SPEAKER IDENTITY PRESERVATION IN DYSARTHRIC SPEECH RECONSTRUCTION BY ADVISERIAL SPEAKER ADAPTATION

Disong Wang\textsuperscript{1}, Songxiang Liu\textsuperscript{1}, Xixin Wu\textsuperscript{1}, Hui Lu\textsuperscript{1}, Lifa Sun\textsuperscript{2}, Xunying Liu\textsuperscript{1}, Helen Meng\textsuperscript{1,3}

\textsuperscript{1}The Chinese University of Hong Kong, Hong Kong SAR, China
\textsuperscript{2}SpeechX Limited, Shenzhen, China
\textsuperscript{3}Centre for Perceptual and Interactive Intelligence, Hong Kong SAR, China

\{dswang, xliu, wuxx, luhui, xylu, hmmeng\}@se.cuhk.edu.hk, lfsun@speechx.cn

\section*{ABSTRACT}
Dysarthric speech reconstruction (DSR), which aims to improve the quality of dysarthric speech, remains a challenge, not only because we need to restore the speech to be normal, but also must preserve the speaker’s identity. The speaker representation extracted by the speaker encoder (SE) optimized for speaker verification has been explored to control the speaker identity. However, the SE may not be able to fully capture the characteristics of dysarthric speakers that are previously unseen. To address this research problem, we propose a novel multi-task learning strategy, i.e., adversarial speaker adaptation (ASA). The primary task of ASA fine-tunes the SE with the speech of the target dysarthric speaker to effectively capture identity-related information, and the secondary task applies adversarial training to avoid the incorporation of abnormal speaking patterns into the reconstructed speech, by regularizing the distribution of reconstructed speech to be close to that of reference speech with high quality. Experiments show that the proposed approach can achieve enhanced speaker similarity and comparable speech naturalness with a strong baseline approach. Compared with dysarthric speech, the reconstructed speech achieves 22.3\% and 31.5\% absolute word error rate reduction for speakers with moderate and moderate-severe dysarthria respectively. Our demo page is released here\textsuperscript{1}.

\textbf{Index Terms}— Dysarthric speech reconstruction, voice conversion, adversarial speaker adaptation, speaker identity

\section{1. INTRODUCTION}
Dysarthria arises from various neurological disorders including Parkinson’s disease or amyotrophic lateral sclerosis, leading to weak regulation of articulators such as jaw, tongue, and lips \cite{1}. Therefore, the resulting dysarthric speech may be perceived as harsh or breathy with abnormal prosody and inaccurate pronunciation, which degrades the efficiency of vocal communication for dysarthric patients. Attempts have been made to improve the quality of dysarthric speech by using various reconstruction approaches, where voice conversion (VC) serves as a promising candidate \cite{2}.

The goal of VC is to convert non-linguistic or para-linguistic factors such as speaker identity \cite{3}, prosody \cite{4}, emotion \cite{5} and accent \cite{6}. VC has also been widely applied in reconstructing different kinds of impaired speech including esophageal speech \cite{7,8}, electrolaryngeal speech \cite{9,10}, hearing-impaired speech \cite{11} and dysarthric speech \cite{2}, where rule-based and statistical VC approaches have been investigated for dysarthric speech reconstruction (DSR). Rule-based VC tends to apply manually designed, speaker-dependent rules to correct phoneme errors or modify temporal and frequency features to improve intelligibility \cite{12,13}. Statistical VC automatically maps the features of dysarthric speech to those of normal speech, where typical approaches contain Gaussian mixture model \cite{14}, non-negative matrix factorization \cite{15,16}, partial least squares \cite{17}, and deep learning methods including sequence-to-sequence (seq2seq) models \cite{18,19} and gated convolutional networks \cite{20}. Though significant progress has been made, previous work generally ignores speaker identity preservation, which loses the ability for patients to demonstrate their personality via acoustic characteristics. Preserving the identities for dysarthric speakers is very challenging since their normal speech utterances are difficult to collect. A few studies \cite{22,23} use a speaker representation to control the speaker identity of reconstructed speech, where the speaker encoder (SE) proposed in our previous work \cite{23} is trained on a speaker verification (SV) task by using large-scale normal speech. However, the SE may fail to effectively extract speaker representations from previously unseen dysarthric speech, which lowers the speaker similarity of reconstructed speech.

This paper proposes an improved DSR system based on \cite{23} by using adversarial speaker adaptation (ASA). The DSR system in \cite{23} contains four modules: (1) A speech encoder extracting accurate phoneme embeddings from dysarthric speech to restore the linguistic content; (2) A prosody corrector inferring normal prosody features that are treated as canonical values for correction; (3) A speaker encoder producing a single vector as speaker representation used to preserve the speaker identity; and (4) A speech generator mapping phoneme embeddings, prosody features and speaker representation to reconstructed mel-spectrograms. The speaker encoder and speech generator are independently trained by using large-scale normal speech data. We term the resulting integrated DSR system using SV-based speaker encoder as the SV-DSR, which can generate the reconstructed speech with high intelligibility and naturalness. To better preserve the identity of the target dysarthric speaker during speech generation, speaker adaptation can be used to fine-tune the speaker encoder by using the dysarthric speech data. However, this approach inevitably incorporates dysarthric speaking patterns into the reconstructed speech. Hence, we propose to use ASA to alleviate this issue, and the resulting DSR system is termed as the ASA-DSR. For each dysarthric speaker, ASA-DSR is first cloned from SV-DSR and then adapted in a multi-task learning manner: (1) The primary task performs speaker adaptation to fine-tune the speaker encoder by using the dysarthric speech data to enhance the speaker similarity; (2) The secondary task performs adversarial training to alternatively optimize the speaker encoder and a system discrimi-

\textsuperscript{1}Audio samples: https://wendison.github.io/ASA-DSR-demo/
2. BASELINE APPROACH: SV-DSR

As shown in Fig. 1, our previously proposed SV-DSR system [23] contains four modules: speech encoder, prosody corrector, speaker encoder and speech generator. The first three modules respectively produce phoneme embeddings, prosody values and speaker representation; and the fourth module, the speech generator, maps these features to reconstructed mel-spectrograms.

**Speech encoder:** To recover the content, a seq2seq-based speech encoder is optimized by two-stage training to infer the normal phoneme duration and pitch features \(p\) and \(v\) as canonical values to replace their abnormal counterparts for generating high-quality speech with enhanced speaker similarity.

**Prosody corrector:** As abnormal duration and pitch are two patterns, rendering the reconstructed speech from ASA-DSR to maintain stable prosody and improved intelligibility.

The main contribution of this paper is the use of proposed ASA approach to effectively preserve speaker identities of dysarthric patients after the reconstruction, without using patients’ normal speech that is nearly impossible to collect. It is noted that our work is different from [24] that aims to achieve robust speech recognition, as the proposed ASA here is used to obtain regularized mel-spectrograms from [24] that aims to achieve robust speech recognition, as the proposed ASA here is used to obtain regularized mel-spectrograms for generating high-quality speech with enhanced speaker similarity.

2. PROPOSED APPROACH: ASA-DSR

The proposed approach of adversarial speaker adaptation (ASA), as illustrated in Fig. 1, aims to enhance speaker similarity, resulting in the proposed ASA-DSR system that shares the same modules as SV-DSR except for the speaker encoder. First, ASA-DSR is cloned from SV-DSR, then a system discriminator \(\varphi\) is introduced to determine whether its input mel-spectrograms are reconstructed by SV-DSR or ASA-DSR systems. Given a dysarthric speaker \(s_d\) with the adaptation data set \(T_{s_d} = \{ (m_j, p_j, v_j) \}\), where each element corresponds to one dysarthric utterance, \(\bar{p}\) are phoneme embeddings extracted by fine-tuned speech encoder \(\Phi_{\overline{e}}\) and expanded with predicted normal duration, \(\bar{v}\) is predicted normal pitch, and \(e^{sv}\) is the speaker representation. Then Parallel WaveGAN (PWG) [27] is adopted as the neural vocoder to transform \(f_g(\bar{p}, \bar{v}, e^{sv}; \theta_g)\) to speech waveform. SV-DSR is a strong baseline as it can generate the speech with high intelligibility and naturalness. However, the speaker encoder is trained on normal speech, which limits its generalization ability to previously unseen dysarthric speech. Therefore, \(e^{sv}\) cannot effectively capture identity-related information of dysarthric speakers. Our experiments found that SV-DSR may even change the gender of speech after the reconstruction.
Then speaker encoder $\theta_{s^{asa}}$ of ASA-DSR and discriminator $\varphi$ are alternatively optimized with remaining networks frozen. On one hand, $\varphi$ is optimized to minimize the discrimination loss $L_{dis}$:

$$L_{dis} = E[(m_k, p_k), \gamma_k] - \gamma_{ad} \left( L_{dis}^{sv} + L_{dis}^{asa} \right)$$

(5)

where $f_d(\cdot, \gamma)$ is the posterior probability of mel-spectrograms reconstructed by SV-DSR. On the other hand, $\theta_{s^{asa}}$ is optimized to minimize the multi-task learning (MTL) loss $L_{MTL}$:

$$L_{MTL} = E[(m_k, p_k), \gamma_k] - \gamma_{ad} \left( L_{adapt} - \lambda L_{dis} \right)$$

(7)

where $\lambda$ is set to 1 empirically. The primary task minimizes the adaptation loss $L_{adapt}$ to force speaker encoder $\theta_{s^{asa}}$ to effectively capture speaker characteristics from the dysarthric speech, so that enhanced speaker similarity can be achieved in reconstructed mel-spectrograms $\hat{z}_k^{asa}$. The secondary task maximizes the discrimination loss $L_{dis}$ to force $\hat{z}_k^{asa}$ to have a similar distribution to $\hat{z}_k^{sv}$ that has high intelligibility and naturalness, which facilitates $\hat{z}_k^{asa}$ to maintain normal pronunciation patterns as $\hat{z}_k^{sv}$. As a result, the proposed ASA-DSR preserves the capacity of SV-DSR to reconstruct high-quality speech, while achieving improved capacity for preserving the speaker identity of the target dysarthric speaker $s_d$.

4. EXPERIMENTS

4.1. Experimental Settings

The datasets used in our experiments contain LibriSpeech [28], VCTK [29], VoxCeleb1 [30], VoxCeleb2 [31], LJSpeech [32] and UASPEECH [33]. Speech encoder $\Phi_p$ is pre-trained by 960h training data of LibriSpeech, prosody corrector is trained by the data of a healthy female speaker from LJSpeech, speech encoder $\theta_{sv}$ is trained by Librispeech, VoxCeleb1 and VoxCeleb2 with around 8.5K healthy speakers, speech generator $\theta_v$ and PWG vocoder are trained by VCTK. For dysarthric speech, two male speakers (M05, M07) and two female speakers (F04, F02) are selected from UASPEECH, where M05/F04 and M07/F02 have moderate and moderate-severe dysarthria respectively. We use the speech data of blocks 1 and 3 of each dysarthric speaker for fine-tuning speech encoder and ASA, and block 2 for testing. The inputs of speech encoder are 40-dim mel-spectrograms appended with deltas and delta-deltas which results in 120-dim vectors, the targets of speech generator are 80-dim mel-spectrograms, all mel-spectrograms are computed with 400-point Fourier transform, 25ms Hamming window and 10ms hop length. $F_0$ is extracted by the Pyworld tool [7], with the 10ms hop length. To stabilize the training and inference of $F_0$ predictor, we adopt the logarithmic scale of $F_0$. All acoustic features are normalized to have zero mean and unit variance.

The speech encoder, prosody corrector, speech encoder and speech generator adopt the same architectures as in [23], where the speech encoder contains 3-layer 256-dim LSTM followed by one fully-connected layer to obtain the 256-dim vector that is L2-normalized as the speaker representation [6]. The pre-training and fine-tuning of speech encoder are performed by Adam optimizer [34] with 1M and 2K steps respectively by using learning rate of 1 and batch size of 8. Both duration and $F_0$ predictors are trained by Adam optimizer [35] with 30K steps by using learning rate of 1e-3 and batch size of 16, speech generator is optimized in a similar way except that the training steps are set to 50K. The training of speaker encoder by using normal speech follows the scheme in [6]. Convolution-based discriminator of StarGAN [36] is used as the system discriminator and alternatively trained with the speaker encoder during ASA for 5K steps. Four DSR systems are compared: (1) SV-DSR; (2) ASA-DSR; (3) SA-DSR, which is an ablation system that performs speaker adaptation similar with ASA-DSR but without adversarial training; and (4) E2E-VC [18], which is an end-to-end DSR model via cross-modal knowledge distillation, where the speaker encoder used in SV-DSR is added to control the speaker identity.

4.2. Experimental Results and Analysis

4.2.1. Comparison Based on Speaker Similarity

Subjective tests are conducted to evaluate the speaker similarity of reconstructed speech, in terms of 5-point mean opinion score (MOS, 1-bad, 2-poor, 3-fair, 4-good, 5-excellent) rated by 20 subjects for 20 utterances randomly selected from each of four dysarthric speakers, and the scores are averaged and shown in Table 1. For E2E-VC and SV-DSR that use the SV-based speaker encoder to control the

Table 1. Comparison Results of MOS with 95% Confidence Intervals for Speaker Similarity.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>M05</th>
<th>F04</th>
<th>M07</th>
<th>F02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>4.93±0.01</td>
<td>4.89±0.02</td>
<td>4.95±0.01</td>
<td>4.96±0.01</td>
</tr>
<tr>
<td>E2E-VC</td>
<td>2.66±0.12</td>
<td>2.50±0.13</td>
<td>2.47±0.16</td>
<td>2.27±0.14</td>
</tr>
<tr>
<td>SV-DSR</td>
<td>2.70±0.14</td>
<td>2.27±0.10</td>
<td>2.55±0.14</td>
<td>1.88±0.13</td>
</tr>
<tr>
<td>SA-DSR</td>
<td>3.26±0.09</td>
<td>3.04±0.12</td>
<td>3.25±0.15</td>
<td>2.99±0.15</td>
</tr>
<tr>
<td>ASA-DSR</td>
<td>3.27±0.10</td>
<td>3.16±0.15</td>
<td>3.20±0.13</td>
<td>2.93±0.15</td>
</tr>
</tbody>
</table>

https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder
Table 2. Comparison Results of MOS with 95% Confidence Intervals for Speech Naturalness.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>M05</th>
<th>F04</th>
<th>M07</th>
<th>F02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>2.37±0.08</td>
<td>2.49±0.09</td>
<td>1.95±0.10</td>
<td>1.79±0.09</td>
</tr>
<tr>
<td>E2E-VC</td>
<td>3.64±0.11</td>
<td>3.40±0.13</td>
<td>3.58±0.12</td>
<td>3.35±0.12</td>
</tr>
<tr>
<td>SV-DSR</td>
<td><strong>3.88±0.11</strong></td>
<td><strong>3.92±0.10</strong></td>
<td><strong>3.80±0.10</strong></td>
<td><strong>3.79±0.09</strong></td>
</tr>
<tr>
<td>SA-DSR</td>
<td>3.56±0.09</td>
<td>3.22±0.14</td>
<td>3.67±0.11</td>
<td>3.38±0.12</td>
</tr>
<tr>
<td>ASA-DSR</td>
<td>3.84±0.09</td>
<td>3.86±0.12</td>
<td>3.79±0.09</td>
<td>3.75±0.11</td>
</tr>
</tbody>
</table>

Table 3. WER(Δ) (%) Results Comparison, Where 'Δ' Denotes the WER Reduction of Different Approaches Compared with Original Dysarthric Speech.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>M05</th>
<th>F04</th>
<th>M07</th>
<th>F02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>69.3(12.4)</td>
<td>72.0(23.9)</td>
<td>70.0(11.7)</td>
<td>69.6(21.4)</td>
</tr>
<tr>
<td>E2E-VC</td>
<td>65.3(30.6)</td>
<td>61.7(29.3)</td>
<td>95.9</td>
<td>62.7(32.9)</td>
</tr>
<tr>
<td>SV-DSR</td>
<td>95.9</td>
<td>69.6(21.4)</td>
<td>81.7</td>
<td>62.5(28.5)</td>
</tr>
<tr>
<td>SA-DSR</td>
<td>95.9</td>
<td>81.7</td>
<td><strong>95.9</strong></td>
<td>62.5(28.5)</td>
</tr>
<tr>
<td>ASA-DSR</td>
<td>95.9</td>
<td>81.7</td>
<td>95.9</td>
<td><strong>62.5(30.1)</strong></td>
</tr>
</tbody>
</table>

speaker identity, lower speaker similarity is achieved. Through our listening tests, the gender of reconstructed speech by E2E-VC and SV-DSR may be changed especially for female speakers, this shows the limited generalization ability of the SV-based speaker encoder to extract effective speaker representations from the dysarthric speech. However, with the speaker adaptation to fine-tune the speaker encoder, both SA-DSR and ASA-DSR can accurately preserve the gender with improved speaker similarity, showing the necessity of using dysarthric speech data to fine-tune the speaker encoder to effectively capture identity-related information of dysarthric speech.

4.2.2. Comparison Based on Speech Naturalness

Table 2 gives the MOS results of naturalness of original or reconstructed speech from different systems. We can see that all DSR systems improve the naturalness of original dysarthric speech, and SV-DSR achieves highest speech naturalness scores for all speakers, which shows the effectiveness of explicit prosody correction to generate the speech with stable and accurate prosody. By using the speaker adaptation without adversarial training, SA-DSR achieves lower naturalness improvements, due to partial dysarthric pronunciation patterns incorporated into the reconstructed speech. This issue can be effectively alleviated by using the proposed ASA to align the statistical distributions of reconstructed speech from ASA-DSR and SV-DSR, which facilitates ASA-DSR to generate high-quality speech that achieves comparable naturalness with SV-DSR.

4.2.3. Comparison Based on Speech Intelligibility

Objective evaluation of speech intelligibility is conducted by using a publicly released speech recognition model, i.e., Jasper [37], to test the word error rate (WER) with greedy decoding, and the results are shown in Table 3. Compared with original dysarthric speech, SV-DSR achieves largest WER reduction for all dysarthric speakers, showing the effectiveness of prosody correction to improve the speech intelligibility. Compared with SV-DSR, the adaptation version of SV-DSR without adversarial training, i.e., SA-DSR, has smaller WER reduction, which is caused by the incorporation of dysarthric speaking characteristics into reconstructed speech. However, with the proposed ASA to alleviate this issue, ASA-DSR outperforms E2E-VC and SA-DSR and matches the performance of SV-DSR, leading to 22.3% and 31.5% absolute WER reduction on average for speakers M05/F04 and M07/F02 that have moderate and moderate-severe dysarthria respectively.

Fig. 3. AB preference test results with 95% confidence intervals for different combinations of phoneme duration and F0, where 'GG' denotes Ground-truth duration and Ground-truth F0, 'GP' denotes Ground-truth duration and Predicted F0, and 'PP' denotes Predicted duration and Predicted F0.

4.2.4. Influence of Phoneme Duration and F0

We also conduct an ablation study to investigate how the phoneme duration and F0 influence the quality of reconstructed speech by the proposed ASA-DSR system. Three combinations of phoneme duration and F0 are used to generate the speech. We perform AB preference tests, where listeners are required to select the utterance that sounds more normal, i.e., more stable prosody and precise articulation, from two utterances generated by two different combinations. The results are illustrated in Fig. 3. For the comparison ‘GG vs. GP’ (i.e., Ground-truth duration and Predicted F0 versus Ground-truth duration and Predicted F0) of different speakers, more reconstructed speech samples are favored by using predicted normal F0 (p-values ≪ 0.05). For the comparison ‘GP vs PP’ (i.e., Ground-truth duration and Predicted F0 versus Predicted duration and Predicted F0), using the predicted normal duration can significantly improve speech quality especially for speakers M05, M07 and F02 who have abnormally slow speaking speed. This shows that both phoneme duration and F0 affect speech normality, and the prosody corrector in ASA-DSR derives normal values of phoneme duration and F0, which facilitate the reconstruction of speech to have normal prosodic patterns.

5. CONCLUSIONS

This paper presents a DSR system based on a novel multi-task learning strategy, i.e., ASA, to simultaneously preserve the speaker identity and maintain high speech quality. This is achieved by a primary task (i.e., speaker adaptation) to facilitate the speaker encoder to capture speaker characteristics from the dysarthric speech, and a secondary task (i.e., adversarial training) to avoid the incorporation of dysarthric speaking patterns into reconstructed speech. Experiments show that the proposed ASA-DSR can effectively achieve dysarthria reductions with improved naturalness and intelligibility, while speaker identity can be effectively maintained with 0.73 and 0.85 absolute MOS improvements of speaker similarity over the strong baseline SV-DSR, for speakers with moderate and moderate-severe dysarthria respectively.

6. ACKNOWLEDGEMENTS

This research is supported partially by the HKSAR Research Grants Council’s General Research Fund (Ref Number 14208817) and also partially by the Centre for Perceptual and Interactive Intelligence, a CUHK InnoCentre.
References


