic models were used. The acoustic feature consisted of 40 dimensional log mel-filterbank coefficients appended with delta and acceleration coefficients; the frame shift was 10 ms. In CLDNN, each static and dynamic component was spliced within 11 frames composed as 3 feature maps. We used 1 convolutional layer with 128 feature maps that used 5x11 frequency-time filters. For pooling, 2x1 frequency-time max pooling was performed. In addition, CNN output was fed into 2 LSTM layers, each of which had 512 cells. LSTM output was fed into a fully connected DNN layer, which had 1,024 hidden units with rectified linear units. The output of CLDNN was a softmax layer with 3072 nodes that correspond to the number of phonetic states. For ASR evaluations, we used a WFST-based decoder [32], [33] and 3-gram approximated hierarchical latent word language models constructed from the training data set [34].

Configurations of the AMDRs are as follows. In generative models of phonetic states, we set phonetic state embedding size and attribute embedding size to 650 and 128, respectively. We employed a single LSTM-RNN layer with 650 units. The LSTM-RNN language models we employed has 3072 output nodes, which corresponds to the number of phonetic states in acoustic models. Optimized was performed by the mini-batch stochastic gradient method. In both regression network based and density network based speech synthesizers, phonetic state embedding size and attribute embedding size were set to 512 and 128, respectively. Both networks employed three LSTM layers with 1024 units. The regression network had 120 output nodes, which corresponds to the dimension of the acoustic features. On the other hand, the density network had 240 output nodes. An Adam optimizer was used for both the regression networks and the density networks.

TABLE I
Experimental data sets.

	# of lectures	# of words	Hours
Train	3,214	7,532,365	506.0
Eval 1	10	26,028	2.3
Eval 2	10	26,661	2.4
Eval 3	10	17,189	1.7

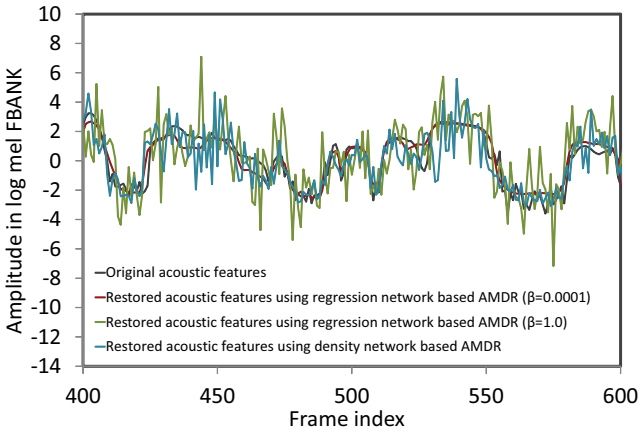

Fig. 5. Acoustic features of original data and restored data.

B. Results

We first analyzed acoustic feature generation methods because they directly impact the quality of acoustic modeling data. We generated acoustic feature sequences using original attribute labels and original phonetic state sequences in training sets.

Fig. 5 shows examples of original acoustic features and restored acoustic features. We rendered 10-th dimension values of the static log mel-filterbank coefficients. In addition, Table II shows root mean squared error (RMSE) between original acoustic features and restored acoustic features in training sets. These results show that acoustic features restored using a regression network based AMDR with $\beta = 0.0001$ were closer to original acoustic features than those yielded by a density network based AMDR with $\beta = 1.0$. We can see that restored acoustic features from the regression network based AMDR with $\beta = 0.0001$ were over-smoothed compared with original data. On the other hand, restored acoustic features from the regression network based AMDR with $\beta = 1.0$ or density network based AMDR exhibited random behavior, i.e.

TABLE III
WER (%) RESULTS OF ACOUSTIC MODELS CONSTRUCTED FROM ORIGINAL DATA AND RESTORED DATA.

		Attributes	Phonetic states	Acoustic features	Hours	Eval 1	Eval 2	Eval 3
(1).	Original data	Original	Original	Original	506.0	15.44	11.84	13.57
(2).	Restored data	AMDR	AMDR	Regression network based AMDR ($\beta = 0.0001$)	50.0	45.56	42.66	45.09
(3).	Restored data	AMDR	AMDR	Regression network based AMDR ($\beta = 1.0$)	50.0	46.44	46.25	46.62
(4).	Restored data	AMDR	AMDR	Density network based AMDR	50.0	46.57	43.32	46.58
(5).	Restored data	AMDR	AMDR	Regression network based AMDR ($\beta = 0.0001$)	500.0	46.14	42.44	45.24
(6).	Restored data	AMDR	AMDR	Regression network based AMDR ($\beta = 1.0$)	500.0	43.24	38.89	42.85
(7).	Restored data	AMDR	AMDR	Density network based AMDR	500.0	42.17	37.10	42.05
(8).	Restored data	AMDR	AMDR	Regression network based AMDR ($\beta = 0.0001$)	5,000.0	45.69	42.37	44.88
(9).	Restored data	AMDR	AMDR	Regression network based AMDR ($\beta = 1.0$)	5,000.0	42.61	38.08	42.24
(10).	Restored data	AMDR	AMDR	Density network based AMDR	5,000.0	**41.28**	**36.28**	**41.40**
(11).	Restored data	Original	Original	Regression network based AMDR ($\beta = 0.0001$)	506.0	35.06	33.49	36.08
(12).	Restored data	Original	Original	Regression network based AMDR ($\beta = 1.0$)	506.0	33.88	30.84	34.79
(13).	Restored data	Original	Original	Density network based AMDR	506.0	32.63	29.06	33.39

TABLE II
RMSE BETWEEN ORIGINAL DATA AND RESTORED DATA.

Acoustic feature generation methods	RMSE
Original acoustic features	0.00
Regression network based AMDR ($\beta = 0.0001$)	0.48
Regression network based AMDR ($\beta = 1.0$)	1.05
Density network based AMDR	0.73

TABLE IV
WER (%) RESULTS OF ACOUSTIC MODELS CONSTRUCTED FROM ORIGINAL DATA AND RESTORED DATA, AND BOTH COMBINATIONS.

	Eval 1	Eval 2	Eval 3
Original data	15.44	11.84	13.57
Restored data	41.28	36.28	41.40
Original and restored data	**15.16**	**11.54**	**13.32**

quite different from the characteristics of the original data.

Next, we conducted an ASR evaluation to assess data restoration methods. We constructed CLDNN acoustic models using original data and various restored data and evaluated them in each test set. Table III shows word error rate (WER) results of each test set. Line 1 shows oracle ASR results where original data was used. Lines 2-10 shows fully data restoration based ASR results when varying the generated data size. Lines 11-13 shows ASR results where acoustic features were generated using only original attribute labels and original phonetic state sequences. The ASR results show that that the restored data did not match original data in terms of ASR performance. Among lines 2-10, line 10 which generated 5,000 hours of restored data using density network based AMDR yielded the highest ASR performance, about 70 % of the ASR performance achieved by the original data. Additionally, regression network based AMDR with $\beta = 1.0$ and density network based AMDR were more effective for acoustic modeling than the regression network based AMDRs with $\beta = 0.0001$. In particular, ASR performance was not improved as the generated data size increased when using the regression network based AMDR with $\beta = 0.001$. On the other hand, restored data from the density network based AMDR yielded ASR performance improvements as the restored data size increased. These results confirmed that over-smoothed acoustic features do not suit acoustic modeling for ASR. In other words, a closeness to original data in terms of an amplitude level is not completely related to effectiveness for ASR performance. Lines 11-13 show that restored acoustic features from neural speech synthesizers were insufficient for accurately constructing acoustic models even if original attribute labels and original phonetic state sequences were introduced for data restoration. This suggests that it is essential to improve acoustic feature generation methods for further improving ASR performance.

We also evaluated CLDNN acoustic models constructed from both the original data and the restored data. Table IV summarizes WER results of each test set where restored data was 5,000 hours data generated from density network based AMDR. The results show that ASR performance yielded by the both the oridinal data and the restored data outperformed that only using the original data. This indicates that the data restoration was useful for data augmentation of acoustic modeling.

V. CONCLUSIONS

In this paper, we presented the first study on data restoration for acoustic modeling. For data restoration, this paper proposed AMDRs that can model the generative process of acoustic modeling data. By constructing AMDRs from the original acoustic modeling data, we can randomly sample artificial acoustic modeling data even if the original data is unavailable. Experiments confirmed that our data restoration method can yield up to 70 % of ideal ASR performance, that achieved by acoustic models constructed from the full original data. In addition, we revealed that a closeness to original data in terms of an amplitude level is not completely related to effectiveness for ASR performance. Furthermore, we demonstrated the restored data could improve ASR performance by leveraging it for data augmentation. We confirm that it is essential to develop improved acoustic feature generation methods to achieve further improvements in ASR performance. In future work, we will introduce generative models of short-time Fourier transform spectrogram instead

of directly generating acoustic features which should improve acoustic feature generation.

REFERENCES

[1] O. Kapralova, J. Alex, E. Weinstein, P. Moreno, and O. Siohan, "A big data approach to acoustic model training corpus selection," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2083–2087, 2014.

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

[3] K. Yamamoto, T. Masatoshi, and S. Nakagawa, "Privacy protection for speech signals," *Procedia Social and Behavioral Sciences*, vol. 2, pp. 153–160, 2010.

[4] S. H. K. Parthasarathi, M. Magimai.-Doss, H. Bourlard, and D. Gatica-Perez, "Evaluating the robustness of privacy sensitive audio features for speech detection in personal audio log scenarios," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4474–4477, 2010.

[5] S. H. K. Parthasarathi, H. Bourlard, and D. Gatica-Perez, "Wordless sounds: robust speaker diarization using privacy preserving audio representations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 85–98, 2013.

[6] D. Yu, L. Deng, and G. E. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," *In Proc. NIPS workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[7] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 437–440, 2011.

[8] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[9] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1045–1048, 2010.

[10] M. Sundermeyer, H. Ney, and R. Schluter, "From feedforward to recurrent LSTM neural networks for language models," *IEEE/ACM Transactions of Audio, Speech and Language processing*, vol. 23, no. 3, pp. 517–529, 2015.

[11] Y. Fan, Y. Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1964–1968, 2014.

[12] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 35, pp. 35–52, 2015.

[13] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent nenural network with recurrent output layer for low-latency speech synthesis," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4470–4474, 2015.

[14] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 8012–8016, 2013.

[15] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3872–3876, 2014.

[16] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," *In Proc. ICML Workshop on Deep Learning for Audio, Speech and. Language*, 2013.

[17] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," *In Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 309–314, 2013.

[18] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 810–814, 2014.

[19] Z. Tuske, P. Golik, D. Nolden, and H. Ney, "Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1420–1424, 2014.

[20] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[21] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3586–3589, 2015.

[22] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[23] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," *In Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 16–23, 2017.

[24] H. Nishizaki, "Data augmentation and feature extraction using variational autoencoder for acoustic modeling," *In Proc. Asia-Pacific Signal and Information Processing Association (APSIPA)*, pp. 1222–1227, 2017.

[25] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," *In Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 301–308, 2017.

[26] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. Astudillo, and K. Takeda, "Back-translation-style data augmentation for end-to-end ASR," *In Proc. Spoken Language Technology Workshop (SLT)*, pp. 426–432, 2018.

[27] M. Mimura, S. Ueno, H. Inaguma, S. Sakai, and T. Kawahara, "Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition," *In Proc. Spoken Language Technology Workshop (SLT)*, pp. 477–484, 2018.

[28] T. Hori, R. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. L. Roux, "Cycle-consistency training for end-to-end speech recognition," *arXiv preprint arXiv:1811.01690*, 2018.

[29] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," *In Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4580–4584, 2015.

[30] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1–5, 2015.

[31] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," *In Proc. International Conference on Language Resources and Evaluation (LREC)*, pp. 947–952, 2000.

[32] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.

[33] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki, "VoiceRex spontaneous speech recognition technology for contact-center conversations," *NTT Technical Review*, vol. 5, no. 1, pp. 22–27, 2007.

[34] R. Masumura, T. Asami, T. Oba, H. Masataki, S. Sakauchi, and A. Ito, "Hierarchical latent words language models for robust modeling to out-of-domain tasks," *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1896–1901, 2015.