Adversarial Data Augmentation for Disordered Speech Recognition

Zengrui Jin1*, Mengzhe Geng1*, Xurong Xie2, Jianwei Yu1, Shansong Liu1, Xunying Liu1, Helen Meng1

1The Chinese University of Hong Kong, Hong Kong SAR, China
2Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

{zrjin, mzgeng, jwyu, ssliu, xyliu, hmmeng}@se.cuhk.edu.hk, xr.xie@siat.ac.cn

Abstract

Automatic recognition of disordered speech remains a highly challenging task to date. The underlying neuro-motor conditions, often compounded with co-occurring physical disabilities, lead to the difficulty in collecting large quantities of impaired speech required for ASR system development. To this end, data augmentation techniques play a vital role in current disordered speech recognition systems. In contrast to existing data augmentation techniques only modifying the speaking rate or overall shape of spectral contour, fine-grained spectro-temporal differences between disordered and normal speech are modelled using deep convolutional generative adversarial networks (DCGAN) during data augmentation to modify normal speech spectra into those closer to disordered speech. Experiments conducted on the UASpeech corpus suggest the proposed adversarial data augmentation approach consistently outperformed the baseline augmentation methods using tempo or speed perturbation on a state-of-the-art hybrid DNN system. An overall word error rate (WER) reduction up to 3.05% (9.7% relative) was obtained over the baseline system using no data augmentation. The final hidden learning unit contribution (LHUC) speaker adapted system using the best adversarial augmentation approach gives an overall WER of 25.89% on the UASpeech test set of 16 dysarthric speakers.

Index Terms: Speech Disorders, Speech Recognition, Data Augmentation, Speaker Adaptive Training

1. Introduction

Despite the rapid progress of automatic speech recognition (ASR) technologies targeting normal speech task domains [1–5] in the past few decades, accurate recognition of disordered speech remains a highly challenging task to date [6–12]. Speech disorders such as dysarthria are caused by a range of neuro-motor conditions including cerebral palsy, amyotrophic lateral sclerosis, Parkinson disease, stroke or traumatic brain injuries [13], leading to weakness or paralysis of muscles used in articulation and reduced intelligibility of speech for human listeners.

Disordered speech presents a range of challenges to current deep neural networks (DNNs) based ASR technologies that are predominantly targeting normal speech. First, a large mismatch between disordered and normal speech is often observed. Such difference systematically manifests themselves in articulatory imprecision, decreased volume and clarity, breathy and hoarse voice, changes in pitch, increased dysfluencies and lower speaking rate [14,15]. State-of-the-art ASR systems designed for normal speech often produce very high recognition error rate above 50% when being applied to impaired speech [12, 16]. Second, the underlying neuro-motor conditions, often compounded with co-occurring physical disabilities and fatigue when speaking, lead to the difficulty in collecting large quantities of disordered speech required for ASR system development. For data intensive deep learning technologies widely used in current speech recognition systems, large quantities of well-matched, in-domain speech data are essential.

To this end, data augmentation techniques play a vital role to address the above data sparsity issue. In the context of normal speech recognition tasks, data augmentation techniques have been widely studied. By expanding the limited training data using, for example, tempo, vocal tract length or speed perturbation [17–20], stochastic feature mapping [21], cross-domain feature adaptation [22], simulation of noisy and reverberated speech to improve environmental robustness [23], and more advanced generative adversarial network (GAN) based augmentation by synthesis [24–26], and end-to-end back translation in end-to-end systems [27], the coverage of the augmented training data and the resulting speech recognition systems generalization can be improved.

In contrast, so far only limited research on data augmentation approaches targeting disordered speech recognition has been conducted. Motivated by the spectro-temporal level differences between disordered speech and normal speech such as slower speaking rates, recent research in this direction has been largely focused on tempo-stretching [9, 28], vocal tract length perturbation (VTLP) [19], and speed perturbation [20] of normal speech recorded from healthy control speakers. The resulting “disordered like” speech carrying a slower speaking rate and modified overall vocal tract spectral shape is then used to augment the limited dysarthric speech training data. Alternative approaches based on cross-domain DNN adaptation [6, 8] and voice conversion [29] have also been investigated. An investigation over a range of speech augmentation approaches for disordered speech recognition reported in [30] suggested the combined use of speaker dependent and independent speed perturbation factors produced the largest performance improvements.

One issue associated with the existing data augmentation approaches, for example, temporal or speed perturbation, is the lack of ability to model fine-grained differences between normal and impaired speech during the augmentation process. Although the overall decrease in speaking rate and speech volume as well as changes in spectral envelope can be characterized using speed perturbation, other prominent features associated with disordered speech including articulatory imprecision, decreased vocal clarity, breathy and hoarse voice as well as increased dysfluencies cannot be fully captured.

In order to address this issue, and inspired by the recent successful application of GAN [31, 32] across a wide range of speech processing tasks including, but limited to, speech enhancement [33, 34], speaker verification [35], speech synthesis and voice conversion [36,37], code-switching sentence genera-
tion [38] and speech emotion recognition [39], speaker dependent GANs are trained to transform tempo or speed perturbed healthy speech spectra into those of individual target dysarthric speakers. The resulting data augmentation process not only allows the overall change in speaking rate, speech volume and spectral envelope to be simulated as in tempo or speed perturbation, but also ensures further fine-grained spectro-temporal characteristics associated with disordered speech including articulatory impairment, decreased clarity, breathy and hoarse voice as well as increased dysfluencies and pauses to be emulated in the final augmented data.

Experiments conducted on the largest publicly available UASpeech dysarthric speech corpus [40] suggest the proposed adversarial data augmentation approach consistently outperformed the baseline augmentation methods using tempo or speed perturbation on state-of-the-art hybrid and end-to-end systems. An overall WER reduction up to 3.05% (9.7% relative) was obtained over the baseline system using no data augmentation. The final LHUC [41]1 speaker adapted system using the best adversarial augmentation approach gave an overall WER of 25.89% on the UASpeech test set of 16 dysphonic speakers. To the best of our knowledge, this is the best performance reported so far on the UASpeech task.

The major contributions of this paper are listed below. To the best of our knowledge, this is the first work to systematically investigate adversarial learning based data augmentation for disordered speech recognition tasks. In contrast, previous application of GANs were primarily studied in normal speech task domains including speech enhancement [33, 34], speaker verification [35], speech synthesis and voice conversion [36, 37], code-switching sentence generation [38] and speech emotion recognition [39]. Furthermore, previous researches on GANs for disordered speech processing were limited to impaired speech restoration [42] and classification [25], but not for ASR system development as considered in this paper. Finally, the final system constructed using the best adversarial augmentation approach in this paper produced the lowest WER of 25.89% on the UASpeech task published so far.

The rest of this paper is organized as follows. Conventional data augmentation techniques based on tempo, speed are presented in Section 2. Section 3 proposes GAN based disordered speech data augmentation. Section 4 presents experiments and results on the UASpeech database. The last section concludes and discusses possible future works.

2. Disordered Speech Data Augmentation

In this section, we present two traditional data augmentation methods, tempo perturbation and speed perturbation for disordered speech recognition. These serve as the baseline augmentation approaches modifying the overall speaking rate, volume and spectral shape, and the necessary time alignment between normal and dysarthric speech utterances to facilitate GAN based augmentation model training in Section 3.

2.1. Tempo Perturbation: Tempo perturbation modifies the duration of the input time-domain signal \( x(t) \), while keeping the overall shape of its spectrum untouched [17, 18]. This is first implemented by decomposing \( x(t) \) into short analysis blocks. These blocks are relocated along the time axis to construct the perturbed output \( y(t) \). This can be obtained using the wave-

1 In order to facilitate adaptation to unseen speakers in practical system deployment, rather than building speaker dependent systems, speaker adaptively trained systems were constructed.

Figure 1: Example spectrogram of (a) control, (b) dysarthric, (c) tempo or (d) speed perturbed control speech, (e) tempo-GAN or (f) speed-GAN generated “dysarthric” speech.

form similarity overlap-add (WSOLA) algorithm [17], which maximizes the similarity between \( x(t) \) and \( y(t) \) by finding the optimal position of each analysis block iteratively.

Suppose the time-domain audio signal \( x(t) \) is decomposed into short analysis blocks \( \tilde{x}_m(r) \). Given the analysis hopsize \( H_m \), these blocks are equally spaced along the time axis. The synthesis blocks \( \tilde{y}(r) \) are relocated along the time axis given the synthesis hopsize \( H_s \) and perturbation factor \( \alpha \) as:

\[
H_s = \alpha \times H_m
\]

WSOLA iteratively updates the positions of \( \tilde{x}_m(r) \). For each \( \tilde{x}_m(r) \), its center is shifted by \( \Delta_m \in [-\Delta_{max}, \Delta_{max}] \) along the time axis. By maximizing the cross-correlation between \( \tilde{x}_m(r) \) and \( \tilde{x}_{m-1}(r) \), WSOLA obtains the optimal value of \( \Delta_m \). This method guarantees that when both frames use the synthesis hopsize \( H_s \), the optimal alignment can be achieved between the periodic structures of the perturbed analysis blocks and those of the previously copied synthesis blocks in the overlapping region. The Hann window function \( \omega(r) \) is then applied on the perturbed analysis block to compute the synthesis block \( \tilde{y}_m(r) \) [17]. After finishing all iterations, the synthesis blocks are processed to reconstruct the actual time-domain perturbed output signal \( y(t) \) using overlap and add (OLA) [43]. The perturbed signal \( y(t) \) has a different duration, for example, representing a slower speaking rate of an impaired speaker, but keeps the same overall spectral shape as \( x(t) \).

2.2. Speed Perturbation: Speed perturbation [20] modifies the input time domain speech signal \( x(t) \) by scaling the sampling resolution via a perturbation factor \( \alpha \). The resulting speed perturbed signal output \( y(t) \) is given as:

\[
y(t) = x(at)
\]

The above time-domain signal modification is equivalent to the following performing in the frequency domain:

\[
X(f) \rightarrow \frac{1}{\alpha} X \left( \frac{f}{\alpha} \right)
\]

where \( X(f) \) and \( \frac{1}{\alpha} X \left( \frac{f}{\alpha} \right) \) denote the Fourier transform of \( x(t) \) and \( y(t) \) respectively. Speed perturbation changes the audio duration and overall spectral shape at the same time [20] to simulate, for example, a slower speaking rate, changes in speech volume and formant position of an impaired speaker.

2.3. Speaker Dependent Data Augmentation: During disordered speech augmentation, a combined use of speaker independent, global tempo or speed perturbation factors applied to disordered speech, for example, \{0.9, 1.0, 1.1\}, in common with the data perturbation methods widely used in normal ASR tasks [44], and impaired speaker dependent (SD) perturbation of
3. GAN Based Data Augmentation

As discussed in Section 1, conventional tempo or speed perturbation of Section 2 only characterize an overall decrease in speaking rate, speech volume and changes in spectral envelope found in disordered speech. In this section, DCGANs are proposed to further inject other prominent features associated with disordered speech such as articulatory imprecision, breathy and hoarse voice, into tempo or speed perturbed normal speech Filter Bank (FBank). For instance, in contrast to the example healthy speech segments FBank of word “burrows” in Fig.1(a), the comparable dysarthric speech spectrogram in Fig.1(b) contains not only incorrectly weakened formants indicating articulatory imprecision, but also some energy distributed across higher frequencies at both the start and end of the utterance due to possible difficulty in breath control when speaking.

3.1. GAN Architecture: The overall architecture configurations of the proposed DCGAN model follow those used in [25], also again shown in Fig.3. The Generator component contains 4 convolutional layers, the first three of which have 8 kernels while the last one has 1 kernel only. All of the kernels in the Generator has a kernel size of $5 \times 3$ and stride of $1 \times 1$. Each of the first three convolutional layers is also immediately connected to ReLU activations. We use Replicate Padding to replicate the edges of the feature map to ensure that the output and input dimensions are the same. The Discriminator component contains 4 convolutional layers of 8, 16, 32 and 64 kernels respectively, all of which use a kernel size of $2 \times 2$ and stride of $2 \times 2$. A flattening operation is applied to concatenate the outputs of convolutional layers, resulting in a 3000-dimensional vector. A fully connected layer with Sigmoid activation is used for binary classification in the Discriminator.

3.2. GAN Training: Prior to GAN model training, pairs of normal and disordered speech utterances of identical word contents but often different lengths need to be formed. In order to facilitate a frame-by-frame comparison between the GAN perturbed normal speech spectrogram against that of the target disordered speech, each normal speech segment is either tempo or speed perturbed to produce a modified duration that matches against that of a target disordered speech utterance, as shown in Fig.2(a). This requires a scaling factor $\alpha$ to be estimated for each such normal and disordered speech segment pair using phone alignment analysis similar to the procedure described in Section 2.3 for speaker level perturbation. The resulting pairs of normal and disordered speech utterances that now have the same durations are further zero mean and unit variance normalized at speaker level, before being used in GAN model training.

The GAN training objective function that both maximizes the binary classification accuracy on target disordered speech and minimizes that obtained on the GAN perturbed normal speech, in the hope that upon convergence the latter is modified to be sufficiently close to the target impaired speech, is given by

$$\min_{G_j} \max_{D_j} V(D_j, G_j) = \mathbb{E}_{f_{Dj} \sim p_d(f)} \left[ \log \left( D_j(f_{Dj}) \right) \right] + \mathbb{E}_{f_{C} \sim p_c(f)} \left[ \log \left( 1 - D_j(G_j(f_{C})) \right) \right]$$

where $j$ represents the index for target dysarthric speaker, $G_j$ and $D_j$ are Generator and Discriminator associate with dysarthric speaker $j$, $f_C$ and $f_{Dj}$ stand for the FBank features of paired control and dysarthric utterances. During GAN model training for each target impaired speaker, the learning rate for both the Generator and Discriminator components is halved every 2500 iterations until convergence.

3.3. Disordered Speech Spectrum Generation: During GAN based data augmentation, as shown in Fig.2(b), the target disordered speaker level tempo or speed perturbed 40-dimensional FBank features obtained from normal speech using the two baseline approaches described in Section 2.3 are fed into the corresponding Generator components to produce the comparable speaker level tempo-GAN” (“TempaC”) or “speed-GAN” (“SpdC”) in short in Table 1) augmented disordered data, before being further zero mean and unit variance normalized for speech recognition system development.
Table 1: Performance on the 16 UASpeech dysarthric speakers of different data augmentation methods. "CTRL" and "DYS" refer to control and disordered speech. "Temp." and "Spd." stand for tempo and speed perturbation, subscript "c" represents GAN based augmentation. "1x" and "2x" stand for the amount of augmented data. "2x" for disordered speech represents speed perturbation using global perturbation factors 0.9 and 1.1. "VL", "L", "M" and "H" denote intelligibility levels (Very Low / Low / Medium / High). "9" denotes a statis. sig. improvement2 is obtained from Sys. 3-5, Sys. 7-9, Sys. 11-19 against Sys. 2, 6, 10 respectively.

<table>
<thead>
<tr>
<th>Data Augmentation</th>
<th>Hrs.</th>
<th>WER % (Unadapted)</th>
<th>WER % (LHUC-SAT Adapted)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>VL</td>
<td>L</td>
</tr>
<tr>
<td>Sys.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>30.6</td>
<td>69.82</td>
<td>52.61</td>
</tr>
<tr>
<td>2</td>
<td>1x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>1x</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>1</td>
<td>1x</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>1x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>1x</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
<td>1x</td>
</tr>
<tr>
<td>9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>2x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>-</td>
<td>2x</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td>2x</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>1x</td>
<td>-</td>
<td>1x</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>-</td>
<td>2x</td>
</tr>
<tr>
<td>16</td>
<td>-</td>
<td>2x</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>-</td>
<td>2x</td>
</tr>
<tr>
<td>18</td>
<td>1x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>-</td>
<td>2x</td>
<td>-</td>
</tr>
</tbody>
</table>

4. Experiments and Results

All experiments are conducted on UASpeech database, which is a dataset for the single word recognition task. It contains 102.7 hours of speech from 16 dysarthric and 13 control speakers with 155 common words and 300 uncommon words. Utterances are divided into 3 blocks, each containing all common words and one-third of the uncommon words. Block 1 and 3 are treated as training set while block 2 of the 16 dysarthric speakers is treated as test set. In the preprocessing stage, we use a GMM-HMM system to implement silence stripping as described in our previous work [11]. Without data augmentation, the final training set contains 99195 utterances, around 30.6 hours. The test set contains 26520 utterances, around 9 hours.

4.1. Experiment Setup: The proposed GAN model is implemented with PyTorch [45]. The ASR System is implemented using an extended Kaldi toolkit [46], featuring the same architecture in our previous work [30]. We use HTK toolkit [47] for phonetic analysis, silence stripping and feature extraction. Both tempo and speed perturbation are implemented using SoX.

4.2. Performance of Data Augmentation: Table 1 (col.8-12) presents the performance of unadapted systems with different data augmentation methods prior to applying LHUC-SAT (last five columns). We can observe several trends from Table 1: 1) Our proposed tempo-GAN approach consistently outperforms the ones with tempo perturbation only (Sys. 3 vs. 2, Sys. 7 vs. 6 and Sys. 11 vs. 10, col. 12) by up to 1.77% absolute (5.77% relative) overall WER reduction (Sys. 11 vs. 10, col. 12); 2) Our proposed speed-GAN approach outperforms the ones with speed perturbation only (Sys. 9 vs. 8, col. 12) by 0.6% absolute (2.05% relative) overall WER reduction; 3) Col. 12 of Sys. 11-19 explore different combinations of GAN generated data and tempo/speed perturbed speech. Sys. 15 gives the largest WER reduction than the baseline system.

Table 2: A comparison between published systems on UASpeech and our system. "DA" stands for data augmentation.

<table>
<thead>
<tr>
<th>Systems</th>
<th>WER %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheffield-2015 Speaker adaptive training [7]</td>
<td>34.80</td>
</tr>
<tr>
<td>CUBH-2018 DNN System Combination [8]</td>
<td>30.60</td>
</tr>
<tr>
<td>Sheffield-2019 Kaldi TDNN + DA [9]</td>
<td>27.88</td>
</tr>
<tr>
<td>Sheffield-2020 Fine-tuning CNN-TDNN speaker adaptation [10]</td>
<td>30.76</td>
</tr>
<tr>
<td>CUBH-2021 LAS + CTC + Meta Learning + SAT [12]</td>
<td>35.00</td>
</tr>
<tr>
<td>DNN + GAN based DA + LHUC-SAT (Table 1, Sys. 13 &amp; 14)</td>
<td>25.89</td>
</tr>
</tbody>
</table>

4.3. Performance after LHUC-SAT: To model the variability in both original and augmented data, we perform LHUC based speaker adaptive training on the systems. With LHUC-SAT, trends can be observed: 1) LHUC-SAT can bring up to 3.96% (Sys. 5, last col.) absolute (12.61% relative) WER reduction compared with those without adaptation (Sys. 5, col. 12); 2) When further combined with LHUC-SAT, the proposed GAN based approaches consistently outperform the ones with perturbation only (Sys. 2-13, last col.) by up to 1.48% absolute (Sys. 3 vs. 2, last col.) (5.06% relative) WER reduction; 3) With LHUC-SAT, the tempo and speed GAN augmented systems 13 & 14 give the lowest WER of 25.89% among all systems in Table 1. This is also the best WER performance reported on UASpeech as summarised in Table 2.

5. Conclusions

This paper presents GAN based data augmentation approaches for disordered speech recognition. Experimental results suggest improved coverage in augmented data and model generalisation were obtained over baseline systems using no augmentation and conventional methods based on tempo or speed perturbation. Future research will focus on improving GAN modelling on richer spectral and temporal characteristics of disordered speech.

6. Acknowledgements

This research is supported by Hong Kong RGC GRF grant No. 14200218, 14200220, TRS T45-407/19N, ITF grant No. ITS/254/19, and SHIAE grant No. MMT-p1-19.

---

2 A matched pairs sentence-segment word error based statistical significance test was performed at a significance level $\alpha = 0.05$. 

---

4806
7. References


