WEAK-SUPERVISED DYSARTHRIA-INVARIANT FEATURES FOR SPOKEN LANGUAGE UNDERSTANDING USING AN FHVAE AND ADVERSARIAL TRAINING

Jinzi Qi[1]*, Hugo Van hamme*[1]

[1]KULeuven, Department Electrical Engineering-ESAT-PSI,
Kasteelpark Arenberg 10, Leuven, Belgium

ABSTRACT

The scarcity of training data and the large speaker variation in dysarthric speech lead to poor accuracy and poor speaker generalization of spoken language understanding systems for dysarthric speech. Through work on the speech features, we focus on improving the model generalization ability with limited dysarthric data. Factorized Hierarchical Variational Auto-Encoders (FHVAE) trained unsupervisedly have shown their advantage in disentangling content and speaker representations. Earlier work showed that the dysarthria shows in both feature vectors. Here, we add adversarial training to bridge the gap between the control and dysarthric speech data domains. We extract dysarthric and speaker invariant features using weak supervision. The extracted features are evaluated on a Spoken Language Understanding task and yield a higher accuracy on unseen speakers with more severe dysarthria compared to features from the basic FHVAE model or plain filterbanks.

Index Terms— Dysarthric speech, FHVAE, adversarial training, weak supervision, end-to-end spoken language understanding.

1. INTRODUCTION

Dysarthric speech data, especially accurately labeled data, are scarce due to difficulties in recruitment, collection and labeling. This data insufficiency problem has been a constant barrier for automatic speech recognition (ASR) model training. Recent ASR models[1, 2, 3] tend to be more complex and employ more parameters, which increases the difficulty to train from limited dysarthric data even further.

Current research mainly solves the dysarthric speech data deficiency problem in three ways. One solution is taking abundant canonical (control) speech data into account during training, e.g. pretraining on control speech and finetuning on dysarthric speech [4, 5, 6]. A second approach is to decrease the model size [7], or to train an inserted small module instead of finetuning the whole model [8, 9], so the number of parameters learned on the dysarthric data is limited. Thirdly and differently from the solutions that work on training strategy or model structure, [10, 11, 12, 13] focus directly on the data and do augmentation to generate more dysarthric speech for use in training.

The features used in most methods preserve the speaker characteristics and the dysarthria information, and thus bound the model generalization ability. However, in practical scenarios, we would like the ASR model to have great generalization ability and work well on unseen speakers (users), to reduce any training effort for dysarthric users. In this work, we focus on improving the automatic recognition model generalization ability with limited dysarthric data. Instead of training and testing the model with speech features that preserve variability between dysarthric speakers, we propose to use dysarthric-invariant and speaker independent speech features for automatic recognition use, which can both boost the model generalization ability on unseen speakers and gives the possibility to involve abundant canonical speech into training.

Factorized Hierarchical Variational Auto-Encoders (FHVAE) [14, 15] have shown their ability to disentangle the content space and the speaker space in input speech. An FHVAE models the generative process of a sequence of segments in a hierarchical structure. It encodes a speech utterance into a segment-related variable (short time scale) and a sequence-related variable (long time scale) via two linked encoders. The segment-related latent variable represents the information that only appears in a single segment in the sequence, such as the content of the segment, and is conditioned on the sequence-related variable. The sequence-related variable reflects the features of the whole sequence, like the acoustic environment or speaker characteristics. The model hence offers a separation between speaker characteristics (sequence) and content (segment). The content variable (segment-related variable) appears to fit the speaker-independence desideratum outlined above. However, our previous work [16] shows that the dysarthria is reflected in both latent variables.

To achieve a greater deal of dysarthria-invariance in the content variable, instead of using forced regularization with phoneme alignment [16], inspired by [17], we introduce adversarial training [18] to the FHVAE model. We feed the FHVAE model with both control and dysarthric speech, regarding the content variable encoder as the generator. We add a discriminator that tries to distinguish control from dysarthric content variables. With adversarial training between the gen-

Copyright 2023 IEEE. This is a preprint version. The original version is to be published in the 2022 IEEE Spoken Language Technology Workshop (SLT) (SLT 2022), scheduled for 09-12 January 2023 in Doha, Qatar. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works, must be obtained from the IEEE. Contact: Manager, Copyrights and Permissions / IEEE Service Center / 445 Hoes Lane / P.O. Box 1331 / Piscataway, NJ 08855-1331, USA. Telephone: + Intl. 908-562-3966.

erator and the discriminator, we can force the content variable encoder (generator) to extract content information from both control and dysarthric speech that shares the same dysarthria-invariant latent space. To make the extracted variable closer to the control data space and improve disentanglement, we also try two additional loss terms called reference loss and disentanglement loss (see section 2). Note that we only need the speaker type label as weak supervision of the training of the complete proposed model. The content encoder is used by both dysarthric speech and canonical speech, which gives opportunities to involve abundant canonical speech features into the downstream model (ASR) training in future work.

For evaluation, we use the trained content encoder to extract the content variables from other datasets and use these variables to train and test a mature End-to-End (E2E) Spoken Language Understanding (SLU) system [19, 20]. The system consists of a recurrent neural network (RNN) encoder, a capsule network [21] and output layers, and can transfer input speech features into an intent (semantic) representation without using an explicit textual representation. E2E SLU evaluation is different from the evaluation on speaker-independent, in the sense that we have (a limited number of) task-specific recordings by dysarthric speakers at our disposal for training the SLU system, whereas ASR systems would typically be trained on task-agnostic data. Because of the domain adaptation inherent in the evaluation, it is harder to show generalization properties on E2E SLU tasks.

Apart from SLU evaluation, we also train and test a binary speaker type classifier (control/dysarthria) using both content variables and sequence-related variables, and use the classification accuracy to show how the dysarthria information distribution changes after the adversarial training.

We introduce the feature extraction methods, the FHVAE model with adversarial training, reference loss and disentanglement loss, in section 2. Section 3 describes the data and the experimental settings. Results and analysis will be provided in section 4, and section 5 gives conclusions.

2. METHOD

In this section, we introduce the feature extraction methods: the FHAVE model extended with adversarial training, reference loss and disentanglement loss.

2.1. FHVAE with adversarial training

Suppose we have I speech data sequences, each sequence $\mathbf{X}^i$ containing N_i segments $\mathbf{x}^{i,n}$, $n = 1, 2, ..., N_i$. The segment $\mathbf{x}^{i,n}$ will be represented by two latent variables $\mathbf{z}_1^{i,n}$ and $\mathbf{z}_2^{i,n}$ via an FHVAE:

$$q(\mathbf{z}_2^{i,n}|\mathbf{x}^{i,n}) = \mathcal{N}(\mathbf{Enc}_{\mu_{\mathbf{z}_2}}(\mathbf{x}^{i,n}), \mathbf{Enc}_{\sigma^2_{\mathbf{z}_2}}(\mathbf{x}^{i,n})) \quad (1)$$

$$q(\mathbf{z}_1^{i,n}|\mathbf{x}^{i,n}, \mathbf{z}_2^{i,n}) = \mathcal{N}(\mathbf{Enc}_{\mu_{\mathbf{z}_1}}(\mathbf{x}^{i,n}, \mathbf{z}_2^{i,n}), \mathbf{Enc}_{\sigma^2_{\mathbf{z}_1}}(\mathbf{x}^{i,n}, \mathbf{z}_2^{i,n}) \quad (2)$$

where $\mathbf{Enc}_{\mu_{\mathbf{z}_2}}(\cdot)$ and $\mathbf{Enc}_{\sigma^2_{\mathbf{z}_2}}(\cdot)$ are the networks encoding input $\mathbf{x}$ to mean and variance of $\mathbf{z}_2$, $\mathbf{Enc}_{\mu_{\mathbf{z}_1}}(\cdot)$ and $\mathbf{Enc}_{\sigma^2_{\mathbf{z}_1}}(\cdot)$ are the encoders of $\mathbf{z}_1$. Following the theory of variational auto-encoders, the latent variables should follow prior distributions:

$$p(\mathbf{z}_1^{i,n}) = \mathcal{N}(\mathbf{0}, \sigma^2_{\mathbf{z}_1}\mathbf{I}) \quad (3)$$

$$p(\mathbf{z}_2^{i,n}|\boldsymbol{\mu}_2^i) = \mathcal{N}(\boldsymbol{\mu}_2^i, \sigma^2_{\mathbf{z}_2}\mathbf{I}), \quad p(\boldsymbol{\mu}_2^i) = \mathcal{N}(\mathbf{0}, \sigma^2_{\boldsymbol{\mu}_2}\mathbf{I}) \quad (4)$$

where $\sigma^2_{\mathbf{z}_1}, \sigma^2_{\mathbf{z}_2}$ and $\sigma^2_{\boldsymbol{\mu}_2}$ are chosen upfront and $\mathbf{z}_1^{i,n}, \mathbf{z}_2^{i,n}$ are i.i.d samples. The prior enforces variables $\mathbf{z}_2^{i,n}$ within the same sequence i to be close.

The $\mathbf{z}_1^{i,n}$, conditioned on $\mathbf{z}_2^{i,n}$ (see equation2), only contains information with the speech segment $\mathbf{x}^{i,n}$, and is regarded as the speaker-independent content variable. The variable $\mathbf{z}_2^{i,n}$ reflects the features of the whole sequence $\mathbf{X}^i$ and can be seen as speaker variable. The mean of sequence-level variable $\boldsymbol{\mu}_2^i$ can be inferred as:

$$\tilde{\mu}_2^i = \frac{\sum_{n=1}^{N} \mathbf{Enc}_{\mu_{\mathbf{z}_2}}(\mathbf{x}^{i,n})}{N + \frac{\sigma^2_{\mathbf{z}_2}}{\sigma^2_{\boldsymbol{\mu}_2}}} \quad (5)$$

Then for the whole sequence $\mathbf{X}^i$, the inference model is written as:

$$q(\mathbf{Z}_1^i, \mathbf{Z}_2^i, \boldsymbol{\mu}_2^i|\mathbf{X}^i) = q(\boldsymbol{\mu}_2^i|\mathbf{x}^{i,n}) \prod_{n=1}^{N} q(\mathbf{z}_1^{i,n}|\mathbf{x}^{i,n}) q(\mathbf{z}_2^{i,n}|\mathbf{x}^{i,n}) \quad (6)$$

where $\mathbf{Z}_1^i = \{\mathbf{z}_1^{i,n}\}_{n=1}^N$, $\mathbf{Z}_2^i = \{\mathbf{z}_2^{i,n}\}_{n=1}^N$. The generative model for the data is then:

$$p(\mathbf{x}^{i,n}|\mathbf{z}_1^{i,n}, \mathbf{z}_2^{i,n}) = \mathcal{N}(\mathbf{Dec}_{\mu_x}(\mathbf{z}_1^{i,n}, \mathbf{z}_2^{i,n}), \mathbf{Dec}_{\sigma^2_x}(\mathbf{z}_1^{i,n}, \mathbf{z}_2^{i,n})) \quad (7)$$

where $\mathbf{Dec}_{\mu_x}(\cdot)$ and $\mathbf{Dec}_{\sigma^2_x}(\cdot)$ form the decoder networks of the FHVAE.

The dysarthria information, as a speaker characteristic, should be extracted to speaker variable $\mathbf{z}_2^{i,n}$ by the FHVAE. However, our previous work [16] found that the FHVAE does not separate the dysarthria and content information and speech impairment is identifiable from $\mathbf{z}_1^{i,n}$.

To obtain dysarthria-invariant features $\mathbf{z}_2^{i,n}$, inspired by [17], we introduce adversarial training into the FHVAE model. The data flow of the proposed model is provided in figure 1.

Repurposing the $\mathbf{z}_1$ mean encoder $\mathbf{Enc}_{\mu_{\mathbf{z}_1}}(\cdot)$ as the generator in adversarial training, we get $\mu_{\mathbf{z}_1}^{i,n}$ as the generator output and feed this into a discriminator $\mathbf{D}(\cdot)$ which is trained to yield the probability $p_D^{i,n}$ that a speech segment $\mathbf{x}^{i,n}$ stems from a control speaker:

$$p_D^{i,n} = \mathbf{D}(\mu_{\mathbf{z}_1}^{i,n}) = \mathbf{D}(\mathbf{Enc}_{\mu_{\mathbf{z}_1}}(\mathbf{x}^{i,n})) \quad (8)$$

Training of $\mathbf{D}(\cdot)$ involves dysarthric and control data and minimizes cross-entropy:

$$L_D = CE_{dys} + CE_{ctrl} \quad (9)$$

where

$$CE_{dys/ctrl} = \mathrm{E}_{\mu_{\mathbf{z}_1} \sim p_{dys/ctrl}(\mu_{\mathbf{z}_1})}[-(l^{i,n}_{dys/ctrl} \times \log p^{i,n}_D) + (1 - l^{i,n}_{dys/ctrl}) \times \log(1 - p^{i,n}_D)] \quad (10)$$

and $l^{i,n}_{dys/ctrl}$ is the label of speech segment $\mathbf{x}^{i,n}$, i.e. $l^{i,n}_{dys/ctrl}$ is 1 when $\mathbf{x}^{i,n}$ is dysarthric and is 0 for control speech.

The loss function of the FHVAE model can be written as:

$$L_{FHVAE} = L_{LB} + L_{z_2,disc} * W_{z_2,disc} + L_{Gen} * W_{Gen} \quad (11)$$

where L_{LB} is the variational lower bound loss of the FHVAE model that guarantees the performance of the FHVAE model as an auto-encoder; $L_{z_2,disc}$ is the discriminative loss of the sequence-related latent variable, which expresses that sequence identity should be inferrable from the sequence-related embedding vector, i.e., it is used to encourage separation in the latent space. L_{Gen} is the cross-entropy between the discriminator output and the opposite dysarthric/control label (control is 0 and dysarthric is 1), similar to equation 9. Unlike a classical Generative Adversarial Network (GAN), the L_{Gen} here acts on the whole FHVAE model instead of only on the generator part.

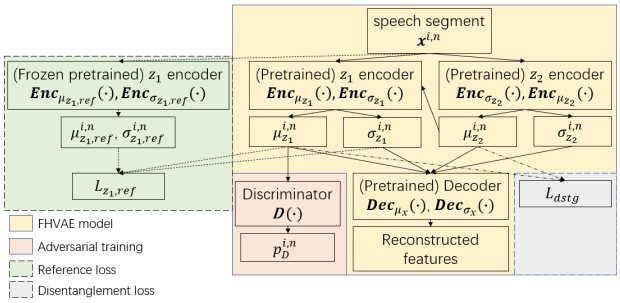


Fig. 1. Dataflow in the proposed model with adversarial training, reference loss and disentanglement loss.

2.2. Extension with reference loss and disentanglement loss

In the model presented so far, the adversarial training will modify the encoder output on both control and dysarthric input, as we want a uniform feature extractor for both types of speech. Modifying the behavior for control data is not desirable since the FHVAE is supposed to be pre-trained well in that part of the input space. Thus we add an additional loss term $L_{z_1,ref}$ called reference loss into the FHVAE model training loss with weight $W_{z_1,ref}$. The reference loss is active on control speech only and is the Kullback-Leibler divergence between the reference z_1 distribution $\mathcal{N}(\mu_{z_1,ref}, \sigma^2_{z_1,ref})$ of the FHVAE pre-trained on control speech and the z_1 distribution during training. We also consider evaluating the L_{Gen} term in Eq. 11 only on the dysarthric data (CE_{dys}), while the other terms are unchanged. In doing so, the content encoder emphasizes changes for dysarthric speech as well.

Finally, to further disentangle speaker and content information, we add the disentanglement loss L_{dstg} with weight W_{dstg}, which minimizes the sum of squares of the entries of the correlation matrix between μ_{z_1} and μ_{z_2}.

The reference loss module and disentanglement loss module are shown in figure 1.

3. EXPERIMENTS

This section explains how we implement and evaluate the proposed model, the datasets and the experimental details.

3.1. Datasets

We use four datasets for pretraining, finetuning and evaluation, respectively:

- ***CGN** dataset* [22]: we use the 'k' and 'o' subsets from the Dutch CGN dataset for FHVAE model pretraining. It contains 35.2-hours of broadcast news and read speech from 326 control Dutch speakers.
- ***Torgo** dataset* [23] is used for model finetuning, containing 6.3 hours of English speech from 7 control speakers (3.8 h) and 8 speakers with dysarthria (2.5 h).
- ***Domotica** dataset* [24] is used for SLU evaluation, containing 8.8 h Dutch speech from 15 speakers with dysarthria (2 speakers have no intelligibility score and are excluded), uttering commands related to home automation, such as "turn on the light in the kitchen". The commands are encoded to 22 slot value labels and 4 categories, and each category command is represented as 1 or 2 active labels. There are in total 3789 speech commands. The severity level (intelligibility score) of dysarthric speakers can be found in [24].
- ***COPAS** dataset* [25] is used for speaker type classification evaluation, containing 7 hours of Dutch speech from 131 control speakers (3 h) and 230 speakers with speech disorders (4 h). 4699 utterances in total with an average duration of 6 seconds.

3.2. Experimental settings

We first pre-train the FHVAE model, then finetune it using the adversarial training module and the two additional loss terms. We also provide a comparison where we finetune the pretrained FHVAE without adversarial training. Then we use the trained encoders $\mathbf{Enc}_{\mu_{\mathbf{z}_1}}(\cdot)$ and $\mathbf{Enc}_{\mu_{\mathbf{z}_2}}(\cdot)$ to extract content and sequence-related variables for each segment. The content-related variables are then used to train and test the SLU system, while the speaker type classifier uses both. The experimental design is illustrated in figure 2.

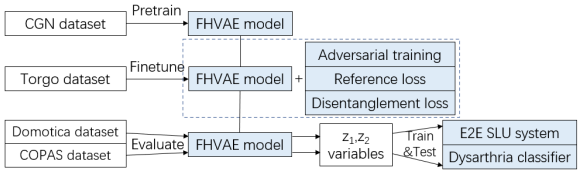


Fig. 2. Experimental design.

3.2.1. Evaluation setup and metrics

Since our SLU data contains only speakers with dysarthria, we perform generalization tests across different severity levels rather than control versus dysarthric speech.

The evaluation on SLU uses two train/test divisions. For the Out-of Domain evaluation, we use speech data (2642 speech utterances) from 11 speakers with mild to medium dysarthria (intelligibility score ≥ 70) to train the SLU model and use 4 speakers with severe dysarthria (intelligibility score < 70) for testing. In this experiment, we evaluate the generalization ability of the model across different severity levels. If the content variable is more dysarthria-invariant, the SLU model should perform better thanks to better generalization.

In the In-Domain type evaluation, we order the speakers by their intelligibility score from high to low, and use the data (1901 speech utterances) from the 8 odd-numbered speakers for training and the rest for testing. In this type, the SLU model sees a wide range of severity levels but tests on unseen speakers within the data domain.

For SLU system evaluation, the micro-averaged F1-score of the detected labels is used as the metric. For each type, we repeat the experiment 5 times and use the average F1 score as the final result.

For the dysarthria classification, we do 6-fold cross-validation experiment that we divide all COPAS speakers into 6 blocks, each with the same number of control and disordered speech utterances. For each fold, we use four blocks to train the classifier, one for validation and one for testing. We repeat the 6-fold experiment 5 times and use the averaged classification accuracy (correct rate) as the final classification results.

3.2.2. Network and training setup

The speech features used in the FHVAE model are 80 MEL filter bank energies with a frame advance of 10 ms. During training, each FHVAE segment is 20 frames with 8 frames shift, i.e., 200 ms of speech. At inference, the shift is one frame.

In the FHVAE network, the encoder consists of 2 layers of bidirectional LSTM layers and 2 parallel dense layers for each latent variable. Each LSTM has 256 units, and the dimension of each latent variable is 32. The decoder consists of 2 bidirectional LSTM layers of the same size and 2 dense layers of 80 units for μ_x and σ_x^2 prediction, respectively. The discriminator consists of two dense layers with 32 hidden nodes and 1 hidden node, respectively. The first layer has $leaky\ ReLU$ activation and a $sigmoid$ is used for the discriminator output.

We choose the Adam optimizer for training with loss weights $W_{z_2,disc} = 10$, $W_{Gen} = 500$, $W_{z_1,ref} = 0.1$, $W_{dstg} = 1$ to match the magnitude of other losses in the final loss function. The hierarchical sampling size is 5000 [26], the batch size is 500 and the maximal number of training epochs is 50. Early stopping with patience of 10 is used during baseline training and simple finetuning. For adversarial training, training stops when it reaches the maximal number of epochs. The default learning rate for the FHVAE model (pre)training and finetuning is 0.001. For the proposed model with adversarial training, we use a learning rate of 0.001 for the FHVAE model and 0.0002 for the discriminator.

We choose the SLU system of [20] with unmodified setting for evaluation. We set the number of epochs to 50, use Adam Optimizer with a learning rate 0.001 and use batch size 16. For dysarthria classification, the classifier consists of 2 dense layers with 100 and 2 hidden nodes, respectively. The first layer has $ReLU$ activation and the second has $Softmax$ activation. We use a batch size of 128. We aggregate the input data over the utterance by taking the mean vector as the classifier input data.

4. RESULTS AND DISCUSSION

4.1. E2E SLU system evaluation

We evaluate how dysarthria and speaker invariant the extracted content variable is.

We first compare SLU performance using filter bank ($Fbank$) versus μ_{z_1}-features extracted by the baselines (pretrained model and simply finetuned model) and proposed models. In table 1, we provide the averaged slot value F1-score for the SLU task using different training sets.

For the Out-of-Domain experiment, the z_1 extracted by the FHVAE model with adversarial training has no advantage over the $Fbank$ features when testing the generalization ability between different dysarthria severity levels. The reason might be that the adversarial training strategy successfully narrows the gap between the dysarthric and canonical speech and it does so by finding common ground in which some of the detail is lost for speakers with high intelligibility, leading to accuracy loss in SLU.

By adding the reference loss and including only CE_{dys} in L_{Gen}, the modification of the control content space is weakened and the dysarthric speech content space moves closer to the control space. Due to the dysarthria invariance of the content variable, the F1-score increases and we obtain the highest F1-score 0.571 from the model with adversarial training, reference loss and CE_{dys} only. In In-Domain SLU, the baseline FHVAE already outperforms the $Fbank$ features by a great margin. The content variable seems to be a powerful

Table 1. F1-scores of SLU tasks (left of the table) and accuracy of dysarthria classifier (right part of the table) using extracted content variables, speaker variables or $Fbank$ features as training and testing data. (* : $p-value \leq 0.05$ compared to results of data extracted by FHVAE simply finetuned on Torgo, † : $p-value \leq 0.05$ compared to results of $Fbank$ features. The $p-value$s are calculated by Wilcoxon signed rank test on recognition accuracy, where the score per utterance is averaged over 5 trials.)

Feature Extraction method	**SLU Task Input**				**Speaker Type Classifier Input**			
	Out-of-Domain		In-Domain		$Fbank$	z_1	z_2	z_{12}
	$Fbank$	z_1	$Fbank$	z_1				
-	0.526	-	0.576*	-	0.719	-	-	-
FHVAE pretrained on CGN	-	0.519	-	0.651†	-	0.734†	0.693*	0.757†
+Simply Finetuned on Torgo	-	0.522	-	0.653†	-	0.742†	0.680†	0.756†
+Adversarial training on Torgo	-	0.520	-	0.638†	-	0.724*	0.701*	0.755†
+Reference loss	-	0.537*	-	0.641†	-	0.714*	0.726*	0.754†
+Only use CE_{dys} in L_{Gen}	-	0.571*†	-	0.651†	-	0.720*	0.726*	0.757†
+Disentanglement loss	-	0.493*†	-	0.600*†	-	0.707*	0.737*†	0.764†

speaker-independent speech representation, a point also made in [14]. For the In-Domain task, it is however less important to narrow the gap between canonical and dysarthric data, since the SLU backend sees both kinds of data in training. The proposed methods for narrowing the gap between canonical and dysarthric speech are hence less relevant in this scenario. However, for the final design with adversarial training, reference loss and restricting the L_{Gen} term to the dysarthric data, there is also no degradation. Hence, the design provides reliable features for SLU in both scenarios.

Finally, the disentanglement loss seems to hurt the content representation and leads to lower accuracy for SLU.

4.2. Dysarthria classification

To further assess the effect of adversarial training in generating a dysarthria-invariant latent variable, we implement a binary classifier that classifies input utterance-level mean vectors into either dysarthric or control type. In the right of table 1, we provide the classification accuracy using the content variable z_1, speaker variable z_2 and z_{12} which concatenates both variables.

From the table, we see that when finetuning on the dysarthric dataset Torgo, although the F1-score increases, the classification accuracy of z_1 increases with an accuracy decrease using z_2. The dysarthria information reassignment shows that when processing dysarthric data, the FHVAE tends to extract dysarthria-related information into the content variable z_1. When we include the adversarial training into the finetuning procedure, we see a drop in accuracy when using z_1 while an increase with z_2. This shows that the adversarial training strategy does increase the invariance of the dysarthric information in z_1, narrows the gap between dysarthric and control latent spaces and promotes disentanglement of the speaker and content space. Meanwhile, when using both content and speaker variable, classification accuracy with z_{12} is higher than using only $Fbank$ features, showing the potential of the FHVAE model in dysarthric speech diagnosis.

When we engage the disentanglement loss into model training, there is even lower classification accuracy with z_1 and higher accuracy with z_2. Hence, the technique is successful in the goal to push dysarthria-related information into the sequence variable. However, this seems to happen at the expense of the quality of the content variable (see section 4.1). Indeed, a lower correlation can also be achieved by making the content variable more noisy.

5. CONCLUSIONS

We propose a dysarthria-invariant and speaker-independent feature extraction method using an FHVAE model and adversarial training. We work in a setting of limited task-agnostic dysarthric speech data for building the feature extractor and evaluate - again on limited task-specific training data - on an SLU task. By repurposing the content encoder in the FHVAE as a generator and adding a discriminator network, we are able to narrow the gap in the latent space for dysarthric and control speech, using only weak supervision. Using the extracted content variable to train and test a mature E2E SLU system, we see significant F1-score improvement on speakers with more severe dysarthria, compared to using filter bank features or representations from a basic FHVAE model. The binary dysarthria/control classifier also yields lower accuracy when using the extracted content variable, showing the effect of the proposed model in disentangling speaker and content information in dysarthric speech. Interestingly, the extra disentanglement loss does improve disentanglement, but it seems to spoil the content information as well.

We test the effectiveness of the extracted dysarthria-invariant and speaker-independent features across different dysarthria severity levels. In future work, we intend to evaluate the proposed features of both dysarthric and control speech in other speech recognition models. Moreover, we see

obvious improvement in In-Domain evaluation when using pretraining and finetuning from our current work.

6. ACKNOWLEDGEMENTS

The research was supported by KU Leuven Special Research Fund grant C24M/22/025 and the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

7. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[3] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

[4] Joel Shor, Dotan Emanuel, Oran Lang, Omry Tuval, Michael Brenner, Julie Cattiau, Fernando Vieira, Maeve McNally, Taylor Charbonneau, Melissa Nollstadt, et al., "Personalizing asr for dysarthric and accented speech with limited data," *Proc. Interspeech 2019*, pp. 784–788, 2019.

[5] Yuki Takashima, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Knowledge transferability between the speech data of persons with dysarthria speaking different languages for dysarthric speech recognition," *IEEE Access*, vol. 7, pp. 164320–164326, 2019.

[6] Pu Wang, Bagher BabaAli, and Hugo Van hamme, "A study into pre-training strategies for spoken language understanding on dysarthric speech," *arXiv preprint arXiv:2106.08313*, 2021.

[7] Pu Wang and Hugo Van hamme, "A light transformer for speech-to-intent applications," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 997–1003.

[8] Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadsy, "Residual adapters for parameter-efficient asr adaptation to atypical and accented speech," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6751–6760.

[9] Fadi Biadsy, Youzheng Chen, Xia Zhang, Oleg Rybakov, Andrew Rosenberg, and Pedro J Moreno, "A scalable model specialization framework for training and inference using submodels and its application to speech model personalization," *arXiv preprint arXiv:2203.12559*, 2022.

[10] Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6009–6013.

[11] Protima Nomo Sudro, Rohan Kumar Das, Rohit Sinha, and SR Mahadeva Prasanna, "Significance of data augmentation for improving cleft lip and palate speech recognition," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 484–490.

[12] Zengrui Jin, Mengzhe Geng, Xurong Xie, Jianwei Yu, Shansong Liu, Xunying Liu, and Helen Meng, "Adversarial data augmentation for disordered speech recognition," *arXiv preprint arXiv:2108.00899*, 2021.

[13] Zengrui Jin, Mengzhe Geng, Jiajun Deng, Tianzi Wang, Shujie Hu, Guinan Li, and Xunying Liu, "Personalized adversarial data augmentation for dysarthric and elderly speech recognition," *arXiv preprint arXiv:2205.06445*, 2022.

[14] Wei-Ning Hsu, Yu Zhang, and James Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Information Processing Systems*, 2017, pp. 1878—-1889.

[15] Suwon Shon, Wei-Ning Hsu, and James Glass, "Unsupervised representation learning of speech for dialect identification," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 105–111.

[16] Jinzi Qi and Hugo Van hamme, "Speech disorder classification using extended factorized hierarchical variational auto-encoders," *Proc. Interspeech 2021*, 2021.

[17] Lars Rumberg, Hanna Ehlert, Ulrike Lüdtke, and Jörn Ostermann, "Age-invariant training for end-to-end child speech recognition using adversarial multi-task learning," *Proc. Interspeech 2021*, pp. 3850–3854, 2021.

[18] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[19] Vincent Renkens, *ASSIST: Assistive Speech Interface for Smart Technologies*, Ph.D. thesis, KU Leuven, 2019.

[20] Vincent Renkens and Hugo Van hamme, "Capsule networks for low resource spoken language understanding," *Proc. Interspeech 2018*, 2018.

[21] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton, "Dynamic routing between capsules," *Advances in neural information processing systems*, vol. 30, 2017.

[22] Nelleke Oostdijk, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat, and R Harald Baayen, "Experiences from the spoken dutch corpus project," in *LREC 2002*. European Language Resources Association, 2002, pp. 340–347.

[23] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya 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.

[24] Bart Ons, Jort F Gemmeke, and Hugo Van hamme, "The self-taught vocal interface," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–16, 2014.

[25] Gwen Van Nuffelen, Marc De Bodt, Catherine Middag, and Jean-Pierre Martens, "Dutch corpus of pathological and normal speech (copas)," Tech. Rep., Antwerp University Hospital and Ghent University, 2009.

[26] Wei-Ning Hsu and James Glass, "Scalable factorized hierarchical variational autoencoder training," *Proc. Interspeech 2018*, 2018.