SYNTHESIS OF NEW WORDS FOR IMPROVED DYSARTHRIC SPEECH RECOGNITION ON AN EXPANDED VOCABULARY

John Harvill[1]*, Dias Issa*[2]*, Mark Hasegawa-Johnson*[1]*, Changdong Yoo*[2]

[1]University of Illinois at Urbana-Champaign, [2]Korea Advanced Institute of Science and Technology

{harvill2,jhasegaw}@illinois.edu, {dias.issa,cd_yoo}@kaist.ac.kr

ABSTRACT

Dysarthria is a condition where people experience a reduction in speech intelligibility due to a neuromotor disorder. Previous works in dysarthric speech recognition have focused on accurate recognition of words encountered in training data. Due to the rarity of dysarthria in the general population, a relatively small amount of publicly-available training data exists for dysarthric speech. The number of unique words in these datasets is small, so ASR systems trained with existing dysarthric speech data are limited to recognition of those words. In this paper, we propose a data augmentation method using voice conversion that allows dysarthric ASR systems to accurately recognize words outside of the training set vocabulary. We demonstrate that a small amount of dysarthric speech data can be used to capture the relevant vocal characteristics of a speaker with dysarthria through a parallel voice conversion system. We show that it's possible to synthesize utterances of new words that were never recorded by speakers with dysarthria, and that these synthesized utterances can be used to train a dysarthric ASR system.

***Index Terms*—** Dysarthric speech, data augmentation, voice conversion, ASR, CTC

1. INTRODUCTION

Dysarthric speech occurs due to a motor disability that impairs the ability to speak clearly. Dysarthric speech is distorted and is therefore much more difficult to recognize compared to normal speech. While state-of-the-art speech recognizers are capable of achieving word error rates (WER) of less than 5% for normal read speech [1], dysarthric speech recognition performance is far behind. Some recent notable work in improving dysarthric ASR includes using realistic language models (LM) for continuous dysarthric ASR [2], applying transfer learning to improve dysarthric ASR across languages [3], and dysarthric speech enhancement [4], [5].

This work was supported by Institute for Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-01396, Development of framework for analyzing, detecting, mitigating of bias in AI model and training data)

Speakers with dysarthria make up a small portion of the population, so there is little training data for dysarthric speech. Popular publicly-available dysarthric speech datasets include UASpeech [6], TORGO [7], and Nemours [8]. Unfortunately, these datasets are limited in vocabulary and hours of speech data. Since characteristics of dysarthric speech vary greatly between and within speakers, large amounts of training data are necessary for improved performance. Data augmentation serves as a natural solution to this problem, and has been explored for both normal and dysarthric speech. For example, SpecAugment [9] is a simple data augmentation technique that improves performance of ASR systems for normal speech. Park et al. show that time-warping, and frequency and time masking of spectrograms act as effective regularizers and improve generalization. Vachhani et al. [10] were the first to suggest dysarthric-specific data augmentation and adjusted the timing and tempo of normal speech by a constant factor to simulate dysarthric speech. Xiong et al. [11] expanded on this work by exploring personalized tempo adjustments to normal speech per dysarthric speaker and adjusting dysarthric speech towards normal speech for recognition by ASR systems trained with normal speech. In contrast, Celin et al. [12] use virtual microphone array synthesis and multi-resolution feature extraction to increase the number of training examples.

While these data augmentation techniques are effective, they only improve recognition performance for words in the training vocabulary of the respective datasets. For practical commercial application, the size of recognizable vocabulary should be much larger than that of the current publicly-available dysarthric speech datasets. In this paper we make two major contributions, described below.

New Task: We propose a new task in dysarthric ASR where we split data into seen and unseen partitions. Seen data simulates all training data in available dysarthric speech corpora, while unseen data simulates utterances of words outside the existing vocabulary in dysarthric speech corpora. The goal is to maximize performance of an ASR system on unseen dysarthric utterances while only having access to seen dysarthric and unseen normal utterances for training.

Synthesis of New Words: We propose a novel voice conversion data augmentation technique that makes it possible to

synthesize new words outside of the seen vocabulary by converting unseen normal speech to artificial unseen dysarthric speech.

We demonstrate that an ASR trained on voice-converted unseen normal speech performs much better when tested on unseen dysarthric speech than an ASR trained on only the seen dysarthric speech or unconverted unseen normal speech. With the method proposed in this paper, it becomes possible to synthesize realistic examples of thousands of new words that could improve dysarthric speech recognition in a practical setting.

2. RELATED WORK

Jiao et al. first proposed the idea of voice conversion as a method of data augmentation for dysarthric speech [13]. The authors use a deep convolutional generative adversarial network (DCGAN) to convert normal speech features to dysarthric speech features. They demonstrate that a classification task between dysarthric and ataxic speech is slightly improved using their augmentation method when compared to a white-noise baseline. They also perform perceptual evaluations with clinical physicians that show that their method transforms normal speech to have perceptual qualities similar to dysarthric speech. In this paper, we use the method of Jiao et al. as a baseline. While we focus on converting normal speech to realistic artificial dysarthric speech, a reverse approach has been applied which also improves dysarthric ASR performance. Chen et al. [4] and Yang et al. [5] train generative adversarial network (GAN) voice conversion systems that convert dysarthric speech to normal speech.

3. DATA

UASpeech: For our experiments, we use the Dysarthric Speech Database for Universal Access Research (UASpeech) [6]. The database contains utterances of 449 unique isolated words consisting of both uncommon and common words such as digits, computer commands, and radio alphabet words. Of the 449 words, 288 are spoken once each, 6 are spoken twice each and 155 are spoken 3 times each. 7 microphones are used to record each utterance. Recordings are provided from both normal speakers and speakers with Cerebral Palsy who self-report a diagnosis of dysarthria. Speech intelligibility ratings are given for each speaker, which are calculated empirically from human annotators' ability to correctly transcribe each person's speech. Transcription accuracy of 0-25% produces an intelligibility rating of "Very Low," 25-50% is rated "Low," 50-75% is rated "Mid," and 75-100% is rated "High." For our experiments, we use 9 control speakers and 4 speakers with dysarthria. Speakers F05 and M14 are female and male speakers with dysarthria, respectively, with intelligibility ratings of "High." Speakers F04 and M05 have intelligibility ratings of "Mid." We would expect to use $(288+6\times2+155\times3)\times(7\text{ microphones})\times(13\text{ speakers})=69615$ total utterances, but due to our preprocessing some files are excluded (69254 total utterances).

Preprocessing and Feature Extraction: There is some stationary noise in the recordings, which we remove using Noisereduce [14]. Next, we trim silence from the beginning and end of each recording. Finally, we extract the mel log spectrogram using Librosa [15]. We use 80 mel frequency bins and a 10ms frame shift. We allow a dynamic range of 120db by clipping anything below -120db, then normalize amplitude in the range zero to one.

	Seen Words		Unseen Words	
	Words	Utts.	Words	Utts.
Normal Train 1	225	25008	224	20794
Normal Train 2	224	22641	225	23152
Normal Val 1	220	1458	224	2016
Normal Val 2	217	1311	225	2024
Dysarthric Train 1	225	10213	224	8338
Dysarthric Train 2	224	9285	225	9360
Dysarthric Val 1	220	1384	224	890
Dysarthric Val 2	217	1258	225	895
Dysarthric Test 1	0	0	224	890
Dysarthric Test 2	0	0	225	895

Table 1. Global partition. Fold indicated as either 1 or 2.

Partition: We split the dataset into seen and unseen sets by word (see Table 1). Of the 449 unique words, half are chosen as "seen," and half as "unseen." We perform two-fold cross validation by swapping the "seen" and "unseen" partitions for a second set of experiments.

Seen data split: We pair normal and dysarthric utterances per speaker pair, from which we create training and validation partitions (98% and 2% of pairs of utterances, respectively). As discussed in Sections 4 and 5 in more detail, the data is split this way because we perform pairwise voice conversion between each normal speaker and each speaker with dysarthria. There are 36 speaker pairs (see Section 5), so the number of validation utterances reported in Table 1 is the size of the union of all 36 validation sets. Some utterances may be used as training examples for some speaker pairs and validation examples for others, but per speaker pair the train and validation partitions are disjoint.

Unseen data split: We have separate partitions for normal and dysarthric speech. For normal speech, we choose one utterance of each unseen word from each speaker for the validation set and use the rest for the unseen normal training set. For dysarthric speech, we choose one utterance of each unseen word for each speaker for the test set, and one for the validation set, and use the rest for the unseen dysarthric training set.

4. ALGORITHMS

Attention-Based Voice Conversion: Since UASpeech is a parallel corpus, we time-align matching normal and dysarthric utterances using dynamic time warping (DTW). DTW is performed using 12 MCEP features per frame, with the same frame shift as the mel log spectrogram features (10ms). Normal and dysarthric utterances are aligned using DTW on MCEP, then the DTW path is used to time-align the mel log spectrogram features.

Our voice conversion model consists of 6 layers of multihead attention [16]. We implement this model using the TransformerEncoderLayer from PyTorch where each multihead attention layer has 8 heads and a model dimension of 80 (number of frequency bins). To train the voice conversion model, we pass the time-aligned normal utterance through the network and apply the mean-squared error (MSE) loss function between the network output and the matching time-aligned dysarthric utterance. We use the Adam optimizer, a batch-size of one, and train for 150,000 iterations.

CTC-Based ASR: The ASR model consists of 4 bi-directional long short-term memory layers (BLSTM) plus two fully-connected layers. The input dimension is 80 (number of frequency bins), and the hidden size of the BLSTM layers is 200. We use a dropout of 0.1 in the BLSTM layers and use a hidden dimension of 500 in the fully-connected layers. We apply the tanh nonlinearity to the first fully-connected layer and the log softmax nonlinearity to the output of the second fully-connected layer. The output has as many frames as the input (no temporal downsampling), but the feature dimension has the size of the number of phones plus the blank, "Start of sentence" (SOS), and "End of sentence" (EOS) symbols. To predict the transcription, we compute the argmax over the output frames, delete blanks, and remove duplicates. All ASR models are trained using datasets augmented by a factor of three using SpecAugment [9]. We train the ASR model with the Adam optimizer, batch size 16 and the CTC loss function [17] until validation loss flattens on a log scale. We then choose the model with lowest validation error for testing.

Language Model: To improve transcription performance for all methods, we use a phone language model to decode the output of the ASR model. Results for all systems are presented both with and without the phone language model, in order to test the benefits that can be obtained for each system by limiting the decoded output to one of the possible 449 unique vocabulary words. Phone N-gram models are trained up to $N = 15$ (longest phone sequence with SOS and EOS added to beginning and end, respectively) using phone sequence representations of all vocabulary words, including both seen and unseen words. Similar to Hannun et al. [18], we perform a beam search over possible transcription sequences, using a beam width of 20, and weighting the ASR model log probabilities equally with the language model log probabilities. To reduce decoding time, we first remove all frames for which the ASR model predicts the blank symbol with highest probability. We then choose SOS as the first symbol by default and decode from the second frame onwards, thereby filtering out sequences that do not correspond to one of the words in the vocabulary.

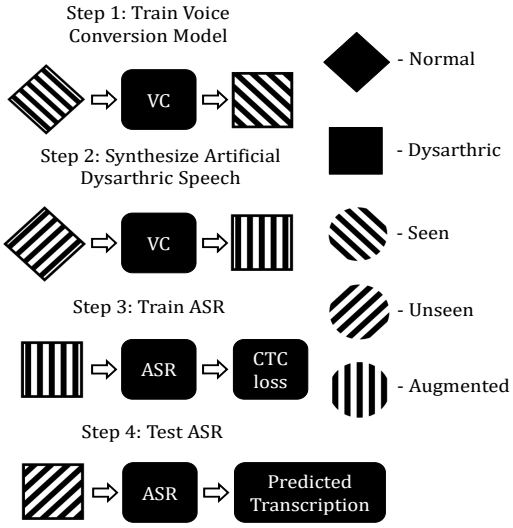

Fig. 1. Proposed Approach: Diamond or square shapes refer to normal or dysarthric speech, respectively. Overlay patterns denote whether data is from the seen or unseen partition, or is augmented (synthetic).

5. EXPERIMENTAL CONFIGURATIONS

The proposed attention-based voice conversion approach, which we refer to as Attention, is experimentally tested by comparison to one Oracle system and three baselines. The experimental Attention approach is described in Figure 1, and in the remainder of this section. The Oracle approach and the three baselines are described in the remainder of this section.

Oracle: Train ASR with Dysarthric unseen train and val data. Test using dysarthric unseen test data. The Oracle baseline simulates the ideal training scenario where we have authentic dysarthric utterances for any word.

Attention: (Proposed system, Figure 1) Normal and dysarthric sentences from the seen partition are paired to train the voice conversion model. For both the Attention and DCGAN voice conversion approaches, we train a separate model for each of the $9 \times 4 = 36$ speaker pairs, each pair including one speaker with and one without dysarthria. The trained voice conversion model is then used to convert data from the normal unseen train and val datasets into artificial dysarthric utterances. We use the converted utterances from all 36 models to train and validate an ASR model and evaluate the ASR model performance on dysarthric unseen test data.

DCGAN: This baseline follows the same procedure as the Attention model, except that the voice conversion model is

our re-implementation of the DCGAN approach proposed by Jiao et al. [13]. Our re-implementation differs from the original approach in one way that may be significant: the original approach trains separate networks for mel cepstral coefficients (MCEP) and band aperiodicity features, but our re-implementation uses mel log spectrogram features, in order to better match the configuration of the proposed Attention system. Our re-implementation of DCGAN uses batch size of 32 as proposed in [13], and is trained for 10,000 iterations.

Lack Baseline: The Lack baseline simulates a training scenario where we lack any dysarthric speech data and train using only normal speech. The ASR is trained using the normal unseen train and validation partitions, then tested using dysarthric unseen test data.

Limited Baseline: The Limited baseline simulates a training scenario where a phone-based ASR is trained using a dysarthric speech corpus with limited vocabulary, then tested using words that were not in the training set. The dysarthric seen partition is used to train and validate the ASR model, which is then tested on dysarthric unseen test data.

6. RESULTS

WER results for each method are shown in Table 2 (using both folds, 890 + 895 = 1785 total test samples). Results are given for each individual speaker, as well as the mean of the scores for all speakers. For each speaker, the best non-Oracle system without the LM and the best non-Oracle system with LM are in bold type. We test statistical significance between pairs of methods using the Gillick test (same test utterances) [19]. We find significance at a level of $p < 10^{-18}$ (max p for all eight tests) between the Attention method and each of the other baselines both with and without LM. When comparing the DCGAN and Lack baselines, we find significance for the LM scenario with $p < 10^{-3}$, but do not find significance for the scenario without the LM ($p < 0.73$).

As expected, the Oracle performs best on unseen test dysarthric speech. The next best method is Attention, which shows significant improvement over both the Lack and DCGAN baselines. This suggests that the Attention voice conversion method proposed in this paper is more effective at capturing properties of the voice of dysarthric speakers that are useful for determining proper transcriptions than the DCGAN voice conversion method. It also demonstrates that the WER of an ASR trained using normal speech can be substantially reduced (by 17 percentage points without a language model) if the normal speech is first converted, using an Attention-based voice conversion system, to sound like dysarthric speech.

CTC training with artificially synthesized dysarthric speech from the DCGAN voice conversion method seems slightly less effective than training with normal utterances. While DCGAN may transform normal utterances such that they are more perceptually similar to authentic dysarthric speech, the transformed features appear less useful for the transcription task. One key difference between the Attention-based and DCGAN voice conversion algorithms, possibly sufficient to explain this finding, is that the generator in DCGAN is a considerably smaller model; the smaller model may reduce its ability to simultaneously capture properties of a dysarthric speaker's voice and preserve linguistic content.

The Limited baseline exhibits the worst performance, both with and without the LM. Our BLSTM ASR learns the phone sequences of words observed during training, and is unable to recognize unseen phone sequences, even with the help of an external LM. The LM improves performance for all methods, by forcing the ASR to output a phone sequence corresponding to one of the (seen or unseen) vocabulary words. The Lack model benefits most from the language model, perhaps because the LM helps overcome the acoustic differences between dysarthric and normal speech.

	F05	M14	F04	M05	Mean
Oracle	4.0	6.9	4.1	6.5	5.4
Attention	**23.6**	**21.4**	**46.8**	**69.5**	**40.3**
DCGAN	46.3	32.1	72.4	80.4	57.8
Lack	43.0	34.3	71.5	80.8	57.4
Limited	98.4	98.4	99.1	98.9	98.7
Oracle + LM	3.1	4.9	1.8	1.8	2.9
Attention + LM	**15.8**	**15.8**	**34.9**	**50.6**	**29.3**
DCGAN + LM	35.6	22.3	58.0	59.9	44.0
Lack + LM	24.7	24.3	50.9	58.4	39.6
Limited + LM	96.7	97.1	98.4	97.3	97.4

Table 2. WER (both folds) of dysarthric ASR models trained using five different data configurations, each with and without LM. "Mean" is the average of results for the four speakers (F05, M14, F04, and M05).

7. CONCLUSIONS

In this paper, we proposed a new task for dysarthric ASR that simulates performance on vocabulary words for which authentic dysarthric speech recordings do not exist. We evaluated the effectiveness of using artificially synthesized dysarthric speech of unseen words to train a CTC-based ASR system. We compared the approach to an ideal, but unrealistic, scenario (Oracle), and to three practical baselines. We find that our proposed approach outperforms all practical baselines. While the collection of much more authentic dysarthric speech data would be ideal, the methods proposed in this paper offer a new way to make the most of existing data and could increase accuracy of dysarthric ASR systems in industry on a much larger vocabulary. This could speed up the process of making ASR-dependent technologies more accessible to people with dysarthria.

8. REFERENCES

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[2] Zhengjun Yue, Feifei Xiong, Heidi Christensen, and Jon Barker, "Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6094–6098.

[3] 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.

[4] Li-Wei Chen, Hung-Yi Lee, and Yu Tsao, "Generative adversarial networks for unpaired voice transformation on impaired speech," *arXiv preprint arXiv:1810.12656*, 2018.

[5] Seung Hee Yang and Minhwa Chung, "Improving dysarthric speech intelligibility using cycle-consistent adversarial training," *arXiv preprint arXiv:2001.04260*, 2020.

[6] 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.

[7] 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.

[8] Xavier Menendez-Pidal, James B Polikoff, Shirley M Peters, Jennie E Leonzio, and H Timothy Bunnell, "The nemours database of dysarthric speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*. IEEE, 1996, vol. 3, pp. 1962–1965.

[9] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[10] Bhavik Vachhani, Chitralekha Bhat, and Sunil Kumar Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition.," in *Interspeech*, 2018, pp. 471–475.

[11] Feifei Xiong, Jon Barker, and Heidi Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5836–5840.

[12] TA Mariya Celin, T Nagarajan, and P Vijayalakshmi, "Data augmentation using virtual microphone array synthesis and multi-resolution feature extraction for isolated word dysarthric speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 346–354, 2020.

[13] 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.

[14] Tim Sainburg, "timsainb/noisereduce: Initial release," Mar 2019.

[15] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[18] Awni Y Hannun, Andrew L Maas, Daniel Jurafsky, and Andrew Y Ng, "First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns," *arXiv preprint arXiv:1408.2873*, 2014.

[19] SJ Cox, "The gillick test: A method for comparing two speech recognisers tested on the same data," Tech. Rep., Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.