INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Variational Auto-Encoder Based Variability Encoding for Dysarthric Speech Recognition

Xurong Xie[1], *Rukiye Ruzi*[1], *Xunying Liu*[2], *Lan Wang*[1]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2]Chinese University of Hong Kong, Hong Kong, China

xr.xie@siat.ac.cn, rkym.rouzi@siat.ac.cn, xyliu@se.cuhk.edu.hk, lan.wang@siat.ac.cn

Abstract

Dysarthric speech recognition is a challenging task due to acoustic variability and limited amount of available data. Diverse conditions of dysarthric speakers account for the acoustic variability, which make the variability difficult to be modeled precisely. This paper presents a variational auto-encoder based variability encoder (VAEVE) to explicitly encode such variability for dysarthric speech. The VAEVE makes use of both phoneme information and low-dimensional latent variable to reconstruct the input acoustic features, thereby the latent variable is forced to encode the phoneme-independent variability. Stochastic gradient variational Bayes algorithm is applied to model the distribution for generating variability encodings, which are further used as auxiliary features for DNN acoustic modeling. Experiment results conducted on the UASpeech corpus show that the VAEVE based variability encodings have complementary effect to the learning hidden unit contributions (LHUC) speaker adaptation. The systems using variability encodings consistently outperform the comparable baseline systems without using them, and obtain absolute word error rate (WER) reduction by up to 2.2% on dysarthric speech with "Very low" intelligibility level, and up to 2% on the "Mixed" type of dysarthric speech with diverse or uncertain conditions.

Index Terms: dysarthric, speech recognition, acoustic variability, variational, auto-encoder

1. Introduction

Dysarthria refers to a set of neuromuscular control that impair the physical production of speech. The underlying causes of dysarthria include neurological conditions such as Parkinson disease [1], cerebral palsy [2], and brain damages due to stroke or head injuries. The consequence of such disorder includes weakness, slowing, incoordination, altered muscle tone and inaccuracy of oral and vocal movements [3], which result in speech with abnormal characteristics in quality as well as reduced intelligibility. Meanwhile, dysarthria is often associated with physical disability, so speech-driven assistive technology can be beneficial for dysarthric people using speech as an interface for communication [4] or for enabling them to control physical devices [5]. However, commercial automatic speech recognition (ASR) systems trained with normal speech are improper to be directly utilized for dysarthric speech [6, 7] due to a large mismatch between training and testing. Sparseness of suitable data is another challenge for ASR system development.

Studies has been carried out for developing dysarthric speech recognition system by constructing new datasets [8, 9], or by using data augmentation techniques to deal with data sparsity [10, 11, 12, 13]. Variability of dysarthric speech has been modeled by different manners. Speech tempo in signal domain [12] and feature domain [14] can be modified to reduce mismatch between dysarthric and normal speech. In [15, 16], articulatory knowledge is used to reduce inter-speaker variability. Visual information can be used to reduce variability in DNN models [17]. Bottleneck features have been extracted using a large amount of data from normal speaker through DNNs [18] or convolutive bottleneck network [19, 20] to improve dysarthric speech recognition. Adaptation techniques are also popular for tackling such variation. Combination of HMM state transition interpolation and maximum a posterior adaptation is exploited to model intra-speaker variability in [21]. Speaker adaptive training [22] is shown to be useful for annihilating inter-speaker variations in dysarthric corpus. Pronunciation lexicon adaptation [23, 24] shows effectiveness in a large vocabulary task for dysarthric speakers. Recently, sequence-to-sequence models including listen, attend and spell (LAS) [25, 26, 27], RNN-transducer [25], transformer [28], and QuartzNet [27] have been used to model dysarthric speech.

Some of the aforementioned studies treat the dysarthric variability as speaker difference, and try to deal with it by employing speaker related information, such as speaker-dependent duration or transformation. However, dysarthric variability is caused by diverse conditions and may be uncertain even for the same speaker. An intuitive idea to deal with dysarthric speech is to explicitly encode the variability and use the encodings as auxiliary features for acoustic modeling. This idea is similar to the speaker aware training [29]. However, unlike the speaker information that can be clearly identified, dysarthric variability is commonly unknown or uncertain, and is difficult to be encoded directly. In contrast to explicitly model the variability, variational auto-encoder (VAE) [30] obtains robust representation by projecting the input features to a low-dimensional latent space, and by applying stochastic gradient variational Bayes (SGVB) algorithm to model uncertainty. Hence variability is suppressed for feature reconstruction. Thanks to this advantage, VAE has been widely used for robust speech recognition [31, 32, 33].

In this work, we propose a variational auto-encoder based variability encoder (VAEVE) to explicitly encode the variability for dysarthric speech. The VAEVE makes use of both phoneme information and low-dimensional latent variable to reconstruct the input acoustic features. In this way, the latent variable is forced to encode the phoneme-independent variability instead of suppressing it. LSTMs are employed in the VAEVE to capture temporal information of dysarthric speech. SGVB algorithm is applied to model the distribution for generating variability encodings. For acoustic modeling, encoder of VAEVE associated only with acoustic features is used to generate variability encodings, which are further used as auxiliary input features of DNN acoustic model. In the ASR task on UASpeech corpus, hybrid DNN systems using the variability encodings consistently outperform the baseline systems, and obtain absolute word error rate (WER) reduction by up to

http://dx.doi.org/10.21437/Interspeech.2021-173

2.2% on dysarthric speech with "Very low" intelligibility level. The variability encodings are shown to be complementary to speaker information modeled by learning hidden unit contributions (LHUC) [34] based speaker adaptive training (SAT), and be effective to the "Mixed" type of dysarthric speech with diverse or uncertain conditions.

The rest of the paper is organized as follows. The proposed VAEVE structure and SGVB algorithm is described in Section 2. Section 3 presents the baseline DNN system adaptively trained with multiple speakers. Section 4 shows the experiment on UASpeech task. The last section draws the conclusion and the future works.

2. VAE based variability encoding

2.1. Model design

Acoustic feature is assumed to manifest the joint effect of phoneme information and acoustic variability. Letting $\boldsymbol{o} = \{\boldsymbol{o}_{t=1:T}\}$ be acoustic feature sequence, $\boldsymbol{c} = \{\boldsymbol{c}_{t=1:T}\}$ be phoneme sequence, and $\boldsymbol{z} = \{\boldsymbol{z}_{t=1:T}\}$ be acoustic variability, the generative model is presented as $\boldsymbol{o} \sim p(\boldsymbol{o}|\boldsymbol{c}, \boldsymbol{z})$. Unlike speaker information that can be clearly identified, variability of dysarthric speech caused by diverse conditions may be unknown or uncertain. Even modeling certain aspects of the variability, e.g., voice, motion, and tempo, it is still difficult to cover the mixed effect in the acoustic feature. To encode the acoustic variability explicitly, $\boldsymbol{z}$ can be regarded as latent variable generated by an encoder over the acoustic feature. When using a low-dimensional latent space, limited information is encoded from the acoustic feature. Meanwhile, the encoded information is jointly used with the phoneme information to generate the acoustic feature. This would force the encoding to retain more information of phoneme-independent variability. The encoder is similar to that in variational auto-encoder (VAE) [30]. This generating process is shown in Figure 1, and it leads to the proposed design of variational auto-encoder based variability encoder (VAEVE).

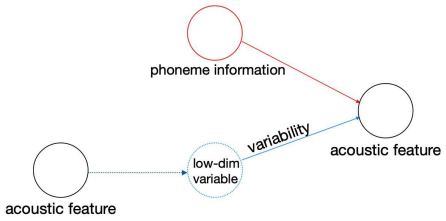

Figure 1: *When using both phoneme information and low-dimensional latent variable associated with acoustic feature to generate the acoustic feature, the latent variable would be forced to retain information of acoustic variability.*

The posterior distribution of $\boldsymbol{z}$ can be approximated by the probabilistic encoder output distribution $q(\boldsymbol{z}|\boldsymbol{o}, \boldsymbol{\phi})$ using SGVB algorithm. The conditional likelihood of $\boldsymbol{o}$ is computed as

$$\log p(\boldsymbol{o}|\boldsymbol{c}, \boldsymbol{\theta}) = \log \int p(\boldsymbol{o}, \boldsymbol{z}|\boldsymbol{c}, \boldsymbol{\theta}) d\boldsymbol{z} \geq \int q(\boldsymbol{z}|\boldsymbol{o},\boldsymbol{\phi}) \log p(\boldsymbol{o}|\boldsymbol{c},\boldsymbol{z},\boldsymbol{\theta}) d\boldsymbol{z} - KL(q||p) \stackrel{\text{def}}{=} \mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{o},\boldsymbol{c}) \quad (1)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ denote the model parameters of decoder and encoder respectively, and $KL(q||p)$ is the KL divergence between $q(\boldsymbol{z}|\boldsymbol{o}, \boldsymbol{\phi})$ and prior distribution $p(\boldsymbol{z})$. By fixing $\boldsymbol{\theta}$, maximizing the lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{o}, \boldsymbol{c})$ is equivalent to minimize the KL divergence between true posterior $p(\boldsymbol{z}|\boldsymbol{o}, \boldsymbol{c})$ and $q(\boldsymbol{z}|\boldsymbol{o}, \boldsymbol{\phi})$. Hence $q(\boldsymbol{z}|\boldsymbol{o}, \boldsymbol{\phi})$ is an approximation to the true posterior.

The probabilistic encoder output distribution $q(\boldsymbol{z}|\boldsymbol{o}, \boldsymbol{\phi})$ can be computed as a Gaussian distribution with diagonal covariance. For the tth instant in an utterance, we have $q(\boldsymbol{z}_t|\boldsymbol{o}, \boldsymbol{\phi}) = \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_t^{(\text{enc})}, (\boldsymbol{\sigma}_t^{(\text{enc})})^2)$. The mean vector and diagonal vector of standard deviation are implemented by an LSTM using the whole utterance as input, such that temporal information can be considered. This is presented as

$$\begin{aligned} \boldsymbol{\mu}_t^{(\text{enc})} &= \boldsymbol{W}_{\mu}^{(\text{enc})} \boldsymbol{h}_t^{(\text{enc})} + \boldsymbol{b}_{\mu}^{(\text{enc})} \\ \boldsymbol{\sigma}_t^{(\text{enc})} &= \exp\{\boldsymbol{W}_{\sigma}^{(\text{enc})} \boldsymbol{h}_t^{(\text{enc})} + \boldsymbol{b}_{\sigma}^{(\text{enc})}\} \\ \boldsymbol{h}_t^{(\text{enc})} &= \text{LSTM}_t^{(\text{enc})} \{\boldsymbol{o}_{t=1:T}\}. \end{aligned} \quad (2)$$

Then, the latent variable $\boldsymbol{z}_t$ is sampled as $\boldsymbol{z}_t = \boldsymbol{\mu}_t^{(\text{enc})} + \boldsymbol{\sigma}_t^{(\text{enc})} \otimes \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$. In practice, the dimension of latent variable $\boldsymbol{z}_t$ should be significantly lower than that of $\boldsymbol{h}_t^{(\text{enc})}$. This encoder structure is shown as the lower part in figure 2.

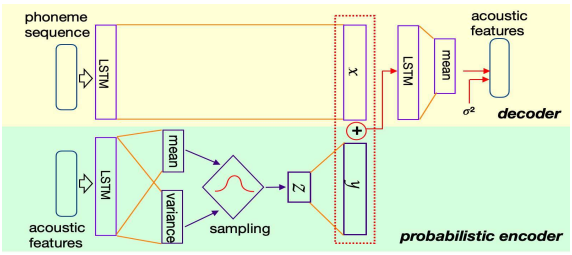

Figure 2: *The structure of variational auto-encoder based variability encoder (VAEVE). The upper part is the decoder; the lower part is the encoder.*

In the decoder, $p(\boldsymbol{o}|\boldsymbol{c}, \boldsymbol{z}, \boldsymbol{\theta})$ is modeled by an isotropic Gaussian distribution. For the tth instant we have $p(\boldsymbol{o}_t|\boldsymbol{c}, \boldsymbol{z}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_t^{(\text{dec})}, \sigma^2 I)$. The mean vector is implemented as

$$\begin{aligned} \boldsymbol{\mu}_t^{(\text{dec})} &= \boldsymbol{W}_2^{(\text{dec})} \boldsymbol{h}_t^{(\text{dec})} + \boldsymbol{b}_2^{(\text{dec})} \\ \boldsymbol{h}_t^{(\text{dec})} &= \text{LSTM}_t^{(\text{dec}_1)} \{\mathbf{f}_{t=1:T}\} \\ \mathbf{f}_t &= \boldsymbol{x}_t + \boldsymbol{y}_t, \quad \boldsymbol{x}_t = \text{LSTM}_t^{(\text{dec}_2)} \{\boldsymbol{c}_{t=1:T}\} \\ \boldsymbol{y}_t &= \boldsymbol{W}_1^{(\text{dec})} \text{Sigmoid}(\boldsymbol{z}_t) + \boldsymbol{b}_1^{(\text{dec})} \end{aligned} \quad (3)$$

where $\boldsymbol{c}_t$ is a one-hot vector representing the tth phoneme label. The variance σ^2 is assumed to be a tunable scalar constant. This decoder structure is shown as the upper part in figure 2.

During training of VAEVE, the prior $p(\boldsymbol{z})$ is assumed to be $\mathcal{N}(0, I)$. The $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{o}, \boldsymbol{c})$ in equation (1) can be rewritten as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{o}, \boldsymbol{c}) \approx \frac{1}{\sigma^2} \sum_{t=1}^{T} \left\{ \frac{\sigma^2}{2} \sum_{d=1}^{D} \left(1 + 2\log(\sigma_{t,d}^{(\text{enc})}) - (\mu_{t,d}^{(\text{enc})})^2 - (\sigma_{t,d}^{(\text{enc})})^2 \right) - \frac{1}{J} \sum_{j=1}^{J} \sum_{d=1}^{D} \frac{1}{2} (o_{t,d} - \mu_{t,d,j}^{(\text{dec})})^2 + \text{constant} \right\} \quad (4)$$

where D denotes the dimension of latent variable, and J is the number of Monte Carlo sampling. The $\mu_{t,d,j}^{(\text{dec})}$ is the dth mean element of decoder computed with the jth sample of $\boldsymbol{z}_t$. During training, the foremost $\frac{1}{\sigma^2}$ can be absorbed by learning rate. The lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{o}, \boldsymbol{c})$ in equation (4) is maximized. The decoder is first pre-trained layer-wise by setting $\boldsymbol{y}_t$ to zero. Then, all parameters of VAEVE are fine-tuned jointly.

2.2. Preventing encoding phoneme information

Although low-dimensional latent space is used by probabilistic encoder, it may also try to model frame-level phoneme information greedily such that the VAEVE degenerates into an identity transform of acoustic features. Some operations can be applied

to prevent the encoder from encoding frame-level phoneme information.

Average Pooling: The acoustic variability is assumed to be stable in a short period. Average pooling over time axis can force the encoder at each time instant to encode the stable information in a short period. During VAEVE training the average pooling is applied to the $\boldsymbol{h}_t^{(\text{enc})}$ in equation (2), which is rewritten as

$$\boldsymbol{h}_t^{(\text{enc})} = \frac{1}{2\tau+1} \sum_{t'=t-\tau:t+\tau} \text{LSTM}_{t'}^{(\text{enc})} \{\boldsymbol{o}_{t=1:T}\} \quad (5)$$

where τ is the radius of time period.

Time Delaying: Time delaying force the decoder at the tth instant to use the latent variable at a preceding instant $t - \Delta t$ as input, where Δt is the delayed time. The latent variable $\boldsymbol{z}_{t-\Delta t}$ is expected to encode similar information to $\boldsymbol{z}_t$. The $\boldsymbol{y}_t$ of decoder in equation (3) is rewritten as

$$\boldsymbol{y}_t = \boldsymbol{W}_1^{(\text{dec})} \text{Sigmoid}(\boldsymbol{z}_{t-\Delta t}) + \boldsymbol{b}_1^{(\text{dec})}. \quad (6)$$

Fixing Decoder: After pre-training, the parameters in the decoder can be fixed. Then, only the encoder is trained in the fine-tuning step. This forces the latent variable to encode complementary information to the phoneme sequence.

Contextual Input: The encoder can use successive frames of acoustic features as input instead of one frame at each instant.

2.3. Variability encoding for DNN acoustic modeling

For acoustic modeling, latent variables generated from the probabilistic encoder of VAEVE is used as variability encodings. Only the acoustic features are required by the encoder, and no additional information is used when making use of the VAEVE for ASR. During acoustic model training, the means and variances computed by the encoder are used to sample latent variables $\boldsymbol{z}$ for each update of DNN parameters. The sampled variables are concatenated with contextual acoustic input features and used as auxiliary input features to train the DNN model. At the tth time instant, this is presented as

$$\boldsymbol{o}'_t = \text{Concatenation}\{\boldsymbol{o}_{t-\tau}, ..., \boldsymbol{o}_t, ..., \boldsymbol{o}_{t+\tau}, \boldsymbol{z}_{t,j}\} \quad (7)$$

where τ is the radius of the contextual window, and $\boldsymbol{z}_{t,j}$ denotes the sampled latent variable at the tth instant for the jth update. However, the latent variables may capture some variability information unrelated to ASR target, such that using $\boldsymbol{o}'_t$ to train DNN model from scratch is prone to over-fitting. An approach to addressing the issue is to use $\boldsymbol{o}'_t$ to retrain a well-trained DNN model by a few steps. In this strategy, the well-trained model parameters are updated by a small learning rate, and the new weights associated with $\boldsymbol{z}_{t,j}$ is updated by a larger learning rate. In the test stage, the latent variable $\boldsymbol{z}_{t,j}$ in equation (7) is replaced with the mean vector $\boldsymbol{\mu}_t^{(\text{enc})}$ and used for decoding.

3. Baseline ASR system description

The baseline hybrid DNN acoustic model in this work is the same as that in [13, 17]. It consists of seven hidden layers and one output layer. Each hidden layer contains a basic set of neural operations performed in sequence, i.e., affine transformation, ReLU activation, and batch normalization. To reduce the number of model parameters, linear bottleneck projections are applied prior to the second to sixth layers. The outputs of the first six layers utilize dropout operation to reduce over-fitting. To accelerate the training process and circumvent the vanishing gradient problem, two skip connections are used to connect the output of the first layer to the third layer, and the output of the fourth layer to the sixth layer respectively.

Multi-task learning (MTL)[35] is used to train the DNN system. The labels for the two tasks are based on frame-level tied tri-phone state alignments and mono-phone alignments respectively. Incorporating frame-level mono-phone alignments in the labels reduces the risk of over-fitting to unreliable alignments of frame-level tri-phone states computed from dysarthric speech.

To model the large variability among dysarthric speakers, LHUC based SAT [34] is employed by the baseline system. Speaker-level LHUC scaling vectors are deployed to the ReLU activation output in the first layer. During training the LHUC vectors are updated once per mini-batch together with the network parameters. Unsupervised LHUC adaptation is performed in the test stage, where the LHUC vectors are updated once per utterance in one epoch.

4. Experiments and results

4.1. Task description

The UASpeech [8] is a corpus of isolated word recognition task consisting of dysarthric speech from 16 dysarthric speakers and normal speech from 13 healthy speakers. The content of speech covers 155 common words and 300 uncommon words. The speakers with dysarthria in types of "Spastic", "Athetoid", and "Mixed" are grouped by their intelligibility levels, which are groups of "Very low", "Low", "Mild", and "High". Silence stripping is performed using a GMM-HMM system to remove redundant silence in the recordings as in [36]. About 30.6 hours of speech from dysarthric and healthy speakers are utilized as training data. About 9 hours of dysarthric speech are used as test data, where 99 words is unseen in training data.

4.2. System settings

The hybrid DNN acoustic model is implemented using the Kaldi toolkit [37]. The contextual input to the DNN acoustic model is 9 successive frames of 80-dimensional filter-bank features with the first order differences. The first six hidden layers contain 2000 neurons each, while the dimension of the linear bottleneck projections is 200 and the dropout rate is 20%. The seventh hidden layer contains 100 neurons. For multi-task learning, the same weight 0.5 is used for both tasks using the 2001 tied tri-phone states and 41 mono-phones. Cross entropies between the task labels and the DNN outputs are minimized by back-propagation based on RMSProp optimizer. A uniform language model is used for decoding. Significance test are based on the matched pairs sentence-segment word error approach.

The VAEVE is also implemented based on the Kaldi toolkit. For the probabilistic encoder in VAEVE, the LSTM contains 128 cells, and the dimension D of latent variable $\boldsymbol{z}_t$ is 39. The acoustic features as encoder input are the same filter-bank features used as DNN acoustic model input. Each LSTMs in the decoder has 256 cells. The mono-phone alignments serve as the phoneme sequences in the decoder. In equation (4) for VAEVE training, we empirically set $\sigma = 0.01$ and $J = 1$. RMSProp optimizer based back-propagation is used for VAEVE training.

When using the generated latent variable $\boldsymbol{z}_t$ for acoustic modeling, the new input features $\boldsymbol{o}'_t$ in equation (7) is used to retrain the well-trained DNN acoustic model with fixed batch normalization. The retraining learning rate of the well-trained parameters is initialized by the final learning rate in DNN training. The retraining learning rate of the new weights associated

with $\boldsymbol{z}_t$ is initialized as 100 times of that of well-trained parameters. Four epochs are performed with the initialized learning rate, which is then halved in the following epochs.

4.3. Results on speaker-independent DNN system

Systems (1) to (3) in Table 1 show the performance of speaker-independent DNN systems with or without using variability encodings generated from VAEVE. In the VAEVE using average pooling, τ in equation (5) is set to 10 frames. Consistent performance improvement is obtained by the DNN systems using variability encodings (Sys (2) and (3)) compared to the baseline system (Sys (1)). Although the VAEVEs with or without using average pooling achieve the same overall WER, they yield different improvements in different intelligibility levels. It seems that the VAEVE with average pooling (Sys (3)) performs better in groups with lower intelligibility than the original VAEVE (Sys (2)) without using operation presented in Section 2.2.

Table 1: *Performance of applying variability encodings generated from VAEVE to the DNN systems on the 16 UASpeech dysarthric speakers. "Origin" refers to VAEVE without using operation presented in Section 2.2, and "Average", "Delay", "FixDec", and "Cntxt" refer to VAEVEs applying average pooling, time delaying, fixing decoder, and contextual input. "Very low", "Low", "Mild" and "High" refer to groups with different intelligibility levels. "†" means the improvement over the comparable baseline system is significant ($P \leq 0.05$).*

Sys	VAEVE	LHUC SAT	Data Aug.	WER(%)				
				Very low	Low	Mild	High	Overall
(1)	-	×	×	69.8	32.6	24.5	10.4	31.5
(2)	Origin	×	×	69.0†	32.6	**23.6**†	**9.8**†	**30.9**†
(3)	Average	×	×	**68.6**†	**32.5**	23.9	10.0†	**30.9**†
(4)	-	√	×	64.4	29.9	20.3	9.0	28.3
(5)	Origin	√	×	63.3†	29.1†	19.3†	8.6†	27.5†
(6)	Average	√	×	**62.2**†	28.8†	19.8	8.6†	27.3†
(7)	Delay	√	×	62.9†	29.0†	19.2†	8.6†	27.4†
(8)	FixDec	√	×	62.9†	28.7†	19.0†	8.6†	27.3†
(9)	+ Delay	√	×	62.6†	**28.6**†	19.0†	8.7†	**27.2**†
(10)	+ Cntxt	√	×	62.8†	28.7†	**18.8**†	**8.5**†	**27.2**†
(11)	-	√	√	62.4	27.6	17.4	**7.9**	26.4
(12)	Average	√	√	**61.2**†	**26.3**†	**16.8**†	8.0	**25.7**†

4.4. Results on DNN system trained with LHUC SAT

Performance of applying variability encodings generated from VAEVE to the DNN systems trained with LHUC SAT is presented as Systems (5) to (10) in Table 1. For the VAEVE using time delaying, Δt in equation (6) is set to 10 frames. For that using contextual input, 9 successive frames of filter-bank features are used as encoder input. Consistent performance improvement is achieved by the DNN systems using variability encodings (Sys (5) to (10)) compared to the baseline system (Sys (4)). The overall improvement is up to 1.1% (Sys (10)) in absolute WER reduction, which is more significant than that for speaker-independent DNN. This implies that the variability encodings generated by VAEVE are complementary to the speaker information modeled by LHUC SAT. Among these systems, the VAEVE applying average pooling (Sys (6)) performs the best on the group with "Very low" intelligibility. The absolute WER reduction is 2.2% over the baseline system. The joint use of fixing decoder and contextual input for VAEVE (Sys (10)) obtains better performance on groups with higher intelligibility, and achieves the best overall performance.

Table 2 shows the performance of using VAEVE based variability encodings on speakers with different dysarthria types. It shows that on the "Mixed" type of dysarthric speech with diverse or uncertain conditions, using the variability encodings gives a statistically significant WER reduction by up to 2.0% (Sys (6)) over the baseline system (Sys (4)). This suggests the effectiveness of the VAEVE method for variability modeling.

Table 2: *Performance of applying VAEVE based variability encodings on the UASpeech speakers with different dysarthria types including "Spastic", "Athetoid", and "Mixed".*

System in Table 1	VAEVE	WER(%)		
		Spastic	Athetoid	Mixed
Sys (4)	-	28.8	21.5	32.3
Sys (6)	Average	27.9†	21.1	**30.3**†
Sys (10)	FixDec+Cntxt	**27.8**†	**20.0**†	31.0†

In addition, the variability encoding is applied to the system trained with data augmentation, which is shown as System (12) in Table 1. The data amount is augmented to 207.5 hours by speed perturbation introduced in [13]. To apply variability encoding, latent variables are sampled for the training data without augmentation. Then, the concatenated features in equation (7) of the training data is employed to retrain the baseline system trained with data augmentation. No augmented data is really used in the retraining applying variability encoding. This enables the retraining process to benefit from data augmentation while without significantly enlarging the retraining time. LHUC adaptation is then applied to test data. Consistent performance improvement is obtained by the system applying variability encoding (Sys (12)) compared to the baseline system (Sys (11)). Table 3 presents the performance of published systems and our system shown as System (12) in Table 1 on UASpeech task.

Table 3: *A comparison on UASpeech task between published systems and our system (Sys (12) in Table 1).*

Systems	WER (%)
Sheffield-2013 Cross domain augmentation [18]	37.5
Sheffield-2015 Speaker adaptive training [22]	34.8
CUHK-2018 DNN System combination [36]	30.6
Sheffield-2019 Kaldi TDNN + Data Aug. [14]	27.9
CNN-TDNN + Speaker adaptation + Transfer learning [38]	30.8
DNN + Data Aug. + LHUC SAT + Cross domain visual [17]	26.8
DNN + Data Aug. + LHUC SAT [13]	26.4
LAS + CTC + Meta-learning + SAT [27]	35.0
QuartzNet + CTC + Meta-learning + SAT [27]	30.5
DNN + Data Aug. + LHUC SAT + AVEVE (ours)	**25.7**

5. Conclusion

This work introduces a variational auto-encoder based variability encoder (VAEVE) to encode the variability for dysarthric speech. Both phoneme information and low-dimensional latent variable are used to reconstruct the input acoustic features in VAEVE, such that the latent variable is forced to encode variability information. SGVB algorithm is used for learning VAEVE. For acoustic modeling, the variability encodings are generated and used as auxiliary features. In the speech recognition task on UASpeech corpus, up to 2.2% absolute WER reduction on dysarthric speech with "Very low" intelligibility is obtained by using variability encodings. The future works may focus on applying VAEVE to acoustic modeling of noisy data.

6. Acknowledgements

This work is supported by Natural Science Foundation of China U1736202, and Shenzhen Fundamental Research Program JCYJ20160429184226930 and KQJSCX20170731163308665.

7. References

[1] S. Scott and F. Caird, "Speech therapy for parkinson's disease." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 46, no. 2, pp. 140–144, 1983.

[2] T. L. Whitehill and V. Ciocca, "Speech errors in cantonese speaking adults with cerebral palsy," *Clinical linguistics & phonetics*, vol. 14, no. 2, pp. 111–130, 2000.

[3] P. Enderby, "Disorders of communication: dysarthria," in *Handbook of clinical neurology*, 2013, vol. 110, pp. 273–281.

[4] M. S. Hawley, S. P. Cunningham, P. D. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neill, "A voice-input voice-output communication aid for people with severe speech impairment," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 21, no. 1, pp. 23–31, 2012.

[5] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. ONeill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Medical Engineering & Physics*, vol. 29, no. 5, pp. 586–593, 2007.

[6] P. C. Doyle, H. A. Leeper, A.-L. Kotler, N. Thomas-Stonell, C. O'Neill, M.-C. Dylke, and K. Rolls, "Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility," *Journal of rehabilitation research and development*, vol. 34, pp. 309–316, 1997.

[7] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[8] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Annual Conference of the International Speech Communication Association*, 2008.

[9] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[10] 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.

[11] 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.

[12] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition." in *Interspeech*, 2018, pp. 471–475.

[13] M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu, and H. Meng, "Investigation of data augmentation techniques for disordered speech recognition." Interspeech, 2020.

[14] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *IEEE ICASSP*, 2019.

[15] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 947–960, 2010.

[16] F. Xiong, J. Barker, and H. Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *Speech Communication; 13th ITG-Symposium*. VDE, 2018, pp. 1–5.

[17] S. Liu, X. Xie, J. Yu, S. Hu, M. Geng, R. Su, S. Zhang, X. Liu, and H. Meng, "Exploiting cross-domain visual feature generation for disordered speech recognition." Interspeech, 2020.

[18] H. Christensen, M. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech." in *INTERSPEECH*, 2013, pp. 3642–3645.

[19] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *ICSP*. IEEE, 2014, pp. 505–509.

[20] Z. Yue, H. Christensen, and J. Barker, "Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition," 2020, pp. 4581–4585.

[21] H. V. Sharma and M. hasegawa Johnson, "State-transition interpolation and map adaptation for hmm-based dysarthric speech recognition," in *Proceedings of the NAACL HLT 2010 workshop on speech and language processing for assistive technologies*, 2010, pp. 72–79.

[22] S. Sehgal and S. Cunningham, "Model adaptation and adaptive training for the recognition of dysarthric speech," in *Proceedings of SLPAT*, 2015, pp. 65–71.

[23] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *IEEE ICASSP*, 2011.

[24] H. Christensen, P. D. Green, and T. Hain, "Learning speaker-specific pronunciations of disordered speech." in *INTERSPEECH*, 2013, pp. 1159–1163.

[25] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt *et al.*, "Personalizing asr for dysarthric and accented speech with limited data," *Proc. Interspeech 2019*, pp. 784–788, 2019.

[26] Y. Takashima, T. Takiguchi, and Y. Ariki, "End-to-end dysarthric speech recognition using multiple databases," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6395–6399.

[27] D. Wang, J. Yu, X. Wu, L. Sun, X. Liu, and H. Meng, "Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization," in *12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021.

[28] Y. Lin, L. Wang, S. Li, J. Dang, and C. Ding, "Staged knowledge distillation for end-to-end dysarthric speech recognition and speech attribute transcription," *Proc. Interspeech 2020*, pp. 4791–4795, 2020.

[29] T. Tan, Y. Qian, D. Yu, S. Kundu, L. Lu, K. C. Sim, X. Xiao, and Y. Zhang, "Speaker-aware training of lstm-rnns for acoustic modelling," in *IEEE ICASSP*, 2016, pp. 5280–5284.

[30] D. P. Kingma and M. Welling, "Stochastic gradient vb and the variational auto-encoder," in *Second International Conference on Learning Representations, ICLR*, vol. 19, 2014.

[31] S. Tan and K. C. Sim, "Learning utterance-level normalisation using variational autoencoders for robust automatic speech recognition," in *IEEE SLT*, 2016, pp. 43–49.

[32] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *IEEE ASRU*, 2017.

[33] W.-N. Hsu and J. Glass, "Extracting domain invariant features by unsupervised learning for robust automatic speech recognition," in *IEEE ICASSP*, 2018, pp. 5614–5618.

[34] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1450–1463, 2016.

[35] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[36] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Lam, X. Wu, K. H. Wong, X. Liu, and H. Meng, "Development of the cuhk dysarthric speech recognition system for the ua speech corpus." in *Interspeech*, 2018, pp. 2938–2942.

[37] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society.

[38] F. Xiong, J. Barker, Z. Yue, and H. Christensen, "Source domain data selection for improved transfer learning targeting dysarthric speech recognition," in *ICASSP*. IEEE, 2020, pp. 7424–7428.