# SPEAKER IDENTITY PRESERVATION IN DYSARTHRIC SPEECH RECONSTRUCTION BY ADVERSARIAL SPEAKER ADAPTATION

*Disong Wang*[1], *Songxiang Liu*[1], *Xixin Wu*[1], *Hui Lu*[1], *Lifa Sun*[2], *Xunying Liu*[1], *Helen Meng*[1,3]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]SpeechX Limited, Shenzhen, China
[3]Centre for Perceptual and Interactive Intelligence, Hong Kong SAR, China

{dswang, sxliu, wuxx, luhui, xyliu, hmmeng}@se.cuhk.edu.hk, lfsun@speechx.cn

## 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[1].

***Index Terms***— Dysarthric speech reconstruction, voice conversion, adversarial speaker adaptation, speaker identity

## 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 [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 [2].

The goal of VC is to convert non-linguistic or para-linguistic factors such as speaker identity [3], prosody [4], emotion [5] and accent [6]. VC has also been widely applied in reconstructing different kinds of impaired speech including esophageal speech [7, 8], electrolaryngeal speech [9, 10], hearing-impaired speech [11] and dysarthric speech [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 [12, 13]. Statistical VC automatically maps the features of dysarthric speech to those of normal speech, where typical approaches contain Gaussian mixture model [14], non-negative matrix factorization [15, 16], partial least squares [17], and deep learning methods including sequence-to-sequence (seq2seq) models [18–20] and gated convolutional networks [21]. 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 [22, 23] use a speaker representation to control the speaker identity of reconstructed speech, where the speaker encoder (SE) proposed in our previous work [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 [23] by using adversarial speaker adaptation (ASA). The DSR system in [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-

[1]Audio samples: https://wendison.github.io/ASA-DSR-demo/

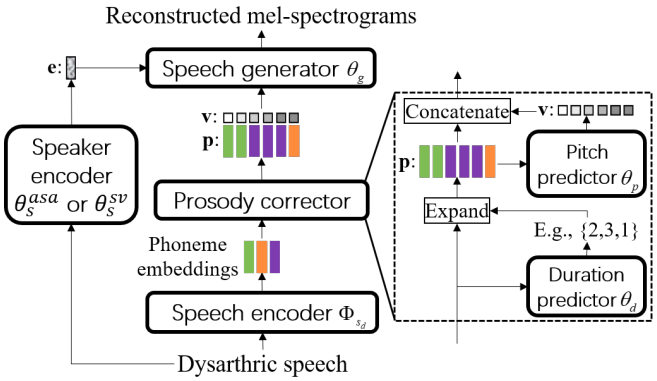


**Fig. 1**. The architecture for the SV-DSR system. The ASA-DSR has the same architecture, except that the speaker encoder $\theta_s^{sv}$ of SV-DSR is trained for a SV task on the normal speech, while $\theta_s^{asa}$ of ASA-DSR is first initialized from $\theta_s^{sv}$ and then fine-tuned by the dysarthric speech via proposed ASA.

nator, by min-maximizing a discrimination loss to classify whether the mel-spectrograms are reconstructed by ASA-DSR or SV-DSR, which forces the reconstructed speech from ASA-DSR to have a distribution close to that of SV-DSR without dysarthric speaking 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 for generating high-quality speech with enhanced speaker similarity.

## 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 phoneme sequence: (1) Pre-training on large-scale normal speech data; (2) Fine-tuning on the speech of a certain dysarthric speaker $s_d$ to achieve accurate phoneme prediction. The outputs of pre-trained speech encoder $\Phi_p$ or fine-tuned speech encoder $\Phi_{s_d}$ are used as phoneme embeddings that denote phoneme probability distributions.

**Prosody corrector:** As abnormal duration and pitch are two essential prosody factors that contribute to dysarthric speech [14], a prosody corrector is used to amend the abnormal prosody to a normal form, it contains two predictors to respectively infer normal phoneme duration and pitch (i.e., fundamental frequency ($F_0$)). The prosody corrector is trained by a healthy speaker's speech with normal prosodic patterns: (1) Given the phoneme embeddings extracted by the speech encoder $\Phi_p$ as inputs, the phoneme duration predictor $\theta_d$ is trained to infer the normal phoneme durations that are obtained from force-alignment via Montreal Forced Aligner toolkit [25]; (2) The ground-truth phoneme durations are used to align phoneme embeddings and $F_0$ as shown in Fig. 1, the expanded phoneme embeddings are denoted as $\mathbf{p}$ and fed into the pitch predictor $\theta_p$ to infer normal $F_0$ that is denoted by $\mathbf{v}$. The prosody corrector is expected to take in phoneme embeddings extracted from dysarthric speech to infer normal values of phoneme duration and $F_0$, which can be used as canonical values to replace their abnormal counterparts for generating the speech with normal prosodic patterns.

**Speaker encoder:** The speaker encoder, $\theta_s^{sv}$, is trained on a SV task to capture speaker characteristics. $\theta_s^{sv}$ takes in mel-spectrograms $\mathbf{m}$ of one utterance with arbitrary length to produce a single vector as speaker representation: $\mathbf{e} = f_s(\mathbf{m}; \theta_s^{sv})$. Following the training scheme in [6], $\theta_s^{sv}$ is optimized to minimize a generalized end-to-end loss [26] by using normal speech data that is easily acquired from thousands of healthy speakers.

**Speech generator:** The speech generator with parameters $\theta_g$ predicts mel-spectrograms as: $\mathbf{z} = f_g(\mathbf{p}, \mathbf{v}, \mathbf{e}; \theta_g)$. To generate normal speech, the speech generator is trained by using normal speech data from a set of healthy speakers $\mathcal{S}$. Each speaker $s_i \sim \mathcal{S}$ has the training data set $\mathcal{T}_{s_i} = \{(\mathbf{m}_j, \mathbf{p}_j, \mathbf{v}_j)\}$, where each sample corresponds to one utterance and contains mel-spectrograms $\mathbf{m}_j$, expanded phoneme embeddings $\mathbf{p}_j$ and pitch features $\mathbf{v}_j$. Then speech generator is optimized by minimizing the generation loss $\mathcal{L}_{gen}^{sv}$, i.e., the L2-norm between the predicted mel-spectrograms $\mathbf{z}_j^{sv}$ and $\mathbf{m}_j$:

$$\mathcal{L}_{gen}^{sv} = \mathbb{E}_{s_i \sim \mathcal{S}, (\mathbf{m}_j, \mathbf{p}_j, \mathbf{v}_j) \sim \mathcal{T}_{s_i}} \left\| \mathbf{z}_j^{sv} - \mathbf{m}_j \right\|_2 \quad (1)$$

$$\mathbf{z}_j^{sv} = f_g\left(\mathbf{p}_j, \mathbf{v}_j, \mathbf{e}_j^{sv}; \theta_g\right), \mathbf{e}_j^{sv} = f_s\left(\mathbf{m}_j; \theta_s^{sv}\right) \quad (2)$$

During the reconstruction phase, the SV-DSR system takes in the dysarthric speech of speaker $s_d$ to generate reconstructed mel-spectrograms as $f_g(\widetilde{\mathbf{p}}, \widetilde{\mathbf{v}}, \mathbf{e}^{sv}; \theta_g)$, where $\widetilde{\mathbf{p}}$ are phoneme embeddings extracted by fine-tuned speech encoder $\Phi_{s_d}$ and expanded with predicted normal duration, $\widetilde{\mathbf{v}}$ is predicted normal pitch, and $\mathbf{e}^{sv}$ is the speaker representation. Then Parallel WaveGAN (PWG) [27] is adopted as the neural vocoder to transform $f_g(\widetilde{\mathbf{p}}, \widetilde{\mathbf{v}}, \mathbf{e}^{sv}; \theta_g)$ 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, $\mathbf{e}^{sv}$ 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.

## 3. PROPOSED APPROACH: ASA-DSR

The proposed approach of adversarial speaker adaptation (ASA), as illustrated in Fig. 2, 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 $\mathcal{T}_{s_d} = \{(\mathbf{m}_k, \mathbf{p}_k, \mathbf{v}_k)\}$, where each element corresponds to one dysarthric utterance, $\mathbf{p}_k$ are phoneme embeddings extracted by $\Phi_{s_d}$ and expanded with dysarthric duration, $\mathbf{v}_k$ is dysarthric pitch, their normal counterparts can be obtained via the prosody corrector as $\tilde{\mathbf{p}}_k$ and $\tilde{\mathbf{v}}_k$, respectively. SV-DSR and ASA-DSR generate reconstructed mel-spectrograms as $\tilde{\mathbf{z}}_k^{sv}$ and $\tilde{\mathbf{z}}_k^{asa}$ respectively:

$$\tilde{\mathbf{z}}_k^{sv} = f_g(\tilde{\mathbf{p}}_k, \tilde{\mathbf{v}}_k, \mathbf{e}_k^{sv}; \theta_g), \ \tilde{\mathbf{z}}_k^{asa} = f_g(\tilde{\mathbf{p}}_k, \tilde{\mathbf{v}}_k, \mathbf{e}_k^{asa}; \theta_g) \quad (3)$$

where $\mathbf{e}_k^{sv}$ and $\mathbf{e}_k^{asa}$ are respectively produced from the speaker encoders $\theta_s^{sv}$ (from SV-DSR) and $\theta_s^{asa}$ (from ASA-DSR) to control the speaker identity. Besides, ASA-DSR predicts dysarthric mel-

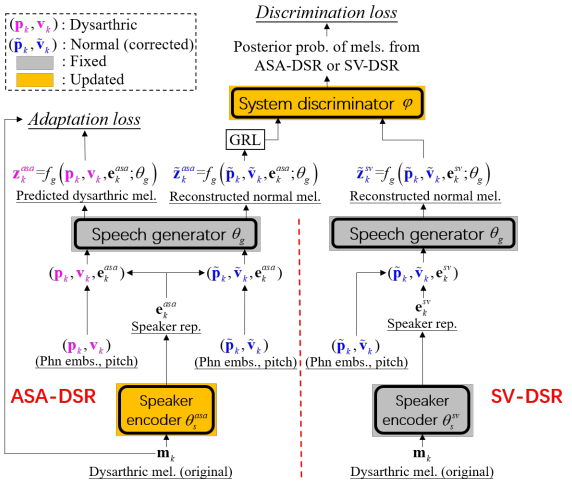


**Fig. 2**. Diagram of ASA. $\mathbf{m}_k$ is the mel-spectrogram of dysarthric speech. $\mathbf{p}_k$ is phoneme embedding expanded with dysarthric duration, $\mathbf{v}_k$ is the pitch of dysathric speech, their normal counterparts are $\tilde{\mathbf{p}}_k$ and $\tilde{\mathbf{v}}_k$ obtained via prosody corrector. GRL is gradient reversal layer that passes the data during forward propagation and inverts the sign of gradient during backward propagation. Only parameters of $\theta_s^{asa}$ and $\varphi$ are updated during the ASA process.

spectrograms as $\mathbf{z}_k^{asa}$ used for adaptation:

$$\mathbf{z}_k^{asa} = f_g(\mathbf{p}_k, \mathbf{v}_k, \mathbf{e}_k^{asa}; \theta_g) \tag{4}$$

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 $\mathcal{L}_{dis}$:

$$\mathcal{L}_{dis} = \mathbb{E}_{(\mathbf{m}_k, \mathbf{p}_k, \mathbf{v}_k) \sim \mathcal{T}_{s_d}} \left\{ \mathcal{L}_{dis}^{sv} + \mathcal{L}_{dis}^{asa} \right\} \tag{5}$$

$$\mathcal{L}_{dis}^{sv} = \log\left(1 - f_d\left(\tilde{\mathbf{z}}_k^{sv}; \varphi\right)\right), \mathcal{L}_{dis}^{asa} = \log f_d\left(\tilde{\mathbf{z}}_k^{asa}; \varphi\right) \tag{6}$$

where $f_d(*; \varphi)$ 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 $\mathcal{L}_{MTL}$:

$$\mathcal{L}_{MTL} = \mathbb{E}_{(\mathbf{m}_k, \mathbf{p}_k, \mathbf{v}_k) \sim \mathcal{T}_{s_d}} \left\{ \mathcal{L}_{adapt} - \lambda \mathcal{L}_{dis} \right\} \tag{7}$$

$$\mathcal{L}_{adapt} = \left\| \mathbf{z}_k^{asa} - \mathbf{m}_k \right\|_2 \tag{8}$$

where $\lambda$ is set to 1 empirically. The primary task minimizes the adaptation loss $\mathcal{L}_{adapt}$ 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 $\tilde{\mathbf{z}}_k^{asa}$. The secondary task maximizes the discrimination loss $\mathcal{L}_{dis}$ to force $\tilde{\mathbf{z}}_k^{asa}$ to have a similar distribution to $\tilde{\mathbf{z}}_k^{sv}$ that has high intelligibility and naturalness, which facilitates $\tilde{\mathbf{z}}_k^{asa}$ to maintain normal pronunciation patterns as $\tilde{\mathbf{z}}_k^{sv}$ . 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, speaker encoder $\theta_s^{sv}$ is trained by Librispeech, VoxCeleb1 and VoxCeleb2 with around 8.5K healthy speakers, speech generator $\theta_g$ 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-spectrogtams are computed with 400-point Fourier transform, 25ms Hanning window and 10ms hop length. $F_0$ is extracted by the Pyworld toolkit[2] 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, speaker encoder and speech generator adopt the same architectures as in [23], where the speaker 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 Adadelta 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.

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

| Approaches | M05 | F04 | M07 | F02 |
|---|---|---|---|---|
| Original | 4.93±0.01 | 4.89±0.02 | 4.95±0.01 | 4.96±0.01 |
| E2E-VC | 2.66±0.12 | 2.50±0.13 | 2.47±0.16 | 2.27±0.14 |
| SV-DSR | 2.70±0.14 | 2.27±0.10 | 2.55±0.14 | 1.88±0.13 |
| SA-DSR | 3.26±0.09 | 3.04±0.12 | **3.25±0.15** | **2.99±0.15** |
| ASA-DSR | **3.27±0.10** | **3.16±0.15** | 3.20±0.13 | 2.93±0.15 |

### 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

[2]https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

**Table 2**. Comparison Results of MOS with 95% Confidence Intervals for Speech Naturalness.

| Approaches | M05 | F04 | M07 | F02 |
|---|---|---|---|---|
| Original | 2.37±0.08 | 2.49±0.09 | 1.95±0.10 | 1.79±0.09 |
| E2E-VC | 3.64±0.11 | 3.40±0.13 | 3.58±0.12 | 3.35±0.12 |
| SV-DSR | **3.88±0.11** | **3.92±0.10** | **3.80±0.10** | **3.79±0.09** |
| SA-DSR | 3.56±0.09 | 3.22±0.14 | 3.67±0.11 | 3.38±0.12 |
| ASA-DSR | 3.84±0.09 | 3.86±0.12 | 3.79±0.09 | 3.75±0.11 |

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

| Approaches | M05 | F04 | M07 | F02 |
|---|---|---|---|---|
| Original | 91.0 | 81.7 | 95.6 | 95.9 |
| E2E-VC | 69.8(21.2) | 69.3(12.4) | 73.1(22.5) | 72.0(23.9) |
| SV-DSR | **61.7(29.3)** | **64.6(17.1)** | **62.7(32.9)** | **65.3(30.6)** |
| SA-DSR | 69.6(21.4) | 70.0(11.7) | 67.8(27.8) | 67.2(28.7) |
| ASA-DSR | 62.5(28.5) | 65.6(16.1) | **62.7(32.9)** | 65.8(30.1) |

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.

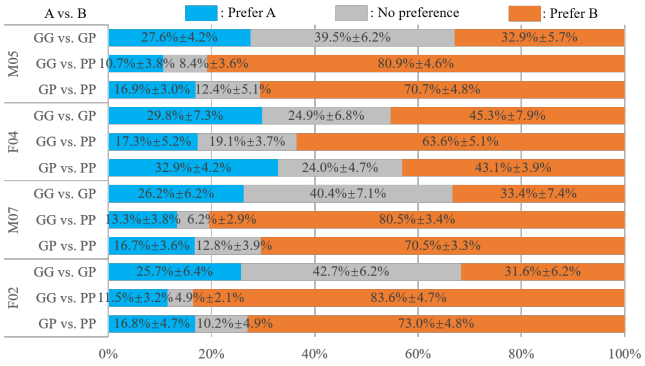


**Fig. 3**. AB preference test results with 95% confidence intervals for different combinations of phoneme duration and $F_0$, where 'GG' denotes Ground-truth duration and Ground-truth $F_0$, 'GP' denotes Ground-truth duration and Predicted $F_0$, and 'PP' denotes Predicted duration and Predicted $F_0$.

#### *4.2.4. Influence of Phoneme Duration and $F_0$*

We also conduct an ablation study to investigate how the phoneme duration and $F_0$ influence the quality of reconstructed speech by the proposed ASA-DSR system. Three combinations of phoneme duration and $F_0$ 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 $F_0$ versus Ground-truth duration and Predicted $F_0$) of different speakers, more reconstructed speech samples are favored by using predicted normal $F_0$ (p-values $\ll 0.05$). For the comparison 'GP vs PP' (i.e., Ground-truth duration and Predicted $F_0$ versus Predicted duration and $F_0$), 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 $F_0$ affect speech normality, and the prosody corrector in ASA-DSR derives normal values of phoneme duration and $F_0$, 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


[1] Y Yunusova, G Weismer, JR Westbury, and MJ Lindstrom, "Articulatory movements during vowels in speakers with dysarthria and healthy controls.," *Journal of Speech, Language, and Hearing Research: JSLHR*, vol. 51, no. 3, pp. 596–611, 2008.

[2] Junichi Yamagishi, Christophe Veaux, Simon King, and Steve Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[3] Seyed Hamidreza Mohammadi and Alexander Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[4] Dimitrios Rentzos, S Vaseghi, E Turajlic, Qin Yan, and Ching-Hsiang Ho, "Transformation of speaker characteristics for voice conversion," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*. IEEE, 2003, pp. 706–711.

[5] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Gmm-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 134–138, 2012.

[6] Songxiang Liu, Disong Wang, Yuewen Cao, Lifa Sun, Xixin Wu, Shiyin Kang, Zhiyong Wu, Xunying Liu, Dan Su, Dong Yu, et al., "End-to-end accent conversion without using native utterances," in *ICASSP*. IEEE, 2020, pp. 6289–6293.

[7] Hironori Doi, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models," *IEICE TRANSACTIONS on Information and Systems*, vol. 93, no. 9, pp. 2472–2482, 2010.

[8] Luis Serrano, Sneha Raman, David Tavarez, Eva Navas, and Inma Hernaez, "Parallel vs. non-parallel voice conversion for esophageal speech.," in *INTERSPEECH*, 2019, pp. 4549–4553.

[9] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[10] Kazuhiro Kobayashi and Tomoki Toda, "Electrolaryngeal speech enhancement with statistical voice conversion based on cldnn," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2115–2119.

[11] Fadi Biadsy, Ron J Weiss, Pedro J Moreno, Dimitri Kanevsky, and Ye Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," *arXiv preprint arXiv:1904.04169*, 2019.

[12] Frank Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech," in *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, 2011, pp. 11–21.

[13] S Arun Kumar and C Santhosh Kumar, "Improving the intelligibility of dysarthric speech towards enhancing the effectiveness of speech therapy," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2016, pp. 1000–1005.

[14] Alexander B Kain, John-Paul Hosom, Xiaochuan Niu, Jan PH Van Santen, Melanie Fried-Oken, and Janice Staehely, "Improving the intelligibility of dysarthric speech," *Speech communication*, vol. 49, no. 9, pp. 743–759, 2007.

[15] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Consonant enhancement for articulation disorders based on non-negative matrix factorization," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.

[16] Aihara, Ryo and Takashima, Ryoichi and Takiguchi, Tetsuya and Ariki, Yasuo, "Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization," in *ICASSP*. IEEE, 2013, pp. 8037–8040.

[17] Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, "Phoneme-discriminative features for dysarthric speech conversion.," in *INTERSPEECH*, 2017, pp. 3374–3378.

[18] Disong Wang, Jianwei Yu, Xixin Wu, Songxiang Liu, Lifa Sun, Xunying Liu, and Helen Meng, "End-to-end voice conversion via cross-modal knowledge distillation for dysarthric speech reconstruction," in *ICASSP*. IEEE, 2020, pp. 7744–7748.

[19] Rohan Doshi, Youzheng Chen, Liyang Jiang, Xia Zhang, Fadi Biadsy, Bhuvana Ramabhadran, Fang Chu, Andrew Rosenberg, and Pedro J Moreno, "Extending parrotron: An end-to-end, speech conversion and speech recognition model for atypical speech," in *ICASSP*. IEEE, 2021, pp. 6988–6992.

[20] Zhehuai Chen, Bhuvana Ramabhadran, Fadi Biadsy, Xia Zhang, Youzheng Chen, Liyang Jiang, Fang Chu, Rohan Doshi, and Pedro J. Moreno, "Conformer Parrotron: A Faster and Stronger End-to-End Speech Conversion and Recognition Model for Atypical Speech," in *INTERSPEECH*, 2021, pp. 4828–4832.

[21] Chen-Yu Chen, Wei-Zhong Zheng, Syu-Siang Wang, Yu Tsao, Pei-Chun Li, and YH Li, "Enhancing intelligibility of dysarthric speech using gated convolutional-based voice conversion system," *INTERSPEECH*, 2020.

[22] Wen-Chin Huang, Kazuhiro Kobayashi, Yu-Huai Peng, Ching-Feng Liu, Yu Tsao, Hsin-Min Wang, and Tomoki Toda, "A preliminary study of a two-stage paradigm for preserving speaker identity in dysarthric voice conversion," *arXiv preprint arXiv:2106.01415*, 2021.

[23] Disong Wang, Songxiang Liu, Lifa Sun, Xixin Wu, Xunying Liu, and Helen Meng, "Learning explicit prosody models and deep speaker embeddings for atypical voice conversion," *arXiv preprint arXiv:2011.01678*, 2021.

[24] Zhong Meng, Jinyu Li, and Yifan Gong, "Adversarial speaker adaptation," in *ICASSP*. IEEE, 2019, pp. 5721–5725.

[25] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *INTERSPEECH*, 2017, vol. 2017, pp. 498–502.

[26] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*. IEEE, 2018, pp. 4879–4883.

[27] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*. IEEE, 2020, pp. 6199–6203.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[29] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[30] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[31] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[32] Keith Ito et al., "The lj speech dataset," 2017.

[33] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S Huang, Kenneth Watkin, and Simone Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[34] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[35] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[37] Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *arXiv preprint arXiv:1904.03288*, 2019.