Interspeech 2018
2-6 September 2018, Hyderabad

Data Augmentation using Healthy Speech for Dysarthric Speech Recognition

Bhavik Vachhani, Chitralekha Bhat, Sunil Kumar Kopparapu

TCS Research and Innovation, Mumbai, India

bhavik.vachhani@tcs.com, bhat.chitralekha@tcs.com, sunil.kopparapu@tcs.com

Abstract

Dysarthria refers to a speech disorder caused by trauma to the brain areas concerned with motor aspects of speech giving rise to effortful, slow, slurred or prosodically abnormal speech. Traditional Automatic Speech Recognizers (ASR) perform poorly on dysarthric speech recognition tasks, owing mostly to insufficient dysarthric speech data. Speaker related challenges complicates data collection process for dysarthric speech. In this paper, we explore data augmentation using temporal and speed modifications to healthy speech to simulate dysarthric speech. DNN-HMM based Automatic Speech Recognition (ASR) and Random Forest based classification were used for evaluation of the proposed method. Dysarthric speech, generated synthetically, is classified for severity level using a Random Forest classifier that is trained on actual dysarthric speech. ASR trained on healthy speech, augmented with simulated dysarthric speech is evaluated for dysarthric speech recognition. All evaluations were carried out using Universal Access dysarthric speech corpus. An absolute improvement of 4.24% and 2% WAS achieved using tempo based and speed based data augmentation respectively as compared to ASR performance using healthy speech alone for training.

Index Terms: Dysarthric speech recognition, Data augmentation, Dysarthria severity

1. Introduction

Dysarthria is a speech disorder resulting from disruption in the execution of speech movements due to neuromuscular disturbances to muscle tone, reflexes, and kinematic aspects of movement. It could be either acquired or developmental. Dysarthric speech is characterized by being slow, slurred, harsh or quiet, or uneven depending on the type of dysarthria. Speech enabled interfaces are gaining popularity, especially in the assisted and smart living domains. Also, speech is a convenient alternative to other machine interfaces such as remote controls, keyboards, or PC mice given that persons with dysarthria are often faced with physical inabilities as well [1]. While traditional, *off-the-shelf* Automatic Speech Recognition (ASR) systems perform well for normal speech, this is not the case with the atypical dysarthric speech owing to the inter-speaker and intra-speaker inconsistencies in the acoustic space as well as the sparseness of data. Several techniques are employed to improve ASR performance for dysarthric speech: acoustic space enhancement, feature engineering, Deep Neural Networks (DNN), speaker adaptation, lexical model adaptation- individually or as a combination thereof [2, 3, 4, 5, 6]. In order to exploit the machine learning techniques for ASR fully, suitable data to build these systems is imperative. However, owing to speaker muscle weakness and fatigue, collection of dysarthric data is tedious, especially for speakers with severe dysarthria. Additionally, since dysarthria can stem from a variety of neurological disorders, the characterization of dysarthric speech is complex, this makes the designing of a data collection process difficult. Thus far, three popular dysarthric speech databases, namely Universal Access (UA) speech corpus [7], Nemours [8] and TORGO [9] exist for American English. Two French corpora, namely the CCM corpus collected by Dr Claude Chevrie-Muller and her team and the Aix-Neurology-Hospital corpus (ANH) have been described in [10]. Authors describe a Dutch dysarthric speech database containing mildly to moderately dysarthric speech from patients with Parkinson's disease, traumatic brain injury and cerebrovascular accident [11]. A Korean dysarthric speech corpus was built as a part of the Quality-of-Life technology (QoLT) project that focuses on the development of speech technologies for people with articulation disabilities [12]. A Cantonese corpus with a focus on investigation of articulatory and prosodic characteristics of Cantonese dysarthric speech is discussed in [13]. German [14], Spanish [15] and Czech [16] corpora were collected with the intent of studying dysarthric speech in patients suffering from Parkinson's disease. While most of the corpora comprise of data collected under clinical settings, [17] describes *the homeService corpus*, a British English corpus of realistic dysarthric data collected in the home environment. Each of the above databases were designed for a specific purpose with a broad perspective of improving the life of people with dysarthria. However, the amount of data is substantially lower than a speech corpus of normal speech used in training the state-of-the-art ASR systems, that use machine learning. To overcome this issue of unavailability of suitable speech data, we adopt data augmentation techniques.

Data augmentation is the process by which we create new synthetic training samples by adding small perturbations on our initial training set. The objective is to make model invariant to perturbations and enhance its ability to generalize. In [18] audio speed was modified using three speed factors and the effectiveness was reported for large vocabulary continuous speech recognition (LVCSR). Different audio data augmentation techniques such as time stretching, pitch shifting, dynamic range compression and mixing with background noise was used for environmental sound classification in a Convolutional Neural Network (CNN) based architecture to significantly improve the classification accuracy [19]. Data augmentation techniques have been used to improve classification tasks such as real life sound classification [20, 21]. In [22] Alzheimers disease (AD) data was augmented using two normative data sets, through minority class oversampling with Adaptive Synthetic sampling (ADASYN), wherein the proposed technique outperformed state-of-the-art results in the binary classification of speech with and without AD.

In this paper we explore how an understanding of the deficits in speech production caused by dysarthria may be used to augment existing data. We present an analysis of phone durations in dysarthric data with bearing on dysarthria severity level. Based on this information, we proceed with data augmentation using temporal and speed modifications to healthy speech to generate synthetic speech that matches the characteristics of

10.21437/Interspeech.2018-1751

dysarthric speech. Further, we classify this synthetic dysarthric speech into four severity levels using Random Forest classifier that is trained on actual dysarthric speech, so as to validate our understanding of the impact of these modifications to healthy speech and how it simulates dysarthric speech. A DNN-HMM based Automatic Speech Recognition (ASR) is trained using healthy speech augmented with simulated dysarthric speech. This ASR system is evaluated for dysarthric speech recognition using Universal Access (UA) dysarthric speech corpus.

The rest of the paper is organized as follows. Section 2 presents an analysis of phone durations in dysarthric speech and motivates the data augmentation process, and discusses the augmentation techniques used, Section 3 describes the experimental setup, In Section 4 we present the results and analysis and we conclude in Section 5.

2. Methodology

2.1. Phoneme duration analysis

In order to modify healthy control speech to emulate dysarthric speech characteristics, we need to first understand the dysarthric speech itself. In our earlier work [23], we modified the tempo of dysarthric speech based on severity to improve the ASR recognition. It was observed that the sonorant regions of dysarthric speech are of longer durations as compared to that of healthy speech. In this work, we further examine the relationship between phone durations of dysarthric speech and the dysarthria severity levels. UA Speech corpus comprises dysarthric speech of 4 severity levels, namely S1, S2, S3, S4 in the increasing order of severity. A total of 3534 utterances of dysarthric speech corresponding computer command words were force-aligned at phone level using Sphinx3 toolkit [24], using Voxforge English acoustic models trained on approximately 35 hours of speech data [25]. The alignment was then manually inspected and corrected for extraction of phone duration. Similar exercise was carried out on TORGO dysarthric speech corpus [9]. TORGO dysarthric speech corpus comprises dysarthric speech of 3 severity levels, namely S1, S2 and S3. A more accurate representation of the relationship between phone durations and severity can be seen for this corpus since it comprises manual annotation of utterances at phone level. We observed that, there is a strong correlation between dysarthria severity and the average duration of a phone. as shown in Figure 1. It was found that average of phone duration is proportional to the severity of dysarthric speech, the higher the severity, longer the phone duration.

Based on this analysis we modify the phone durations of healthy control speech to generate synthetic dysarthric speech data. We use this modified speech along with the healthy control speech to augment the ASR training data.

2.2. Synthetic dysarthric data generation

Healthy control speech was modified using two different time domain perturbations, namely (1) Time (Speed) perturbation and (2) Tempo perturbation. *Rubberband – an audio time-stretching and pitch-shifting utility program* was used for this purpose and is described below [26]. Healthy control speech modified in this manner amounts to synthetically generated dysarthric speech data. To the best of our knowledge data augmentation in the context of dysarthric speech recognition has not been reported in literature previously.

2.2.1. Time (Speed) perturbation based data augmentation

Speed perturbation is achieved by re-sampling the input signal by a factor R1. If R1 < 1, signal duration is increased and for R1 > 1, signal duration is reduced. In this work we use different values of R1 as R1 $\in \{1.2, 1.4, 1.6, 1.8, 2.0, 2.2\}$ to modify the durations of healthy control speech. Below command will stretch the given input signal duration to R1 times original duration in the Rubberband toolkit.

```
rubberband -t R1 <infile.wav> <outfile.wav>
```

2.2.2. Tempo perturbation based data augmentation

The tempo of the signal is modified by factor R2 while ensuring that the pitch and spectral envelope of the signal do not change. If R2 > 1, signal duration reduces and R2 < 1 signal duration increases, making the healthy control speech slower. In this work we use R2 as R2 $\in \{0.4, 0.6, 0.8\}$ to modify the healthy control speech. Below command will modify the given input signal duration to R2 times original duration.

```
rubberband -T R2 <infile.wav> <outfile.wav>
```

The parameters R1 and R2 were selected empirically based on the severity classification provided by the Random Forest classifier for various values of R1 and R2 as discussed in Section 3.2.

3. Experimental setup

3.1. Database

Data from Universal Access (UA) speech corpus [7] was used for both training and testing. UA speech corpus comprises data from 13 healthy control (HC) speakers and 15 dysarthric (DYS) speakers with cerebral palsy. The recording material consisted of 455 distinct words with 10 digits, 26 international radio alphabets, 19 computer commands, 100 common words and 300 uncommon words that were distributed into three blocks. Audio data was recorded using a 7-channel microphone array, fitted to the top of a computer monitor. Speech intelligibility ratings for each dysarthric speaker, as assessed by five naive listeners is also included in the corpus. Speakers were divided into four different categories based on the intelligibility. We use this information to analyze the performance of our recognition systems at different dysarthria severity level. In this paper we have used 19 computer command words from 13 healthy control (HC-CC) and 15 dysarthric (DYS-CC) speakers.

Long silence regions at the start and end of both the healthy control (HC) data used for training and dysarthric speech (DYS) used for testing of the ASR are trimmed using energy based method using PRAAT tool [27]. Initial experiments were conducted to understand the effect of silence removal at the start and end of the HC and DYS speech. Traditional DNN- HMM based system using standard MFCC features as discussed in Section 3.3. Table 1 shows the ASR performance in terms of word error rate (WER) for training and test data with and without silence pre-processing. An absolute improvement of 15% (48.47% to 33.11%) was achieved by using fMLLR transform, further improvement of 4% (33.11% to 29.06%) was achieved using silence pre-processing. We use the best WER, wherein both training and test data were pre-processed with fMLLR-based ASR configuration as baseline for reporting our current work.

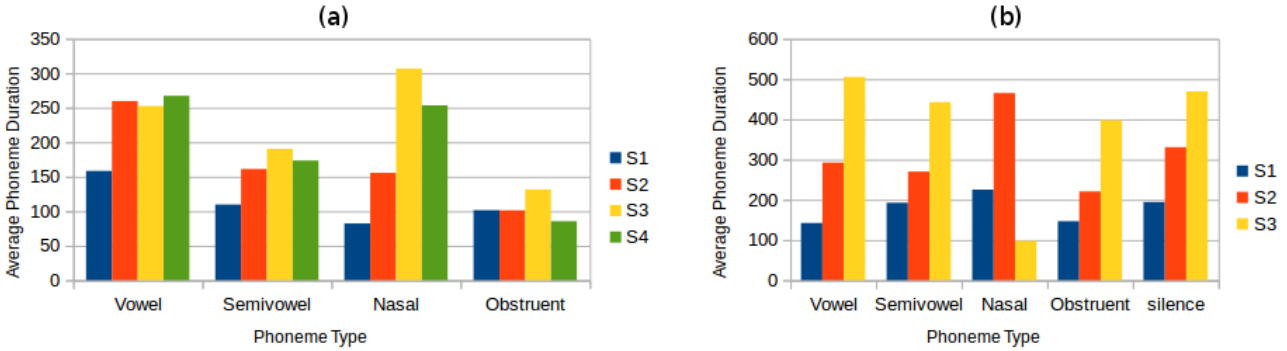

Figure 1: *(a) Average phone duration (ms) for UA dysarthric speech Corpus (b) Average phone duration (ms) for TORGO dysarthric speech corpus*

Table 1: *Effect of data Pre-processing on WER*

Training Total 3458 utt	Testing Total 3534 utt	WER w/o fMLLR	WER with fMLLR
HC-CC	DYS-CC	48.47	33.11
SIL trimmed HC-CC	SIL trimmed DYS-CC	37.32	29.06

3.2. Dysarthria severity classification on augmented data

Validity of using the synthetically generated dysarthric speech to augment the ASR training data for dysarthric speech recognition needs to be ascertained. Synthetically generated dysarthric speech is automatically classified using Random Forest classifier trained on actual dysarthric speech. Classifier was trained using the feature set suggested by *Intespeech 2009 emotion challenge*, extracted using openSMILE toolkit [28]. A total of 3534 dysarthric utterances were used for training the classifier using 5 fold cross validation using WEKA toolkit [29]. An accuracy of 96% was achieved for dysarthric speech classification into 4 classes, based on the intelligibility score provided in the UA Speech corpus. A total of 3458 healthy control (HC) utterances modified using various tempo and speed perturbation parameters described in Section 2 were classified using this framework into 4 severity classes.

3.3. DNN-HMM based ASR framework

Kaldi toolkit [30] was used for DNN-HMM based dysarthric speech recognition. GMM-HMM system was trained using a maximum likelihood estimation (MLE) training approach along with 100 senones and 8 Gaussian mixtures. Cepstral mean and variance normalization (CMVN) was applied on 23 dimension MFCC features. Dimensionality reduction was done using Linear Discriminant Analysis (LDA), wherein LDA builds HMM states using feature vectors with a reduced feature space. We use a context of 6 frames (3 left and 3 right) to compute LDA. The feature vector size post LDA is set to 40.

The input layer of DNN has 360 (40×9 frames) dimensions using a left and right context of 4 frames. The output layer has a dimension of 96 (number of senones available in the data). Two hidden layers with 512 nodes in each layer were used. Feature-space Maximum Likelihood Linear Regression (fMLLR) transformed features are used as input to the DNN training, making it a feature normalization technique. In decoding process we use configurations with and without fMLLR transformed features as a input [31]. DNN training was carried out using 15 epochs for all experiments. Dysarthric speech recognition was carried out using a constrained Language Model (LM), wherein we restrict the recognizer to give one word as output per utterance.Performance of each of the recognition systems is reported in terms of word error rate (WER).

Training configurations for the DNN-HMM based ASR is as shown in Table 2. A total of 3534 dysarthric speech utterances corresponding to 19 computer command words from blocks B1 and B3 have been used for testing purpose.

Table 2: *Training data for different systems*

Training Set	Info	Total no.utterances
A	No augmentation	3458
B	Time stretching $R1 \in \{1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2\}$	24206
C	Tempo stretching $R2 \in \{0.4, 0.6, 0.8, 1.0\}$	13832

4. Experimental Results and Analysis

Synthetically generated dysarthric speech was classified into 4 classes as discussed in Section 3. Table 3 shows the classification of 3458 healthy control utterances modified into dysarthric utterances using various augmentation parameters. Synthetically generated dysarthric speech is classified into 4 classes, namely S1, S2, S3 and S4 in increasing order of severity. It can be seen for both speed and tempo modifications that the synthetically generated dysarthric utterance classification is closely correlated to the duration of the utterance. Table 4 shows the performance of the ASR for training configurations mentioned in Table 2, examined at individual severity level.

Tempo based and speed based augmentation techniques give an absolute improvement of 4.24% and 2% respectively. Higher improvement was observed for higher severity (S4), approximately 3% and 12% absolute improvement over baseline for systems speed and tempo augmentation respectively. Table 5 shows the effects of each data augmentation parameter on 4

Table 3: *Severity classification of synthetically generated dysarthric data - %Accuracy*

Classifier System	Augmentation parameter	S1	S2	S3	S4
A	None	93.97	4.29	1.74	0.00
B1	R1 = 1.2	88.35	8.00	3.30	0.35
B2	R1 = 1.4	81.05	12.51	4.98	1.45
B3	R1 = 1.6	65.24	21.55	9.79	3.42
B4	R1 = 1.8	46.87	33.43	14.83	4.87
B5	R1 = 2.0	33.72	39.86	21.15	5.27
C1	R2 = 0.4	4.06	39.98	38.18	17.79
C2	R2 = 0.6	59.62	23.99	12.05	4.35
C3	R2 = 0.8	86.96	8.98	3.48	0.58

different severity levels. It can be seen that the proposed method gives improvement at all severity levels.

Table 4: *Severity wise WER for testing data*

Training Set	S1	S2	S3	S4	Overall WER
A	1.05	17.89	44.73	78.51	**29.06**
B	0.98	19.73	36.44	75.43	**27.05**
C	1.28	15.52	37.36	66.96	**24.82**

In order to attribute the improvement in the ASR performance to the synthetically generated dysarthric speech data, we look into the ASR performance for data augmentation parameters R1 and R2 separately. 8 separate ASR systems were trained as seen in Table 5, each with 3458 synthetically generated dysarthric utterances. Table 5 shows the effect of individual augmentation parameters on ASR performance. No healthy control data was used in the training of the ASR. Correlation between the ASR performance for actual dysarthric speech and the durations of the synthetic dysarthric speech data is seen for both speed and tempo perturbations. From Table 5 and Table 3, it is seen that increasing the phone durations using augmentation, degrades the ASR performance for low severity dysarthric speech (S1 and S2).

Table 5: *Effect of data augmentation on WER for individual severity level*

Training Set	Augementation parameter	S1	S2	S3	S4	Overall WER
A	None	1.05	17.89	44.74	78.51	29.06
B1	R1=1.2	0.98	17.63	42.11	72.95	27.33
B2	R1=1.4	0.90	18.82	38.42	73.68	26.91
B3	R1=1.6	1.35	21.32	36.45	69.30	26.34
B4	R1=1.8	1.20	22.63	37.76	67.25	26.46
B5	R1=2.0	1.58	21.18	34.87	69.01	26.00
C1	R2=0.4	2.33	18.95	39.87	70.47	27.16
C2	R2=0.6	1.05	20.79	37.37	77.34	27.87
C3	R2=0.8	0.75	18.55	34.61	74.56	26.15

Based on the ASR performance for synthetically generated dysarthric data, optimal values of augmentation parameters R1 and R2 to generate dysarthric data of different severity levels is as shown in Table 6.

Table 6: *R1 and R2 recommendation for optimal ASR recognition*

Severity	R1	R2
S1	1.4	0.8
S2	1.2	0.8
S3	2	0.4
S4	1.8	0.4

5. Conclusions

Given that speech is an attractive interface to control the devices used in assisted living and smart homes, it is imperative that we look into improving the ASR performance for pathological speech. Due to lack of suitable data to train the ASRs, machine learning techniques are not fully exploited for pathological speech recognition. In this paper, we address the data challenge for dysarthric speech using data augmentation to synthetically generate dysarthric speech data using healthy control speech. An understanding of the deficits in speech production caused by speech pathology has been used to augment existing data using speed and tempo modifications to the healthy control speech. A DNN-HMM based Automatic Speech Recognition (ASR) system and Random Forest based classification system have been used for evaluation of proposed method. Synthetically generated dysarthric speech is classified into 4 different severity levels using Random Forest classifier trained on actual dysarthric speech. ASR system trained using healthy control speech augmented using synthetically generated dysarthric speech is evaluated for dysarthric speech utterances. All evaluations were carried out on Universal Access dysarthric speech corpus computer command words. An absolute improvement of 15% was achieved by using fMLLR transform as compared to our previous work [6]. Additionally, ASR peformance improved by 4% using silence pre-processing. We use this WER (29.06%) as baseline to report our current work. An absolute improvement of 4.24% and 2% was achieved using tempo based and speed based data augmentation system over baseline system.

6. References

[1] F. Rudzicz, "Learning mixed acoustic/articulatory models for disabled speech," in *Proc. NIPS 2010, – Workshop on Machine Learning for Assistive Technologies at the 24th annual conference on Neural Information Processing Systems*, 2010, pp. 70–78.

[2] ——, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163 – 1177, 2013, special Issue on Speech and Language Processing for Assistive Technology.

[3] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data," in *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2014, pp. 254–259.

[4] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *12th International Conference on Signal Processing (ICSP)*, Oct 2014, pp. 505–509.

[5] E. Yılmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Multistage dnn training for automatic recognition of dysarthric speech," in *INTERSPEECH*, 2017, pp. 2685–2689.

[6] B. Vachhani, C. Bhat, B. Das, and S. Kopparapu, "Deep autoencoder based speech features for improved dysarthric speech recognition," in *INTERSPEECH*, 2017, pp. 1854–1858.

[7] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research." in *INTERSPEECH*, 2008, pp. 1741–1744.

[8] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, Oct 1996, pp. 1962–1965 vol.3.

[9] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2011.

[10] C. Fougeron, L. Crevier-Buchman, C. Fredouille, A. Ghio, C. Meunier, C. Chevrie-Muller, J.-F. Bonastre, A. Colazo-Simon, C. D. Looze, D. Duez, C. Gendrot, T. Legou, N. Lévêque, C. Pillot-Loiseau, S. Pinto, G. Pouchoulin, D. Robert, J. Vaissière, F. Viallet, and C. Vincent, "The despho-apady project: Developing an acoustic-phonetic characterization of dysarthric speech in french," in *LREC*, 2010.

[11] E. Yilmaz, M. Ganzeboom, L. Beijer, C. Cucchiarini, and H. Strik, "A dutch dysarthric speech database for individualized speech therapy research," in *LREC*, 2016.

[12] D. L. Choi, B. W. Kim, Y. J. Lee, Y. Um, and M. Chung, "Design and creation of dysarthric speech database for development of qolt software technology," in *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, Oct 2011, pp. 47–50.

[13] K.-H. Wong, Y. T. Yeung, E. H. Y. Chan, P. C. M. Wong, G.-A. Levow, and H. M. Meng, "Development of a cantonese dysarthric speech corpus," in *INTERSPEECH*, 2015.

[14] S. Skodda, W. Grnheit, and U. Schlegel, "Intonation and speech rate in parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission," *Journal of Voice*, vol. 25, no. 4, pp. e199 – e205, 2011.

[15] J. R. Orozco-Arroyave, J. D. Arias-Londoo, J. F. Vargas-Bonilla, M. C. Gonzalez-Rtiva, and E. Nth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[16] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinsons disease," *The journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.

[17] M. Nicolao, H. Christensen, S. Cunningham, P. Green, and T. Hain, "A framework for collecting realistic recordings of dysarthric speech - the homeservice corpus," May 2016, © European Language Resources Association, 2016. The LREC 2016 Proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

[18] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[19] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017.

[20] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sept 2015, pp. 1–6.

[21] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6440–6444, 2016.

[22] Z. Noorian, C. Pou-Prom, and F. Rudzicz, "On the importance of normative data in speech-based assessment," *ArXiv e-prints*, Nov. 2017.

[23] C. Bhat, B. Vachhani, and S. Kopparapu, "Improving recognition of dysarthric speech using severity based tempo adaptation," in *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings*, 2016, pp. 370–377.

[24] CMU Sphinx:, "The Carnegie Mellon Sphinx Project." [Online]. Available: http://cmusphinx.sourceforge.net/, Mar 2018

[25] VoxForge, http://www.voxforge.org/home/downloads, viewed March 2018.

[26] C. Cannam, "Rubber band: A library and utility program for changing tempo and pitch of an audio recording." [Online]. Available: http://breakfastquay.com/rubberband/

[27] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]. version 6.0.37." [Online]. Available: http://www.praat.org/

[28] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838.

[29] E. Frank, M. A. Hall, and I. H. Witten, "The weka workbench. online appendix for data mining: Practical machine learning tools and techniques." Morgan Kaufmann, Fourth Edition, 2016.

[30] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[31] S. H. K. Parthasarathi, B. Hoffmeister, S. Matsoukas, A. Mandal, N. Strom, and S. Garimella, "fmllr based feature-space speaker adaptation of dnn acoustic models," in *INTERSPEECH*, 2015.